Proceedings of

SMAC 03

Stockholm Music Acoustics Conference

August 6-9, 2003

Volume II

Edited by Roberto Bresin

2nd edition

Copyright © 2003 by the authors. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

The website of this book is `http://www.speech.kth.se/smac03`

Cover by Kjetil Falkenberg Hansen & Roberto Bresin

Contents
Volume I: Opening Session - The Sundberg Session – Violins – Guitars – Pianos – Brasses – Woodwinds – Organs – Percussions - Physics-Based Modelling
Volume II: Singing Voice - Music Performance - Music Perception - Instrument Control - Miscellanea

Volume I: ISBN 91-7283-558-3
Volume II: ISBN 91-7283-559-1

TRITA-TMH 2003:8
ISSN 1104-5787

CONTENTS

VOLUME I

Scientific advisory committee *xvi*

Preface *xvii*

Sponsors *xix*

OPENING SESSION

Recent developments in musical sound synthesis based on a physical model (*invited paper*) 3
J O Smith III

From musical acoustics to everyday acoustics: a physical modeling route (*invited paper*) 7
D Rocchesso

THE SUNDBERG SESSION

A marriage of the Director Musices program and the Conductor Program (*invited paper*) 13
M V Mathews, A Friberg, G Bennett, C Sapp, J Sundberg

Expressiveness in music performance: analysis and modelling (*invited paper*) 17
G De Poli

The voice as a musical instrument: fundamental differences between man-made and biological designs (*invited paper*) 21
Ingo R. Titze

Why music does not produce basic emotions: A plea for changing the paradigm in studying the emotional effects of music (*invited paper*) 25
K R Scherer

VIOLINS

The bowed string as we understand it today (*invited paper*) 31
J Woodhouse

Generalized normal mode violin acoustics (*invited paper*) 35
G Bissinger

An acoustical comparison of flat- and round-backed double basses 39
A W Brown

Operating deflecting modes of five conventional and
two unconventional violins recorded with TV-holography 43
A Buen, O J Løkberg

Numerical optimization of violin bows with varying densities
of the wood material 47
P Carlsson, M Tinnsten

Modeling Savart 's trapezoidal violin using a digital waveguide mesh 51
F Fontana, S Serafin

Experiments with an automatic bowing machine 55
P M Galluzzo, J Woodhouse

Psychoacoustic investigations on the possibility of aurally identical violins 59
P Geissler, O Martner, C Zerbs, M Schleske

The influence of vibrato and noise on the assessment of violin tone 63
C Gough

Bow speed or bowing position - which one influences spectrum the most? 67
K Guettler, E Schoonderwaldt, A Askenfelt

The BH-hill and tonal quality of the violin 71
E V Jansson

Inferring decision rules from jurys' ranking of competing violins 75
J Jelonek, E Łukasik, A Naganovski, R Slowinski,

AMATI: A multimedia database of violin sounds 79
E Łukasik

Unsupervised machine learning methods in timbral violin
characteristics visualization 83
E Łukasik, R Susmaga

Logarithmic damping decrement as the characteristic of
physical acoustics for quality estimation of violin wedges 87
E Rajčan, S Urgela, M Čulík

Effect of the width of the bow hair on the violin string spectrum 91
E Schoonderwaldt, K Guettler, A Askenfelt

Bowed string simulation using an elasto-plastic friction model 95
S Serafin, F Avanzini, D Rocchesso

Bowed string physical model validation through use of a bow controller
and examination of bow strokes 99
S Serafin, D Young

Determination of important wood properties for blanks
for violin tops by the use of numerical optimisation 103
M Tinnsten, P Carlsson

The admittance matrix of a cello 107
J Woodhouse, P E Courtney

Wireless sensor system for measurement of violin bowing parameters 111
D Young

GUITARS

Guitar models for makers (*invited paper*) 117
B E Richardson

Physical model of a complete classical guitar body 121
R Bader

A comparative study of the vibro-acoustical behaviour
of electric guitars produced in different decades 125
E Esposito

Modal radiation from classical guitars: experimental measurements
and theoretical predictions 129
T J W Hill, B E Richardson, S J Richardson

New Designs for the kantele with improved sound radiation 133
J Pölkki, C Erkut, H Penttinen, M Karjalainen, V Välimäki

The transient behaviour of guitar strings 137
J Woodhouse

PIANOS

Modeling the longitudinal vibration of piano strings 143
B Bank, L Sujbert

Quality of treble piano tones 147
A Galembo

Phase randomisation in piano bass tones 151
A Galembo, A Askenfelt

Measurement and reproduction accuracy
of computer-controlled grand pianos 155
W Goebl, R Bresin

The piano action as the performer's interface:
Timing properties, dynamic behaviour and the performer's possibilities 159
W Goebl, R Bresin, A Galembo

The pretransient of the harpsichord sound 163
T Gäumann

Sonological analyses of Harpsichord Sounds 167
A Schneider, A Beurmann

The piano soundboard behaviour in relation with its mechanical admittance 171
J S Skala

Experimental and theoretical studies of piano hammer 175
A Stulov

Analysis of piano tones with soft and hard touches 179
H Suzuki, J Kanno, S Mikimoto

BRASSES

Brass instruments as we know them today (*invited paper*) 185
D M Campbell

Harmonics: what do they do in the didjeridu? 189
N Amir

Bridging instrument control aspects of brass instruments with physics-based parameters 193
M Bertsch

Reproducibility and control of the embouchure of an artificial mouth for playing brass instruments 197
S R Bromage, O F Richards, D M Campbell

Australian Aboriginal musical instruments -- the gum-leaf 201
N H Fletcher

Studying lip oscillators of brass instruments: A distributed two dimensional lip model and its electrical equivalent circuit 205
W Kausel

Physical modeling of the trombone player's lips 209
D O Ludwigsen

Trumpet bell vibrations and their effect on the sound of the instrument 213
T R Moore, E T Shirley, A E Daniels

Corrections to the plane-wave approximation for acoustic waves in rapidly flaring horns 217
C J Nederveen

Gradient based optimisation of brass instruments 221
J O D Noreland

Measurement of the force applied to the mouthpiece during brass instrument playing 225
J F Petiot

Differentiation of trumpets' sound: experimental study with an adaptable depth mouthpiece 229
J F Petiot, E Poirson, J-L Le Carrou, J Gilbert

Modelling the lip reed - computational and experimental investigations of two-mode inward/outward striking behaviour 233
O F Richards, D M Campbell, J Gilbert

WOODWINDS

From sound synthesis to instrument making: an overview of recent researches on woodwinds and organs (*invited paper*) 239
B Fabre, A Hirschberg

Experimental research on double reed physical properties and its application to sound synthesis 243
A Almeida, C Vergez, R Caussé, X Rodet

The 'Virtual Flute': acoustic modelling at the service of players and composers 247
A Botros, J Smith, J Wolfe

Relationship between blowing pressure, pitch, and timbre of a Scottish bellows blown bagpipe 251
S Carral, D M Campbell, T D Rossing

The free reed coupled to a pipe resonator 255
J P Cottingham

Waveguide modelling of the Panpipes 259
A Czyzewski, B Kostek

Playing frequency shift due to the interaction between the vocal tract of the musician and the clarinet 263
C Fritz, J Wolfe, J Kergomard, R Caussé

Influence of losses on the saturation mechanism of single reed instruments 267
J Gilbert, J-P Dalmont, M Atig

Vibrato of single reed instruments 271
J Gilbert, L Simon, J Terroir

Acoustical and psychoacoustical investigations of the effect of crook bore profile on the playability of bassoons 275
T J W Hill, D B Sharp

RIAM (Reed Instrument Artificial Mouth): A computer controlled excitation device for reed instruments 279
A Mayer

Experimental investigation of clarinet reed operation and its consequence on the non-linear characteristics of the mouthpiece 283
S Ollivier, J-P Dalmont

Resonance curves and frequencies in multi-holed ocarinas 287
D R Peterson

Modeling vocal-tract influence in reed wind instruments 291
G P Scavone

Head joint, embouchure hole and filtering effects
on the input impedance of flutes 295
J Smith, J Wolfe, M Green

Flute performology; Statistical analyses and
numerical simulation of flute tone excitation. 299
Jan Tro, Aslak Bjerkvik, Ulf Kristiansen

Relating the harmonic-rich sound of the Chinese flute (dizi)
to the cubic non-linearity of its membrane 303
C G Tsai

Some effects of the player's vocal tract and tongue
on wind instrument sound 307
J Wolfe, A Z Tarnopolsky, N H Fletcher, L C L Hollenberg, J Smith

ORGANS

CFD analysis of air jet deflection - Comparison with Nolle's measurements 313
S Adachi

Breathing modes and sound radiation of metallic organ pipes 317
S Bergweiler, A Bergner, T Görne, M Wegener, R Gerhard-Multhaupt

Particle image velocimetry (PIV) measurements of velocity fields
at an organ pipe labium 321
E-L Johansson, L Benckert, P Gren

Space and air economy in the organ low register 325
J Liljencrants

Free reed instruments: clues for a physical model 329
L Millot

Subjective evaluation of organ pipe timbre in the standard listener positions 333
V Syrový, Z Otčenášek, J Štěpánek

PERCUSSIONS

Making of a computer carillon 339
M Karjalainen, P A A Esquef, V Välimäki

Normal modes of the elephant bell 343
R Perrin, B Deutsch, A Robinson, R Felce, T R Moore, G M Swallowe

Rayleigh's bell model revisited 347
R Perrin, G M Swallowe

The HANG: A hand-played steel drum 351
T D Rossing, U J Hansen, F Rohner, S Schärer

Vibrational modes of five-octave concert marimbas 355
J Yoo, T D Rossing, B Larkin

PHYSICS-BASED MODELLING

Physics-based modeling of musical instruments (*invited paper*) 361
V Välimäki

A power normalized non-linear lossy piano hammer 365
J Bensa, S D Bilbao, R Kronland-Martinet, J O Smith III

Sound synthesis for three-dimensional objects:
dynamic contact between two arbitrary elastic bodies 369
J Bensoam, R Caussé, C Vergez, N Misdariis, N Ellis

Musical applications of multichannel generalized digital waveguides 373
C Burns, S Serafin, M Burtner

Synthesis of voiced sounds by means of waveform adaptive physical models 377
C Drioli

The development of a modular software paradigm
for the physical modelling of musical instruments 381
I A Drumm

The principle of closed wavetrains, resonance and efficiency:
past, present and future 385
G Essl, P R Cook

Real-time synthesis models of wind instruments based on physical models 389
P Guillemain, J Kergomard, T Voinier

Time-domain physical modeling and real-time synthesis
using mixed modeling paradigms 393
M Karjalainen

A new method for the calculation of self-sustained oscillations:
the perturbation of the Helmholtz motion 397
J Kergomard, S Divoir, S Farner, C Vergez

Simulating the mechanism of sound generation in flutes
using the lattice Boltzmann method 401
H Kühnelt

Methods for synthesizing very high Q
parametrically well behaved two pole filters 405
M V Mathews, J O Smith III

Modeling and real-time synthesis of the kantele
using distributed tension modulation 409
J Pakarinen, M Karjalainen, V Valimaki

The estimation of birdsong control parameters
using maximum likelihood and minimum action 413
T Smyth, J S Abel, J O Smith III

Author index *417*

GNU Free Documentation License *425*

VOLUME II

Sponsors *xvii*

SINGING VOICE

Singing voice modeling - as we know it today (*invited paper*) 431
M Kob

Using imaging and modeling techniques to understand the relationship of vocal tract shape to acoustic characteristics (*invited paper*) 435
B H Story

Sample-based singing voice synthesizer by spectral concatenation 439
J Bonada, A Loscos, H Kenmochi

Acoustical effects of the core principles of the 'Bel Canto' method on choral singing 443
L Fagnan, X Rodet

The influence of the practice of basso continuo on the intonation of a professional singer in the time of Monteverdi 447
G Faraone, S Johansson, P Polotti

Acoustic pulse reflectometry for vocal tract measurement 451
C D Gray, D M Campbell, C A Greated

On the use of electroglottography for characterisation of the laryngeal mechanisms 455
N Henrich, B Roubeau, M Castellengo

Vocal fold resonances at low and high pitch tuning 459
S Hertegård, H Larsson, S Granqvist

A capella SATB quartet in-tune singing: evidence of intonation shift 463
D M Howard

Larynx closed quotients in a capella SATB quartet singing 467
D M Howard

The effect of the hypopharyngeal and supra-glottic shapes for the singing voice 471
H Imagawa, K-I Sakakibara, N Tayama, S Niimi

The effect of vocal vibrato on blending individual singer characteristic with hall acoustics 475
K Kato, D Noson, Y Ando

Singer identification through dynamic modeling of vocal fold and vocal tract parameters 479
Y E Kim

Spectral modeling of the singing voice using
asymmetric generalized Gaussian functions 483
M E Lee, M J T Smith

Comparison of directional sources in simulating a soprano voice 487
L Parati, F Otondo

Neck muscle activation in professional classical singing 491
V Pettersen, R H Westgaard

The laryngeal flow model for pressed-type singing voices 495
K-I Sakakibara, H Imagawa, S Niimi, N Osaka

Acoustic and aerodynamic analyses of support and resonance
demonstrated by an elite baritone singing teacher 499
R C Scherer, N Radhakrishnan, A Poulimenos, F Alipour

Perceptual Relevance of the 5kHz spectral region
to sex identification in children's singing voices 503
P J Sjölander

MUSIC PERFORMANCE

Large-scale performance studies with
intelligent data analysis methods (*invited paper*) 509
G Widmer

Studies of music performance: a theoretical analysis
of empirical findings (*invited paper*) 513
P N Juslin

Estimation of the note durations in a polyphonic piano transcription system 517
A M Barbancho, I Barbancho, L J Tardón

Expressive Director: a system for the real-time control
of music performance synthesis 521
S Canazza, A Rodà, P Zanon, A Friberg

Expressiveness analysis of virtual sound movements and
its musical applications 525
A de Götzen

A statistical approach to expressive intentions recognition
in violin performances 529
R Dillon

Melodic characterization of monophonic recordings for
expressive tempo transformations 533
E Gómez, M Grachten, X Amatriain, J L Arcos

Measurements and models of musical articulation 537
J Jerkert

The percussion instrument operation synchronized
with the visual and/or auditory stimuli 541
T Kamitani, K Kouzuki, M Matsuda

Aspects on time modification in score-based performance control 545
M Laurson, M Kuuskankare

Interaction among physical acoustic knowledge and
performance in the accordion and its educational side 549
R Llanos-Vazquez, J Alonso-Moral, M J Elejalde-Garcia, E Macho-Stadler

Time domain note average energy based music onset detection 553
R Liu, N J L Griffith, J Walker, P Murphy

Application of Bayesian Networks to automatic recognition
of expressive content of piano improvisations 557
L Mion

ESP - an interactive collaborative game using non-verbal communication 561
M L Rinman, A Friberg, I Kjellmo, A Camurri, D Cirotteau,
S Dahl, B Mazzarino, B Bendikesen, H McCarthy

Mapping a physical correlate of loudness into
the velocity space of MIDI-controlled piano tones 565
T Taguti

On the relation between performance cues and emotional engagement 569
R Timmers, A Camurri, G Volpe

Analysis of jazz drummers' movements in performance
of swing grooves - a preliminary report 573
C H Waadeland

A singing transcription system using melody tracking algorithm
based on Adaptive Round Semitone (ARS) plus music grammar constraints 577
C K Wang, R Y Lyu, Y C Chiang

Learning to recognize famous pianists with machine learning techniques 581
P Zanon, G Widmer

MUSIC PERCEPTION

Automatic transcription of music (*invited paper*) 587
A P Klapuri

Brain activity in perception and retention in memory
of the pitch of music and speech 591
E Castro-Sierra, H A Poblano

Looking at perception of continuous tempo drift - a new method
for estimating Internal drift and Just Noticeable Difference 595
S Dahl, S Granqvist

What can the body movements reveal about a
musician's emotional intention? 599
S Dahl, A Friberg

Time estimation: isochronous versus accelerated audio sequences 603
D Dall'Osto

Timbral semantics and the pipe organ 607
A C Disley, D M Howard

Hardness recognition in synthetic sounds 611
B L Giordano, K Petrini

Categorical perception of microtonal intervals 615
E Huovinen

Towards a virtual singer: contribution to the development of the teaching of sung diction through the intermediary of a functionalised system 619
M Nahas, H Huitric, J Kiss

Musical sound parameters revisited 623
B Kostek, P Zwan, M Dziubinski

Pitch measurements versus perception of South Indian classical music 627
A Krishnaswamy

Relationships between musical audio, perceived qualities and motoric responses 631
M Leman, V Vermeulen, L De Voogdt, A Camurri, B Mazzarino, G Volpe

User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology 635
M Lesaffre, M Leman, K Tanghe, B De Baets, H De Meyer, J P Martens

Consistency in listeners' ratings as a function of listening time 639
G Madison, B Merker

Verbal description of musical sound timbre in czech language 643
O Moravec, J Štěpánek

Development of a language for specifying saxophone timbre 647
A J Nykänen, Ö Johansson

Influence of duration of tone stationary part on perception of starting transient 651
Z Otčenášek, J Štěpánek, V Syrový

Perceived influence of changes in musical instrument directivity representation 655
F Otondo, B Kirkwood

Effects of the grouping cue on the pitch shift of a pure tone induced by other tones 659
T Shirado, M Yanagida

Listeners common and group perceptual dimensions in violin timbre 663
J Štěpánek, Z Otčenášek

Pattern recognition of musical instruments using Hidden Markov Models 667
R Ventura-Miravet, F Murtagh, J Ming

Influence of rhythmic, melodic, and semantic violations in language and music on the electrical activity in the brain 671
S Ystad, C Magne, S Farner, G Pallone, C Astesano, V Pasdeloup, R Kronland-Martinet, M Besson

INSTRUMENT CONTROL

Remutualizing the Instrument:
Co-design of synthesis algorithms and controllers (*invited paper*) 677
P Cook

Interactive music systems for everyone:
Exploring visual feedback as a way for creating more intuitive,
productive and learnable instruments (*invited paper*) 681
S Jordà

Controlling the virtual bodhran - the vodhran 685
R Bresin, S Dahl, M Rath, M Marshall, B Moynihan

The Radio Baton as configurable musical instrument and controller 689
R Bresin, K F Hansen, S Dahl

DJ scratching performance techniques: Analysis and synthesis 693
K F Hansen, R Bresin

Gesture-tactile controlled physical modelling music synthesis 697
D M Howard, S Rimell

MISCELLANEA
Including* OPTICAL AND ACOUSTICAL MEASUREMENTS *and
TEACHING MUSIC ACOUSTICS TO MUSICIANS

Visualization of sound fields with application
to musical instruments (*invited paper*) 703
N E Molin

Breathing modes of woodwind bodies:
Experimental detection with ring-shaped piezoelectric-polymer sensors 707
M Wegener, R Gerhard-Multhaupt, W Wirges, A Bergner, S Bergweiler

Leak detection in musical wind instruments using
acoustic pulse reflectometry 711
V Chilekwa, D B Sharp, T J W Hill

An EDS modelling tool for tracking and modifying the musical signals 715
B David, G Richard, R Badeau

Didactic resources for the teaching and
popularisation of the physics of the music 719
M J Elejalde-García, E Macho- Stadler, A Franco-García,
J Janariz-Larumbe

Quantitative assessment of air flow from professional bass reflex systems
ports by Particle Image Velocimetry and Laser Doppler Anemometry 723
E Esposito, M Marassi

Computer-animated illustrations of vibrations and waves 727
D E Hall

Musical acoustics courses in the education of music students in the U.S.A. 729
D E Hall

Unusual motions of a nonlinear asymmetrical vibrating string 731
R J Hanson, H K Macomber, A C Morrison

Reducing the source tube to improve the bandwidth of acoustic pulse reflectometry 735
A Li, D Sharp

An apparatus for measuring string vibration using electric field sensing 739
J Pakarinen, M Karjalainen

An ensemble method for automatic alignment of sound and its transcription 743
K Sjölander

Measurement of the termination impedance of a tube using particle image velocimetry 747
D J Skulina, D M Campbell, C A Greated

Direct measurement of the 3-D acoustic impedance in wind instruments near field 751
D Stanzial, D Bonsi

Sound and vibration from a special musical instrument: the Mazzacorati theatre 755
L Tronchin, E Esposito, V Tarabusi

Large-bandwidth measurement of acoustic impedance using two microphones and four calibrations 759
M O van Walstijn, D M Campbell

Refractoscopic visualisation of sound in musical instruments 763
L Zipser, H Franke

Author index *767*

GNU Free Documentation License *775*

SPONSORS

Stockholm Music Acoustics Conference, August 6 – 9, 2003 (SMAC 03) SMAC 03 was made possible due to generous grants from:



Royal Institute of Technology (KTH)
www.kth.se



The Wenner-Gren Foundations
www.wenner-grenstift.a.se



The Bank of Sweden Tercentenary Foundation
www.rj.se



The Swedish Research Council
www.vr.se

Helge Ax:son Johnsons Stiftelse



Royal Swedish Academy of Music
www.musakad.se



Catgut Acoustical Society
www.catgutacoustical.org



City of Stockholm
www.stockholm.se



The Italian Embassy
www.itemb.se/science/



The Italian Institute of Culture in Stockholm
http://www.italkult.com/



The Swedish Radio
www.sr.se

SINGING VOICE

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

SINGING VOICE MODELING – AS WE KNOW IT TODAY

Malte Kob

Dep. of Phoniatrics, Pedaudiology and Communication Disorders
University Hospital, Aachen University
mkob@ukaachen.de

ABSTRACT

The human voice can be described as the joint product of several functional components of the voice organ. Several approaches follow the beaten path of source-filter separation of the complex system. With such models the interaction between the components cannot be taken into account for real-time calculations of the voice signal. These missing feedback effects can be included by using a time-domain model for simulation of the voice generation process, including a multiple-mass vocal fold model, noise generation, wave propagation through the vocal tract, and radiation at the mouth. An example program is demonstrated that includes a graphic user interface allowing the modification of parameters like lung pressure, glottis configuration and vocal tract shape during calculation. The application of the model to singing voice synthesis (e.g. overtone singing) and generation of pathologic vocal fold movement is demonstrated.

1. INTRODUCTION

A comprehensive overview of the numerous approaches of voice modeling is hard to give within the scope of this four-page article. Luckily, the review written in 1998 by Perry R. Cook [1] addresses many of nowadays methods related to singing voice synthesis. A number of successful implementations of singing voice programs is presented (MUSSE DIG [2], CHANT [3], SPASM [4], Pavarobotti [5], Lyricos [6]), and aspects of face, lip and voice perception are discussed. Whereas that article focuses rather on the voice modulation and shaping due to the vocal tract function, it does not address the relevant physical aspects of the voice generation process at the voice source, i.e. the vocal fold function. Another, recent approach is to use morphed voice samples for resynthesis as in speech synthesis. In this article, only models that are based upon the singer's physiology will be discussed.

So-called physical models try to imitate the voice generation process by mathematically or physically mimicking the physical laws that reign in the voice organs. This approach is very different from wave modeling where the aim is an approximation of a given waveform or spectrum by additive synthesis or filtering of selected signals. A satisfactory real-time modeling of the singing voice based upon the physiology of an individual singer is not yet possible, even though the advantages of such an approach are tempting:

- the synthesis of individual voices could be based upon the actual physiology,
- register changes or timbre were the result of complex muscle and tissue interaction with the air flow,
- voice pathologies in professional singers could be studied *in vitro*, and critical situations could be simulated and investigated by monitoring the forces within the tissues.

Limitations of physical models available today are the computational costs for an accurate simulation of the physiology of the voice organs. But not only the computer speed needs to be improved. Still, some important knowledge is missing as the three-dimensional description of the flow in the glottis, which is crucial for an accurate modeling of the vocal fold movement [7]. More restrictions occur in the availability of geometrical and physical properties of the voice organ. This work presents a model implemented in MATLAB for easy implementation and extension of algorithms and visualisation of their results.

2. MODEL COMPONENTS

For singing voice synthesis a combined model, named VOX, was used, consisting of separate models for vocal folds (VF), vocal tract (VT) and noise generation. The degree of interaction between the models and the choice of parameters can be controlled in a graphical user interface (GUI) with buttons and sliders that change parameters such as:

- Reflected flow from glottis back into VT, noise generation, radiation impedance
- Adjustment of lung pressure, properties of vocal folds and vocal tract

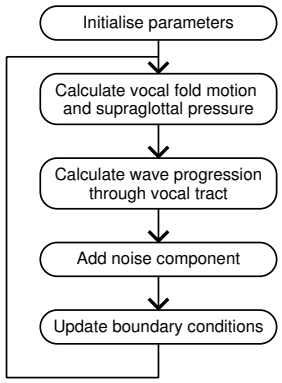


Figure 1: *Signal flow of the combined model VOX*

The signal flow of the combined model is given in Figure 1. After initialization of the global parameters the vocal fold signal is calculated sample-wise. The sound pressure wave is fed into the vocal tract model and the resulting pressure wave at the mouth is stored. All user-defined parameters are accessible during calculation and allow an arbitrary change of the boundary conditions.

2.1. Vocal Folds

The described model is based on the 16-mass model developed by I. R. Titze [8] but includes some more recently published modifications. The values of parameters and the equations are discussed in [9]. The implemented model offers the following features:

- Separate modeling of each vocal fold
- Modeling of modal, head and falsetto register
- Variation of individual masses or spring stiffness in an arbitrary number of sections for simulation of e. g. singer's nodules
- All parameters can be changed and monitored during calculation

Figure 2 illustrates the geometry of the vocal fold model. The arrangement of n masses $m_{m,i}$, with $i = 1..n$ represents the vocalis muscle, $m_{v,i}$ represents the mucosa membrane. Each of the big masses $m_{v,i}$ is connected to the boundary by a spring with stiffness k^b and damping D^b, and to the small mass $m_{m,i}$ by a spring with stiffness k^m and damping D^m.

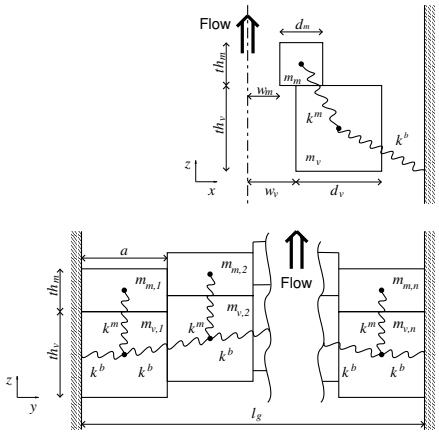


Figure 2: Sectional view (top) and side view (bottom) of the implemented vocal fold model

The most important difference between this model and classical two-mass models is the segmentation of the VF in n segments. The sum of the arbitrary number of n segments of width a is the longitudinal length l_g of the VF. Further, the assumption of jet separation in the divergent glottis was made [10].

2.2. Vocal Tract

Two models have been implemented: a waveguide approach based on the model of J. L. Kelly and C. C. Lochbaum (KL) [11] and a model using a multiconvolution algorithm (CTIM).

2.2.1. Reflection type line analog

The implemented waveguide model takes into account frequency-independent propagation losses by multiplication of the transmitted pressure wave by a damping factor. The radiation impedance, however, is calculated using the model of a baffeled piston. In Figure 3 the signal flow for the waveguide calculation is shown. For each sample, the user-defined parameters are imported from the main program *VOX*. Depending on the spacing of the vocal tract data, the sample rate is calculated. When the noise generation module is active, the sound pressure in appropriate segments is augmented by turbulent noise. Then the vectors are shifted and the loop is repeated until the time limit is reached.

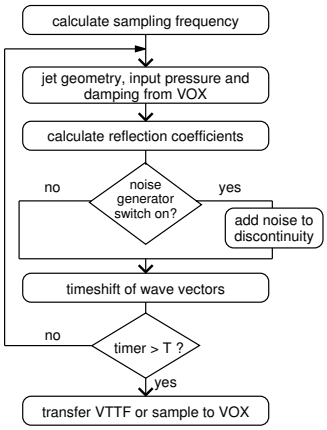


Figure 3: *Algorithm of the waveguide model*

The algorithm can be used in interactive mode for the generation of single pressure samples or in stand-alone mode for the calculation of vocal tract transfer functions (VTTF). A more detailed description and the MATLAB code of the algorithm is given in [9].

2.2.2. Multiconvolution technique

The algorithm of A. Barjau *et al.* [12] is an improved multiconvolution technique based on the work of J. Martínez *et al.* [13, 14]. Whereas the time-domain algorithm described by Martínez requires a spacing of the discontinuities that is restricted to multiples of phase velocity times time step, the interesting aspect of the CTIM (continuous-time interpolated multiconvolution) algorithm is its independence of the spatial VT discretization and the sample rate. Originally, the algorithm has been invented for the impedance calculation of woodwind instruments with toneholes and moderately changing cross-sections. The calculation scheme differs from the waveguide approach in the following aspect: An update of the tube segment geometry is necessary after changes in the vocal tract geometry only. Therefore, the constant parameters that are used for the convolution routine can be calculated once before the loop is entered. As a consequence, the algorithm is much faster than the waveguide approach. However, for certain geometries the CTIM algorithm does not produce stable results [15].

The calculation of the tube segments is done with the help of a graphic user interface and allows the instantaneous interpretation of the results during construction.

2.3. Noise Generation

During voice production noise is generated either when a stream of air is directed against an obstacle (e.g. fricatives) or when a divergent configuration or a discontinuity of the bounding walls causes a laminar flow to separate from the boundary (aspiration noise). In both cases the laminar flow is disturbed and vortices are

shed that travel some distance and then break down and dissipate their energy in turbulent air movement.

The noise generation algorithm used here is based on the work of D. J. Sinder [16]. For each vortex, position and diameter are calculated from the actual flow parameters. The vortex production rate increases when the flow increases or when the diameter of the opening decreases.

$$P_{noise} = \frac{-\pi d \rho_0}{A}[\vec{\Gamma} \times \vec{v} \cdot \vec{U}]. \quad (1)$$

In equation (1) $\Gamma = \frac{1}{2}v^2 T$ is the vortex rotation, $\vec{U}$ denotes the normal component of the surrounding flow, and T is the vortex shedding interval that depends on the jet diameter d, the Strouhal number and the jet velocity v. According to equation (1) the contribution of the vortex to the turbulent sound is calculated.

During propagation the object properties are updated until the conditions for the life of the vortex are expired. Different from Sinder's algorithm the assumption of energy loss of the vortex during propagation has been added. It is assumed that the vortex energy is reduced exponentionally with growing distance at an estimated rate of 40 dB for a travel distance of $4\,D$.

The propagation of vortices through the vocal tract can be visualised for different vowels. It was found that the aspiration noise contribution is significantly affected by the vocal tract geometry. This findings are in agreement with measurement results [9].

3. APPLICATIONS

Two examples for the application of the above model are presented here: the modeling of the VT transfer function of an overtone singer and the modeling of the vocal fold movement of a singer with vocal fold nodules. More results for other configurations of the voice organ have been obtained such as sustained vowels in different registers [9].

3.1. Modeling of Overtone Singing

The synthesis of overtone singing aims at a verification of the modeled process of sound generation. In a first step the vocal tract geometry is modeled using MRI data that were obtained by Adachi and Yamada [17] from a Xöömij overtone singer for four different melody pitches F_6, G_6, A_6, C_7 which equal 1397 Hz, 1568 Hz, 1760 Hz and 2093 Hz. In Figure 4 the comparison of the original and the modified area functions of the overtone A_6 as well as the corresponding vocal tract transfer functions (VTTF) are shown. The area functions taken from S. Adachi (thin solid, visible at the mouth) have been modified at the mouth opening (thick solid). The VTTF calculation has been carried out using the CTIM algorithm.

An improvement of the second resonance by about 15 dB was achieved by matching two resonance frequencies [9, 18]. Similar results can be obtained using the KL model, although the amplification is less impressive compared to the CTIM model. An auralization of the sound using the VF model in modal register configuration and the above configuration of the VT is clearly identified as a sound with two pitches.

3.2. Modeling of Voice Pathologies

Singer's nodules are characterized by a local increase of tissue, in most cases on opposite locations of the vocal folds. In the set-up

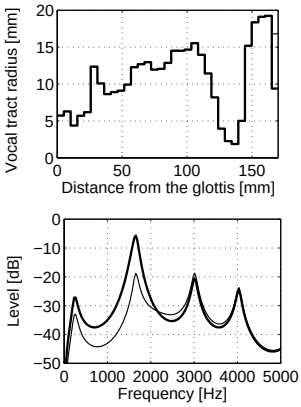


Figure 4: *Area functions (top) and VTTF (bottom) of overtone* A_6

for the calculation of the voice with a singer's nodule (see Figure 5) one mass of one segment out of 11 vocal fold segments is increased. A factor of the value 3 has been chosen, both for vocal

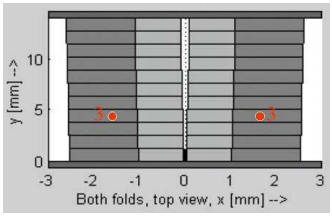


Figure 5: *Set-up for fold movement with nodule*

fold and mucosa segments. The observation of the animation of the masses reveals an irregular VF movement that starts up regular with the first (1,0) mode. After a few cycles, the oscillation shifts to the second (longitudinal) (2,0) mode and stays in a periodic but rather complex movement pattern. The fundamental frequency is divided by two compared to the simulation of a healthy voice. The sound pressure and the spectrum of the resulting signals can be visualized and auralized. The signal exhibits a significantly lower

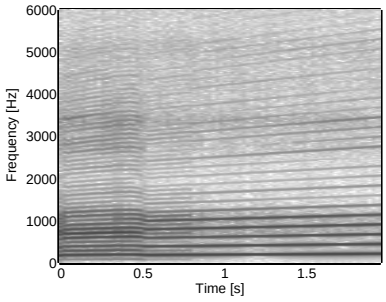


Figure 6: *Register transition due to increase of vocalis stress*

fundamental frequency that is due to the irregular VF movement pattern in the (2,0) mode. In contrast to the simulations the *in vivo* observation of the vocal folds in patients suffering from singer's nodules does not necessarily exhibit an irregular VF movement.

This might be caused by an automatic compensatory parameter change. It was found that the irregular movement can be turned into a regular one by increasing the active stress $T_{v,act}$ on the *vocalis* muscle. With increasing $T_{v,act}$, the fundamental frequency rises and, at a certain value, the oscillation falls back to the first (1,0) mode. Figure 6 depicts a spectrogram of the transition that occurred during the increase of the active stress. It was found out by successive increasing of $T_{v,act}$ that the change between the two states of oscillation takes place when the active stress is increased by ca. 5%. Sound examples, animations of the vocal fold movement for different configurations, pictures of the graphic user interface, and the complete source code of VOX can be found on the Internet [19].

4. CONCLUSIONS

In this article an overview has been given of components that constitute a physical model for voice synthesis. Applications of the model have shown that it can serve both as a research device and as an educational tool for modeling, visualization and auralization of the human voice. The challenge of building such a model is to find a trade-off between a detailed description of the physiology and the physical relations that are involved on one hand and, on the other hand, the implementation of fast and stable algorithms. Due to the complexity of the voice generation process only the most important models for generation of sustained vowels have been implemented so far. The focus was laid on the mathematical description of the main components of voice generation with an emphasis on singing voice synthesis. In a pure physical model no attempt is made to "improve" the generated signals by means of post-processing as filtering to obtain a more realistic sound. For all models, algorithms described in literature were modified, implemented in MATLAB and checked with respect to their suitability for singing voice synthesis.

These models are coupled in several ways. The influence of their coupling has been investigated and it could be shown that the simple oscillator-filter approach is of limited accuracy [9]. The model includes generation of noise that is inserted into the vocal tract at appropriate locations with a sound pressure that depends on the aeroacoustic conditions of vortex shedding during phonation. The result is a more natural sound than just the sum of a noise signal and the harmonic voice signal.

The application of a physical model to the synthesis of overtone singing gives insight into the physics of this amazing singing style. The modeling of a voice with pathologic vocal fold configuration demonstrates the relation between changes in the vocal fold parameters and the resulting sound wave.

One aspect that is most difficult when modeling singing voice is the control of the numerous parameters. The accurate control of e.g. vibrato is now a complex issue: it is rather easy to derive the statistics of a waveform modulation that perceptually gives a good vibrato but it is very difficult to know what physical parameters must be altered to give a satisfactory result. Literature on this subject is rather rare, and interdisciplinary research is necessary to approach this question.

A current project focuses upon the inclusion of all intra- and most extralaryngeal muscles into the sound organ. The goal is a natural definition of the boundary conditions (tensions, length, position) of the vocalis and mucosa tissues.

I am confident that we will we hear the sound of a satisfactory real-time physical singing voice model at SMAC 2013!

5. REFERENCES

[1] Perry R. Cook (1998): Toward the Perfect Audio Morph? Singing Voice Synthesis and Processing.

[2] J. Zera, J. Gauffin and J. Sundberg (1984): Synthesis of Selected VCV-Syllables in Singing. Proc. Int. Comp. Music Conf. IRCAM, Paris 83-86.

[3] M. Mathews and J. Pierce (1989): Some Current Directions in Computer Music Research. Cambridge MA. MIT Press 57-63.

[4] Perry R. Cook (1992): SPASM: a Real-Time Vocal Tract Physical Model Editor/Controller and Singer: the Companion Software Synthesis System. Computer Music Journal **17** (1) 30-44.

[5] Brad Story and Ingo Titze (1995): Voice Simulation With a Body-Cover Model of the Vocal Folds. J. Acoust. Soc. Am. **97** (2) 3416-3413.

[6] M. Macon, L. Jensen-Link, J. Oliverio, M. Clements and E. George (1997): A Singing Voice Synthesis System Based on Sinusoidal Modeling. Proc. ICASSP Vol. 1 439-442.

[7] X. Pelorson *et al.* (1995): Description of the flow through in-vitro models of the glottis during phonation. acta acustica **3** 191-202.

[8] I. R. Titze (1973): The Human Vocal Cords: A Mathematical Model Part I. Phonetica **28** 129-170.

[9] M. Kob (2002): Physical modeling of the singing voice. Dissertation, University of Technology Aachen, Logos Berlin. Download at `http://www.akustik.rwth-aachen.de/~malte/papers`

[10] A. Hirschberg (1992): Some fluid dynamic aspects of speech. Bulletin de la Communication Parlée **2** 7-30.

[11] J. L. Kelly and C. C. Lochbaum (1962): Speech synthesis. Proceedings of the 4^{th} International Congress on Acoustics, 1-4.

[12] A. Barjau, D. H. Keefe and S. Cardona (1999): Time-domain simulation of acoustical waveguides with arbitrarily spaced discontinuities. J. Acoust. Soc. Am. **105** (3) 1951-1964.

[13] J. Martínez and J. Agulló (1988): Conical bores. Part I: Reflection functions associated with discontinuities. J. Acoust. Soc. Am. **84** (5) 1613-1619.

[14] J. Martínez, J. Agulló and S. Cardona (1988): Conical bores. Part II: Multiconvolution. J. Acoust. Soc. Am. **84** (5) 1620-1627.

[15] , J. Agulló, A. Barjau and J. Martínez (1992): On the time-domain description of conical bores. J. Acoust. Soc. Am. **91** (2) 1099-1105.

[16] D. J. Sinder (1999): Speech Synthesis Using an Aeroacoustic Fricative Model. Ph.D. Thesis, University of New Jersey.

[17] S. Adachi and M. Yamada (1999): An acoustical study of sound production in biphonic singing, Xöömij. J. Acoust. Soc. Am. **105** (5) 2920-2932.

[18] M. Kob (2001): Untersuchung der Eigenschaften des Obertongesangs. Fortschritte der Akustik – DAGA 01, 436-437.

[19] URL: `http://www.akustik.rwth-aachen.de/~malte/vox`

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

USING IMAGING AND MODELING TECHNIQUES TO UNDERSTAND THE RELATION BETWEEN VOCAL TRACT SHAPE TO ACOUSTIC CHARACTERISTICS

Brad H. Story

Department of Speech and Hearing Sciences
University of Arizona
bstory@u.arizona.edu

ABSTRACT

The respiratory, laryngeal, and vocal tract systems are precisely orchestrated by humans as an instrument of speech and song. This presentation will focus primarily on the role of the vocal tract as a versatile acoustic device capable of producing a wide range of possible vowel and vowel-like sounds. In particular, recent data obtained in vocal tract imaging studies will be used to demonstrate some relations between the area function and resulting acoustic characteristics. Formant clustering and vowel modification will be specifically addressed.

1. INTRODUCTION

The system of airways that courses through the chest, neck, and head of the human body can be considered part of an instrument of music for singing or an instrument of communication for speech. In either case, the detailed shape of each subcomponent of the airway system acts as an acoustic filter that transforms the sounds generated by one or more possible sources into the acoustic signal radiated from a speaker or singer.

Figure 1 shows the trachea, vocal tract, nasal tract, and various sinuses as an idealized set of interconnected tubes. When the vocal folds vibrate and produce a periodic train of glottal flow pulses or turbulence is produced somewhere in the airway system, each of the tubes is potentially involved in filtering the sound. The tracheal section is known to support a series of acoustic resonances or "subglottal formants" ([1],[2]) that may have some influence on the sound produced by a speaker or singer, especially for breathy voices [3]. It has also been suggested that tracheal resonances contribute to involuntary vocal register transitions based on the relationship of acoustic pressures inferior to the vocal folds and pressures that drive them intraglottally [4]. The resonances of the nasal tract [5] will affect the output sound as well, however, the strength of their presence depends on the degree of acoustic coupling provided by the velopharyngeal port (represented by the hinged door in Fig. 1).

While their acoustic effects can be significant, only minor intentional adjustments to the shape of either the trachea or nasal tract are possible. In contrast, through positioning and movement of the tongue, jaw, and lips, a speaker or singer can exert a large amount of control over the *vocal tract* shape. Thus, "playing the instrument" primarily involves knowing how to optimally influence the shape of the vocal tract tube of Figure 1. Two-tube and three-tube approximations for vowels are well known [6] and reasonably explain their phonetic qualities in terms of the frequency values of the first two formants (F1 & F2). However, the shaping of the vocal tract that gives rise to each person's unique sound,

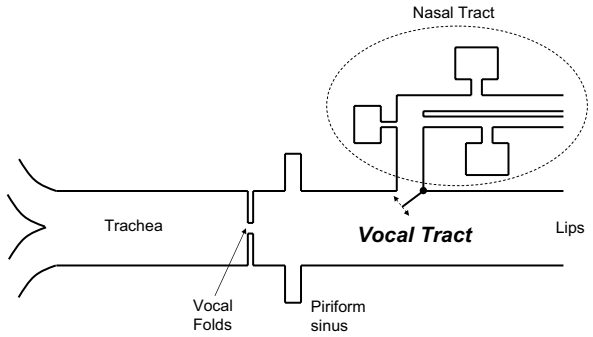


Figure 1: *Idealized tubular representation of the airway system. Included are the trachea, vocal tract, nasal tract and various sinuses.*

the enhancement of spectral properties for singing, or production of various voice qualities, typically involves subtle modifications to the tract shape. These may not be important phonetically, but nonetheless make large differences in the characteristics of the output sound.

A familiar route toward understanding this complex relation between the vocal tract shape and its acoustic filtering consequences, is to combine accurate measurements with acoustic modeling. The aim of this paper is to first give a brief overview of volumetric imaging of the vocal tract to measure area functions. Then, results of some imaging experiments will be used to discuss two types of vocal tract modifications that are relevant to singing.

2. VOCAL TRACT IMAGING

In recent years various imaging techniques, have become viable means for measuring the three-dimensional vocal tract shape (e.g. [7],[8].[9],[10]). Because of its relative safety, magnetic resonance imaging (MRI) is typically the technique of choice. But x-ray computed tomography (CT) is also an option if the speaker/singer being imaged is allowed only very low doses of radiation. CT has the advantage of higher resolution, greater speed of image acquisition, and imaging of teeth. Either type of imaging allows collection of a contiguous set of image slices across the volume of the head and neck. From this a 3-D representation of the vocal tract can be reconstructed.

As an example, Figure 2 shows a 3-D reconstruction of a male *spoken* vowel /a/ based on images collected with CT [9] (note that only airspace is visible in this picture – all articulatory structures

have been removed). The area function for this vowel, shown in Figure 3, was obtained by measuring the cross-sectional area within oblique sections perpendicular to a centerline extending from the glottis to the lips. In a gross sense, both the image reconstruction and area function show the typical pharyngeal constriction and widened oral cavity that would be expected with this vowel. Other features that may have acoustical significance are also apparent, such as the epilaryngeal tube, piriform sinuses, and coupling to the nasal tract (making this a nasalized vowel).

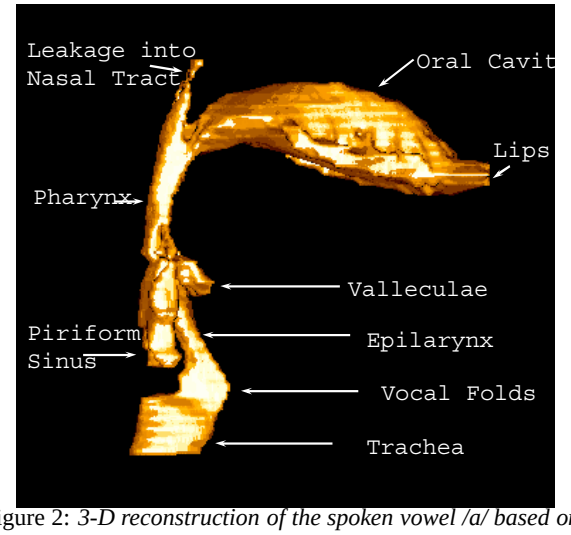


Figure 2: *3-D reconstruction of the spoken vowel /a/ based on CT imaging ([9]).*

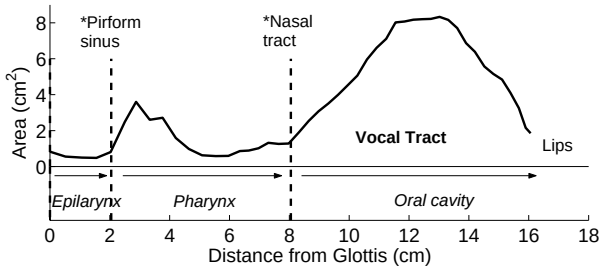


Figure 3: *Area function of a spoken vowel /a/ shown in Fig. 2. The *'s denote the locations at which the piriform sinuses or nasal tract would be coupled to the main vocal tract (Based on data in [9]).*

3. SINGER'S FORMANT

The epilarynx tube, located just above the vocal folds (see Fig. 2), has been suggested as a primary contributor to a clustering of formant frequencies around 3000 Hz, collectively called the "Singer's Formant" [11]. This tube is thought to create a nearly independent resonator if its cross-sectional area is roughly six times less than the area in the lower pharynx [11]. The area function in Fig. 3 does meet this criteria and a frequency response function calculated for it (Fig. 4 – solid line) indicates a clustering of F3 and F4, but in the frequency range of about 4300 Hz rather than 3000 Hz (the nasal coupling indicated in Fig. 2 for this vowel was ignored for this calculation). However, the length of the epilarynx for this particular *speaker* is only 2 cm, which does correspond to a closed-open tube resonance frequency of

$$F = \frac{c}{4L} = \frac{35000 \text{ cm/s}}{(4)(2 \text{ cm})} = 4375 \text{ Hz}. \quad (1)$$

A *singer* may lower the larynx and extend this tube length or just be naturally endowed with a longer epilarynx in the first place. The frequency response function shown by the dashed line in Fig. 3 was calculated for the same area function as before except that the epilarynx length was changed to 3 cm. It is apparent that the clustering of F3 and F4 is now centered around 3000 Hz which is nearly the calculated resonance frequency for the longer 3 cm tube. Also noteworthy is that the first two formants (F1 and F2) have been left undisturbed by modification of the epilaryngeal length, suggesting that the vocal tract shape in the region near the glottis primarily contributes to voice quality and not phonetic quality.

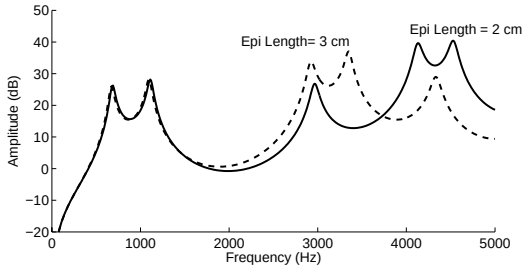


Figure 4: *Frequency response functions for the area function shown in Fig. 3. Solid line = 2 cm epilarynx length, dashed line = 3 cm epilarynx length.*

4. VOWEL MODIFICATION

Voice production at high fundamental frequencies by female singers is often accompanied by modifications to the vocal tract configuration of a particular vowel [12],[13]. The singer's desire seems to be to avoid phonation at frequencies exceeding the first formant (F1) of the vocal tract. This is particularly an issue for the vowels /i/ and /u/, which both have an F1 around 300 Hz. For example, if the demands of the music require production of the note D5 (587 Hz) with the vowel /i/, the fundamental frequency of phonation would be above F1. That is, unless the vocal tract shape is modified to move F1 upward in frequency. The reasons for vowel modification seem to be related to maintaining an appropriately high intensity. When the fundamental frequency is maintained just below F1, the gain in intensity could be as much as 30 dB [13].

In a study of voice quality [14], the vocal tracts of two professional vocalists (one male and one female) were imaged with MRI during the production of the vowels /i/,/ae/,/a/, and /u/ under conditions of normal speech, twangy, and yawny voice qualities. In addition, similar vocal tract data for vowels /i/, /a/, /u/ were collected under the conditions of high-pitched singing [15].

Figure 5 shows the area functions for the female's normal speech-like /i/ vowel along with the /i/ shape she produced while phonating on note D5 (587 Hz). A significant increase in cross-sectional area is observed in the oral cavity (9–14.5 cm from the glottis) for the sung vowel. The calculated frequency response functions for both area functions are shown in Figure 6. Also, vertical lines are placed at the location of the fundamental frequency

used for each vowel production. The frequency response functions show that the increase in oral cavity area moved F1 upward in frequency so that it is just above the fundamental frequency.

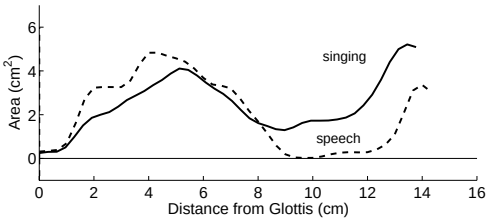


Figure 5: *Area functions for /i/ vowels spoken (dashed line) and sung (solid line) by a female singer.*

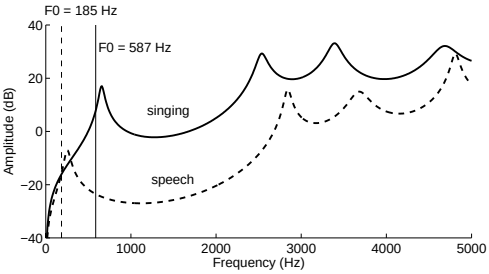


Figure 6: *Frequency response functions for the /i/ area functions shown in Fig. 5. The vertical lines indicate the fundamental frequencies for each vowel production. (Dashed = speech, Solid = sung.)*

Results for the vowel /a/ are shown in Figures 7 & 8. In this case the area functions show only slight differences between the spoken and sung versions. One likely reason is that the F1 for an /a/ vowel is well above a fundamental frequency of 587 Hz, even in the normal speech case. Thus, there is no need to modify the vocal tract shape for purposes of increasing F1. In fact, the sung version of the vowel has a slightly lower F1 than does the speech-like version. Perhaps the singer was attempting to tune F0 to F1 to increase the intensity.

As mentioned previously, image data for an /u/ vowel were also collected. The results were quite similar to that observed for the /i/. That is, the vocal tract shape was modified in the sung case such that the primary constriction for the /u/ was widened in order to move F1 upward and keep it higher than the fundamental frequency.

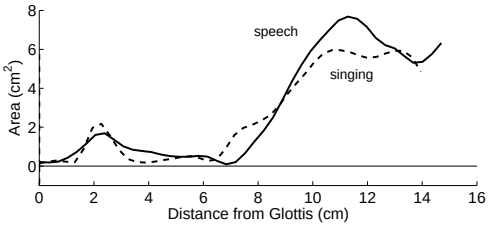


Figure 7: *Area functions for /a/ vowels spoken (dashed line) and sung (solid line) by a female singer.*

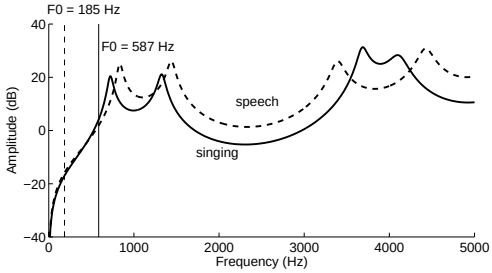


Figure 8: *Frequency response functions for the /a/ area functions shown in Fig. 5. The vertical lines indicate the fundamental frequencies for each vowel production. (Dashed = speech, Solid = sung.)*

4.1. Vowel space

To further explore vowel modification, vowel space plots were generated with a technique in which a principal components analysis is performed on a given set of area functions [16]. The components, called "modes", are used to construct thousands of new area functions that did not exist in the original set. For each new area function, an F1,F2 are calculated and plotted. Figure 9 shows two such vowel space plots. The large vowel space (dots) was generated by including all of the vowels across each of the four voice qualities (normal, twang, yawn, singing) for the female subject in [14],[15]. The smaller vowel space (solid) is based on only the sung vowels /i/, /a/, and /u/. A vertical line is drawn across the vowel space at a frequency of 587.33 Hz corresponding to the F0 used for the sung vowels.

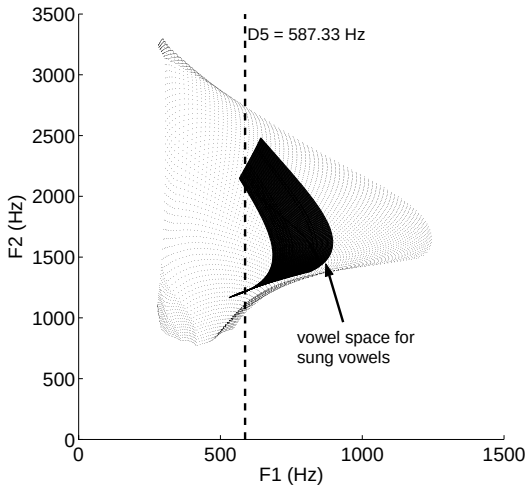


Figure 9: *Vowel space plots for a female speaker/singer based on vocal tract area functions from [14],[15].*

It is observed that nearly the entire vowel space for the sung vowels maintains an F1 that is above the fundamental frequency of 587 Hz. Thus, the vowel space roughly shows the possibilities for vowel production at this particular F0. However, if the F0 were increased, the vowel space would be expected to move to the right in order to keep F1 above F0 and would probably be further compressed. Conversely, the vowel space would be expected to move

to the left and expand for a decrease in F0. When the fundamental frequency is below the lowest F1 in the large vowel space (about 300 Hz), the entire space would be "available."

5. REFERENCES

[1] Ishizaka, K., Matsudaira, M., and Kaneko, T. (1976). Input acoustic-impedance measurement of the subglottal system, J. Acoust. Soc. Am., 60, 190–197.

[2] Cranen, B., and Boves, L. (1987). On subglottal formant analysis, J. Acoust. Soc. Am., 81, 734–746.

[3] Fant, G., Ishizaka, K., Lindquist, J., and Sundberg, J. (1972). Subglottal formants, R. Inst. Tech., Speech Transmission Lab., QPSR 1, 1–12.

[4] Titze, I.R. (1988). A framework for the study of vocal registers, J. Voice, 2(3), 183–194.

[5] Dang J., Honda K., Suzuki H. (1994). Morphological and acoustical analysis of the nasal and the paranasal cavities, J. Acoust. Soc. Am., 96(4), 2088-2100.

[6] Fant, G. (1960). *The Acoustic Theory of Speech Production*, Mouton, The Hague.

[7] Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W., (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels, J. Acoust. Soc. Am., 90, 799-828.

[8] Narayanan, S. S., Alwan, A. A., and Haker, K., (1995). An articulatory study of fricative consonants using magnetic resonance imaging, J. Acoust. Soc. Am., 98(3), 1325-1347.

[9] Story, B. H., Titze, I. R., and Hoffman, E. A., (1996). Vocal tract area functions from magnetic resonance imaging, J. Acoust. Soc. Am., 100(1), 537-554.

[10] Story, B.H., Titze, I.R., and Hoffman, E.A., (1998). Vocal tract area functions for an adult female speaker based on volumetric imaging, J. Acoust. Soc. Am., 104(1), 471-487.

[11] Sundberg, J. (1974). Articulatory interpretation of the singing formant, J. Acoust. Soc. Am., 55, 838–844.

[12] Johansson, C., Sundberg, J., and Wilbrand, H. (1985). X-ray study of articulation and formant frequencies in two female singers. In A. Askenfelt, S. Felicetti, E. Jansson, & J. Sundberg (Eds), SMAC 83, Proceedings of the Stockholm Intl. Music Acoust. Conf., Vol. 1, Stockholm, Royal Swedish Acad. Music., 46, 203–218.

[13] Sundberg, J. (1987. *The Science of the Singing Voice*, Northern Illinois University Press, Dekalb, IL.

[14] Story, B.H., Titze, I.R., and Hoffman, E.A., (2001). The relationship of vocal tract shape to three voice qualities, J. Acoust. Soc. Am., 109, 1651–1667.

[15] Story, B.H., Titze, I.R., and Hoffman, E.A., (1998). Contributions of vocal tract shape to voice quality, J. Acoust. Soc. Am., 104(3) pt. 2, 3aSC7, 1805, Norfolk, VA.

[16] Story, B.H., and Titze, I.R., (1998). Parameterization of vocal tract area functions by empirical orthogonal modes, J. Phonetics, 26(3), 223-260.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

SAMPLE-BASED SINGING VOICE SYNTHESIZER BY SPECTRAL CONCATENATION

Jordi Bonada, Alex Loscos
Music Technology Group, Audiovisual Institute, Universitat Pompeu Fabra
Barcelona, Spain
{jordi.bonada, alex.loscos}@iua.upf.es
http://www.iua.upf.es/mtg

Hideki Kenmochi
Advanced System Development Center, YAMAHA corporation
Hamamatsu, Japan
kenmochi@beat.yamaha.co.jp

ABSTRACT

The singing synthesis system we present generates a performance of an artificial singer out of the musical score and the phonetic transcription of a song using a frame-based frequency domain technique. This performance mimics the real singing of a singer that has been previously recorded, analyzed and stored in a database. To synthesize such performance the systems concatenates a set of elemental synthesis units. These units are obtained by transposing and time-scaling the database samples. The concatenation of these transformed samples is performed by spreading out the spectral shape and phase discontinuities of the boundaries along a set of transition frames that surround the joint frames. The expression of the singing is applied through a Voice Model built up on top of a Spectral Peak Processing (SPP) technique.

1. INTRODUCTION

The voice is considered to be the most flexible, rich and powerful of all instruments and it has always been an inexorable source of inspiration for musicians, sound designers and audio programmers. Music computer people have been interested in vocal sound manipulation and synthesis ever since the appearing of the very first techniques that allowed such sort of processing. Nowadays this interest has spread to most of the music communities reaching a point in which it is hard to find a popular musical production without any vocal alienation.

The synthesis of voice has been approached from many different directions and though it is a slack categorization, we can split synthesis models into two groups: spectral models, which are based on the perception of the sound, and physical models, which are based on the production of the sound.

Both spectral and physical models have their own benefits and drawbacks [1]. Physical models such as acoustic tube models use the control parameters humans use to control their own vocal system and can generate time-varying behaviors by the model itself. On the other hand some parameters might not have intuitive significance and might interact in non-obvious ways. Spectral models such as those based on phase vocoders or sinusoidal tracks have precise analysis / synthesis methods and work close to the human perception but their parameterization is not that suitable for study, manipulation, or composition since they don't match any auditory system structure.

Models such as the one used in formant synthesizers are considered to be pseudo-physical models because even though these are mainly spectral models they make use of the source / filter decomposition. The singing synthesizer we present would be part of this model group.

2. SYSTEM DESCRIPTION

The system is composed of two modules: the expression module, which is in charge of giving expressiveness and naturalness to the input melody, and the synthesis module, which is in charge of the synthesis process itself.

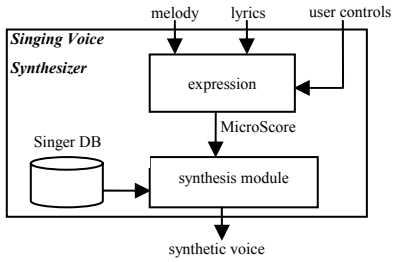


Figure 1: *General system diagram*

The indispensable inputs of the expression module are the lyrics and the musical score of the voice melody, that is to say, the phonetics and the fundamental frequency, dynamics and quantized duration values of each note. Out of these, the synthesizer can generate a virtual performance with default expression parameters. However, user control parameters can also be inputted to modify the expression through specifications on different types of note attacks, note transitions, vibratos, etcetera. The output of the expression module is a detailed musical score that characterizes the expressivity of the virtual singer performance through an ordered list of temporal events, each of which describes the local expression of the performance through the control parameters. This score is called Microscore.

The inputs of the synthesis module are the Microscore and the singer database. The synthesis module reads the Microscore information and synthesizes the virtual performance by taking the corresponding samples from the database, transforming them according to the inputted score and concatenating them [2].

3. VOICE AND SPECTRAL MODELING

In this section we will introduce both EpR Voice Model and Spectral Peak Processing (SPP) technique on which the system is based.

3.1. Voice Model (EpR)

Our Voice Model is based on an extension of the well known source/filter approach [3] we call EpR (Excitation plus Resonances). The EpR filter can be decomposed in two cascade filters. The first of them models the differentiated glottal pulse frequency response and the second the vocal tract (resonance filter).

The EpR Source filter

The EpR source is modeled as a frequency domain curve and one source resonance. The curve is defined by a gain and an exponential decay as follows:

$$Source_{dB} = Gain_{dB} + SlopeDepth_{dB}\left(e^{Slope \cdot f} - 1\right)$$

It is obtained from an approximation to the harmonic spectral shape (*HSS*) determined by the harmonics identified in the SMS analysis

$$HSS\left(f\right) = envelope_{i=0..n-1}\left[f_i, 20\log\left(a_i\right)\right]$$

where i is the index of the harmonic, n is the number of harmonics, f_i and a_i are the frequency and amplitude of the i^{th} harmonic.

On top of the curve, we add a resonance in order to model the low frequency content of the spectrum below the first formant. This source resonance is modeled as a symmetric second order filter (based on the Klatt formant synthesizer [4]) with center frequency *F*, bandwidth *Bw* and linear amplitude *Amp*, which is relative to the source curve.

The EpR vocal tract filter

The vocal tract is modeled by a vector of resonances plus a differential spectral shape envelope. It can be understood as an approximation to the vocal tract filter. These filter resonances are modeled just like the source resonance, where the lower frequency resonances are somewhat equivalent to the vocal tract formants.

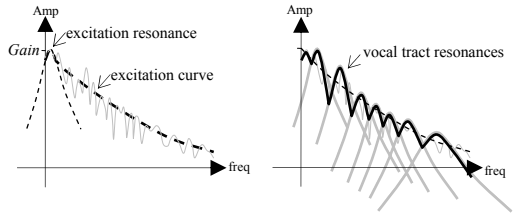


Figure 3: *The EpR model*

The differential spectral shape envelope actually stores the differences (in dB) between the ideal EpR model and the real harmonic spectral shape (*HSS*) of a singer's performance. We calculate it as a 30 Hz equidistant step envelope.

$$DSS\left(f\right) = envelope_{i=0..}\left[30i, HSS_{dB}(30i) - iEpR_{dB}(30i)\right] \qquad (1)$$

Thus, the original singer's spectrum can be obtained if no transformations are applied to the EpR model.

3.2. Spectral model (SPP)

Spectral Peak Processing (SPP) is used as the sample transformation tool. It performs a frame based spectral analysis of audio, giving as output the STFT, the harmonic peaks and the pitch. This technique is based on the phase-locked vocoder [5], but spectral peaks are calculated by parabolic interpolation [6] and a pitch analysis is performed. SPP considers the spectrum as a set of regions, each of which belongs to one harmonic peak and its surroundings. The goal of such technique is to preserve the local convolution of the analysis window after transposition and equalization transformations. A basic diagram can be seen in Figure 3.

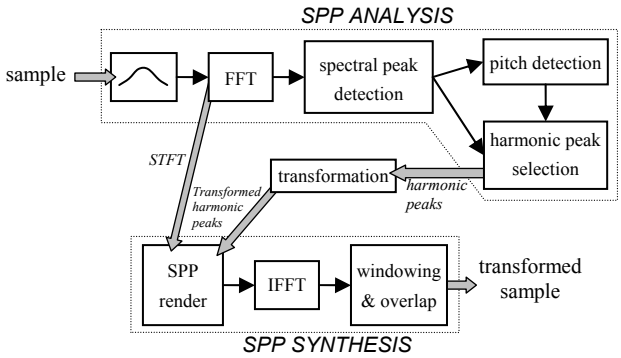


Figure 2: *SPP sample transformation*

4. SINGER DATABASE

The singer databases are created from dry and noiseless recordings of real singers. These databases hold both phonetic and expression information about the singer.

The phonetic database represents and stores the time varying phonetic characteristics of the singer so it stores the singer timbres in most representative phonetic contexts. It contains steady-states and articulations at different pitches.

The expression database characterizes the low level expression of the singer so it tries to retain the singer expressivity in different musical contexts. This database includes vibratos and note templates, which model attacks, releases and note transitions at different pitches. They are considered phoneme independent and organized into a musical meaningful classification [7].

5. TRANSFORMING SAMPLES

Basically, three kinds of transformation are applied to the samples: transposition, equalization and time-scaling. Time-scaling issues are described in detail in [8].

5.1. Transposition

Transposition means to multiply the harmonic's frequencies by a constant value. In terms of SPP, this operation can be done by shifting SPP regions in frequency. The linear frequency displacement for all the bins of each region will be the same. In most cases, this frequency displacement will be a non integer value. Thus we interpolate the spectrum with a 3rd order spline. This might not be the best method but it is a good compromise between quality and computational cost.

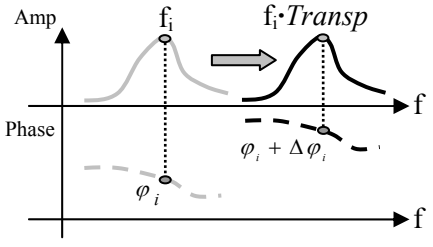


Figure 4: *Phase shift in SPP transposition*

In Figure 4, we can see an example of transposition to a higher pitch. In this case, the SPP regions will be separated in the resulting spectrum by gaps. Whenever transposing to a lower pitch, the SPP regions overlap in the resulting spectrum. This overlapping is computed by adding the complex values at each spectral bin.

Phase correction

When a SPP region is shifted in frequency in a frame by frame process, the phase needs to be corrected in order to continue the harmonics. For the i[th] harmonic, assuming linear frequency, the ideal phase increment between two consecutive frames is

$$\Delta\varphi_i = 2\pi f_i \Delta t$$

where the time increment between frames is Δt . If we transpose a sample by a factor *transp*, then the ideal phase increment should be

$$\Delta\varphi_i = 2\pi f_i \cdot transp \cdot \Delta t$$

Therefore, the amount of phase that should be added to the spectrum phase in order to continue the i[th] harmonic is

$$\Delta\varphi_i = 2\pi f_i \cdot transp \cdot \Delta t - 2\pi f_i \Delta t = 2\pi f_i \left(transp - 1\right)\Delta t$$

This phase increment is added to all the bins that belong to the i[th] region (see Figure 4). This way we can preserve the phase consistency across bins, i.e. the vertical phase coherence [5].

The phase increment applied to each harmonic is added to an accumulated phase frame by frame. This accumulated phase is the phase we finally add to each SPP region. However, this implementation results in the loss of the harmonic phase alignment after several frames because the frequencies do not follow a perfect harmonic scale. This loss of phase alignment produces some phasiness and loss of presence in the synthesized voice, especially for low pitches. To solve such problem, we consider the voice as perfectly harmonic. In that case, the equation of the phase increment will be

$$\Delta\varphi_i = 2\pi \cdot pitch \cdot \left(i+1\right)\left(transp - 1\right)\Delta t \qquad (2)$$

where *i* is the index of the harmonic (*0* for the fundamental) and *pitch* is the estimated fundamental frequency (different from f_0).

5.2. Equalization

Equalization means timbre change. On one hand the SPP analysis outputs a spectrum with the harmonic peaks and the SPP regions. On the other hand we have an envelope that defines the desired timbre. The SPP equalization is done by calculating for each region the amplitude amount needed to match the timbre envelope, and adding it to all the bins that belong to the region. Therefore, the spectrum amplitude of a region will be just shifted up or down and the phase will not be changed.

6. CONCATENATING SAMPLES

When connecting transformed samples there will appear spectral shape and phase discontinuities. In order to minimize them, phase and amplitude corrections can be spread out along a set of transition frames that surround the joint frame.

6.1. Phase concatenation

In order to avoid phase discontinuities at the segment boundaries, we have come out with a phase continuity condition that takes care of the boundary phases and the possibly different

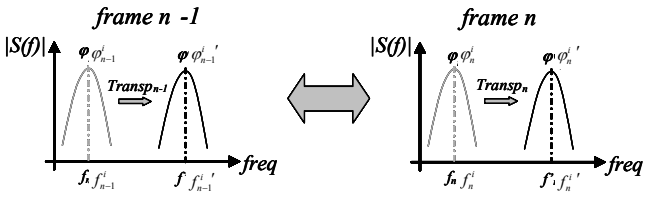


Figure 5: *Concatenation boundary*

transposition factors applied to each segment. In Figure 5 we can see the joint frames, where *frame n-1* is the last frame of the left segment and *frame n* is the first frame of the right segment. f_{n-1}^i and f_n^i refer to the frequencies of the *i*[th] harmonic.

The basic condition for phase continuity comes from the assumption that the frequency of each harmonic varies linearly between these two consecutive frames. In that case the phase relation between both frames should be

$$\varphi_n^{i\,\prime} = \varphi_{n-1}^{i\,\prime} + 2\pi \frac{f_{n-1}^{i\,\prime} + f_n^{i\,\prime}}{2}\Delta t$$

where $\varphi_n^{i\,\prime}$ and $\varphi_{n-1}^{i\,\prime}$ are the phases of the *i*[th] harmonic at the right and left frame respectively. Thus, the desired phase for the left frame $\phi_{n-1}^{i\,\prime}$ should be

$$\phi_{n-1}^{i\,\prime} = \varphi_n^{i\,\prime} - 2\pi \frac{f_{n-1}^{i\,\prime} + f_{n-1}^{i\,\prime}}{2}\Delta t$$

But in fact we are not just concatenating two segments, but also transposing them with different transposition factors ($transp_{n-1}$ and $transp_n$). We should distinct then between the original frequency of each segment (f_{n-1}^i and f_n^i) and the transposed ones ($f_{n-1}^{i\,\prime}$ and $f_n^{i\,\prime}$), where $f_{n-1}^{i\,\prime} = f_{n-1}^i \cdot transp_{n-1}$ and $f_n^{i\,\prime} = f_n^i \cdot transp_n$. The basic condition should be applied to the frequencies and phases before transposition. This can be expressed as

$$\phi_{n-1}^i = \varphi_n^i - 2\pi \cdot \frac{f_{n-1}^i \cdot transp_{n-1} + f_n^i \cdot transp_n}{2} \cdot \Delta t + 2\pi\left(i+1\right)\Delta t \cdot pitch_n \cdot \left(transp_n - 1\right)$$

by using the transposition phase correction equation (2) and taking into account that the phase correction due to transposition is accumulated frame by frame. Finally, we can calculate the phase correction $\Delta\varphi_c^i$ that guarantees the phase continuation of the *i*[th] harmonic as the difference between the phase at the left frame and the desired phase.

$$\Delta\varphi_c^i = \phi_{n-1}^i - \varphi_{n-1}^i \qquad (3)$$

This correction is applied either to the left or to the right segment around the boundary, and spread along several frames in order to get a smooth transition. In Figure 6 we can see an example where this phase correction is spread along 5 frames on the left part of the boundary.

Since the impulsive periods of left ant right segments are not aligned, we will often have big phase correction values. Therefore it's better if we calculate how much we should move the right segment (Δt_{sync}) in order to align the beginning of both periods, so that the phase correction to be applied is minimized. We could approximate this time shifting by assuming that the

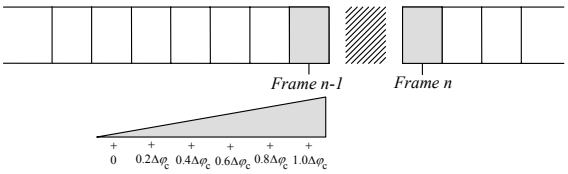


Figure 6: *Phase spreading in concatenation*

beginning of the periods will be aligned if the phase of the fundamental is continued. This can be expressed as

$$\Delta t_{sync} = \frac{-\Delta\varphi_c^0}{2\pi\, pitch_n \cdot transp_n} \qquad (4)$$

where $\Delta\varphi_c^0$ is calculated as in equation (3) for the particular case of i=0 (the fundamental). Finally, if we combine equations (3) and (4) we can obtain the desired phase at left frame as

$$\phi_{n-1}^i = \varphi_n^i + 2\pi\,(i+1)\, pitch_n \cdot transp_n \cdot \Delta t_{sync} - 2\pi \cdot \frac{f_{n-1}^i \cdot transp_{n-1} + f_n^i \cdot transp_n}{2} \cdot \Delta t + 2\pi\,(i+1)\,\Delta t \cdot pitch_n \cdot (transp_n - 1)$$

6.2. Spectral shape concatenation

In order to avoid spectral shape discontinuities at the segment boundaries we make use of the EpR model. First we estimate the EpR of the boundary frames. The left EpR then is stretched using the resonance frequencies as mapping points ($F0_{left}$ *to* $F0_{right}$, $F1_{left}$ *to* $F1_{right}$,…) and it is subtracted from the right EpR. The differential envelope obtained accounts for the spectral shape differences between the two joint frames.

For each one of the transition frames at the left of the boundary, a frequency mapping function is obtained from the interpolation between the above resonance frequency mapping and the identity mapping (y=x) with a factor *1-SSIntp* which stands for the distance from the frame to the boundary (*SSIntp* is 0 at the beginning of the interpolation zone and 1 at the boundary). This mapping is applied to each transition frame spectral shape and finally the differential envelope (weighted by *SSIntp*) is added to it (see Figure 7).

Notice that the spectral shape interpolation is spread along several frames in a similar way to the phase concatenation, but with the addition of the spectrum stretching.

7. CONCLUSIONS

The singing synthesizer we have presented in this paper has proven to comprise suitable methods and techniques for the generation of synthetic vocal tracks.

The SPP has brought a significant improvement in comparison to our previous analysis / synthesis technique [7] based on sinusoidal plus residual decomposition [6]. SPP avoids the tricky sinusoids / noise decomposition and preserves accurately and consistently the voice quality of the original singer. The preservation of the voice quality is a very valued feature when it comes to create databases of singers with vocal disorders such as hoarseness or breathiness.

Even tough the system is in a stage in which a synthetic voice is distinguishable from a real voice the quality is good enough to use it for backing vocals. Anyway we believe we will not have to wait that long to hear synthesized lead vocals that sound indistinguishable from a human singer.

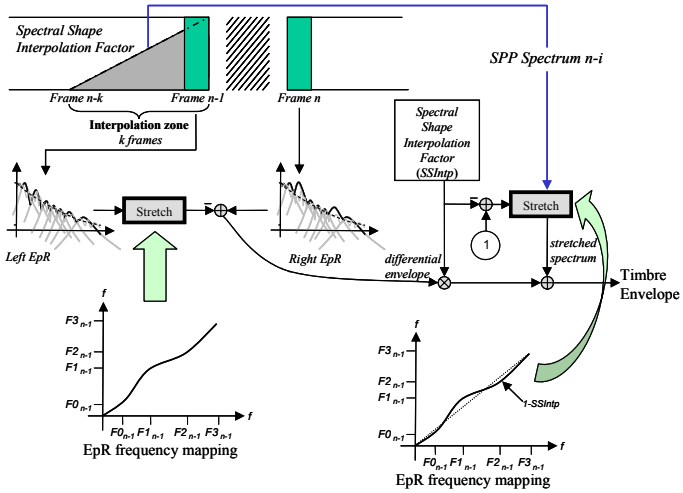


Figure 7: *The sample transformation*

8. REFERENCES

[1] Cook, R., *Toward the Perfect Audio Morph? Singing Voice Synthesis and Processing*, Proceedings of the Digital Audio Effects Workshop DAFX, Barcelona, 1998.

[2] Bonada, J., Loscos, A., Cano, P., and Serra, X, *Spectral Approach to the Modeling of the Singing Voice*, Proceedings of the 111th AES Convention, New York, USA, 2001.

[3] Childers, D. G., *Measuring and Modeling Vocal Source-Tract Interaction*, IEEE Transactions on Biomedical Engineering, 1994.

[4] Klatt, D. H., *Software for a cascade/parallel formant synthesizer*, Journal of the Acoustical Society of America, pp. 971-995, 1980.

[5] Laroche, J. and Dolson, M., *Improved phase-vocoder. Time-Scale Modification of Audio*, IEEE Transactions on Speech and Audio Processing, vol. 7, no. 3, pp. 323-332, May 1999.

[6] Serra, X, *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*., PhD thesis, CCRMA, Dept. of Music, Standford University, 1989.

[7] Bonada, J., Celma, O., Loscos, A., Ortolà, J., and Serra, X., *Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models*, Proceedings of the International Computer Music Conference, Havana, Cuba, 2001.

[8] Bonada, J., *Automatic Technique in Frequency Domain for Near-Lossless Time-Scale Modification of Audio*, Proceedings of the International Computer Music Conference, Berlin, Germany, 2000.

THE ACOUSTICAL EFFECTS OF THE CORE PRINCIPLES OF THE *BEL CANTO* METHOD ON CHORAL SINGING

Laurier Fagnan, Xavier Rodet

Faculté Saint-Jean, University of Alberta, Canada; IRCAM, Paris
lfagnan@telus.net; Xavier.Rodet@ircam.fr

ABSTRACT

The vocal principles of the *bel canto* method of singing have been endorsed for centuries for solo singing but rarely for ensembles. The study undertaken at Ircam has shed light on benefits they hold for choral singing.

A variety of choral ensembles was chosen. Each ensemble's own conductor initially led the ensemble in a series of exercises. The choir then worked on learning the *bel canto* principles and applying them to the same exercises. Finally, the acoustical comparison of these two "before and after" versions was undertaken.

The validity of Manuel Garcia's theory on *le coup de glotte* and how it affects vocal vibration and spectral energy was examined. Substantial increases in energy in the 2.5 to 4kHz band have been consistently observed. This principle's influence on precision of attack and release was also studied. It has been observed that allowing the vocal folds to begin their vibratory cycle from the closed (adducted) position improves the exactitude of pitch at onset. The paper also explores how the maxim of *chiaroscuro* (light-dark) resonance acts upon the intrinsic pitch of vowels and the effect on typical articulatory *F0* perturbation patterns. Finally, we tested how the maintenance of *bel canto* vibratory and resonance energy affected the acoustic energy and pitch of ensembles during diminuendos and soft singing. Surprising improvements in both upper harmonic energy and intonation were observed.

1. INTRODUCTION

It has not been the aim of this study to compare choral singing technique with solo singing technique, nor to have choristers sing as though they were soloists. The goal has rather been to have choirs implement the time-proven principles of the *bel canto* method while respecting the norms of choral practice, i.e. blend, group intonation, etc., and to subsequently test the acoustic impact that these principles exercise on choral singing.

1.1. Vocal Vibration

The first analysis demonstrates how the *coup de glotte* proposed by Manuel Garcia can have a positive effect on the regularity of vibration as well as the stability of the fundamental frequency upon onset in choral singing (fig.1.1). This principle has unfortunately been one of the most widely misunderstood of all the *bel canto* vocal tenets, in large part because of the use of the term *coup* (stroke, hit). The original meaning of this term simply implied that the initial vibration of the vocal folds was to begin from a closed (adducted), not an open, position.[1] To implement this principle, choristers are encouraged to lightly close the glottis at the end of inspiration, just prior to the attack which is to begin in a controlled fashion from this closed position. As phonation continues, they are to maintain a feeling of consistent and concentrated vibration, the theory being that properly-toned vocal folds which do not allow superfluous breath to spread them unnecessarily lead to a more stable vocal tone.

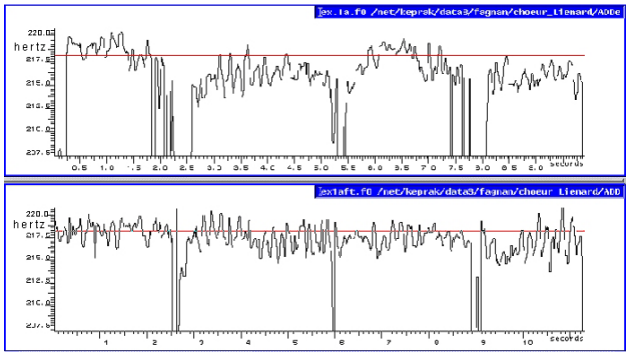


Figure 1.1: *Choir singing [i], [a], [i], [a] to A3 (220Hz). 'x' axis=time, 'y'=frequency in Hertz. Top window is before the ensemble has worked the 'closure' principle, bottom is after.*

As is evidenced in this example, the choir is more capable of successfully beginning a voiced sound on a targeted pitch after mastering the *coup de glotte*, and of accurately sustaining that pitch for the duration of a sung sound.

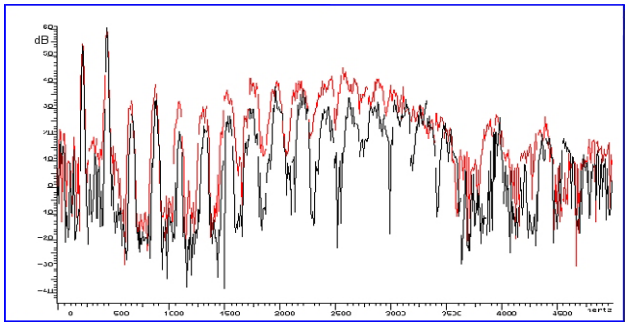


Figure 1.2: *FFT of Male chorus singing an [i] portion from figure 1.1. Spectrum in red is after the application of the coup de glotte, black is before.*

The increase in spectral energy which results from a more complete and efficient source vibration is immediately noticeable. Of interest in this analysis (fig.1.2) is the fact that the first two partials are equal in amplitude, hence the choir sang at the same volume in both versions. On the other hand, starting from this point and especially in formant regions, there is a

remarkable increase in the spectral energy of the more vibrant version. It is in the area of the singer's formant (2.3-3.5kHz), a spectral band in which our hearing is the most sensitive, that the difference between the two examples is most remarkable. This suggests that increased acoustic efficiency is available to a choir which applies the principle of controlled and regular vibration espoused by the *bel canto* masters.

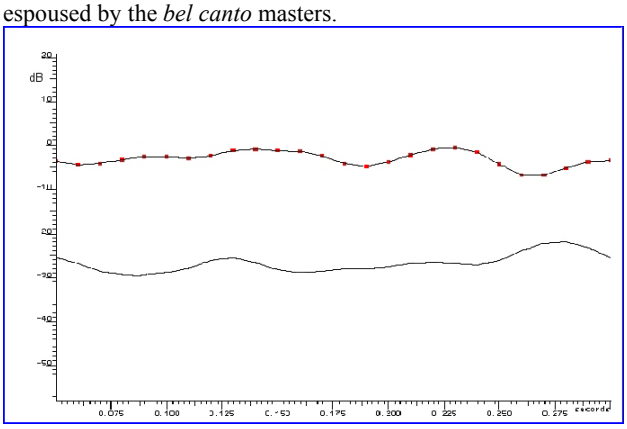


Figure 1.3: *Comparison of energy in the spectral band of 2.3 to 3.5kHz (singer's formant region). Y axis is amplitude in dB, x axis is time. 11dB gain in the 'after' version in red.*

Much has been written concerning the importance of harmonic energy in the singer's formant region and on its influence on vocal timbre as well as vowel intelligibility. There is also much documentation on the lack of energy in this range in choral singing. The ability to effectively extract and measure the spectral energy in this band using the *EnerBande* program was very useful to this study. It enabled us to objectively quantify the energy in this region both before and after *bel canto* principles were implemented and assisted in critically assessing much of what has been written on the vocal technique of choirs and, in particular, on how proper vocal vibration can help to provide adequate energy in this band. This added energy will be of special interest to choirs that sing with orchestra as it has been shown that orchestral sound is not as strong in this region and that voices possessing strength in these upper frequencies can more easily be heard and clearly discerned over an orchestra's sound.

1.2 *Chiaroscuro* ("bright-dark") Resonance Balancing

It is a well-known acoustical fact that all vowels were not created equal. Individual vowels do not resonate in quite the same way, or at the same frequency. To effect a vowel change, the articulators must change position and the vocal tract must alter its shape. These movements bring with them a plethora of muscular adjustments which often have a pejorative effect on the stability of the fundamental frequency. This phenomenon has often been observed and commented upon by vocal acousticians under the headings of "intrinsic pitch of vowels" and "articulatory F0 disturbances." For example, several experiments have shown that the change from [i] (heed) to [E] (head), is often accompanied by a rather significant drop in fundamental frequency. There are various theories which attempt to explain this correlation, one of them being the "tongue-pull" hypothesis which suggests that the higher tongue position required for frontal vowels might also cause a longitudinal stretching of the vocalis muscles which would tend to raise the fundamental frequency. [2] Whatever the precise reason(s) might be, the relationship between changes in resonance and alterations in vibration of the vocal cords (frequency) seems to be quite strong, as has been shown by Ternström and Sundberg.[1]

Our research indicates, however, that the pejorative influence that necessary articulatory changes exercise on vocal vibration can be reduced, and in some cases eliminated, when the *bel canto* principle of *chiaroscuro* resonance is observed. Though certain vowels are incontestably more *chiaro* (bright) or *oscuro* (dark) than others by nature, this principle seeks to establish an equal proportion of brightness and darkness across the complete vowel spectrum. Concentrating on equalizing the consistency of this "bright-dark" quality seems to even out the upper harmonic activity among the different vowels. It also helps to remove many of the extraneous and unnecessary articulatory modifications made when changing vowels, thereby providing a more stable environment for the vibrating vocal folds. These elements of *chiaroscuro* resonance tend to play an important role in reversing the pejorative effect that articulatory changes can have on fundamental frequency, as demonstrated by the following analyses.

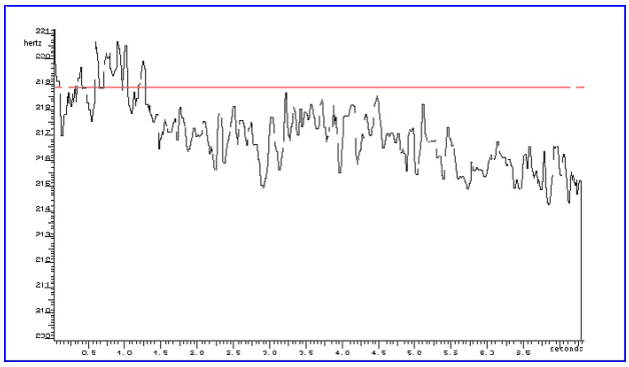


Figure 1.4: *Choir singing the vowel series [i]-[E]-[y]-[a] at approximately 1.5 second intervals along the 'x' axis,* ***before*** *chiaroscuro (compare with figure 1.5).*

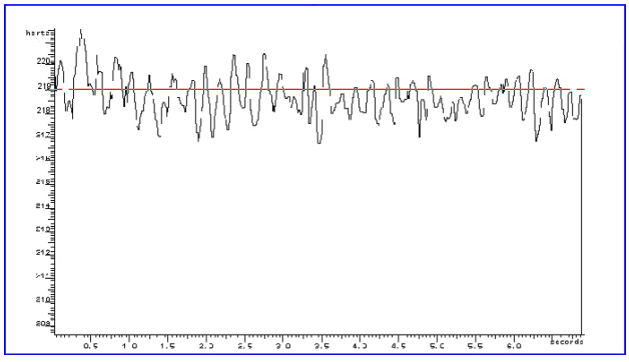


Figure 1.5: *Same choir as in figure 1.4 singing the same vowel series* ***after*** *having worked the principle of chiaroscuro resonance balancing.*

In the exercise leading to figure 1.5, choristers were asked to maintain, as much as possible, an equal amount of brightness and darkness in all of the four vowels. They experimented with the ability to maintain this consistency in vocal timbre while allowing the articulators to make the necessary movements to realize the different vowel sounds. Their attention was in no way drawn to the fact that there was an *intonation* problem, but rather a *resonance* problem. After having worked on the principle of *chiaroscuro* resonance for approximately fifteen minutes, it is evident that the articulatory perturbations which were exercising a negative influence on fundamental frequency stability in figure 2.1 are all but gone in 2.2. The drop in pitch from the first to the last vowel was of 3.22Hz in figure 2.1 and is reduced to 0.29Hz in figure 2.2. These findings were substantiated in multiple experiments with choirs of different size and ability.

1.3 *La nota mentale* (the mental note)

This *bel canto* vocal tenet incorporates the previous two principles (efficient vocal vibration and *chiaroscuro* resonance) and adds to them the notion of effective preparation of mental concept of vocal tone during inhalation, before a sound is ever made. The name derives from the practice of imagining one is already singing during the inhalation process. The preparation of the proper environment in the torso which leads to effective breath management, or *appoggio*, is also to be incorporated into each inhalation.

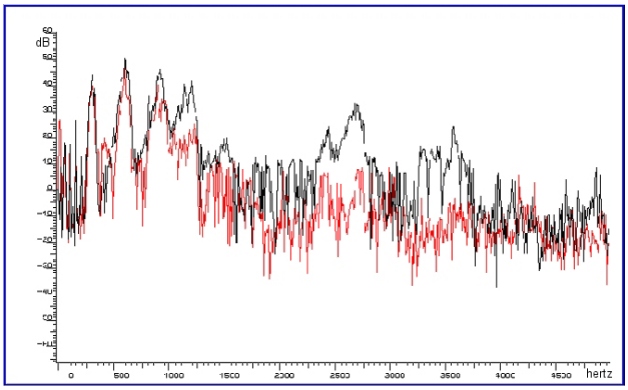


Figure 1.6: *Male chorus singing Eb4, first 0.5sec. of attack. Red is before nota* mentale *work, black after.*

The sound files which led to the Fourier Transform analysis in figure 1.6 are extracted from the very beginning of a phrase of the well-known American song "Shenandoah" ("**A**way you rolling river"). Although the amplitude of the fundamental frequency is the same in both versions, it is evident that there is very little upper harmonic activity in the 'before' (red) version. Conversely, with the implementation of the technique of *la nota mentale* (black spectrum), upper formant peaks are readily apparent within this very brief beginning (first 0.5sec. of the attack). Studies have shown that these upper formants enable the sound to carry in a large hall and add colour to the vocal timbre while increasing vowel intelligibility.

A closer examination of the evolution of harmonic energy in the singer's formant region for this very brief segment provides another important observation. At the very beginning of the window (figure 1.7), there is a difference of 9dB which separates the intensity of upper harmonic activity in the two versions. After 0.4 seconds, this gap has widened to 24dB, demonstrating that the singers' vocal preparation has very rapidly produced acoustical dividends.

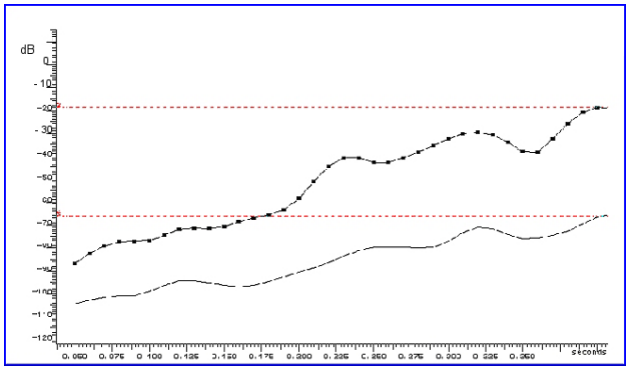


Figure 1.7: *A measure of the evolution of harmonic energy in the region of the singer's formant for the first 0.4 seconds of a phrase. Based on the same sound excerpt as figure 1.6. Dotted line is after nota mentale work.*

1.4 Messa di voce

Just as the principle of *la nota mentale* assures a consistency of vocal energy and accuracy of intonation at the beginning and end of phrases, that of *messa di voce* is designed to maintain this same consistency within the context of changing dynamics. Of special interest to the early *bel canto* masters was the art of performing the perfect *diminuendo,* one in which the voice effortlessly diminishes in volume to the softest point while still maintaining all of its colour and carrying power. It has often been remarked, in both the concert setting as well as in scholarly journals, that vocal ensembles tend to relinquish vocal intensity and stability as they diminish in volume.[3] The ideal proposed by the principle of *messa di voce* is that the singer must maintain exactly the same vibration along the full course of the *diminuendo* process without allowing any breath flow to compromise this vibration. This is easier to accomplish if one applies the breath management principle of *inalare la voce,* which advocates the maintenance of a feeling of openness in the body similar to that felt during inhalation. This process seeks to delay the collapsing action of the intercostal muscles which help to expel carbon dioxide from our lungs. With superfluous breath flow held in check and singers concentrating on maintaining consistent vocal vibration (which will aid in stabilizing the fundamental frequency), choristers have only to add to this the sustaining of complete *chiaroscuro* resonance to effect a successful *diminuendo.*

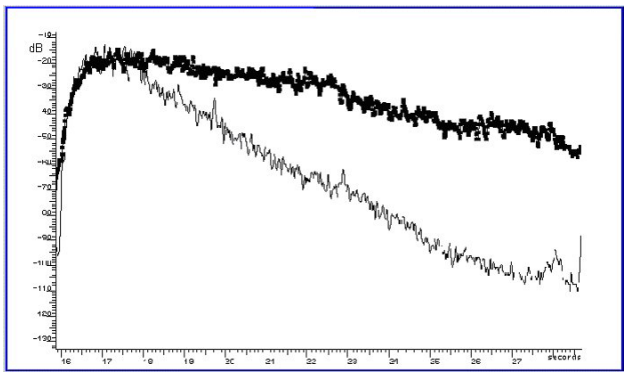


Figure 1.8: *A comparison in singer's formant energy during a decrescendo sung by a mixed choir. Grey=before work on messa di voce, black=after.*

It is evident in figure 1.8 that as time progresses and volume diminishes, the gap in the volume/upper harmonic energy ratio widens and the *bel canto* version displays its acoustic advantages. As in previous examples, one can note that the two versions begin at the same amplitude. However, the ensemble relinquishes vocal energy much more rapidly in diminishing volume when not employing techniques outlined in the principle of *messa di voce.*

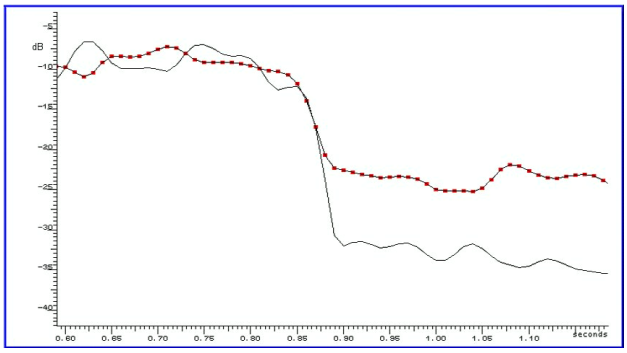


Figure 1.9: *Comparison of upper harmonic energy during abrupt shift from forte to pianissimo (at 0.9s). Dotted line=after bel canto work.*

In figure 1.9, the middle section of the diminuendo of figure 1.8 has been removed to further highlight the difference in upper harmonic energy between the *forte* and *piano* portions of the *diminuendo.* The preservation of an additional nine decibels of upper harmonic energy in the 'after' version while singing softly speaks to the overall effectiveness of this principle. In the sound samples which generated this comparative analysis, a noticeable drop in pitch and tone quality is easily discernable in the 'before' version.

2. CONCLUSION

Although the indiscriminate application of *bel canto* vocal principles can lead to occasional problems with blend, these techniques can be very helpful in enhancing choristers' vocal production when they are applied with discretion and properly worked into a choir's sound. Further, they can improve a choir's overall acoustical efficiency as an ensemble. The foregoing samples demonstrate that the acoustic benefits provided by the implementation of these principles within the choral context may far outweigh any disadvantages. It is clear that encouraging an ensemble to improve its ability to sing with consistent and collected vocal vibration will do much to improve the group's sound and harmonic spectrum. It has also been demonstrated that using consistent *chiaroscuro* resonance can have a very positive effect on group intonation while improving overall timbre. The application of *messa di voce* practices in softer dynamics has been established as a useful tool in assisting ensembles to sing with greater efficiency and a more consistent sound in softer dynamics. It is hoped that more choral conductors will look to these principles in the future when developing the vocal sound and technique of their ensembles.

3. REFERENCES

[1] Garcia, M., *Hints on Singing.* New York: Patelson's, 1982.

[2] Ternström, S., Sundberg, J. and Colldén, A., "Articulatory *F0* Perturbations and Auditory Feedback", *Journal of Speech and Hearing Research*, 31: 187-192, 1988.

[3] Ternström, S., "Physical and Acoustic Factors that Interact with the Singer to Produce Choral Sound." *Journal of Voice*, 5,2: 128-143, 1991.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

THE INFLUENCE OF THE PRACTICE OF BASSO CONTINUO ON THE INTONATION OF A PROFESSIONAL SINGER IN THE TIME OF MONTEVERDI

G Faraone[1], S Johansson[2], P Polotti[1,3]

[1]International Academy of Music, IRMus, Milano, Italy
[2]International Academy of Music, Department of Early Music, Milano, Italy
[3]Politecnico di Milano, Dipartimento di Elettronica e Informazione, Milano, Italy

ABSTRACT

The goal of this work is to investigate the influence of the practice of basso continuo, realized at the harpsichord according to the temperaments in use in the period of Monteverdi and its affect on the intonation of a professional singer. Our investigation focuses on the tuning relationship between the voice and the accompaniment. In one selected piece, several cadences, habitually adopted by the basso continuo players of the late 1500's and the early 1600's, according to the indications in the treatises, are analyzed. The considered cadence is realized in several different ways by the musicians and a spectral analysis is made to evaluate the tonal consonance between the voice and the accompaniment. The same material was presented to a group of expert listeners for a subjective evaluation. Finally, we compare the spectral measurement to both the subjective test results and the rules of the early treatises. Our idea is that spectral analysis can be a good measurement tool for the effectiveness of a cadence. The goals of such work are to see if the rules and suggestions of the treatises, aside from relating to musical aesthetics, can also have a foundation in acoustics and therefore furnish additional advise to the singer relative to correct intonation in singing.

1. INTRODUCTION

The main motivations of this work come out of the considerations that commonly are made by those who practice and listen to so-called "early music". What they affirm is that the very limited dynamics, which distinguish this gender of music, force the performer and listener not to undervalue acoustic phenomena and aspects such as beatings produced by the harmonics of chords in a harpsichord or the accurate intonation of a singer whose voice can be very exposed in such a repertory.
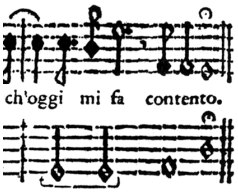


Figure 1: *Final cadence of "Rosa del ciel"*

The music of the 16th and early 17th centuries in Europe was performed in intimate spaces where there was a fine line between sound and silence. For this reason, we have recorded the sound samples in a typical room in a sixteenth century villa. In order to evaluate the influence of the practice of basso continuo on the intonation of a singer, we have chosen to examine, in one selected piece of this period, one particular cadence. A cadence is that melodic/harmonic progression associated with the ending of a section or a phrase. In a cadence we usually find, if "diminutions" are not adopted, sustained notes in the vocal line and a vertically clear sequence of chords built upon the bass line. Thus cadences prove to be ideal sections to examine the tuning relationship between the voice and the accompaniment. For this purpose we have selected the final cadence of the famous aria «Rosa del ciel» from «L'Orfeo favola in musica» by Claudio Monteverdi (Cremona 1567-Venezia 1643) performed for the first time in Mantova in 1607. We have recorded this cadence performed by a baritone and a harpsichordist in several different realizations, according to the practical examples gathered from the treatises and the repertoire of the late 1500's and the early 1600's. These recordings were submitted to 6 "expert listeners", all experienced instrumentalists skilled in the music of the time of Monteverdi for a subjective judgement on the tonal consonance between the two performers. The term "consonance" implies the ideas of pleasantness, smoothness and absence of beats [1]. This concept has evolved over the centuries [2]. We considered a spectral and pitch analysis of all the above-mentioned perceivable elements (beats and intonation), taking into account that in the 16th and 17th centuries musicians and listeners were endowed only with an unique spectrum analyser: the human ear. Finally, starting from harmonic thoughts on the realizations of continuo, referring to the early treatises, we have compared the subjective data with the respective spectra and detected pitches, searching for an experimental corroboration of the listening test results and an explanation of these perceptual sensations, such as roughness [3]. For an accurate analysis of the singer's intonation we resorted to a pitch detection based on the computation of the autocorrelation. The structure of the paper is the following. In Section 2 we provide some musicological remarks, in order to clarify the musical terminologies adopted in the paper. In Section 3 we present recorded material, technical equipment, recording methodology and the tuning adopted for the harpsichord. In Section 4 , analysing the harmony of the basso-

continuo realizations, we discuss the listening test results comparing them with the spectral data. In Section 5 we draw our conclusions and we give suggestions to the singer relative to correct intonation in singing this repertory.

2. MUSICOLOGICAL REMARKS

When the music performed "a cappella" by several voices migrates to the practice of solo voice accompanied by some kind of instrument, it is clear that we also have to take into consideration the temperament (the way of tuning the instrument). In the time of Monteverdi there were many different instruments used for accompanying the voice and they had all the same limits. Having only twelve notes per octave on the keyboard does not allow the use of certain tonalities, nor does it allow one to have both perfectly tuned thirds and pure fifths at the same time. One had to choose in which tonality to play and then tune the instrument according to what would be the optimal choice for that particular piece. Apparently this would not work on larger instruments such as church organs and therefore a kind of standard was established. The most evident testimonies we have regarding the temperaments from that time are the organs, whose pitch for obvious reasons was not manipulated that often. This brings us to the conclusion that quarter comma mean tone [4] was actually the most frequent used temperament at that time and therefore used in this paper. A second major element in the starting phase of this work was to consider the so called "contrapunto osservato" [5]. According to the "contrapunto osservato" we know that the octave and fifth are *perfect consonances* [6]. The third and the sixth are considered *imperfect consonances*, while the fourth and the seventh are considered the dissonances, which have to be prepared and resolved. As we will see in detail in the next section this was one of the main criteria for the choice and the organization of the different cadences, wich is the object of our investigation. In fact what we have realized in the sequence of cadences is a kind of a journey begining with the “purest” realization with only perfect consonances (octaves and fifths) and continuing with examples with imperfect consonances, dissonances and passing tones. As described in the following sections this leads to interesting results concerning the relation between the voice and the different types of realization of the basso continuo.

3. MATERIAL: 8 VERSIONS OF THE CADENCE

The example we considered is the final cadence of “Rosa del ciel” from Monteverdi’s “Orfeo” shown in Figure 1**.** In our experiment this cadence was played by a harpsichordist and sung by a baritone. We have utilized 8 different versions of the basso continuo, recording each version three times consecutively. We agreed on recording the whole of the material in a single session, simulating a typical rehearsal of today. We decided also to record all of the cadences played by the harpsichordist alone, in order to check out any occurrence of beats due to the tempered instrument. We have made use of a professional hard disk recorder, an unidirectional microphone for the singer and an omnidirectional microphone close to the harpsichord. The two signals were recorded on to two separate channels with a sampling rate of 44.1 kHz and a resolution of 16 bits. The italian style harpsichord was tempered in ¼ sintonic comma meantone.

4. EXPERIMENTAL RESULTS

In this section we provide an overview of the final cadence (in modern notation) with the 8 recorded realizations of the basso continuo. The Greek letters **α**, **β** and **γ** mark the most notable points in the singing voice where the group of listeners have found clashes in intonation. The letter **α** denotes the part of the sustained D right after the occurrence of the second chord in the bass line. The letter **β** marks the two consecutive Es after the rest. The letter **γ** denotes the sustained D in the final bar.The chords, marked by the rectangles with bold letters (**a**, **b**, **c**, **d**,…) in the score, indicate the points that we have taken into consideration for our tests.
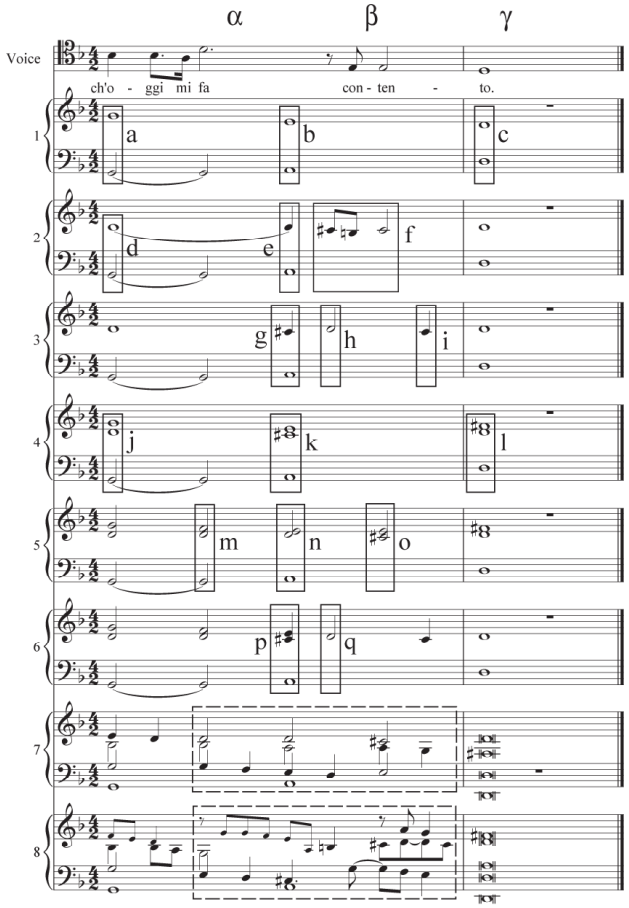


Figure 2: *the 8 versions of the cadence*

In the listening tests the 6 expert listeners were asked to judge the intonation of the baritone in each repetition of each cadence, focusing on the pitch of the sung notes and not on the realizations of basso continuo. The listeners were also called for detecting possible points of “roughness”, even in sections where the singer is not notably out of tune.

1. The keyboardist has chosen to realize perfect consonances (octave and fifth) in **a, b** and **c** according to the so-called "contrapunto osservato" in order to present the skeletal structure of the harmony. The expert listeners reported an irregular, not well tuned, vibrato in the sustained D (see **α**), right after the chord in **b**. They noticed that the consecutive two Es in **β** are not sung at the same pitch and they perceived a slight instability during the final D in **γ**. In **b** we have a narrow fifth, due to the mean tone temperament that produces beatings disturbing the vibrato of **α**. The term "*beating*" referred to in this work means, more generically, any perceivable *amplitude modulation* however generated. Thus the following two Es are not equal. Furthermore, the singer in **γ** needs to address one slow "*beating*", due to the instrument used for this recording, that will become double starting from **l.**

2. In order to obtain a dissonance (see **e**) on the beat one has to prepare it (see **d**) in the previous chord and resolve it (see **f**). In this cadence **α** is perceived as less steady. In fact, in **e** we find that the beatings are less disturbing, because no new chord interferes. Actually, the isolated note A of the basso continuo, by means of its partials, helps the singer to correctly tune the following E. Besides the ornamented resolution of the dissonance (see **f**) seems helpful to the singer.

3. An imperfect consonance (see **g**) going to the dissonance (as a leading tone) in **h** resolves to the first again in **i**. This is a very typical example of a final cadence and in fact in Italian it is

called "Cadenza Composta" which means composed cadence. Consisting of more than one element (not only the fourth going to the third) it concludes a phrase. Even in this case **α** is perceived by the expert listeners with an uneven vibrato. The dissonance between C# and D in **α** emphasizes this sensation of *roughness*. As expected **g** + **h** disturb **α** and **β** that both result in being out of tune. In **γ** now the listeners judge the singer's voice to be less stable, but the absence of roughness is the general sensation. Actually, as clearly shown in the sonograms, the baritone's voice is getting accustomed to the slow beating encountered for the first time in **c** and heard repeatedly until this moment. It seems that the singer tends to follow the beat by means of a slow modulation of amplitude in his voice as if he's trying to simulate the effect of beating, even after it is finished.

4 The perfect consonances of the first example are now presented with a fifth in **j** and a third in **k**. In all the three repetitions of this cadence **α** is newly perceived as rough and **β** is out of tune. The spectral analysis of the accompaniment reveals a rapid double beating producing a "wa-wa" effect, barely perceivable in the recording, which could have disturbed the singer. This double beating probably could have been heightened by the acoustical response of the room. In **l** the listeners noticed more roughness. Actually, the singer is now approaching a new pulsing rate originated by more beatings within the bar.

5. In the point corresponding to **d** of the second example we add an octave descending to the seventh in **m** and then to the fifth in **n** of the dominant that resolves to the third on the tonic in **o**. Now, in **n** the dissonance D against E disrupts the singer in **α** as well affecting **γ** too. Both of these effects are readily perceived by the listeners along with the above-mentioned unstableness of **l.** In point **n** the baritone naturally respondes by means of an unconscious fading in **α.**

6.This harmonic situation is the same as the third example but with the addition of the octave descending to the seventh and finally to the fifth in **p**. The expert listeners noticed a rough point in **α** and they judged the singer to be out of tune in **β** in all the three repetitions of this cadence. Harmonically speaking, we could expect these phenomena, because of the presence of a tempered D in **q** which destabilizes the singer just before the onset. The pitch analyses corroborate this musical insights.

7. - 8. These two cadences have been realized with an internal moving voice which produced the following results. The expert listeners judged all the repetitions of these consonances as pleasant (not rough). In addition, the singer's intonation in **α** and in **β** was judged to be in tune with respect to all the previous versions.

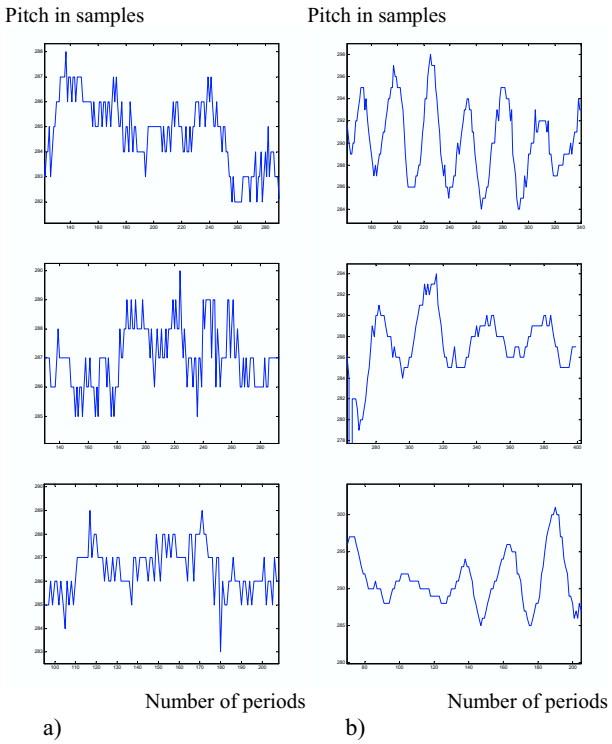


Figure 3: *pitch detection of point ß in the 3 repetitions of cadence 8 (a) and in the three repetition of cadence 4 (b), respectively.*

As an example, in Figure 3 a) we compare the pitch detection of point *ß* (the two consecutive Es) in all three repetitions of cadence 8 (8-1, 8-2, 8-3) compared with the pitch detection in Figure 3 b) of the same point in all three repetitions of cadence 4 (4-1, 4-2, 4-3). In table 1 we report the standard variations of the mean pitch of the two cadences in point **β**.

It is clearly evident how the pitch variations of cadence 4 are considerably greater than those in cadence 8. Listening to various realizations of the basso continuo with particular attention to the differences between very few notes and "full harmony", the expert listeners and the pitch analysis confirm that the "simpler" examples are less desirable than the more complex and ornamented ones.

Cadence	Mean pitch (in samples)	Standard deviation
8-1-**β**	284.8327	1.7639
8-2- **β**	286.8706	1.1718
8-3- **β**	286.5545	1.0342
4-1- **β**	290.2928	3.3145
4-2- **β**	288.0180	2.2278
4-3- **β**	290.9917	3.4046

Table 1: *Mean and standard deviation of the data coming out from the pitch detection shown in Figure 3*

The common judgement is that, with a reduced realization, the singer has to make a bigger effort to sing perfectly in tune. Playing the "full harmony" often creates dissonances with the singing part, which brings the harpsichordist to the point where he has to move the parts and ornament them in order to maintain the sound but still avoiding crucial interference with the singer's notes. According to these data, we can say that, with the exception of the points readily out of tune, whenever an expert listener individuates a 'roughness', this corresponds to a variation of the pitch of the voice or better said to an irregular vibrato. Our idea is that this vibrato arises as a natural and unconscious response of the singer to the amplitude modulation due to the beatings of the chords of the harpsichord itself. By using moving parts in the realization of basso continuo, the harpsichordist prevents the singer from perceiving the beatings which could destabilize the intontation of the voice. This practice corresponds with the suggestions of the early treatises concerning the necessity of moving the parts. These considerations cast light on a practice which, aside from relating to musical aesthetics, could also have a foundation in acoustics.

5. CONCLUSIONS

In this work we have analyzed the final cadence of the aria "Rosa del ciel" from Monteverdi's "Orfeo", performed adopting 8 different realizations of its basso continuo. We asked a group of expert listeners to detect any clash and/or roughness, in order to evaluate the tonal consonance between the voice and the accompaniment. Afterwards, we compared the listening test results with the results of spectral analysis and pitch detections. In this way we were able to detect how the intonation of a voice reacts to different situations in the accompaniment, i.e. to different versions of a single cadence. The data coming from the spectral analysis and the pitch detection confirm both the listening test results performed by the group of expert listeners and the suggestions given by the treatises of the age of Monteverdi. By a simple spectral and pitch analysis. (even performed by his/her ears after a proper training!) a singer can be fully aware of how the continuo 'realization' can influence her/his intonation, because of the possible occurrence of beatings originated by the chords of a tempered harpsichord. Thus we suggest the singer, in preparing the music, check out the proposed realizations of continuo, together with the harpsichordist, by means of a careful listening. In doing so they could be able to locate certain points where beatings occur. They subsequently can decide to move the parts within the selected harmony in order to reduce their effect as much as possible. This work is just a beginning of a wider project that we'll carry on in the future. Therefore this investigation is not intended to be exhaustive with regard to all aspects that have been here considered. The conclusions we have just drawn need to be enforced by means of a broader research that ought to involve more singers and more examples of cadences from the repertoire. Selected audio materials and resources referred to this work along with further investigation will be available at the following websites:

http://space.tin.it/scienza/faraone.giuseppe/SMAC03/index.html

http://www.voicescience.org

6. AKNOWLEDGMENTS

We are grateful to Gabriele Pola for his invaluable help during the recordings, as well as Giovanni Togni for the musicological advice, and to Christa Patton for the proofreading.

7. REFERENCES

[1] Sundberg, J., The science of musical sounds, Academic Press, Inc., 1991.

[2] Tenney, J., A Hystory of "Consonance" and "Dissonance", Excelsior Music Pub. Co., NY, 1988.

[3] Helmholtz, H. von, On the sensations of tones as a physiological basis for the theory of music (1877), Dover New York, 1954.

[4] Aaron, Pietro, Thoscanello de la musica. Impressa per Bernardino et Mattheo de Vitali, Venezia, 1523, 1529, 1539, 1562.

[5] Diruta, Girolamo, Il Transilvano, appresso Giacomo Vincenti, Venezia, 1593, 1609.

[6] Zarlino, Gioseffo, Istitutioni harmoniche. Pietro da Fino, Venezia, 1558, 1562, 1573, 1589.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

ACOUSTIC PULSE REFLECTOMETRY FOR VOCAL TRACT MEASUREMENT

C.D. Gray, D.M. Campbell, C.A. Greated

Acoustics and Fluid Dynamics Group
University of Edinburgh
cgray@ph.ed.ac.uk

ABSTRACT

As a musical instrument, the human vocal tract can be reconfigured in a variety of shapes through the use of articulators forming cavities which filter the complex glottal waveform into what we perceive as voice sounds.

In model-based voice synthesis, models of the vocal tract are often based on geometric data obtained using MRI (Magnetic Resonance Imaging) scans from which a piecewise cylindrical section model can be devised. The disadvantages associated with using MRI for this purpose include the extended period for which the subject must remain immobile during scanning, the post processing of images and the limit on access to such expensive equipment.

As an alternative to MRI, an experimental attempt using acoustic pulse reflectometry for vocal tract measurement is described here.

The time domain technique of acoustic pulse reflectometry is used to measure the input impulse response of a tubular object, such as the vocal tract, from which both the bore profile and input impedance can be calculated. Recent developments have allowed a combined measurement and reconstruction time of the order of 1 second.

Results from modelling the bore profile of the vocal tract during vowel production are presented, and applications for this technique in voice analysis and synthesis systems and in clinical usage are discussed.

1. INTRODUCTION

The measurement of human airway dimensions has been a frequent objective in the field of speech research. Early work was carried out in the frequency domain, but in 1971, Sondhi et al. [1,2] began using time-domain reflectometry at Bell Telephone Laboratories. Sondhi was first to suggest that the reflections of sound within tubular acoustic structures (arising at changes of cross-sectional area) could be used to determine the internal area profile of the mouth cavity.

Experiments were performed on vocal tracts and lungs by Jackson et al. [3]. A spark was used as a sound source, with the resulting pulse travelling along a section of cylindrical tubing called a source tube, then into the airway under test. The resulting reflections were measured by a microphone in the source tube wall. The source tube was present to physically separate the in-going pulse signal from the impulse response. Measurements on human patients were performed by Fredberg et al. [4] and clinical use followed [5]. e.g. In Sleep Centres to detect areas of constriction (along axial distance from mouth to glottis) in apnoea patients.

2. THE PRINCIPLE OF ACOUSTIC REFLECTOMETRY

The theory of acoustic pulse reflectometry (including bore reconstruction) has been reviewed in detail by [6,7,8] and will not be presented here.

The principle of the technique is appealingly simple [9]. A schematic diagram of the apparatus used is shown in figure 1.

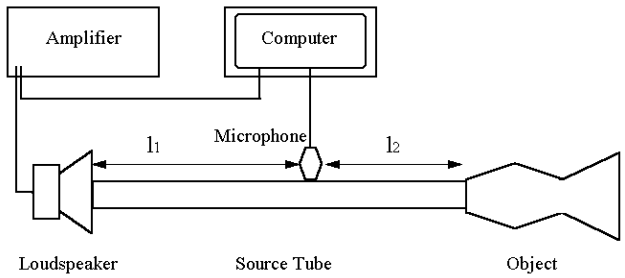


Figure 1: Schematic diagram of pulse reflectometer

An acoustic pulse is emitted from a loudspeaker and passes down a source tube into the object being measured. A microphone embedded in the wall of the source tube records the impulse and the reflection from the object, and this signal is sampled by a PC. Appropriate analysis of the reflected waveform leads to a reconstruction of the impedance (and hence the cross-sectional area) profile of the object.

2.1 Maximum Length Sequences [MLS]

The most prevalent method used for measuring the input impulse response involves injecting a pulse into the source tube and measuring the reflections. By using a signal that continues over a longer time interval we may put more energy into the system, improving the signal to noise ratio and removing the need to average over large number of measurements (necessary in previous reflectometers).

One solution is to use a pseudo-random binary signal called a maximum length sequence (MLS). An MLS signal consists of an apparently random sequence of 0's and 1's that has a flat frequency spectrum for all frequencies up to the Nyquist frequency with the exception of the dc value. This method is not

only computationally efficient to generate on computer, but is deterministic and therefore repeatable; the input impulse response of a system can be extracted by a cross-correlation procedure. This technique has been reviewed in detail in the thesis of Kemp [8].

3. EXPERIMENTAL METHODS

The apparatus used is similar to that developed by Sharp et al. [7] (figure 1) using the MLS excitation signal employed by Kemp [8].

The diameter of the source tube (although not critical) was chosen to achieve an approximate impedance match with a relaxed mouth opening. Commercial grade plastic tubing with a bore of 20.2mm and 3mm wall thickness was used. The source tube has dimensions l_1 = 2.80m and l_2 = 0.7m, where placement of the embedded microphone is 0.7m away from end of source tube.

Data acquisition was performed in Matlab for Windows with a soundcard sample rate (F_S) of 44.1kHz.

3.1. Procedure

The computer sends an MLS signal to the loudspeaker of order m=18, which has a time period:

$$\frac{2^m - 1}{F_S} = 5.944 \text{ seconds} \qquad (1)$$

and is directed down a cylindrical source tube and into the object to be measured.

The MLS signal is played through twice; the response of the first run through is ignored in order to ensure that the measured response conforms to the periodicity condition of the signal.

A calibration pulse is first measured with a closed source tube. This measurement is then repeated with a test object coupled to the end of the source tube. The resulting reflections from the test object are deconvolved with the calibration pulse to obtain the internal profile of the object.

To check the reliability of the reflectometer, a series of measurements was made using a stepped tube (of known geometry) as the test object before measuring the vocal tract.

3.2. Vocal tract measurement for vowel production

5 American English vowels (table 1) were studied.

Vowel	Example Words
/a/	*Ask, half, past*
/e/	*Ate, made, they*
/i/	*Each, free, keep*
/o/	*Obey, note, go*
/u/	*Ooze, too*

Table 1: *Vowels with example words (southern British English pronunciation)*

The subject (adult male) was asked to phonate the vowel sound for a sustained period and then halt phonation but hold the phonatory position (vocal tract and articulators). The subject's mouth was coupled to a mouthpiece (100mm in length) attached to the end of the source tube (figure 2). The subject was asked to remain still, refraining from breathing for ~12seconds whilst a measurement was taken.

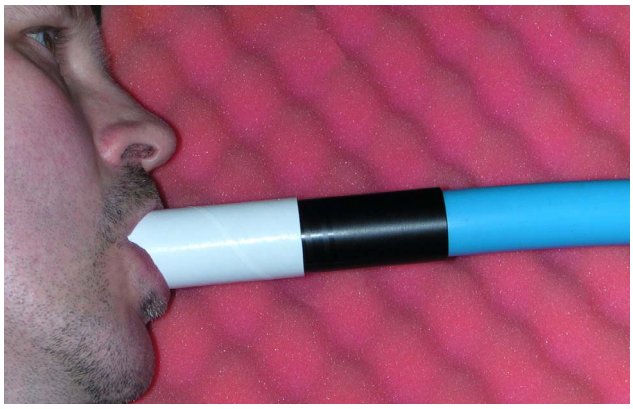


Figure 2: *Subject's coupling with mouthpiece and source-tube.*

6 measurements of the vocal tract were taken for each of the 5 vowels and the data for each vowel sound was averaged.

4. RESULTS & DISCUSSION

The reconstruction of the vocal tract for each vowel is shown in figures 3(i-v). For clarity of presentation, error bars have been omitted from each graph, although it was evident that errors were encountered more frequently between data sets for the reconstruction of pharyngeal regions where the largest discontinuities between segments can be seen (figures (ii), (iii) and (v)).

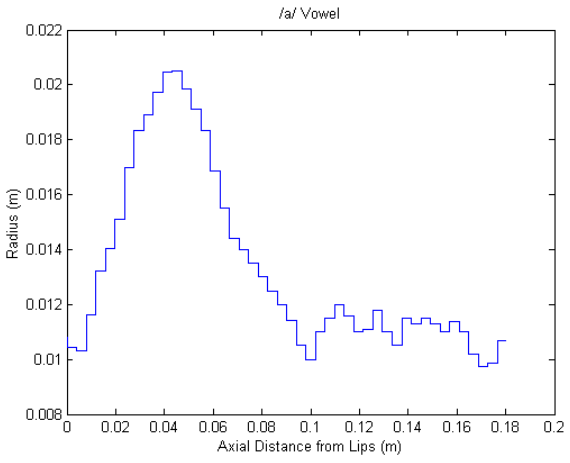


Figure 3(i): *Vocal tract area reconstruction for vowel /a/*

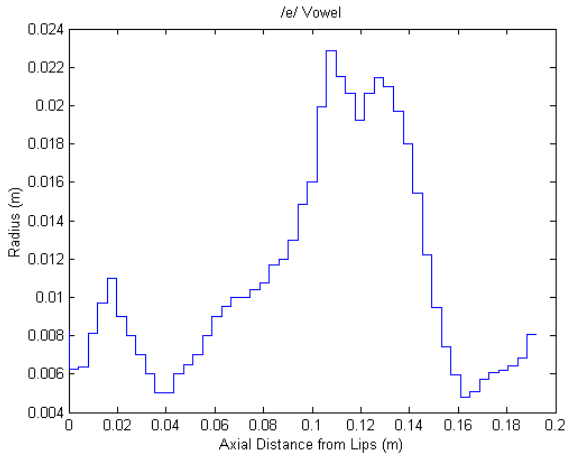


Figure 3(ii): *Vocal tract area reconstruction for vowel /e/*

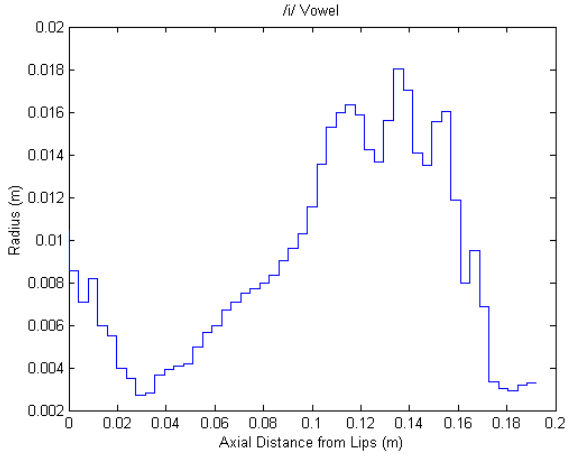


Figure 3(iii): *Vocal tract area reconstruction for vowel /i/*

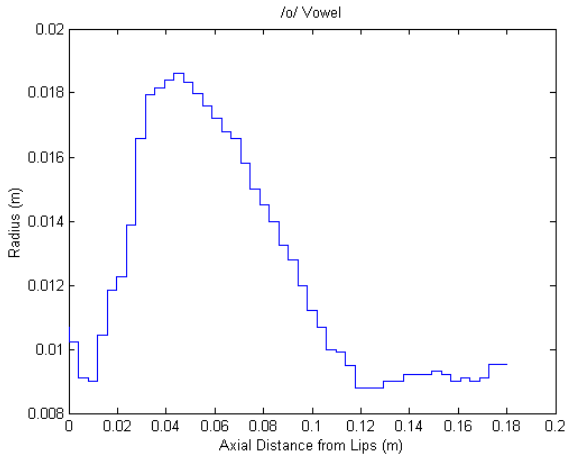


Figure 3(iv): *Vocal tract area reconstruction for vowel /o/*

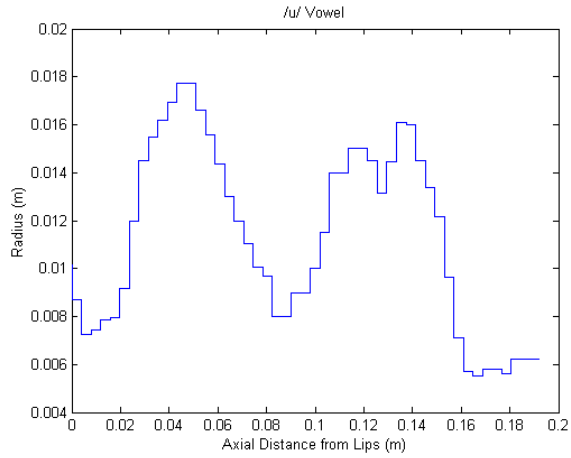


Figure 3(v): *Vocal tract area reconstruction for vowel /u/*

The results show good agreement in overall profile with those of [10] which were obtained using MRI measurement, although no direct comparisons can be made because of differing methods and subjects used in these experiments. An accurate validation of the results in the form of a comparative study between data obtained by MRI and acoustic reflectometry for the same subject is planned.

4.1 The vowel /i/: A High Front Vowel

For the production of the vowel /i/ (a high front vowel), the tongue assumes a position that is forward and up in the oral cavity. This results in a constriction forming a narrow tube which, in combination with the back cavity, forms a Helmholtz resonator with a particularly low resonant frequency (below 300Hz).

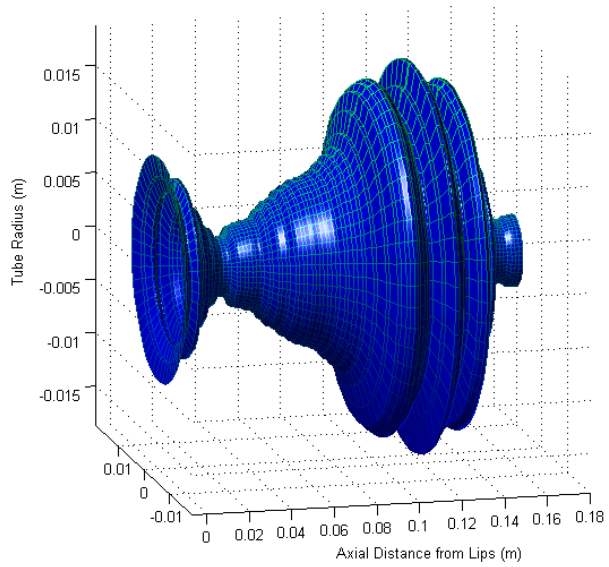


Figure 4: *3D area equivalent (tube) model for /i/ vowel.*

4.2 Accuracy & Repeatability

A problem with the current reflectometer is the way in which the mouth is coupled to the source tube. The mouthpiece used in these experiments required the subject to round their lip shape to form a tight seal. This has the disadvantage that all vowels measured will have a rounded lip shape during reconstruction. Also, having the one fixed mouthpiece does not allow for large changes in jaw height which will provide either an over- or under-estimation in the internal area reconstruction for some vowel sounds. For example, lip aperture for vowel /o/ should be smaller than measured in figure 3(iv). A redesign of a mask/mouthpiece system allowing an unimpeded, natural coupling is currently being undertaken.

Through repeated measurements, involuntary movement of the articulators during the required measurement period was observed. A relatively small shift in vocal articulators can sometimes produce major differences in resonation characteristics between vowels. Lower order MLS sequences for quicker measurement times are being assessed.

5. CONCLUSIONS

Due to the important biomechanical functions used in breathing, swallowing and speaking, it is important that experiments are carried out to obtain results speedily and to be non invasive. Acoustic pulse reflectometry has been employed in clinical studies of the vocal tract for several years, normally in the function of airway measurement. Our interest in this method is the ability to capture vocal tract profiles in the phonation of vowel sounds for the purposes of sound synthesis. By making an equivalent tube model of the vocal tract, the propagation of an acoustic wave can be readily calculated using techniques such as waveguide modelling, which will in turn allow us to synthesise sound and form the basis of a physical model of the voice.

The attractions of this technique for vocal tract measurement are many: it is non-invasive, safe, repeatable and inexpensive.

6. REFERENCES

[1] Sondhi, M. M. and Gopinath, B., “Determination of vocal-tract shape from impulse response at the lips”, J. Acoust. Soc. Am., 49(6):1867-1873, 1971.

[2] Sondhi, M. M. and Resnick, J. R., “The inverse problem for the vocal tract: numerical methods, acoustical experiments and speech synthesis”, J. Acoust. Soc. Am., 73(3):985-1002, 1983.

[3] Jackson, A. C. and Olsen, D. E., “Comparison of direct and acoustical area measurements in physical models of human central airways”, J. Appl. Physiol., 48(5):896-902, 1980.

[4] Fredberg, J. J., Wohl, M. E. B., Glass, G. M. and Dorkin, H. L., “Airway area by acoustic reflections measured at the mouth”, J. Appl. Physiol., 48(5):749-758, 1980.

[5] Brooks, L. J., Castile, R. G., Glass, G. M. and Griscom, N. T., “Reproducibility and accuracy of airway area by acoustic reflection”, J. Appl. Physiol., 57(3):777-787, 1984.

[6] Marshall, I., Acoustic reflectometry for airway measurement, PhD thesis, University of Edinburgh, 1992.

[7] Sharp, D. B., Acoustic pulse reflectometry for the measurement of musical wind instruments, PhD thesis, University of Edinburgh, 1996.

[8] Kemp, J. A., Theoretical and experimental study of wave propagation in brass musical instruments, PhD thesis, University of Edinburgh, 2002.

[9] Marshall, I., Rogers, M., and Drummond, G., “Acoustic reflectometry for airway measurement. Principles, limitations and previous work”, Clin. Phys. Physiol. Meas., Vol. 12, No. 2, 131-141, 1991.

[10] Story, B.H., Titze, I.R., Hoffman, E.A. “Vocal tract area functions from magnetic resonance imaging” J. Acoust. Soc. Amer., Vol. 100 (1), 537-554, 1996.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

ON THE USE OF ELECTROGLOTTOGRAPHY FOR CHARACTERISATION OF THE LARYNGEAL MECHANISMS

Nathalie Henrich, Bernard Roubeau and Michèle Castellengo

Laboratoire d'Acoustique Musicale (CNRS - UPMC - MCC), Paris, France
Service ORL, Hôpital Tenon, Paris, France
nathalie.henrich@lam.jussieu.fr
bernard.roubeau@tnn.ap-hop-paris.fr
castel@ccr.jussieu.fr

ABSTRACT

In order to produce a wide range of pitches, the human vocal production is characterized by the use of four distinct laryngeal configurations. The evidence of such division is given by noticeable transitions in the audio signal, the electroglottographic signal and its derivative, during the production of a glissando. So as to avoid the confusion linked to the notion of register in singing and to bring out the difference between change in the glottis configuration (laryngeal mechanism) and adjustment of the vocal tract (resonance register), the following terminology is proposed: mechanism 0 (”vocal fry”), mechanism I (”modal” or ”chest” register), mechanism II (”falsetto” for male voice and ”head” register for female voice) and mechanism III (”whistle” register). A shape description of the electroglottographic signal (EGG) and its derivative (DEGG) is provided in the case of the four laryngeal mechanisms and the strong correlation between electroglottography-based open quotient measurements and laryngeal mechanisms is illustrated. The possibilities and limitations of using the electroglottography in order to detect a given laryngeal mechanism are discussed.

1. INTRODUCTION

From the lowest tone (about 60 Hz) to the highest one (around 1500 Hz), the production of a glissando can sometimes be clearly divided into four distinctive parts, as illustrated in Figure 1. The evidence of such division is given by the noticeable jumps in fundamental frequency which can be detected at the transitions from one part to another, suggesting that these phenomenons would be related to some adjustments in the vocal folds' vibratory movement. For clarity, these four distinctive parts will be called: “mechanism 0 (m0)”, “mechanism I (mI)”, “mechanism II (mII)” and “mechanism III (mIII)”. A rationale will be formulated later in this article to explain this terminology.

In order to define these mechanisms and understand more about them, the electroglottography (EGG) is chosen as a very simple and non-invasive method for indirect exploration of the vocal folds' vibratory process. In which way can electroglottography help for characterisation and identification of these mechanisms ? What is the relation between these observations and the notion of vocal register ?

To answer these questions, the conclusions resulted from recent and more ancient research studies are brought up. We refer to two major studies for which audio and EGG signals were recorded simultaneously :

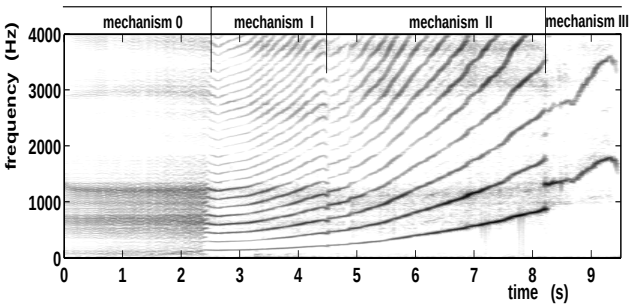


Figure 1: *Illustration of the four laryngeal mechanisms on an ascending glissando sung by a soprano.*

- **study S1** (Roubeau *et al.* [1, 2, 3, 4], 1993) :
  19 subjects, male and female, trained and untrained singers, participated in this study, which purpose was to explore the voice registers. The subjects were asked to produce ascending and descending glissandos, and sustained vowels with a noticeable register transition.

- **study S2** (Henrich *et al.* [5, 6], 2001) :
  A corpus of sung items designed for the purpose of open quotient measurements was recorded by 18 classically trained singers (7 baritones, 2 tenors, 3 counter-tenors, 3 mezzo-sopranos and 3 sopranos). The subjects were asked to sing a few sentences, sustained vowels, crescendos, decrescendos and glissandos.

The most significant outcomes of these experiments will be summed up here and illustrated on a few typical examples, with a view to shed light on the notion of vocal register. In the first part, we will define the laryngeal mechanisms on the basis of observations made on the EGG signal and its derivative (DEGG) at the instants of transition. In the second part, the four laryngeal mechanisms will be described within a glottal period. The notion of register will be discuss with regard to these observations. We will finally conlude in discussing whether an EGG signal and its derivative could be a relevant indicator of the laryngeal mechanism.

2. THE LARYNGEAL MECHANISMS

2.1. The contribution of electroglottography

Electroglottography, which was invented by Fabre in 1956 [7], is nowadays a widespread technique that enables the investigation of vocal folds contact area in phonation, in an easy and non-invasive way (for a detailed review, see Childers & Krishnamurthy [8], Colton & Conture [9], Orlikoff [10]). Two electrodes are placed on the neck of the subject, and they generate a high-frequency modulated current which admittance varies with the vibratory movements of the glottis : it increases as the glottis closes. In comparing this signal with other means of exploration, it has been shown that the EGG signal is related to the vocal folds contact area : the larger the contact surface, the larger the measured admittance.

A few studies have shown that the EGG shape over a glottal period is dependent on the vocal registers, i.e. "chest" versus "falsetto" register (Lecluse & Brocaar [11], Kitzing [12]). The EGG glottal period is more asymmetric in chest than in falsetto register, with a more rapid closure phase and a gradual opening phase. These differences reflect the differences observed on a glottal flow period (Childers & Lee [13]) and can be explained in terms of vertical glottal contact behavior by the use of a mathematical model of vocal folds contact area (Titze [14]).

Nevertheless, the observation of an EGG glottal period seems not sufficient enough to characterize a voice production in terms of vocal register. Indeed, in studying in detail the EGG shape in the case of 19 subjects (study S1), Roubeau [3] observed a large amount of variation across subjects and within a given subject. In order to understand more about the stable phases of a given vocal production, we shall first look at the transition phenomenons which could occur during this production and which may convey much information about the production before and after the transition phase.

2.2. Illustration of the transition phenomenons

Two typical examples have been chosen to illustrate the observations of transition phenomenons : an ascending-descending glissando produced by a baritone singer (see Figure 2) and a descending-ascending glissando produced by a soprano (see Figure 3).

As illustrated in Figure 2, the transition between mechanism 0 and mechanism I is clearly visible on the sound spectrum. The change of voice quality is audible and it is thus easy to determine whether a given voice production is done in mechanism 0 or not. The transition is easily detectable on the EGG signal and its derivative, as illustrated on the bottom panel in Figure 2.

The transition between mechanism I and mechanism II is illustrated in both Figure 2 and Figure 3. It goes with an abrupt amplitude modification in the EGG and DEGG signals. This transition may go with a characteristic jump in frequency (Roubeau *et al.* [1, 3], Svec *et al.* [15], Miller [16]) but the EGG and DEGG amplitude change does not depend on this frequency jump : as illustrated on the sound spectrogram in Figure 2, the transition mI → mII goes with a jump in frequency and the transition mII → mI is smooth. It must be emphasized that the frequency region of the jump do not vary much and is not dependent on the sexe or the tessitura of the singer. In both examples, it takes place around 250 - 300 Hz.

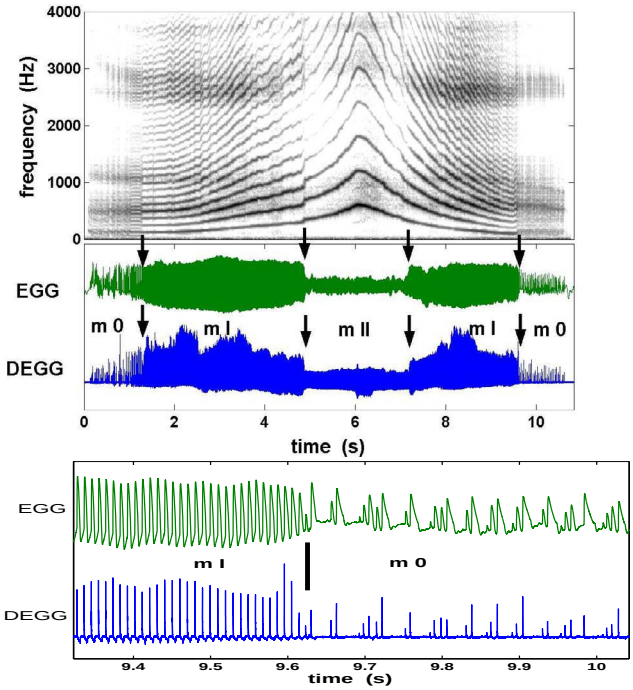


Figure 2: *Ascending-descending glissando produced by a baritone singer, which can be divided into 5 parts : m0 - mI - mII - mI - m0. The bottom panel presents a zoom on the last transition (mI - m0). (example extracted from study S2)*

The transition between mechanism II and mechanism III is illustrated in Figure 3. Similar to transition mI ↔ mII, the transition mIII → mII is very sudden and it goes with a great amplitude change in the EGG and DEGG signals. This example illustrates also a peculiar behavior. The transition mII → mIII goes with a jump in fundamental frequency, but the great amplitude change occurs 200 ms before the transition. In this case, one can not rely on the EGG or DEGG global amplitude change to detect the transition. The change of shape from one period to another should be taken into account, as we will see in part 3.

To summarize, during a transition from one mechanism to another, we observe a change in the amplitude of the EGG and DEGG signals. It can go with a characteristic jump in fundamental frequency, but it occurs also when the transition is produced smoothly. It must be emphasized that the amplitude change is a characteristic of the transition phenomenon and it does not give any information which would be reliable for mechanism detection during the stable part. Indeed, similar signal amplitude can be observed in the different mechanisms, as illustrated in Figures 2 and 3.

2.3. Definition of the laryngeal mechanisms

As the EGG signal is related to the vocal folds contact area, an abrupt amplitude change in the EGG and DEGG signals provide clear evidence that the transition phases go with great modification in the vocal folds vibratory process (Hirano [17]). On the basis of these transition phenomenons, we define the laryngeal mechanisms, which correspond to different physiological, mechanical

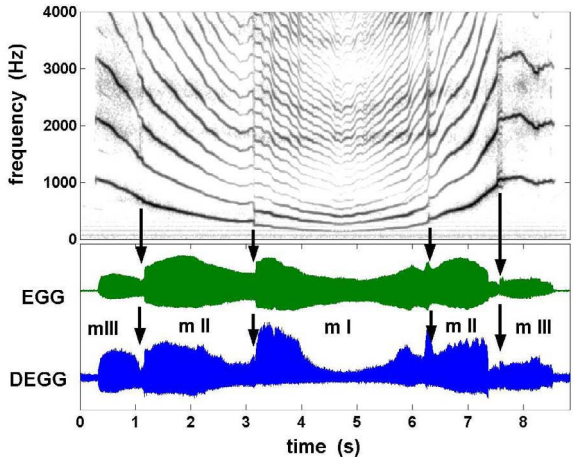


Figure 3: *Descending-ascending glissando produced by a soprano singer, which can be divided into 5 parts : mIII - mII - mI - mII - mIII. (example extracted from study S1)*

and vibratory glottal behaviors.

The laryngeal mechanisms can be related to the physiological knowledge that we have of the glottal vibratory process and to the notion of vocal register, a notion which is widespread in the literature but rather confused and badly defined. As matters stand, the notion of register combines the acoustical phenomenons which are related to the laryngeal configuration and its vibratory properties, and the acoustical phenomenons which depend only on the resonance properties of the vocal tract. So as to avoid this confusion and to bring out the difference between change in the glottis configuration (laryngeal mechanism) and adjustment of the vocal tract (resonance register), we have chosen to use the term *laryngeal mechanism* instead of *register*, and to number them.

3. ELECTROGLOTTOGRAPHIC DESCRIPTION OF A GLOTTAL CYCLE

In the previous section, we have based the definition of the laryngeal mechanisms on the noticeable transitions which can be found on an EGG or DEGG signal. We will now provide an electroglottographic description of the glottal cycle within a given laryngeal mechanism and relate this terminology to our knowledge of the laryngeal physiology and vocal registers.

3.1. Laryngeal mechanism 0

Laryngeal mechanism 0 corresponds to the so-called ”vocal fry”, ”pulse” or ”strohbass” register (Hollien [18]). It is the way to produce the lowest tones. It is much used in speech, especially by male voice. It is hardly used in western operatic singing.

A few EGG and DEGG periods are shown on the bottom panel in Figure 2. In this mechanism, the vocal folds are short, very thick and slack (Hollien [18]). The vibratory process is characterized by a very long closed phase, with respect to the fundamental period.

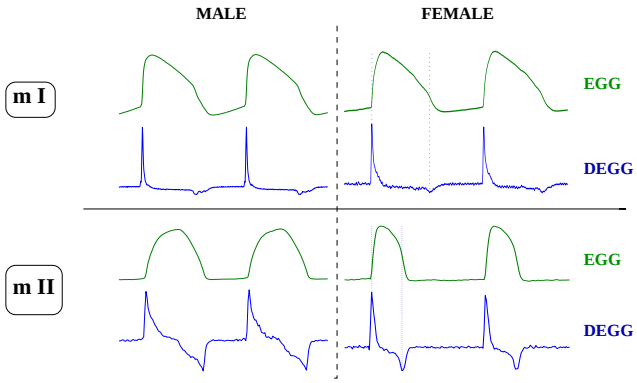


Figure 4: *Typical examples of EGG and DEGG glottal periods in mechanism I and II, in the case of male and female singers.*

3.2. Laryngeal mechanisms I and II

Laryngeal mechanism I corresponds to the so-called ”chest” or ”modal” register for male and female and male ”head” register. Laryngeal mechanism II corresponds to the ”falsetto” or ”loft” male register, and to the ”head” female register. Both mechanisms are commonly used in speech and singing.

Typical electroglottographic signals are shown in Figure 4. In mechanism I, the vocal folds are thick and they vibrate over their whole length with a vertical phase difference, whereas in mechanism II, the mass and vibratory length is reduced (Vennard [19]) and there is no vertical phase difference. The EGG shape is thus more asymmetric in mI than in mII. The DEGG present a strong glottal closure peak and a weak glottal opening peak in mI, whereas both peaks could present similar amplitude in mII.

3.3. Laryngeal mechanism III

Laryngeal mechanism III corresponds to the so-called ”whistle”, ”flageolet” or ”flute” register (Miller & Schutte [20]). It is used to produce the highest tones and is hardly used in speech and singing.

In this mechanism, the vocal folds are thin, very tensed and the vibratory amplitude is much reduced as compared to mechanism II. The aperture between the vocal folds is very thin, and there could be no contact between the vocal folds. Either there is no EGG signal, or the EGG glottal period is very symmetric in shape.

3.4. Open quotient

The observations made about the difference in shape of EGG signal between mI and mII can be quantified in measuring the open quotient (O_q), defined as the ratio between glottal open time and fundamental period. The measurements conducted on a singing database (study S2) confirm that the open quotient is dependent on the laryngeal mechanism used during the vocal production : open quotient values are usually lower in mechanism I than in mechanism II, as illustrated in Figure 5. Thus, the open quotient can be seen as a good indicator of the laryngeal mechanism within a given voice production. Yet, one should not rely solely on the values of open quotient to determine which laryngeal mechanism is used, as there is a degree of overlap between the O_q ranges corresponding to laryngeal mechanisms I and II. Listening to the corresponding sound samples can provide additional information and help to

characterize a given vocal production. Nevertheless, the ear can sometimes be tricked by the vocal technique of the singer. For this reason, a combination of both analysis, listening and measuring of open quotient and other acoustical and EGG parameters, is usually required to find out which laryngeal mechanism is being used.

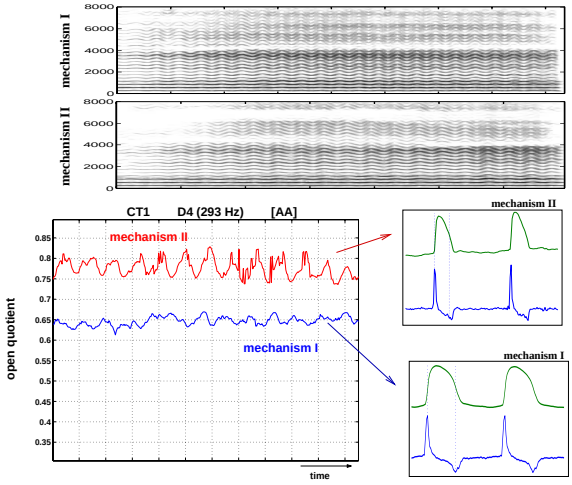


Figure 5: *Measurement of open quotient in the case of vowel [a] sung by a counter-tenor (study S2) on the same pitch (D4) in mechanism I and II (bottom left panel). The corresponding EGG and DEGG signals have also been plotted (upper and lower bottom right panels). The spectrograms of the two sound samples are shown on the two top panels.*

4. CONCLUSION

In this article, we have pointed out that the human male and female vocal production is characterized by the use of four laryngeal mechanisms. The evidence of such division is given by the abrupt amplitude change which can be observed on the EGG and DEGG signals during a transition. These laryngeal mechanisms are similar for male and female. They correspond to different glottal configurations and different vibratory process, which are not fully understood yet.

The electroglottography can be used as a reliable indicator of a laryngeal mechanism. In some cases, the EGG and DEGG features are typical of a given laryngeal mechanism and we can easily conclude. In other cases, the distinction remains unclear. It is then necessary to explore the vocal production within a transition phenomenon to assess the validity of a laryngeal mechanism estimation. An abrupt change in the EGG and DEGG signals gives a clear clue to a laryngeal change.

Such a characterization of the laryngeal mechanisms should help to clarify the notion of vocal register. It lays the foundation for further investigation on resonance registers, like the ”middle” voice.

5. REFERENCES

[1] Roubeau B., Chevrie-Muller C., and Arabia-Guidet C., “Electroglottographic study of the changes of voice registers,” *Folia Phoniatr.*, vol. 39, pp. 280–89, 1987.

[2] Roubeau B., Chevrie-Muller C., and Arabia C., “Control of laryngeal vibration in register change,” in *Vocal Fold Physiology : Acoustic, perceptual, and physiological aspects of voice mechanisms*, J. Gauffin and B. Hammarberg, Eds., pp. 279–86. Singular Publishing Group, San Diego, 1991.

[3] Roubeau B., *Mécanismes vibratoires laryngés et contrôle neuro-musculaire de la fréquence fondamentale*, thèse, Université Paris XI, Orsay, 1993.

[4] Roubeau B. and Castellengo M., “Revision of the notion of voice register,” *XIXth International CoMeT Congress, Utrecht*, 1993.

[5] Henrich N., *Etude de la source glottique en voix parlée et chantée*, Ph.d. thesis, Université Paris 6, France, 2001.

[6] Henrich N., d’Alessandro C., Castellengo M., and Doval B., “Open quotient measurements in singing,” *Submitted to J. Acous. Soc. Am.*, 2003.

[7] Fabre P., “Un procédé électrique percutané d’inscription de l’accolement glottique au cours de la phonation : glottographie de haute fréquence,” *Bull. Nat. Med.*, pp. 66–69, 1957.

[8] Childers D. G. and Krishnamurthy A. K., “A critical review of electroglottography,” *CRC Critical Reviews in Biomedical Engineering*, vol. 12, pp. 131–161, 1985.

[9] Colton R. H. and Conture E. G., “Problems and pitfalls of electroglottography,” *J. Voice*, vol. 4, pp. 10–24, 1990.

[10] Orlikoff R. F., “The uses and abuses of electroglottography,” *Phonoscope*, vol. 1, no. 1, pp. 37–53, 1998.

[11] Lecluse F. and Brocaar M., “Quantitative measurements in the electroglottogram,” *17th International Congress of Logopedics and Phoniatrics*, 1977.

[12] Kitzing P., “Photo- and electroglottographical recording of the laryngeal vibratory pattern during different registers,” *Folia Phoniatr.*, vol. 34, pp. 234–41, 1982.

[13] Childers D. G. and Lee C. K., “Vocal quality factors: Analysis, synthesis, and perception,” *J. Acous. Soc. Am.*, vol. 90, pp. 2394–2410, 1991.

[14] Titze I.R., “Interpretation of the electroglottographic signal,” *J. Voice*, vol. 4, no. 1, pp. 1–9, 1990.

[15] Svec J. G., Schutte H. K., and Miller D. G., “On pitch jumps between chest and falsetto registers in voice : Data from living and excised human larynges,” *J. Acous. Soc. Am.*, vol. 106, no. 3, pp. 1523–31, 1999.

[16] Miller D. G., *Registers in singing*, Ph.d. thesis, Rijksuniversiteit Groningen, the Netherlands, 2000.

[17] Hirano M., “The role of the layer structure of the vocal folds in register control,” *Vox Humana (University of Jyväskylä, Finland)*, 1982.

[18] Hollien H., “On vocal registers,” *Journal of Phonetics*, vol. 2, pp. 125–43, 1974.

[19] Vennard W., *Singing : the mechanism and the technic*, Carl Fischer Inc., New York, 1967.

[20] Miller D. G. and Schutte H. K., “Physical definition of the flageolet register,” *J. Voice*, vol. 7, no. 3, pp. 206–12, 1993.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

VOCAL FOLD RESONANCES AT LOW AND HIGH PITCH TUNING

-preliminary results

Stellan Hertegård[1] and Hans Larsson[1], Svante Granqvist [2]

[1]Karolinska Institute, Department of Logopedics and Phoniatrics
Huddinge University Hospital,

[2]Department of Speech Music and Hearing
Royal Institute of Technology.
Stockholm Sweden

stellanhertegard@hotmail.com

ABSTRACT

Four trained singers were examined with high-speed recording during exernal acousic excitation of the vocal tract with the vocal folds kept in a phonatory position, but without phonating (so-called Kaneko manoeuvre). Automatic glottal area analysis was made and spectral analysis was performed on the obtained area waveform curves. The results showed resonance frequencies close to habitual phonation frequency in three out of the four subjects. Intended pitch at phonation seems to affect the resonance frequency as well as the voice category of the subjects.

INTRODUCTION

Previous stroboscopic examinations indicate that the mechanical properties of the vocal folds differ between high- and low pitched singing. There may also be an interaction with the vocal tract resonances. The mechanical properties of the vocal folds are pertinent to both the speaking and singing voice. These characteristics are usually investigated directly with endoscopy during phonation or indirectly by means of acoustical methods. The mode of phonation itself might affect the vibratory modes found at laryngoscopy. An alternative approach was suggested by Kaneko et al [1] and by Švec et al [2] who used external stimulation for excitation of the voice source when the vocal folds were kept in a phonatory position (a so-called Kaneko manoeuvre) during a frequency sweep. A similar experiment was performed by Granqvist et al [3] on a normal male speaker using acoustic excitation. In this experiment, a loudspeaker and a sweep tone was utilised for the excitation and the vocal fold response was studied with high-speed imaging. The results showed three different resonances for a relaxed phonatory setting. The aim of the present study is to examine the vocal fold resonances at low and high pitch tuning. A second aim was to study differences between 3 different voice categories, soprano, mezzo-soprano and barytone from the stimulation experiment.

Material and Task

Four subjects, one soprano, one mezzo-soprano and two barytones (age between 40 and 47 years) participated in the study. They were all healthy non-smokers with normal larynges. Three of the subjects were trained professional opera singers, whereas one subject (barytone) was an amateur singer with some voice training.

The task was to perform the Kaneko manoeuvre, i.e to phonate followed by a cease in phonation but keeping the phonatory vocal fold setting during the frequency sweep. The subjects were provided live feedback from the camera monitor in order to complete this rather difficult task. Two recordings were made of each subject, first phonations / manoeuvres at habitual pitch (chosen by the subject) and later phonations / manoeuvres at a higher pitch (approximately one octave higher).

Experiment Apparatus

For the experiment a loudspeaker was connected to a plastic tube with an inner diameter of 28 mm. A laryngeal endoscope (Storz 70 degree) was passed through a hole in the tube (which was sealed tight with plasticine to avoid losses) and connected to a Weinberger high-speed camera (image resolution 256x64pixels at 1904 frames/s). A custom made software was used to create the sound stimulus: For the females a sinusoidal sweep tone from 500-100 Hz was used and for the male subjects a tone from 250-50Hz. The sound pressure level was typically in the range 140-150 dB inside the tube and vocal cavity. The duration of the sweeps was selected to 3 seconds in order to fit within the 4 second capacity of the high-speed camera. The sweep tone was controlled from one computer and the high speed images during the sweep was recorded on another computer. The sound pressure level (SPL) inside the tube was measured by means of a Brüel & Kjaer (B&K) condenser microphone (4192) and a B&K microphone amplifier (2669) and recorded on the same computer that was used for the sweep generation. SPL was calibrated by a B&K calibrator, which generated a constant SPL of 94 dB.

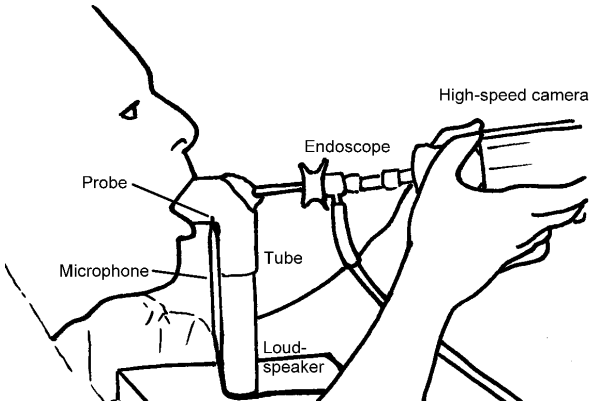


Figure 1: The experimental apparatus. The subject is phonating with the mouth closed around a tube where the acoustic excitation is fed from a loudspeaker. An endoscope connected to the high speed camera is passed through an airtight hole in the tube

Analyses

The software High speed toolbox was used for analysing the high speed recordings of the vocal fold vibration. The program can automatically detect the edges of the vocal folds and generate an area waveform, This waveform will correspond to the motion of the vocal folds as excited by the frequency sweep. The automatic area extraction sometimes had to be manually corrected. A special software was used for FFT analysis of the area and sound waveforms The software used a simple algorithm to detect F0 in the waveform files; the highest peak in the FFT spectrum was assumed to correspond to the F0. While this is not generally a good F0 extraction method for voice, it is sufficiently good for the purpose of this experiment, where the exciting signal was a sinusoid. Using quadratic interpolation, the F0 values could be detected with high precision even though the sampling rate was as low as 4 times the highest frequency of interest (1904 Hz and 500 Hz, respectively) The program used semi-automatic detection of the frequency sweeps in the area and sound waveforms. Since the sweeps were exponential, exponential regression was utilised on the part of the extracted F0 curve that contained the frequency sweep. Thus, two equations describing the location of the F0 within the two waveforms could be obtained These equations was then utilised to select a number of positions in the waveform files, re-do the FFTs for these positions and detect the height of the peak at F0. These heights correspond to the response of the sound (upper curve) and area (middle curve), respectively, and are plotted in the lower panel of the computer program. Also, the difference between these two curves is plotted (lower curve), and this curve corresponds to the transfer function from exciting sound pressure to glottal area. (Figure 2)

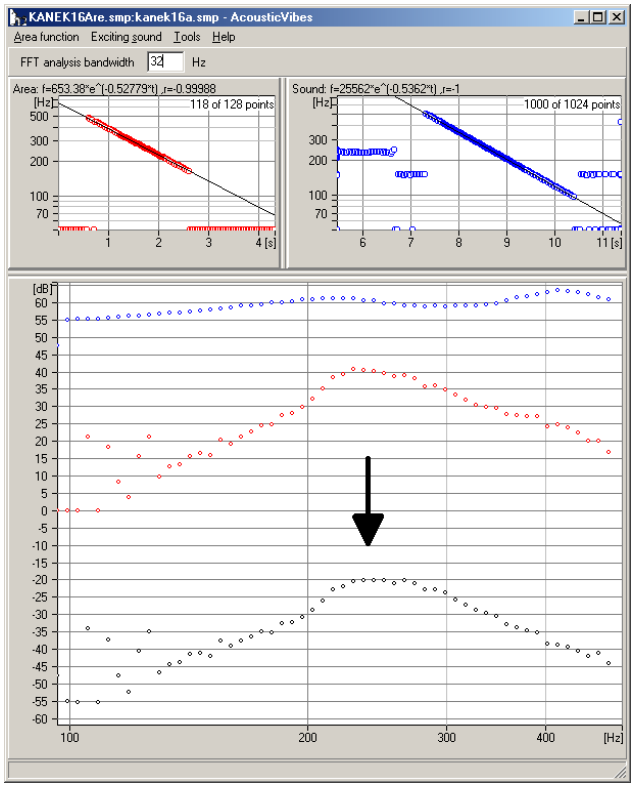


Figure 2 Analysis of resonance frequency in the glottal area variation during the sinusoid sweep. The sweeps are shown in the upper panels, and the response curves are shown in the lower panel. The lower curve in the bottom panel shows the compensated response curve of the glottal area; note the resonance at around 230Hz (arrow).

Results

The preliminary results show that all subjects had a resonance frequency close to the F0 of the phonations before the Kaneko manoeuvre. The subject barytone 1 also had a lower resonance at 80 Hz. The resonances of the high-pitched phonatory settings were difficult to analyse but seemed to be lower than the F0 of the phonations before the Kaneko manoeuvres. There were also clear differences between the different voice categories, as expected with the soprano and mezzo-soprano showing the highest habitual setting resonances. The preliminary results are summarised in table 1.

Subject	**Phonation 1** Resonance (habitual pitch setting)	**Phonation 2** Resonance (higher pitch setting)
Soprano	230Hz (230Hz)	300Hz (345Hz)
Mezzosoprano	200 Hz (210Hz)	300Hz (420Hz)
Barytone 1 (pro)	75Hz (115Hz)	80Hz and 180Hz (220-225Hz)
Barytone 2 (amateur)	120Hz (120-140Hz)	180Hz (190Hz)

Table 1: *Resonances found for the habitual phonatory settings and for the phonatory settings with higher pitch, within brackets:F0 for the phonations before the Kaneko manoeuvre.*

DISCUSSION and CONCLUSIONS

The method of exciting the vocal folds by means of an acoustic signal seems to work well, and thus it seems possible to study vocal fold resonances with the technique used. The area analysis is rather time-consuming, in particular when the automatic detection needs manual correction. The technique also requires subjects who are able to perform the Kaneko manoeuvre. The results indicate that the method can be used for studying the vocal fold's small-signal resonances. These resonances are probably important factor in the description of vocal fold mechanics and tissue characteristics.

In the future, it may be worthwhile to utilise some other signal than of the glottal area, e.g. a single point on the edge of the vocal fold. In this way, it would be possible to analyse the properties of the vocal folds separately, and also to analyse the response at different locations along the vocal folds.

REFERENCES

[1] Kaneko, T, Masuda, T, Shimada, A., et al. Resonance characteristics of the human vocal fold in vivo and in vitro by an impulse excitation. Laryngeal function in phonation and respiration, in Vocal Fold Physiology Eds Bar, Sasaki, Harris., Collge Hill Press. Boston MA pp 349-365, 1987.

[2] Švec, J, Horacek, J, Sram, F., et al. Resonance properties of the vocal folds: in vivo laryngoscopic investigation of the externally excited laryngeal vibrations. J Acoust Soc Am. 2000 Oct 108(4):1397-407.

[3] Granqvist, S, Hertegård, S. In vivo examination of resonances in the larynx using external acousitc stimulation. Presentation at Pevoc 4 Stockholm Sweden August 23-28 2001

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

A CAPELLA SATB QUARTET IN-TUNE SINGING: EVIDENCE OF INTONATION SHIFT

David M Howard

Media Engineering Research Group
Department of Electronics, University of York, Heslington, York, YO10 5DD, UK
dh@ohm.york.ac.uk

ABSTRACT

This paper sets out to test the hypothesis that if an a capella SATB quartet tends towards the use of a non-equal tempered intonation, then their absolute pitch reference should drift away from its starting point with modulation. In order to test this hypothesis, special SATB exercises were written that incorporated a number of modulations between the starting and finishing chords, which themselves were note for note either identical or exactly an octave apart. The intonation difference between these two chords (after allowing for the octave difference where appropriate) was measured relative to mean tone and equal tempered tuning. Four electrolaryngographs were employed, one per singer, set up so their radio frequency outputs did not interfere with each other. Mean fundamental frequency measurements were made on a note-by-note basis to an accuracy of 1microsecond. Pitch shifts were observed with respect to equal tempered tuning. The results suggest that conductors who worry about overall intonation drift in a capella performance of musical material that has significant modulation may be misguided, since this may be in direct contradiction with being able to remain in-tune. It is further suggested that achieving consistent equal tempered tuning when singing a capella is an even greater challenge for choirs, which has even less chance of being accurately and consistently achieved.

1. INTRODUCTION

Each live musical performance is different, as a number of important aspects are subject to subtle changes on each occasion. Much of the overall nature of a performance is governed by the emotional state of the performer, her/his stage confidence and familiarity with the score as well as the performance situation itself. Two elements of musical communication that are subject to such variation between individual performances are timing and pitching. The singer has wide ranging opportunities for varying both of these, and in particular, very fine degrees of pitch variation.

This paper is concerned with the relative and absolute tuning between members of an SATB vocal quartet. In particular, it sets out to investigate whether or not choral singers tend to tune away from equal temperament, as might be predicted if there is a tendency to make use of an intonational strategy that seeks minimum dissonance. The methodology makes use of specially written singing exercises that involve a number of chord changes that visit different keys, but each adjacent chord pair has one note in common, which is held to provide a local tuning reference. Thus there is an intonation link as the chords change. If singers display a tendency to tune towards pure intonation [e.g. 1], then the pitch of a piece of a capella music has the possibility of drifting in absolute terms, while remaining accurate in relative terms. This suggestion is investigated.

2. FUNDAMENTAL FREQUENCY MEASUREMENT

Fundamental frequency (f0) measurement for speech signals has been the subject of research since the 1920s, and to date, there is no one algorithm that will measure f0 reliably for any speaker in any acoustic situation. The choice of an f0 analysis technique should be made with respect to the nature of the input signal, the acoustic conditions in which the system will be used and the analysis output errors that are acceptable and those that are not. An extensive review can be found in [2, 3].

The operation of fundamental frequency (f0) measurement algorithms can be subdivided into those that make use of:

- the input pressure waveform - time domain techniques
- the spectrum of the input signal - frequency domain techniques
- hybrid time and frequency domain technique
- direct measurement of larynx activity.

Time domain f_0 estimation techniques exploit the fact that the waveform of a periodic sound repeats, and they detect features which occur once per cycle during voiced sounds, such as the major positive (or negative) peak, positive- (or negative-) going zero crossings, the slope changes associated with major peaks or any readily identifiable feature on the waveform which can be identified as occurring once per cycle. Potential problems with time domain f0 estimation can occur for example, when the shape of the speech pressure waveform changes as the formant positions move during the articulation of different sounds. Such changes, as well as background acoustic noise can affect the definition of the repeating

features in the waveform, such as peaks, troughs and zero crossings.

Frequency domain f0 estimation techniques take advantage of the fact that the frequency spectrum of a periodic signal has a series of regularly spaced peaks, or 'harmonics', and a non-periodic signal has a continuous spectrum with no regularly spaced peaks. F0 estimation is typically based on finding and tracking either: (a) the f0 component (first harmonic) itself; or (b) the frequency difference between any two adjacent harmonics, which is equal to f0; or (c) the f0 which best 'fits' all the harmonics present. Particular problems can occur with each of these method as follows: (a) when the first harmonic is very weak in amplitude compared to the other harmonics; then the estimated f0 jumps up an octave to the second harmonic, (b) when individual harmonics are weak due to the frequency relations between formant positions and adjacency tends to become obscured, and (c) in establishing a best fit f0 depending on the algorithm employed.

Of rather greater significance is the observation that frequency domain techniques cannot be used when an very accurate estimate of f0 is required, particularly if cycle-by-cycle f0 values are required, because of the need for a finite length of the input signal, known as a time 'window', that is needed for the spectrum computation. The window must be at least as long as one cycle at the lowest f0 expected. This results in averaging of the output f0 value, particularly when f0 is changing rapidly and when f0 is high and the analysis window will be encompassing a number of cycles. Frequency domain techniques tend do however, tend to be more resilient in the presence of acoustic noise.

Hybrid techniques make use of a combination of time and frequency domain features such that some merits of one domain might override some drawback of the other domain.

Direct measurement of larynx activity can be based on the use of a throat microphone, or electrical impedance measurements systems such as the electrolaryngograph or electroglottograph [e.g. 4]. These techniques are particularly appropriate practically, due to their immunity to both extraneous acoustic noise (that is any sound other than the desired signal), and environmental acoustic modifications such as the early sound and reverberant sound. These devices also generally give a more rigorous estimation than many microphone based devices of the instant at which the boundaries between silence, voiced and voiceless sections of an utterance occur. The electrolaryngograph or electroglottograph are therefore widely used as reference f0 estimation systems [5, 6].

A practical f0 analysis technique will often incorporate output smoothing in order to enable a reasonable output to be obtained in the presence of background noise. The use or otherwise of f0 smoothing must also be considered with reference to the requirements of the analysis and the nature of the input data and acoustic conditions. If highly accurate cycle-by-cycle f0 results are required, then smoothing should probably not be employed.

For the purposes of this experiment, a highly accurate cycle-by-cycle measurement of f0 is required in order to provide a basis for subtle intonation variation in choral singing to be tracked. In addition, four singers are to sing together in the same space, so the use of an acoustically based f0 estimation technique is not an option. Therefore four electrolaryngographs were employed to enable f0 to be monitored non-invasively. These devices are not affected by local acoustic signals (such as the other three singers), but they have the potential to interfere with each other since they operate in the radio frequency range. The four electrolaryngographs were specially set up by Laryngograph Ltd. [7] to ensure that their operating frequencies were widely spaced.

Two electrodes are placed superficially on either side of the neck of the subject at the level of the larynx and held in place with an elastic neckband. A constant amplitude high frequency voltage is applied enabling the electrolaryngograph to monitor the electrical impedance between the electrodes as the vocal folds vibrate by detecting the current flowing between them. The electrolaryngograph output waveform (Lx) therefore represents current flow between the electrodes, and this will be greater as the vocal fold contact area increases and less as they part. The Lx waveshape remains relatively static during speech or singing, so f0 can be readily measured using a time domain technique.

3. EXPERIMENTAL DETAILS

A soprano, alto, tenor, bass (SATB) vocal quartet took part in the experiment. All members were experienced singers, and three sang regularly together. The fourth (soprano) sang regularly in other choirs. At the start of the recording session, some time was spent singing together in order to become familiar with each other's sound, as well as the local acoustic (a domestic living room).

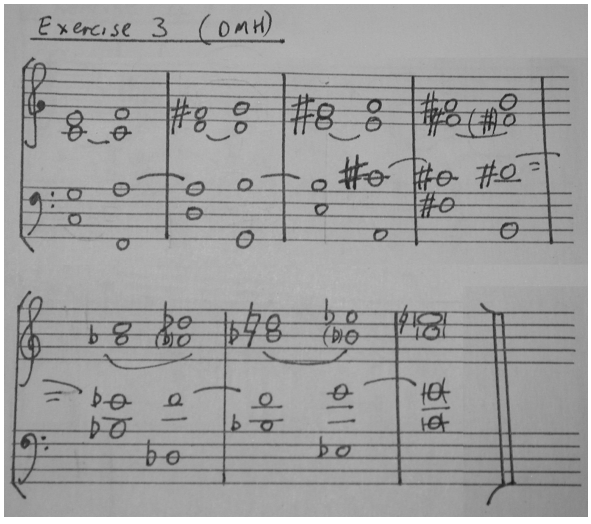


Figure 1 The score for exercise 3 used in the analysis.

The electrodes of each electrolaryngograph were placed on each singer's neck and the output Lx amplitude was checked with an oscilloscope to ensure optimum placement as indicated in [8]. Four cardioid microphones

were used (two AKG C480B with cardioid capsules, and two AKG C414 B-ULS set to cardioid), each positioned close to one of the singers. The eight signals were captured simultaneously on a Korg D1600 hard disk multi-track recorder.

The singers sang specially prepared exercises that were designed to change between chords using a pivot note; a note common to both chords that was held by one singer to provide a tuning reference for the other three singers. The chord-by-chord changes were therefore reasonably musically founded, since adjacent chords were constrained to having one note in common, but a number of different chords was visited during the relatively few bars of each exercise. Each exercise had a starting and ending chord that were either identical in note layout, or one octave apart in note layout. Figure 1 shows exercise3 which was used in this analysis.

Analysis was carried out using the 'spead' program from Laryngograph Ltd. [7]. This enables cycle-by-cycle f0 measurements to be made. The mode employed for these data was to plot f0 on the screen, delimit the appropriate note with the mouse, and right clicking provided access to a 'voice profile' analysis that includes mean f0 and its standard deviation for the selection made.

4. RESULTS

The f0 values obtained from 'spead' were input to an excel spreadsheet. In order to carry out an intonation analysis, one note is selected (in this case, the first note sung by the alto) to which all the others are related by their frequency ratio. In particular, the ratio between the target mean tuned and target equal tempered tuned notes are calculated for each note of the exercise. These results are converted to cents using (from [13]):

$$C = 3986.3137 \log_{10}(F1/F2) \qquad (1)$$

Where: c = cents, and F1 and F2 are the frequencies of the notes.

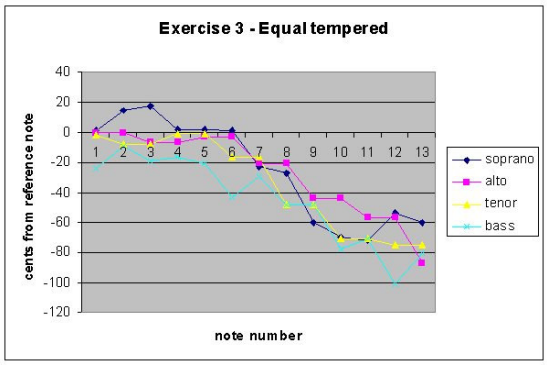


Figure 2: F0 variation with reference to equal tempered tuning in cents from the first alto note.

Figures 2 and 3 show the number of cents from the reference for all notes in the exercise based on equal tempered and mean tuning respectively. It can be seen that the equal tempered plot indicates that the exercise becames flat after the fifth chord and continues to flatten to the end, where the overall tuning has drifted by about three quarters of a semitone. When the relationship is made to mean tuning (figure 3), the intonation is slightly sharp throughout by a different amount in each part, but in part order, which suggests a stretching of the chords. The stability of the notes that serve to join the chord can be observed, particularly in the region of chords 5-8. Overall, there is some tendency to stabilise to a sharper intonation by approximately a quarter of a semitone.

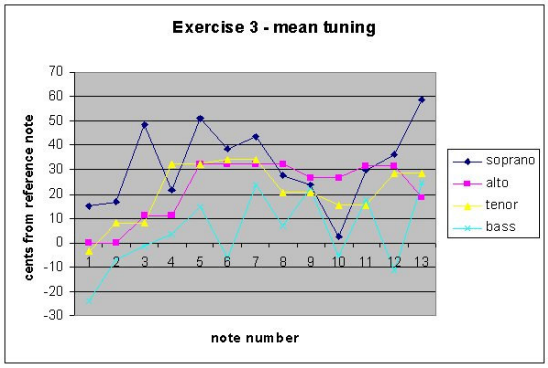


Figure 3: F0 variation with reference to mean tuning in cents from the first alto note.

5. DISCUSSION AND CONCLUSIONS

The notion that the underlying pitch of a piece does not waver is one commonly suggested in the literature and frequently cited by choral conductors and musical critics. For example, Padgham [9] suggests that singers must adjust their intervals to arrive back at the same note, and Pierce [10] good singers will "*maintain the initial key through using unjust intervals where these are acceptable*".

Bohrer [11], whose thesis finds evidence of variations in reference frequencies during a capella 4-part choral singing, cites a statement by Lloyd and Boyle [12] on the difficulty of following an equal temperament in performance: "*it is a superhuman feat to play 'off the note' with exactly the mistuning required for equal temperament, for we may be sure that the player has no physical means of reproducing equal temperament with the accuracy with which a good string quartet can play in tune, Even a piano tuner with an ear trained to measure the desired dissonance, and with ample time to listen to the beats he produces, does not tune the instrument perfectly in equal temperament. ... If the piano tuner does not succeed in attaining a perfect equal temperament how can the artist, playing a momentary note on an instrument of free intonation estimate exactly a dissonance which, as Helmholtz showed, must lack definition?" [9].* This puts another slant on the issue in terms of the task difficulty of maintaining equal tempered intonation in the absence of suitable accompaniment.

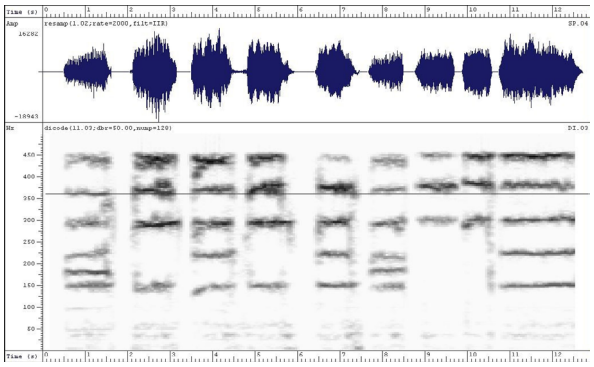


Figure 2: Waveform (downsampled to 1kHz) and narrow band (10Hz) spectrogram for the D major chords sang during the carol "The first nowell". A horizontal line has been plotted to enable comparison between the pitches of the chords.

A key question here is the extent to which it might happen in practice. One indication comes from the result of a recording of an a capella 4-part Christmas Carol ("The first nowell") that was made during the same recording session. Figure 4 shows the waveform (downsampled to 1kHz) for all of the D major chords sang during the carol and a narrow band (10Hz) spectrogram. The carol drifted sharp during the verse, and this can be observed with reference to the horizontal line plotted on the spectrogram. There are a number of potential reasons why a capella singers might drift in pitch of course, including wayward intonation, competing acoustic sounds, the acoustics of the space, and the preference of a strong singer to move toward particular pitch centres. Clearly, more work needs to be carried out to more fully understand the contribution, if any, that there might be from intonation strategies towards overall pitch drift in a capella singing.

6. ACKNOWLEDGEMENTS

The author thanks the members of the quartet for giving up their time to make these recordings.

7. REFERENCES

[1] Hagerman, B. and Sundberg, J. (1980). Fundamental frequency adjustment in barbershop singing, J Research in singing, **4**, 3-17.

[2] Hess, W. (1983). *Pitch determination of speech signals: Algorithms and devices*, Berlin: Springer

[3] Howard, D.M. (1998). Practical voice measurement, In: *The voice clinic handbook*, Harris, T., Harris, S., Rubin, J.S., and Howard, D.M. (Eds.), London: Whurr Publishing Company, 323-380.

[4] Baken, R.J. (1987). *Clinical Measurement of Speech and Voice*, London: Taylor and Francis.

[5] Hess, W., and Indefrey, H. (1984). Accurate pitch determination of speech signals by means of a laryngograph, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, **ICASSP-84**, 1-4.

[6] Howard, D.M. (1989). Peak-picking fundamental period estimation for hearing prostheses, *Journal of the Acoustical Society of America.*, **86,** (3), 902-910.

[7] www.laryngograph.com

[8] Abberton, E.R.M., Howard, D.M., and Fourcin, A.J. (1989). Laryngographic assessment of normal voice: A tutorial, *Clinical Linguistics and Phonetics*, **3**, (3), 281-296.

[9] .Padgham, C.A. (1986). The well-tempered organ, Oxford: Positif Press.

[10] Pierce, J.R. (1992). The science of musical sound, New York: Freeman.

[11] Bohrer, J.C.S. (2002). Intonational strategies in ensemble singing, University of London: PhD Thesis.

[12] Lloyd, L.S. and Boyle, H. (1978). Intervals, scales and temperaments, London: MacDonald and Jane's.

[13] Howard, D.M., and Angus, J.A.S. (2001). Acoustics and psychoacoustics, Oxford: Focal Press.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

LARYNX CLOSED QUOTIENTS IN A CAPELLA SATB QUARTET SINGING

David M Howard

Media Engineering Research Group
Department of Electronics, University of York, Heslington, York, YO10 5DD, UK
dh@ohm.york.ac.uk

ABSTRACT

The percentage of each vocal fold cycle for which the folds remain in contact, or larynx closed quotient (CQ), has been documented for trained and untrained adults and children. In all cases, patterned variations have been observed with singing training and experience. This paper describes an experiment to investigate larynx closed quotient distribution between each of the members of a four part (soprano alto tenor bass - SATB) vocal quartet.

Each member of an SATB a capella group was simultaneously recorded whilst each member wore an electrolaryngograph when singing carols and other items. The individual plots of CQ against the logarithm of fundamental frequency (f0) for each singer were combined into one. Whilst the plots for each of the four singers were distinct in both their CQ value as well as their pitch range (f0), the outputs for the four singers lay in a line in SATB part order. Any f0 separation is a function of the scoring, but ordering in terms of CQ values is not a result that would suggest itself from knowledge of the patterned variation of CQ and f0 with singing experience and training. One conjecture might be that this provides evidence of a strategy being employed for choral blend, differentiating the parts evenly in terms of their area of operation in terms of f0 and CQ

1. INTRODUCTION

The traditional teaching of singing is essentially a qualitative process that makes use of a technique that has been handed down from one generation of practitioners to the next over many decades. The success of this process relies primarily on the ears and perceptual skills of the teacher, who must guide the student singer towards a posture and production technique to enable the appropriate sound to be produced.

This process is supported by the knowledge and personal experience of the teacher, usually communicated during lessons through the use of imagery [e.g. 1]. This imagery may be in the form of 'psychological hooks' or concepts designed to promote the use of particular postural gestures such as: "sing on the point of the yawn", or "sing as if there is an orange stuck in the throat", or "sing through an imaginary hole in the forehead".

Whilst such psychological hooks clearly benefit many students in terms of enabling them to produce an appropriate vocal output, the physical reality of the voice production process itself is rarely being described by the hooks themselves.

Our quantitative knowledge of human voice production has made rapid advances during the latter part of the 20th century. These advances have accelerated in more recent times due to the widespread availability of office and home PC machines, which provide the level of computational facilities that were once only found in specialist speech science laboratories. Highly complex real-time audio processing is now possible on a standard multi-media PC machine, and the use of analysis tools as well as real-time visual displays is becoming increasingly widespread.

The usefulness of such displays for voice training depends on their being designed to provide data that are appropriate to the developing voice user, and on these data being presented in a form that is clear and readily understood by users. Quantitative research is therefore required to establish which parameters change during voice development and what the typical nature of that change might be.

One parameter which has been shown to vary with singing training/experience as a function of fundamental frequency (f0) is larynx closed quotient (CQ), which is defined as the percentage of each vocal fold vibrational cycle for which the folds are in contact. The experimental data gathered for CQ have suggested trends for adult males [2], adult females [3] and children [4], and a number of differences are observed. The purpose of this paper was to investigate the nature of CQ between different singers when they sing as a vocal quartet.

2. LARYNX CLOSED QUOTIENT MEASUREMENT

The measurement of CQ is based on the output from the electrolaryngograph [5], which enables vocal fold contact area to be monitored non-invasively. Two electrodes are placed superficially on either side of the neck of the subject at the level of the larynx and held in place with an elastic neckband. A constant amplitude high frequency voltage is applied enabling the electrolaryngograph to monitor the electrical impedance between the electrodes as the vocal folds vibrate by detecting the current flowing between them. The electrolaryngograph output waveform (Lx) therefore represents current flow between the

electrodes, and this will be greater as the vocal fold contact area increases and less as they part (see figure 1). Four phases can be defined during each cycle as indicated the figure. CQ is calculated as:

$$CQ = (cp/Tx)*100)\% \quad (1)$$

where: cp = closed phase; Tx = fundamental period.

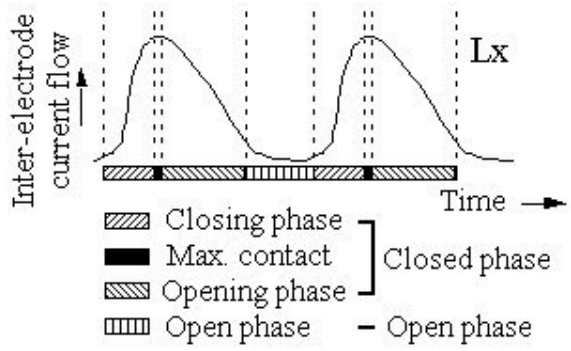


Figure 1: *A few idealised of the output waveform from the electrolaryngograph (Lx) indicating its main phases: closing, maximum contact, opening and open.*

CQ is measured on a cycle-by-cycle basis for each and every cycle of the Lx waveform, and the presentation of CQ data is normally in the form of a scattergram plot of CQ against the logarithm of f0 [6], known as Qx. No averaging or smoothing is applied and values for every Lx cycle are included in the resulting Qx plot. An example is shown in figure 2 for a two octave ascending and descending G major scale sung by a girl chorister. These plots have been used to provide the basis for the analysis of CQ variation with singing training.

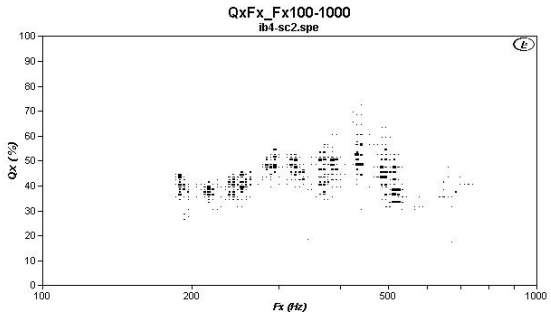


Figure 2: *Example CQ against log(f0) scattergram plot (Qx) for a two octave ascending and descending G major scale sung by a girl chorister.*

2.1. CQ variation with singing training

A highly statistically significant increase has been found [2] in the measured CQ with increasing numbers of years singing training/experience for adult male singers. These measures were based on a two octave ascending and descending G major scale. In all cases, CQ remained essentially constant with f0 (see summary plots in figure 3), and it was further noted that trained and experienced adult male singers exhibit CQ values when singing that are higher than their habitual speaking CQ, whereas untrained adult males display CQ values when singing that are lower than their customary speech values.

A later study [3] revealed the nature of CQ variation with f0 for adult female trained and untrained singers, which was very different to that obtained for adult males (see figure 2). Overall, the gradients of a plot of CQ against log(f0) for the adult females increase with singing training/experience. All the gradients tend to be positive in the lower octave and those for untrained subjects tend to be negative in the higher octave, with turning point occurs around G#4. Overall, CQ tends to be reduced for pitches below D4 and increased for pitches higher than B4 with training, and the CQ:log(f0) gradient within the pitch ranges: G3 to G#4, and B4 to G5 tends to correlate positively with the number of years singing training/experience.

More recent work has extended this study to include children [4], where the trained group consists of girl and boy choristers who sing the daily offices in English cathedrals. Their data exhibited very similar trends to that for adult females (see figure 2).

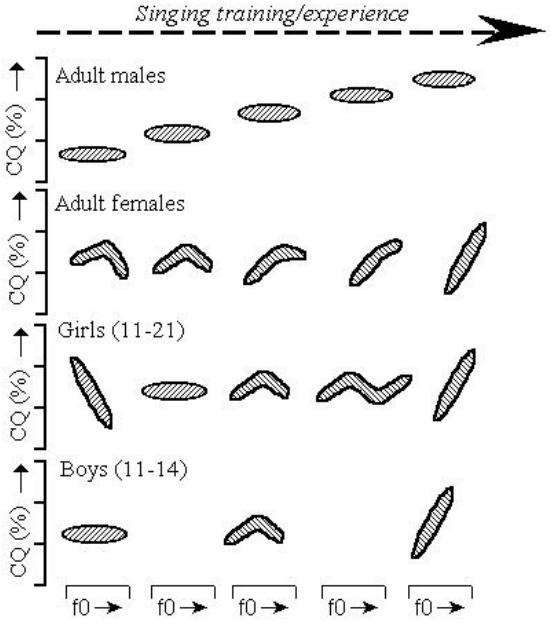


Figure 3: *Shape of variation of CQ against f0 with singing training/experience. (Each shape represents the region occupied by values on a CQ against log(f0) scatter plot.)* {Adult data from Howard et al., (1990) and Howard (1995), children data from Barlow (1999)}

2. EXPERIMENTAL DETAILS

For this experiment, a simultaneous recording of the Lx waveform from four electrolaryngographs along with the outputs from four microphones was required to capture acoustic and voice source data for each member of the SATB vocal quartet. Four field laryngographs from Laryngograph Ltd. [7] were employed that had been

specially prepared by checking that their operating radio frequencies were well separated, to ensure that they did not interfere with other electromagnetically. The four microphones (two AKG C480B with cardioid capsules, and two AKG C414 B-ULS set to cardioid) were each positioned close to one of the singers. The eight signals were captured simultaneously on a Korg D1600 hard disk multi-track recorder.

Prior to the recording, the Lx amplitudes were checked using an oscilloscope to ensure appropriate electrode positioning as described in [6], and the recording levels were set and monitored throughout the recording.

A soprano, alto, tenor, bass (SATB) vocal quartet took part in the experiment. All members were experienced singers, and three sang regularly together. The fourth (soprano) sang regularly in other choirs. At the start of the recording session, some time was spent singing together in order to become familiar with each other's sound, as well as the local acoustic (a domestic living room). For this recording, they sang two carols in four part harmony from [8]: "O little town of Bethlehem" (arranged by Ralph Vaughan Williams) and "The first nowell" (arranged by John Stainer). Both are well known and the harmonies are those customarily associated with the carols.

3. RESULTS

The Lx waveforms for each singer and each carol were analysed in terms of CQ and f0 using the "spead" and "QAnalysis" software from Laryngograph Ltd. [7]. The first point to notice is that the number of cycles analysed is different for each part, based on each part's pitch height (bass < tenor < alto < soprano). This is illustrated in the pie charts in figure 4. It is interesting to note that the percentages are very close for each carol, which suggests that the average musical interval (frequency ratio) between each part is similar.

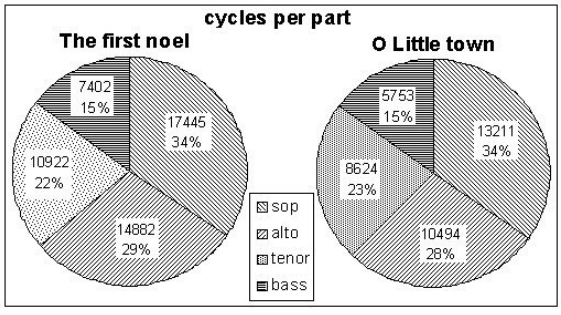


Figure 4: *Pie charts showing the number of cycles analysed for each part (SATB) for the two carols.*

The CQ values were analysed for each part on a cycle-by-cycle basis, and to facilitate comparison for b/w display, these are plotted in figure 5 as a separate Qx plot for each vocal part for both carols. It can be seen that the CQ values are highest in both cases for the bass, next highest for the tenor, lower for the alto, and lowest for the soprano. When these changes in CQ are coupled with the changes in f0 appropriate for each part, the four groups of points (one per singer) lie along a diagonal. Indeed, this is the way the data were first observed in a colour plot. Figure 6 shows these complete plots for each carol with lines to demark the plots for each vocal part.

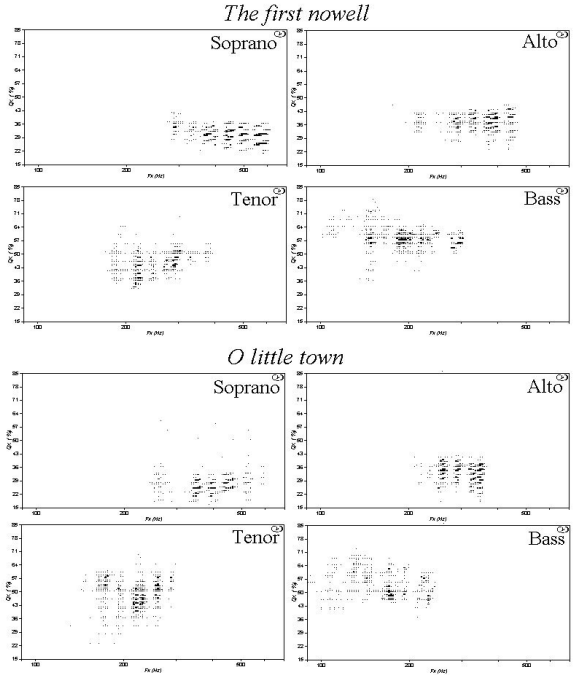


Figure 5: *Qx plots for each of the four parts (SATB) for "The first nowell" (upper) and "O little town" (lower).*

4. DISCUSSION AND CONCLUSIONS

Singing in a vocal quartet demands the ability to listen both to one's own output as well as that from each of the other members of the group. One of the key consequences of this process is in-tune (or otherwise!) singing (e.g. 9], but a vital aspect of a successful quartet sound is blend between the singers.

Choral blend is a function of many aspects of both voice production as well as the effects of the local acoustic. Ternström [10] brings out a number of issues relating to choral blend, including: the need for singers to match their loudness, pitch and voice timbre to the group average. He provides quantitative indicators as to the main ingredients of each as follows. Loudness matching relates to the ability to monitor appropriately the acoustic pressure level difference between one's own vocal output and that of the rest of the group. Pitch matching relates to control of f0 such that the output is heard as being "in-tune", but along with this is an observed scatter in f0 amongst singers singing in unison, which is an important contributor to choral blend. Timbre matching is based on vowel uniformity, and evidence suggests that singers modify the frequencies of their lower formants to be reasonably close.

The data provided by this experiment gives an indication of the values of CQ being employed by this quartet. The overall manner in which the points in the CQ against log(f0) scattergrams for each of the four singers

seem to 'fit' together is somewhat compelling, and indicative of some structural underpinning. One suggestion might be blend, in the sense that the individual singers are working together through acoustic feedback to set appropriately their vocal source characteristic alongside their matching of loudness, pitch and timbre.

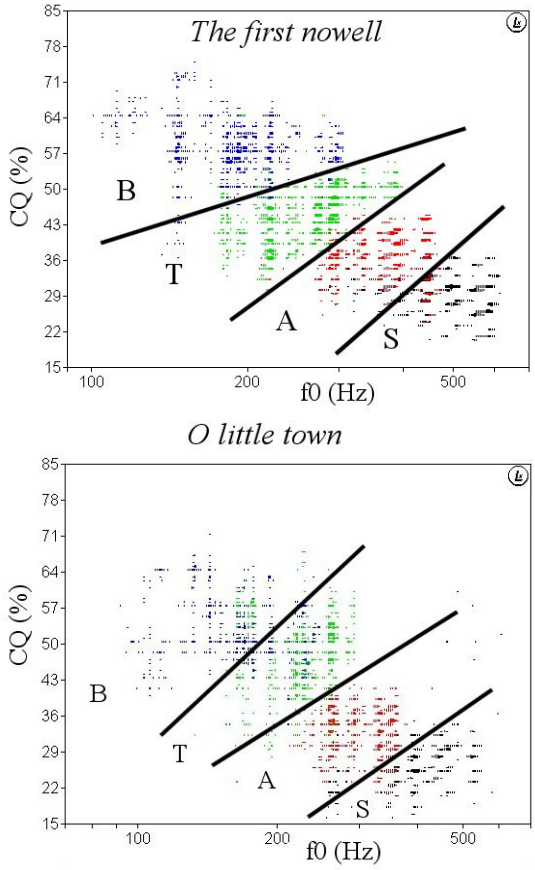


Figure 6: *Qx plots for carol with lines to demark the data points for each part for "The first nowell" (upper) and "O little town" (lower).*

It has been shown that CQ varies as a function of singing training [1-3], and that this variation is different for adult males and adult females. It has also been found that CQ tends to increase as a function of vocal intensity increase [11, 12]. These findings suggest that CQ can be rather readily varied, and while we may not have explanations that fully describe the acoustic effects of such CQ changes, it could be the organisation of CQ with f0 as indicated by these data is another facet of choral blend.

Considerable additional work needs to be carried out to support such a claim and to understand more fully the contribution that varying CQ makes to the acoustic output in different singing (and speaking) situations.

5. ACKNOWLEDGEMENTS

The author would like to thank the singers who gave their time to take part in this experiment.

6. REFERENCES

[1] Moorcroft, L. (2002). Embracing alternative methodologies: Science and imagery in the teaching and performance of singing, In: Proc 7th Int Conf Music Perception and Cognition, Stevens, C., Burnham, D., McPherson, G., Schubert, E., and Renwick, J. (Eds.), Adelaide: Casual Publications, 561-654.

[2] Howard, D.M., Lindsey, G.A. & Allen, B. (1990). Toward the quantification of vocal efficiency, J Voice, 4, (3), 205-121. [See also Errata (1991). J Voice, 5, 93-95.]

[3] Howard, D.M. (1995). Variation of Electrolaryngographically derived closed quotient for trained and untrained adult female singers, J Voice, 9, (2), 163-172.

[4] Howard, D.M., Barlow, C., and Welch, G.F. (2001a). Vocal production and listener perception of trained girls and boys in the English cathedral choir, Bul the Council for Research in Music Education, 147, 81-86.

[5] Fourcin, A.J., and Abberton, E.R.M. (1971). First applications of a new laryngograph, Medical and Biological Illustrated, 21, 172-182.

[6] Abberton, E.R.M., Howard, D.M. & Fourcin, A.J. (1989). Laryngographic assessment of normal voice: A tutorial, Clinical Linguistics and Phonetics, 3, 281-296.

[7] www.laryngograph.com

[8] Jenkins, N. (Ed.), (1993). Carol singer's handbook, Kevin Mayhew Ltd.

[9] Howard, D.M. (2003). A capella SATB quartet in-tune singing: evidence of intonation shift, Proc. Stockholm Music Acoustics Conference, **SMAC-03**, these proceedings.

[10] Ternström, S. (1991). Physical and acoustic factors that interact with the singer to produce the choral sound, J Voice, **5**, (2), 128-143.

[11] Orlikoff, R.F. (1991). Assessment of the dynamics of vocal fold contact from the electroglottogram: data from normal male subjects, J of speech and hearing research, **34**, 1066-1072.

[12] Sundberg, J., Cleveland, T.F., Stone, R.E., and Iwarsson, J. (1999). Voice source characteristics in six premier country singers, J Voice, **13**, (2), 168-183.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

THE EFFECT OF THE HYPOPHARYNGEAL AND SUPRA-GLOTTIC SHAPES ON THE SINGING VOICE

Hiroshi Imagawa [*1], *Ken-Ichi Sakakibara* [*2], *Niro Tayama* [*3], *Seiji Niimi* [*4]

[*1] Department of Speech Physiology, The University of Tokyo
7-3-1, Hongou, Bunkyo-ku, Tokyo, 113-0033, Japan
imagawa@m.u-tokyo.ac.jp

[*2] NTT Communication Science Laboratories
3-1, Morinosato Wakamiya, Atsugi-shi, Kanagawa, 243-0198, Japan
kis@brl.ntt.co.jp

[*3] International Medical Center of Japan
1-21-1, Toyama, Shinjuku-ku, Tokyo, 162-8655, Japan
ntayama@imcj.hosp.go.jp

[*4] Speech and Hearing Center, International University of Health and Welfare
2600-6, Kitakanemaru, Ohtawara, 324-0011, Japan
niimi@iuhw.ac.jp

ABSTRACT

The timbre of the singing voice is strongly affected by the shapes of the laryngeal tube and hypopharynx. We propose a physical model of the vocal tract and larynx, including the vocal and ventricular folds, for synthesis. We study the effect of the shapes of the laryngeal tube and hypopharynx on the synthesized voice using the proposed model.

We synthesized normal phonation, operatic singing, and throat singing voices by changing the hypopharyngeal and supra-glottic shapes and evaluated the acoustical effects of the various shapes of the hypopharynx and laryngeal tube. The results show that all the shapes of the hypopharynx, larynx tube, piriform fossa, and laryngeal ventricle play an important role in determining voice quality in singing.

1. INTRODUCTION

The voice quality in singing is determined by due to both the laryngeal voice and vocal tract shape. In most cases, the laryngeal voice is characterized by the vibratory pattern of the vocal folds. These vibratory patterns vary depending on the vocal registers (whistle, falsetto, modal, and vocal fry) and singing style (belting, opera, and so on). In throat singing (Khöömei, Köömij, Kai, etc.), the ventricular folds (also referred to as the false vocal folds) vibrate as well and their vibration is essential for the special timbres of the drone and kargyraa voices [3, 8, 11, 12].

The vocal tract shape also contributes to other aspects of voice quality besides phonetics. In particular, the contributions of the lower vocal tract (the larynx tube, piriform fossa, and hypopharynx) seem to be important in generating the special voice quality in singing. Previous fibrescopic observation have shown differences in the the lower vocal tract shape in various styles of singing. In operatic singing the hypopharynx and piriform fossa are relatively wide, and larynx tube is relatively narrow. The volume ratio matching of the larynx tube to the hypopharynx produces "the singer's formant" [13]. The belting voice is characterized by a narrow hypopharynx and piriform fossa and a narrow larynx tube [14]. Japanese Min-yoh is characterised by a narrow larynx tube and ventricular fold constriction [7].

In this paper, we investigate the production mechanisms of various timbres in singing voices using a physical model that allows the vibration of the ventricular folds. The model is realized by acoustical coupling of the tow-by-two-mass model and vocal tract with the piriform fossa and infraglottic cavity. We use the model to study the acoustic effect of the shape of the larynx tube (the laryngeal ventricle, ventricular fold, and upper part) and hypopharynx on the voice quality.

2. SYNTHESIS MODEL

The mechanism of the physical model for the synthesis of singing voices is depicted in Fig. 1. The laryngeal part is describedy by the 2×2-mass model, which represents the ventricular folds in a self-oscillating model as well as the vocal folds. The 2×2-mas model was is obtained by improving the two-mass model [2, 4, 10].

The vocal tract is represented as a lossless transmission line. The length of each section is set to $0.4\ \mathrm{cm}^2$. The ventricular folds and laryngeal ventricle are also assumed to be parts of the vocal tract. In the case that the ventricular folds vibrate, the area of the ventricular-fold section varies.

The effect of the nasal cavity and change of intra-pressure of the lungs are neglected for simplicity. Fig. 2 shows the equivalent circuit of our proposed model.

2.1. Subglottic region

The cross-sectional areas of the subglottal system are determined based on the anatomical data reported in [5]. We roughly approximate the subglottic region by 66 cylindrical sections each $0.4\ \mathrm{cm}$ long and calculate the acoustic characteristics by using the equivalent circuit as in Fig. 2. Let $y\ \mathrm{cm}$ be the distance from the glottis. Then, the areas $A\,\mathrm{cm}^2$ is determined as follows: $A = 2.5$ if $0.00 \leq y < 12.0$, $A = 2.0$ if $12.0 \leq y < 20.0$, $A = 5.0$ if $20.0 \leq y < 23.2$, and $A = 15.0$ if $23.2 \leq y < 26.4$.

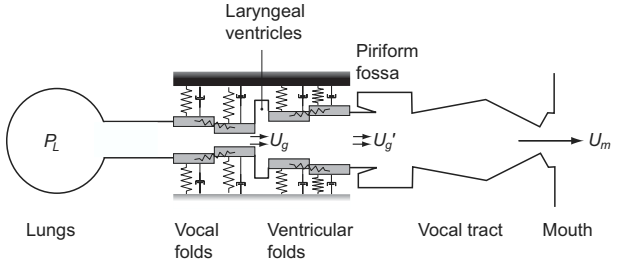


Figure 1: Physical model for synthesis of singing voices.

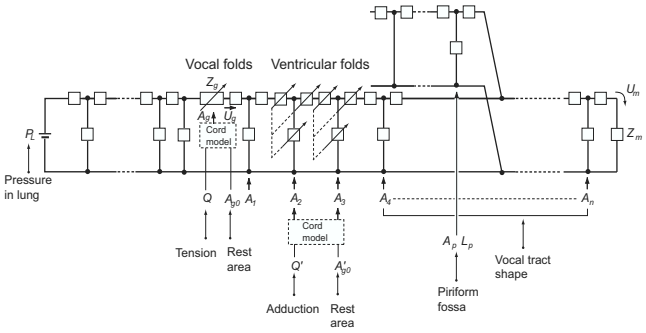


Figure 2: Equivalent circuit of the synthesis model.

2.2. Vocal folds

The vocal folds are represented as the two-mass model proposed in [4]. The F_0 of vocal-fold oscillation is controlled by changing the tension parameter Q [4]. For the initial setting of physical parameters, we use all of the normal values described in [4] except the rest glottal areas $A_{g01} = A_{g02} = 0.02\ \mathrm{cm}^2$.

2.3. Laryngeal ventricle

The laryngeal ventricle as a cylindrical section (A_1, l_1) such that $l_1 = 0.4$ cm. A_1 is set depending on the phonation type, but to 1.5 cm as the normal value. Note that even if the ventricular folds strongly constrict and contact as in throat singing, the space of the laryngeal ventricle is observed [9].

2.4. Ventricular folds

The ventricular folds contain few muscle fibres and, unlike the vocal folds, their physical properties essentially do not change. Therefore, it is meaningless to define a tension parameter for the ventricular folds. Hence, some other parametrization is necessary.

It is a physiological fact that the ventricular folds are adducted by the action of certain laryngeal muscles, such as the cricoepiglottic muscle and thyroepiglottic muscle [6], but it is unclear whether their physiological properties, such as mass and stiffness, are changed or not by the adduction.

We take into account the changing shapes of the ventricular folds and introduce an *adduction* parameter Q' for the ventricular folds, which is one possible parametrization for the stiffness, mass, and the false glottal area at rest [10]. We set the initial values of the parameters for the ventricular folds and laryngeal ventricles as follows: $d_1' = d_2' = 0.40$ cm, $m_1' = m_2' = 0.13$ g, $k_1' = k_2' = 50$ kdyn/cm, $\zeta_1 = \zeta_2 = 0.4$, $k_c = 80$ kdyn/cm.

The two different larygeal voice in throat singing, drone and kargyraa, are generated by coupling of the vocal and ventricular vibrations. In the drone voice, the ventricular folds vibrate in the same period as the vocal folds, and in the kargyraa voice, the ventricular fold vibrate in the integer multiple (usually double or triple) period of the vocal fold. The constriction of the ventricular folds at rest is strong for both cases. However, the constriction in the case of kargyraa is relatiely loose. Therefore, by changing the area between the ventricular folds at rest, the vibratory patterns of normal, drone, and kargyraa phonations can be simulated by using the 2×2-mass model [10].

In this paper, we also change the false glottal areas at rest A'_{g01} and A'_{g02} depending on the phonation type. The longitude of the ventricular fold is as follows: $L'_g = \min(0.7, \sqrt{A'_{g02}})$ cm.

2.5. Piriform fossa

The bilateral piriform fossa is assumed to be symmetric and, thereby, implemented as one cavity. From the result of the preliminary experiment, the acoustic characteristics are not significantly different between one-cavity and two-cavity cases.

The piriform fossa is implemeted as follows. First, according to the MRI data in [1], we assume each piriform fossa to be a cone whose depth is 2.0 cm and volume is $1.5\ \mathrm{cm}^2$ for each side. We represent the piriform fossa by five cylindrical sections $(A_1^p, l_1^p), \ldots, (A_5^p, l_5^p)$ with $l_i^p = 0.4$ cm. We determine $\{A_i^p\}_i$ to approximate the cone by $\sum A_i^p l_i^p = 3.0\ \mathrm{cm}^2$. We adjust the length of the cylinder portion above the arytenoid apex plane according to [1], and use the value 0.75 for the end correction coefficient. Finally, if these adjustments require us to extend the number of sections, we add necessary sections such that $A = A_5^p, l = 0.4$.

2.6. Vocal tract

The vocal tract is represented as a transmission line of n cylindrical hard-walled sections (A_n, l_n) with cross-sectional area A_n and length l_n. We assume that (A_1, l_1) is the region of the laryngeal ventricle with $l_1 = 0.4$ cm and that $(A_2, l_2), (A_3, l_3)$ are the spaces between the ventricular folds with $l_2, l_3 = 0.4$ cm. The ventricular folds are able to vibrate. We set $n = 43$. Hence, the length of vocal tract is 17.2 cm. For $k = 3, \ldots, 43$, each section has length $l_k = 0.4$ cm and variable A_k.

3. ACOUSTIC MEASUREMENT USING THE SYNTHESIS MODEL

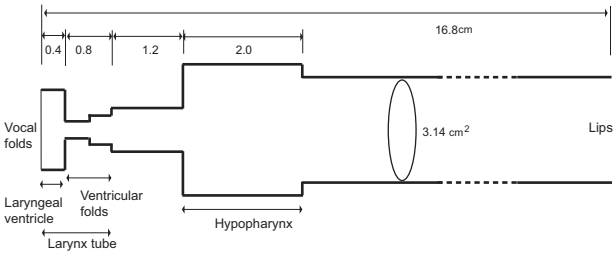


Figure 3: Sketch of the vocal tract model

We set the length of each part of the vocal tract as shown in Fig. 3. The length of the larynx tube is 2.4 cm, which includes the ventricular folds sections with the length of 8 cm and laryngeal ventricle section with the length of 4 cm. We also set the length of hypopharynx to 2.0 cm. We set $P_L = 5\ \mathrm{cm\,H_2O}$ in default and

attached the piriform fossa as shown in Fig. 2. The default areas were set to $1.5\,\mathrm{cm}^2$ for the laryngeal ventricles, $0.5\,\mathrm{cm}^2$ for the ventricular folds and remaining larynx tube section, and $3.14\,\mathrm{cm}^2$ for the hypopharynx and the upper vocal tract.

3.1. Effect of the larynx tube

We changed the areas of the larynx tube $A_2, \ldots, A_6$, excluding the laryngeal ventricle section. The spectral envelopes of the synthesized sounds are shown in Fig. 4 (LPC analysis, $p = 24$). With decreasing larynx tube area, F_2, F_3, F_4 move lower. If the area of larynz tube is greater than $0.05\,\mathrm{cm}^2$, the ventricular folds do not contact. In addition, F_5 moves slightly lower and gradually disappear and the power in the range greater than 4000 Hz decreases.

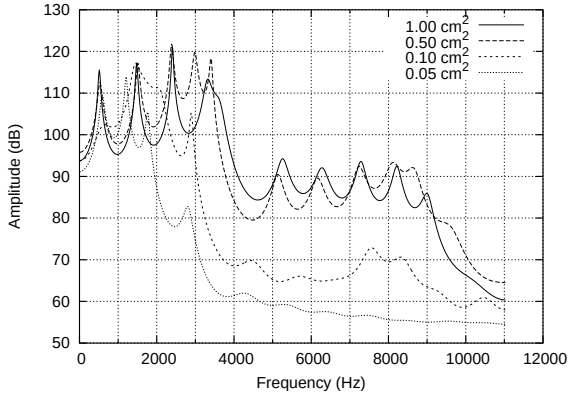


Figure 4: Acoustic effect of changing larynx tube area

3.2. Effect of the hypopharynx

Fig. 5 shows the spectral envelopes of the synthesized sounds when we changed the hypopharynx area. The are shown in F_3 and F_5 move close to F_4. In some range, $A = 3$ or $6\,\mathrm{cm}^2$ in Fig. 5 , the formant cluster of F_3, F_4, F_5 is observed around 3000 Hz.

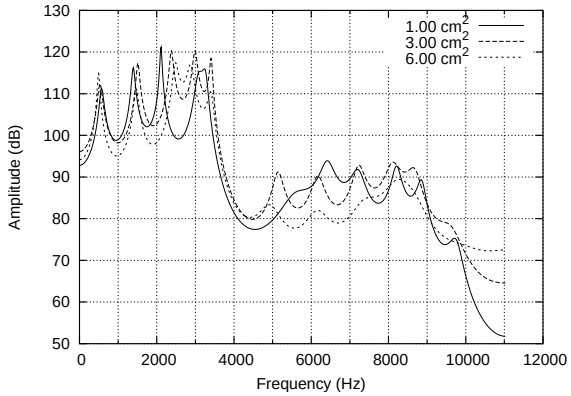


Figure 5: Acoustic effect of changing hypopharynx area.

3.3. Effect of the laryngeal ventricle

The results are shown in Fig. 6, when the laryngeal ventricle area was changed. As the area increases, F_3, F_4 move lower and are pushed close to F_2 and F_2 is sharpened. The effect of a zero in the range from 4000 to 5000 Hz also becomes larger.

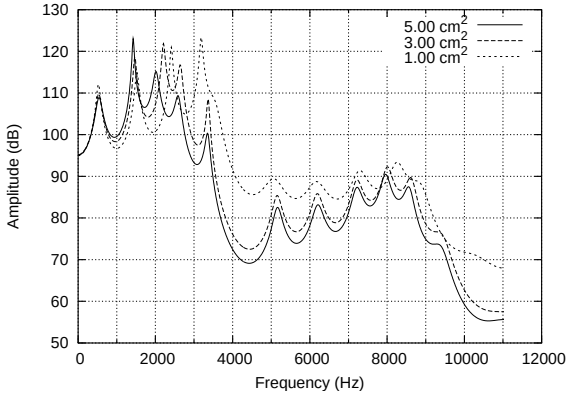


Figure 6: Acoustic effect of changing laryngeal ventricle area.

3.4. Effect of the piriform fossa

The spectral envelope of synthesized sounds are shown in Fig. 7 for changes in piriform fossa volume. As the volume is increased, the effect of a zero around $4500\,\mathrm{Hz}$ becomes larger. As pointed out in [15], inceasing the volume of the piriform fossa repels the formants, i.e. F_1, F_2, F_3, and F_4 are pushed lower and so is F_5.

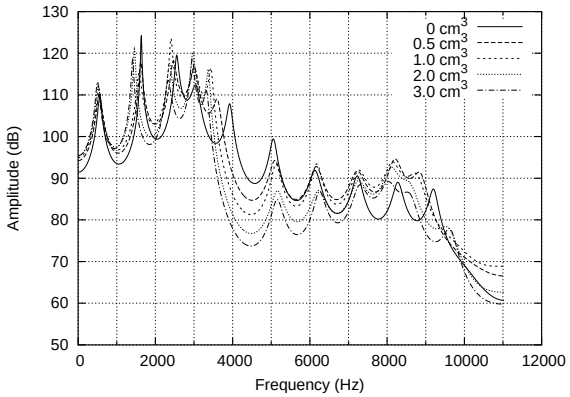


Figure 7: Acoustic effect of changing piriform fossa volume.

4. SINGING VOICE SYNTHESIS

Based on the results of the acoustic measurements of the effects of varying the shape of the larynx tube, hypopharynx, laryngeal ventricle, and piriform fossa, we chose the nominal settings of the parameters for normal phonation, operatic singing, and throat singing (Table 1). In the case of operatic singing, we assume that the lowest four sections in the hypopharynx have constant areas $A_7 = A_8 = 5.0\,\mathrm{cm}^2$ and $A_9 = A_{10} = 5.0\,\mathrm{cm}^2$ to maintain the large volume ratio of the hypopharynx to the larynx tube. In throat singing, by controlling the adduction parameter Q', we synthesized both drone and kargyraa phontaions. We set the initial value of the A_2, A_3 to $0.04\ \mathrm{cm}^2$, $Q' = 1$ for drone, and $Q' = 0.55$.

Table 1: Areas of the sub-glottic region. LV: laryngeal ventricle, VTF: ventricular fold, ULT: upper larynx tube

Style	LV: A_1 cm^2	VTF: A_2, A_3 cm^2	ULT: A_4, A_5, A_6 cm^2
Normal	1.5	0.5	1.0
Opera	1.5	0.5	0.5
Throat	2.0	Given by Q'	0.4

No signifi cant differences were observed between the spectral envelopes of the two phonations. In Fig. 8, the spectral envelope of the drone voice is shown.

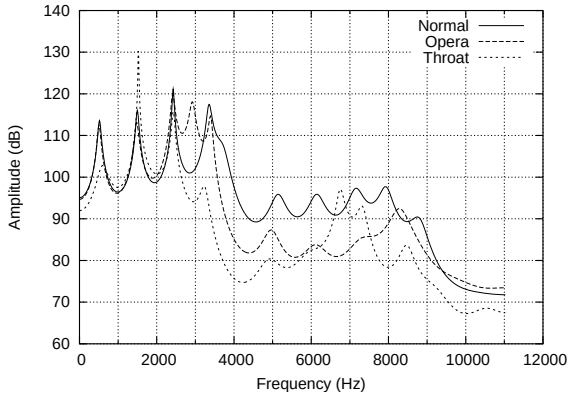


Figure 8: Singing voices with neutral vocal tract.

The spectral envelopes are show in Fig. 8. The synthesized operatic singing voice have the singer's formant around 3000 Hz. In the synthesized throat singing voice, F_2 is sharp and the power in the range from 4000 to 6000 Hz is relatively low. The sharpness of the F_2 is the effect of the laryngeal ventricle resonance and contributes to the generation of the whistle-like tone.

5. CONCLUSIONS

We studied the acoustic effect of the shape of the larynx tube and hypopharynx. The acoustic characteristics of the synthesized singing voices reveal good accordance with known results for operatic and throat singing. Our results show that the dimensions of the laryngeal ventricle, larynx tube, hypopharynx, and piriform fossa play an important role in determining voice quality in singing.

For operatic singing, relatively narrow larynx tube and wide hypopharynx effect to creat singing formant as reported previous studies [13, 15]. The piriform fossa also contribute to the cluster of F_3, F_4. For throat singing, strong constriction of the ventricular folds sharpens F_2 and produces the special laryngeal voices by vibration of the ventricular folds as well.

Most previous studies assumed that the larynx tube has uniformal area. However, our results reveal that the laryngeal ventricle volume and the ventricular fold constriction also affect the acoustic characterisitcs of the singing voices.

In this study, we did not consider the mechanical interaction of each space in the lower vocal tract. However, the physiological mechanism determining the shape of the lower vocal tract is very important for fi nding appropriate parameters to synthesize various voice qualities. In particular, layngeal height is very important in determining the confi guration of the lower vocal tract. The activation of the thyropharyngeal muscle, cricopharyngeal muscle, and paratopharyngeal muscle are also important for regulating the shape of the hypopharynx and the intrisic laryngeal muscles are important for constricting the larynx tube [6]. Further investigations by using EMG and MRI are needed in order to understand their physiology and, thereby, controll the parameters of the synthesis model.

Acknowledgments

We would like to thank Seiji Adachi, Kiyoshi Honda, Emi Z. Murano, and Sotaro Sekimoto for their helpful discussions.

6. REFERENCES

[1] J. Dang and K. Honda. Acoustic characteristics of the piriform fossa in models and humans. *J. Acoust. Soc. Am.*, 101(1):456–465, 1997.

[2] J. L. Flanagan, K. Ishizaka, and K. L. Shipley. Synthesis of speech from a dynamical model of the vocal cords and vocal tract. *Bell Systems Tech. J.*, 54:485–506, 1975.

[3] L. Fuks, B. Hammarberg, and J. Sundberg. A self-sustained vocal-ventricular phonation mode: acoustical, aerodynamic and glottographic evidences. *KTH TMH-QPSR*, 3/1998:49–59, 1998.

[4] K. Ishizaka and J. L. Flanagan. Synthesis of voiced soudns from a two-mass model of the vocal cords. *Bell Systems Tech. J.*, 51(6):1233–1268, 1972.

[5] K. Ishizaka, M. Matsudaira, and T. Kaneko. Input acoustic-impedance measurement of the subglottal system. *J. Acoust. Soc. Am.*, 60(1):190–197, 1976.

[6] M. Kimura, K.-I. Sakakibara, H. Imagawa, R. Chan, S. Niimi, and N. Tayama. Histological investigation of the supra-glottal structures in human for understanding abnormal phonation. *J. Acoust. Soc. Am.*, 112:2446, 2002.

[7] N. Kobayashi, Y. Tohkura, S. Tenpaku, and S. Niimi. Acoustic and physiological characteristics of traditional singing in Japan. *Tech. Rep. IEICE*, SP89-147:39–45, 1990.

[8] T. C. Levin and M. E. Edgerton. The throat singers of tuva. *Scientific America*, Sep-1999:80–87, 1999.

[9] V. T. Maslov. Functional peculiarities of the larynx during the vocal formation in Tuva two-voice singing. *Vestn. Otorinolaringol.*, Mar.–Arp.:58–61, 1979. in Russian.

[10] K.-I. Sakakibara, H. Imagawa, S. Niimi, and N. Osaka. Synthesis of the laryngeal source of throat singing using a 2×2-mass model. In *Proc. ICMC 2002*, pages 5–8, 2002.

[11] K.-I. Sakakibara, T. Konishi, H. Imagawa, E. Z. Murano, K. Kondo, M. Kumada, and S. Niimi. Observation of the laryngeal movements for throat singing — vibration of two pairs of the folds in human larynx. *Acoust. Soc. Am. World Wide Press Room*, 144th meeting of the ASA, 2002. `http://www.acoustics.org/press/`.

[12] K.-I. Sakakibara, T. Konishi, K. Kondo, E. Z. Murano, M. Kumada, H. Imagawa, and S. Niimi. Vocal fold and false vocal fold vibrations and synthesis of khöömei. In *Proc. ICMC 2001*, pages 135–138. ICMA, 2001.

[13] J. Sundberg. *The science of the singing voice*. Nothern Illinois Univ. Pr., 1989.

[14] J. Sundberg, P. Gramming, and J. Lovetri. Comparisons of pharynx, source, formant and pressure characteristics in Operatic and musical theatre singing. *J. Voice*, 7(4):301–310, 1993.

[15] I. R. Titze and B. H. Story. Acoustic interactions of the voice source with the lower vocal tract. *J. Acoust. Soc. Am.*, 101(4):2234–2243, 1997.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

THE EFFECT OF VOCAL VIBRATO ON BLENDING INDIVIDUAL SINGER CHARACTERISTICS WITH HALL ACOUSTICS

Kosuke KATO, Dennis NOSON**, and Yoichi ANDO*

Graduate School of Science and Technology Kumamoto University,
2-39-1 Kurokami, Kumamoto, 860-8555, JAPAN.
kkato@kobe-u.com

BRC Acoustics Inc., 1741 First Ave. S., Suite 401, Seattle, WA 98134. U.S.A.
dnoson@brcacoustics.com

ABSTRACT

New approaches are being made in the study of recital and opera singers and the response of the listener and performer to the acoustics of the performing environment. Significant progress into characterizing the optimum performing environment have followed from singer studies which utilize the effective duration (τ_e) of the running autocorrelation function (ACF) - i.e. the 10-percentile decay of the ACF envelope- as a key parameter for evaluating the musical source signal. A close correlation of the minimum value of the running τ_e, $(\tau_e)_{min}$, with the subjective response of both listeners and performers to time-varying sound fields has been confirmed [Y. Ando, Architectural Acoustics, AIP/Springer-Verlag, New York (1998)]. The investigation of the relation between the values of τ_e and musical performing styles provides a basis for supporting mutual understanding of singers and acousticians. For example, Kato and Ando reported values of $(\tau_e)_{min}$ for *falsetto* which are longer than that for operatic singing style in vocal performances [Journal of Sound and Vibration **25** , 463-472 (2002)]. This new study attempts to evaluate individual singer characteristics by analyzing running τ_e of the singing voice of twelve singers (3 sopranos, 3 altos, 3 tenors, and 3 bases). Significant findings are (1) that the values of $\log_{10}(\tau_e)$ differ greatly for each singer and can be said to be expressive of the individual singer s character , and (2) that the values of $\log_{10}(\tau_e)$ have a negative correlation with individual vibrato extent, presumably guiding the singer s preference for stage sound field conditions, and listener s preferences as well.

1. INTRODUCTION

Opera houses and concert halls have unique visual, spatial, and temporal characteristics. Architects and acousticians may design a hall which expresses personal design concepts, thereby leaving a message in the largest of all musical instruments, the sound field. As for sound sources, musical instruments are characterized by construction methods of their builders, the styles of the instrument tuners, or by inherent nature, and all are distinctive. In particular, singers are quite unique instruments, using their natural body as a musical instrument. Performers have individual interpretations of styles of music, layered over the intent of musical composers, and communicate their own musical sensitivity in performance. In order to respect and integrate the individual contributions to musical performance, it is essential to support the mutual understanding of architects, acousticians, manufacturers, tuners, performers, and composers with each other. Significant effort has been expended in the last 50 years to identify the acoustical characteristics of concert hall and opera house sound fields for the *audience* which enhances the listening experience for any given musical program. More recently, increased attention has been paid to the needs of performers and to the nature of the musical sound source itself, which can influence listener and performer preference for one sound field over another.

The decay characteristic of the autocorrelation function (ACF) of a sound source has been shown to be useful in an analysis of the subjective preference of sound fields [1]. The effective duration of running ACF (abbreviated as r-ACF hereafter), symbolized as τ_e, is a temporal parameter of a given sound source that is strongly related to the preferred temporal conditions for both listeners [1] and performers [4]. In the context of these studies, it is important to understand the τ_e characteristics of music signals from different musical instruments. Taguti and Ando (1997) have shown that tempo, articulation, and damper pedaling are the main elements for determining τ_e of piano music [2]. However, no systematic attempt has been made to examine variations in τ_e for the singing voice and changes due to singing style.

Studies of the human singing voice have been mainly focused on the mechanism of voice source production, voice directivity, and power spectrum of the voice source. A small number of studies have dealt with the preferred sound fields for vocal performers [3, 4]. Singer preference studies that consider time domain parameters of the sound source are rare. A single study has examined the relationship of singer sound field preference to the temporal parameter, τ_e, of the voice source signal [4].

In order to investigate τ_e for vocal music sound sources, our previous study dealt with singing styles [5]. Using simple motifs, with five solo singers (tenor), this study revealed that the singing style affects the minimum value of τ_e, i.e. $(\tau_e)_{min}$. Values of $(\tau_e)_{min}$ for voice sources for singing *falsetto* or medium *falsetto* are much longer than those for operatic singing. Many other variations to the singing voice are possible: differing motif lyrics, stretching note value or adjusting music tempo, varying the degree of music interval, changes in dynamics, such as crescendo and decrescendo, various kinds of articulation, such as *staccato*, *legato* (or *tenuto*), *marcato*, *sfor ando*, etc., and others such as vibrato or *tremolo*. The nature of vocal source is also be affected by the performer s voice classification and the feedback from the sound field. This study focuses on (1) motif s lyrics, (2) music tempo, (3) the degree of variation of music interval in the melodic line, and (4) voice classification, all of which are considered to have likely potential for producing a larger variation in the value of τ_e.
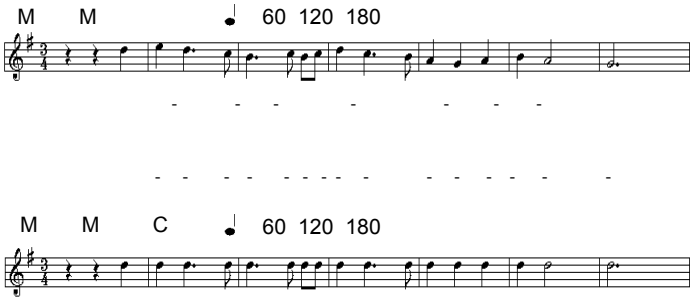


Figure 1: *Music scores of the two motifs used in the experiment.*

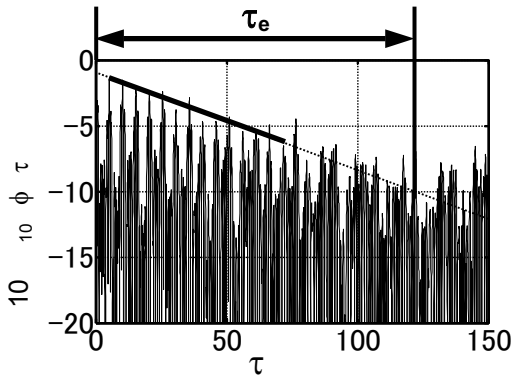


Figure 2: *Example of determining the effective duration of the running ACF.*

2. PROCEDURE

The power spectrum of the voice differs with the voice classification. Since the autocorrelation function and the power spectrum mathematically contain the same information, we expected that the voice classification may affect the value of τ_e. To examine the effect of the voice classification, we selected 12 singers (3 sopranos: KH, YT, CT 3 Altos: KT, AT, YA 3 tenors: RK, HS, KK and 3 basses: NH, DT, AH).

Figure 1 shows the two musical scores used in this experiment. Motif I (FP) from a traditional Italian song *ia il sole dal gange* consists of eighth notes and longer, arranged in horizontal intervals of major and minor seconds, with no indication of expression, such as *staccato*, stress accent, etc. which was also used in our previous study [5]. In order to examine the difference in τ_e when the motif is sung in different lyrics, we conducted an analysis of τ_e of the sound source with the 12 singers for a music motif. Singers were asked to perform motif I (FP) in *legato* style with three different lyrics: 10 times with real lyrics, 10 times with la syllables, and 10 times with humming.

Motif I (FP) was arranged to motif I (CP), featuring no horizontal pitch interval (constant pitch). In order to systematically examine the difference of the τ_e when the motifs were sung in different music tempos, and to examine the effect of the presence of the music interval in the melodic line we conducted an analysis of τ_e of the sound source, 12 singers for two music motifs. Singers were asked to sing both motif I (FP) and motif I (CP) in *legato* style with la syllables. Singers sang the two motifs with three different music tempos: ♩ = 60, 120, and 180. For the ease of performing, we adjusted the range of notes for each singer. The soprano singer s start note was D5, the alto singer G4, tenor singer D4, and bass singer G3.

The music tempo was maintained by the aid of a visual metronome located at a 1.0 m distance in front of the singer. The singer s voice was received by a none-directional condenser microphone located approximately 25 cm in front and 5 cm to the side of the singer s mouth. Signals were sampled at 44.1 kHz. Singers performed in an anechoic room.

The r-ACF is obtained through a chosen length of time window that may be considered to correspond to the psychological present. The integration interval $2T$ is recommended to be set at around $30(\tau_e)_{min}$ [6]. For the purpose of comparing the results obtained through different integration intervals, we selected three integration intervals ($2T$ = 0.5 s, 1.0 s, and 2.0 s) around the recommended interval. As shown in Figure 2, the envelope decay of the initial part of the r-ACF can be fitted by a straight line in the 5 dB range from the first major peak. The effective duration τ_e is obtained from the decay rate extrapolated at 10 dB.

3. RESULTS

Figure 3 shows examples of (a) real waveform (b) measured τ_e values of the r-ACF of a music signal, after passing through an *A*-weighted network with a 100-ms stepping interval as a function of time and (c) normal probability plots for their values of τ_e. The values of $\log_{10}(\tau_e)$ were distrubuted approximately as a normal probability distribution, and the mean and variance were unique for each singer.

In order to examine the results statistically, $(\tau_e)_{min}$, $(\tau_e)_{-1SD}$, $(\tau_e)_{MEAN}$, $(\tau_e)_{+1SD}$, and $(\tau_e)_{MAX}$ were extracted. Let $(\tau_e)_{MAX}$ be the maximum value of a running τ_e. Since the values of $\log_{10}(\tau_e)$ are approximately normally distributed, the $(\tau_e)_{-1SD}$, $(\tau_e)_{MEAN}$, $(\tau_e)_{+1SD}$ are written as

$$(\tau_e)_{-1SD} = 10^{\bar{} - \sigma}, (\tau_e)_{MEAN} = 10^{\bar{}}, (\tau_e)_{+1SD} = 10^{\bar{} + \sigma} \qquad (1)$$

where = $\log_{10}(\tau_e)$, $\bar{}$ is the mean value of , and σ is the standard deviation of . These statistics are shown in Figure 4.

The changes of $\log_{10}(\tau_e)_{min}$ due to the integration intervals at each sample were on average 12 of the mean values. The values of $(\tau_e)_{-1SD}$ were more independent of the integration interval than the other values. This study mainly focuses on $\log_{10}(\tau_e)_{min}$, which is the most active part of music and which has a close correlation with subjective responses to the sound field [1, 4, 6]. The figures for the sessions using different lyrics (Figure 4 - (a)-(c)) indicate that values of $\log_{10}(\tau_e)_{min}$ for hum singing at each integration interval are significantly longer than those for singing with la syllables or singing with lyrics ($p < 0.01$, using a two sample t-test) and that the difference in $\log_{10}(\tau_e)_{min}$ at each integration interval between the case of using lyrics and the case with la syllables was not significant. The results for the session with different music tempos (Figure 4- (d)-(i)) indicate that the difference in $\log_{10}(\tau_e)_{min}$ of each singer at each integration interval, among three music tempos, are not significant (using a two sample t-test) and that the values of $\log_{10}(\tau_e)_{min}$ for motif I (CP) were significantly longer than those for motif I (FP) ($p <$ 0.01, using a two sample t-test).

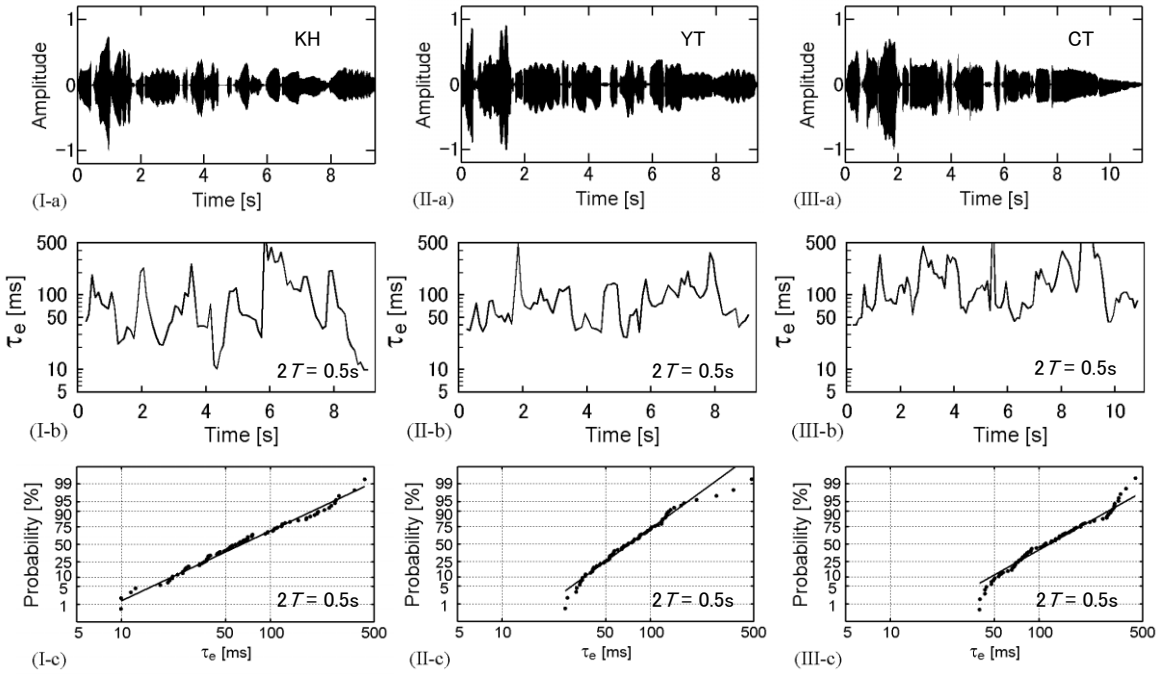


Figure 3: *Examples of (a) real waveform (b) measured τ_e values of the r-ACF of a music signal, after passing through an A-weighted network with a 100-ms stepping interval as a function of time for Motif (FP), with ♩ = 120, with real lyrics and (c) normal probability plots for their values of τ_e: 2T 0.5s performed by three sopranos:) Singer ,) Singer T,) Singer CT.*

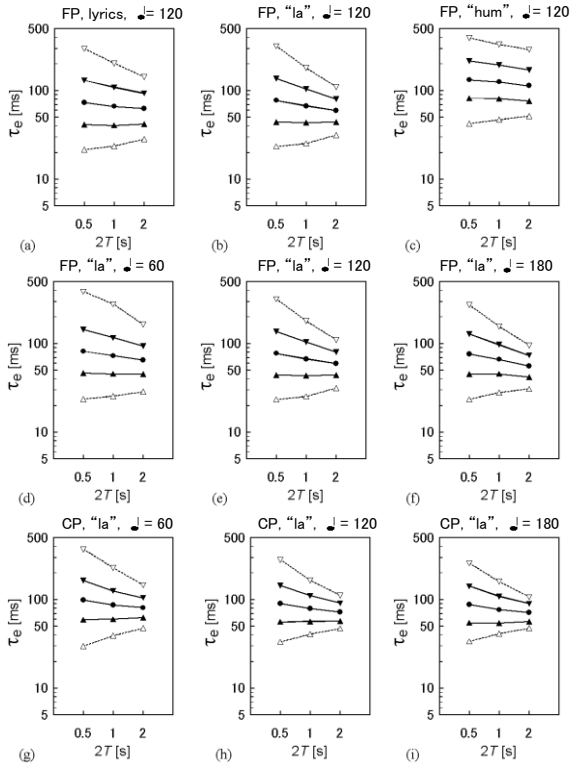


Figure 4*: eometric mean values of the τ_e statistics obtained from 120 vocal sources (12 singers x 10 trials) for motif (FP) and motif (CP) sung by different lyrics or music tempo.*

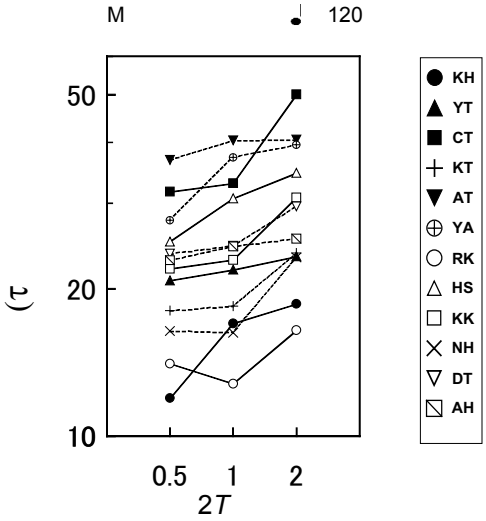


Figure 5*: eometric mean values of (τe)min of 10 trials of individual singers for motif (FP) sung by real lyrics as a function of 2T.*

Figure 5 shows the geometric mean values of $(\tau_e)_{min}$ obtained from 10 trials of each singer for motif I (FP) sung with lyrics Values of $(\tau_e)_{min}$ greatly varied with individual singers, instead of depending on voice classifications. The average standard deviation of $\log_{10}(\tau_e)_{min}$ in 10 trials of each singer, which expresses the reproducibility of his or her performance, ranged between 5.4 and 9.8 of the mean values.

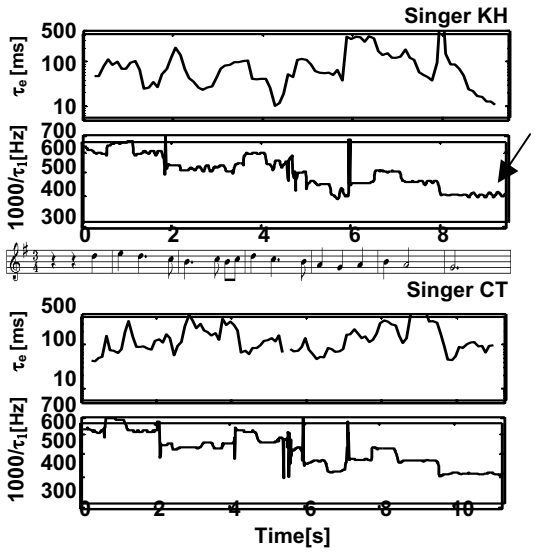


Figure 6: *Comparisons of (a) measured τ_e values of the r-ACF with 100-ms stepping interval (2T 0.5 s), (b) measured τ_1 values of the r-ACF with 10-ms stepping interval (2T 0.1 s) for two sopranos.*

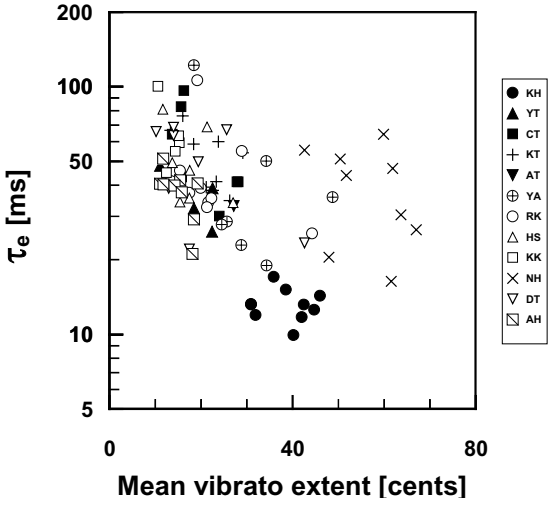


Figure 7: *Relationship between τ_e and vibrato extent. Negative correlation between $\log_{10}(\tau_e)$ and vibrato extent. (r 0. , p 0.001)*

4. DISCUSSION

The results of Figure 5 showed that the values of $(\tau_e)_{min}$ vary greatly for each singer, rather than varying in relation to the voice classification of the singers. This is an interesting, since there remains potential for singers to control the values of $(\tau_e)_{min}$ by adjusting his or her own individual interpretation styles, even after the music program (score) is confirmed.

Here, we hypothesize that vibrato may be one of the vocal expressions which determine individual singer characteristics, and which may have a strong relationship to the τ_e of the singer s vocal style. This study defines vibrato as the fluctuation of the delay time (τ_1) of the first major peak of the r-ACF ($2T$ = 0.1s). Figure 6 shows comparisons of (a) measured τ_e values, (b) measured τ_1 values for two sopranos. The smaller τ_e values and strong fluctuations of τ_1 were typically observed at the end of the performance of singer KH. For evaluating these fluctuations of τ_1, we calculated the vibrato extent using Prame s method [7].

Figure 7 shows the relationship between the measured τ_e values and the measured mean vibrato extent, obtained from the last 0.5 s portion of the session, singing with lyrics. After eliminating data with a small vibrato extent (less than 10 cent), a negative correlation between $\log_{10}(\tau_e)$ and mean vibrato extent was observed. Therefore, the control of vibrato extent may contribute to changes of temporal impression of sound field, such as echo perception [1]. It is noteworthy, for example, that the values of τ_e for singer KH are much shorter than those for the other singers. This indicates that a professional singer KH produces an esthetic vibrato with a high degree of expressive information.

We hope that improving the understanding of a singer s vocal signal will help communication between musicians and acousticians. For example, acoustical designers may provide for adjustable reflection panels, allowing singers an additional control of their performance, to enhance the uniqueness of the performance sound field, all to the benefit of the aesthetics of music program, and the individual interpretation style of singers.

5. CONCLUSIONS

The present investigation showed the followings:

(1) The values of $\log_{10}(\tau_e)_{min}$ for humming singing are longer than those for singing with la syllables or singing with real lyrics ($p<0.01$).

(2) Variations in music tempo do not have significant effect on the value of $\log_{10}(\tau_e)_{min}$.

(3) The values of $\log_{10}(\tau_e)_{min}$ for passages with no music interval sung with la syllables are longer than those with musical (variable) pitch intervals ($p<0.01$).

(4) The values of $\log_{10}(\tau_e)_{min}$ vary greatly for each singer, rather than varying in relation to the voice classification of the singers.

(5) The values of $\log_{10}(\tau_e)$ have a negative correlation with vibrato extent at the end of a long tone ($r = 0.47$, $p < 0.001$).

. REFERENCES

[1] Ando, Y., Architectural Acoustics-Blending Sound Sources, Sound Fields, and Listeners. AIP/Springer-Verlag, New York, 1998.

[2] Taguti, T. and Ando, Y., Characteristics of the short-term autocorrelation function of sound signals in piano performances , in Music and Concert Hall Acoustics., Chapter 23. Academic Press, London, 1997.

[3] Ternst m, S., Preferred self-to-other ratios in choir singing , J. Acoust. Soc. Am., Vol. 105, 3563-3574, 1999.

[4] Noson, D., Sato, S., Sakai, H., and Ando, Y., Melisma singing and preferred stage acoustics for singers , Journal of Sound and Vibration, Vol. 258, 473-485, 2002.

[5] Kato, K. and Ando, Y., A study of the blending of vocal music with the sound field by different singing styles , Journal of Sound and Vibration, Vol. 258, 463-472, 2002.

[6] Mouri, K. and Akiyama, K., Preliminary study on recommended time duration of source signals to be analyzed, in relation to its effective duration of the auto-correlation function , Journal of Sound and Vibration, Vol. 241, 87-95, 2001.

[7] Prame, E., Vibrato extent and intonation in professional Western lyric singing , J. Acoust. Soc. Am., Vol. 102, 616-621, 1997.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

SINGER IDENTIFICATION AND TRANSFORMATION THROUGH DYNAMIC MODELING OF VOCAL FOLD AND VOCAL TRACT PARAMETERS

Youngmoo E. Kim

MIT Media Lab
Cambridge, MA USA
moo@media.mit.edu

ABSTRACT

Oftentimes when we listen to a familiar singer, the unique qualities of that performer's voice allow us to establish the singer's identity with relative ease. It is believed that the unique acoustic qualities of an individual singer's voice arise from a combination of innate physical factors (e.g. vocal tract and vocal fold physiology) and individual characteristics of performance and expression (e.g. pronunciation and accent). In this research, we jointly estimate pole-zero filter and LF glottal waveform model parameters to model the shape of the vocal tract and the glottal excitation, respectively over short time periods. These time-varying parameters, corresponding to the physical characteristics of the singer, are used to train a Hidden Markov Model (HMM). The HMM is used to model the dynamic behavior of the source-filter parameters, corresponding to some of the expressive characteristics of the singer. We propose a system that is able to identify singers based upon the model of greatest likelihood among the individually trained HMMs. We also explore the use of individual HMM states as a method of mapping from the parameters of one singer to another in a preliminary attempt at singing voice transformation. The data used in this analysis was recorded from four conservatory-level classically trained singers.

1. INTRODUCTION

Individual voices tend to be easily distinguished and thus reflect the identity of an individual. We are undeniably sensitive to the unique qualities of voices and can perceive those qualities after listening to a voice for just a short time. This is because the vocal apparatus, while being extremely complex and flexible, is also highly self-consistent. The distinctive properties of the voice are believed to be a combination of physiological factors (e.g. vocal fold stiffness and vocal tract size) and expressive factors (e.g. pronunciation and accent). In this research, we attempt to model both types of factors in order to establish and simulate singer identity.

A good deal of research has been performed on features for speaker (talker) identification. Much of this work has focused on spectral features used in speech recognition systems, such as Mel-frequency cepstral coefficients (MFCCs), which correlate to the shape of the vocal tract [1]. Some research has also investigated glottal excitation features and their potential usefulness in the determination of speaker identity. Both approaches have proven to be moderately successful, and performance improves when they are used together [2].

The system presented in this paper is specific to the classically-trained singing voice and takes advantage of certain assumptions that discriminate classical singing from speaking. For example, in classical singing there is a much higher degree of voicing and longer vowel durations than in speech. The ultimate goal of this research is not only the identification of a particular singer's voice, but also the parameterization of features unique to that voice. Our system also has possible applications in singing voice transmission (coding), evaluating voice similarity, and singing voice synthesis.

2. OVERVIEW OF SYSTEM

This section contains a summary of the major components of our analysis system. Figure 1 shows a block diagram of the framework.

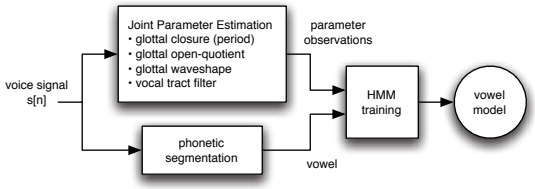


Figure 1: **Flow diagram of framework components**

2.1. Source-Filter Parameter Estimation

The source-filter model reflects the physical mechanism of vocal production and is commonly used for voice analysis/synthesis applications. For the source component, rather than modeling the glottal flow directly it is common to model the glottal derivative wave, which conveniently accounts for the effect of lip radiation (a differentiation). In our representation, the glottal derivative wave, $g[n]$ is filtered by the vocal tract impulse response, $h[n]$, to produce the voice ouput, $s[n]$.

$$s[n] = g[n] * h[n] \tag{1}$$

Historically, the estimation of source and filter parameters has been performed independently in order to simplify the analysis. Though convenient, this approximation is also inaccurate since vocal fold oscillation and vocal tract shape are actually somewhat dependent. For example, singers frequently modify their vowels (vocal tract shape) in order to more easily sing a high pitch. These dependencies may also be distinctive features of an individual voice. To account for these variables, we derive our source-filter model parameters jointly (simultaneously) from the acoustic data.

In our analysis, we initially use the KLGLOTT88 model [3] to represent the glottal derivative source, and a fixed-order filter

(derived via linear prediction) to model the vocal tract filter. These models lend themselves to a particularly efficient solution for joint parameter estimation, via convex optimization. Using the jointly-derived filter estimates, the excitation is then re-parameterized using the more complex LF model, which more accurately reflects the waveshape of the glottal derivative. Effects not represented by these models (such as turbulence) result in a residual noise signal, which is modeled seperately. The following estimation procedure was first suggested in [4] and a brief summary follows.

The KLGLOTT88 glottal derivative model is defined as follows:

$$\hat{g}[n] = \begin{cases} 2an - 3bn^2, & 0 \leq n < T \cdot OQ \\ 0, & T \cdot OQ \leq n < T \end{cases} \quad (2)$$

T corresponds to the pitch period (in samples) and OQ is the open-quotient, or the fraction of the period for which the glottis is open. The parameters a and b are further related as follows:

$$a = b \cdot OQ \cdot T \quad (3)$$

Simultaneous estimation of the glottal derivative and vocal tract parameters involves de-convolution of the source and filter functions from the voice signal. To accomplish this, we attempt to minimize the distance between the KLGLOTT88 source model, $\hat{g}[n]$ and linear prediction residual, $g[n]$, over one period given values of the period T and open-quotient OQ. This error function is convex and therefore has a guaranteed optimal solution. The solution is calculated using quadratic programming, resulting in simultaneous estimates for the KLGLOTT88 parameters and polynomial filter coefficients of the LP filter.

Our representation differs from [4] in the following ways: 1) Joint parameter estimation is performed on a warped frequency scale, to more accurately model the frequency sensitivity of human perception, 2) Glottal closure instants are not calculated *a priori*, but are optimized from the data given the assumed models for source and filter, and 3) The residual noise is modeled using a stochastic codebook, individually trained for each singer. These extensions are described in greater detail in the sections that follow.

2.1.1. Parameter Estimation with Warped Linear Prediction

Standard *Linear Prediction* (LP) estimates a signal from a linear combination of previous samples [5].

$$s[n] = \sum_{k=1}^{p} \alpha_k s[n-k] + g[n] \quad (4)$$

From the source-filter relation (Eq. 1), we can derive the transfer function, $H(z)$, which is an all-pole filter.

$$H(z) = \frac{S(z)}{G(z)} = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} = \frac{1}{A(z)} \quad (5)$$

A disadvantage of standard LP is that all frequencies are treated equally on a linear scale while the frequency sensitivity of the human ear is closer to logarithmic. As a result, LP analysis sometimes places additional poles at higher frequencies while neglecting closely spaced formants at lower frequencies where the ear is more sensitive. Instead, we use *Warped Linear Prediction* (WLP) to nonlinearly warp the spectrum of a signal to increase resolution at lower frequencies. This can be accomplished by replacing each standard delay with the following all-pass filter [6].

$$z^{-1} \rightarrow D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \quad (6)$$

We use a parameter value of λ=0.4, which provides a moderate degree of frequency warping. Higher values were found to cause more instability in the filter estimates.

To use WLP in the joint source-filter parameter estimation, we must reformulate the LP residual as a WLP residual. From Equation (5), we obtain the z-transform of the LP residual, $G(z)$:

$$G(z) = \frac{S(z)}{H(z)} = S(z)A(z) = S(z)\left(1 - \sum_{k=1}^{p} \alpha_k z^{-k}\right) \quad (7)$$

To take advantage of frequency warping, we must replace each delay with the allpass filter $D(z)$ of Equation (6).

$$G(z) = S(z)\left(1 - \sum_{k=1}^{p} \alpha_k D(z)^k\right) \quad (8)$$

In the time domain, $D(z)^k$ is the k-fold convolution of $\delta[n]$ (the impulse response of $D(z)$) with the original non-delayed signal. We denote this using the generalized shift operator $d_k\{\cdot\}$. [6].

$$\begin{aligned} d_1\{s[n]\} &\equiv \delta[n] * s[n] \\ d_2\{s[n]\} &\equiv \delta[n] * \delta[n] * s[n] \\ &\vdots \end{aligned} \quad (9)$$

Thus, we obtain the following relation between the WLP residual and the voice signal:

$$g[n] = s[n] - \sum_{k=1}^{p} \alpha_k d_k\{s[n]\} \quad (10)$$

We want to determine the parameter values that minimize the distance between $g[n]$ and our KLGLOTT88 model, $\hat{g}[n]$:

$$e[n] = \hat{g}[n] - g[n] = \begin{cases} 2an - 3bn^2 - s[n] + \sum_{k=1}^{p} \alpha_k d_k\{s[n]\}, & 0 \leq n < T \cdot OQ \\ 0 - s[n] + \sum_{k=1}^{p} \alpha_k d_k\{s[n]\}, & T \cdot OQ \leq n < T \end{cases} \quad (11)$$

This gives us the error at each sample, but we want to minimize the L_2-norm (squared error) over an entire period.

$$\min \sum_{n=0}^{T} (e[n])^2 = \min \sum_{n=0}^{T} (\hat{g}[n] - g[n])^2 \quad (12)$$

As in [4], this modified constrained optimization problem can be solved efficiently using quadratic programming. The result is a simultaneous estimate of the excitation parameters a and b, and the warped LP filter coefficients (α_k) for each analysis frame. An example of parameter estimation from one period is shown in Figure 2. Minimizing Equation (11), however, assumes that T and OQ are known. The following section (2.1.2) describes the estimation of these parameters.

The warped analysis results in an all-pole filter in the warped frequency domain, but when transformed to the linear frequency

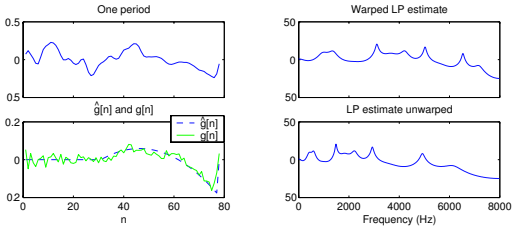


Figure 2: **Joint source-filter parameter estimation of vowel [e]**

domain it is actually a pole-zero filter. Fortunately, the analysis is conducted entirely in the warped domain, which maintains the simplicity of the all-pole representation (requiring fewer coefficients).

Once the filter parameters α_k are estimated, the WLP residual $g[n]$ can then be fitted to the more sophisticated LF model, $\tilde{g}[n]$, which more accurately describes the shape of the wave.

$$\tilde{g}[n] = \begin{cases} E_o e^{\alpha n T_s} \sin(\omega_g n T_s), & 0 \le nT_s < T_e \\ -\frac{E_o}{\epsilon T_a}\left[e^{-\epsilon(nT_s - T_e)} - e^{-\epsilon(T_c - T_e)}\right], & T_e \le nT_s < T_c \\ 0, & T_c \le nT_s < T_o \end{cases} \quad (13)$$

$E_o, \alpha, \omega_g, \epsilon, T_a, T_e$, and T_c are all free parameters, though there are dependencies between some of them. T_s is the sampling period and T_o is the pitch period. The parameters are estimated using constrained nonlinear minimization, which is described in detail in [4].

2.1.2. Glottal Closure Instant and Period Detection

The parameter estimation from the previous section is pitch-synchronous, where each analysis frame is time-aligned with each period of the output waveform. This requires the estimation of period boundaries which are aligned to the glottal closure instants (GCIs). Several techniques have been proposed for GCI detection (e.g. [7]), but each technique has some deficiencies which can result in GCI errors, leading to poor model fitting and poor source-filter parameter estimates. Instead of attempting to calculate the GCIs *a priori*, we perform a search for the period resulting in the best model fit.

Likewise, it is necessary to search for the appropriate value of the open quotient, OQ. Therefore, we simultaneously search for the optimal values of T and OQ that will minimize the overall error (Eq. 12). We obtain an initial estimate of the first several GCIs from a fixed-frame size LP residual. The residual has large peaks at moments of least linear predictability, which are usually close to the instants of glottal closure. The search is initialized by performing a linear search over all reasonable values of OQ (0.4 to 0.9) and values of T close to (within a few samples) the initial GCI estimate. Since neither T nor OQ will not vary greatly from one period to the next, we need only search a small range of values around the current values for each successive frame.

2.1.3. Stochastic Component Estimation

There are several stochastic sources that contribute to the overall vocal output, such as glottal aspiration noise and air turbulence in the vocal tract. These other noise-like sources prevent a perfect match to the vocal wave using the LF and WLP parameterization. Glottal aspiration noise is strongly correlated to the instants of glottal opening and closure, and certain vocal tract shapes are also more susceptible to air turbulence. Thus, the stochastic components must be modeled in a very specific way to reflect these dependencies. Glottal aspiration noise has been previously modeled statistically using wavelet de-noising [4].

We use a stochastic codebook approach, where the codebook is determined via *Principle Components Analysis* (PCA) of the glottal derivative residuals. PCA involves calculating the eigenvectors and eigenvalues from a matrix of input vectors. The eigenvectors corresponding to the highest eigenvalues capture the greatest amount of variance, statistically, of the input data. We want to find the eigenvectors of our set of residual noise vectors, $r[n]$, where

$$r[n] = \tilde{g}[n] - g[n] \quad (14)$$

PCA requires that all input vectors be the same length for statistical analysis, but in our case the length of each residual noise vector $r[n]$ varies according to the period. So we transform each residual to the frequency domain using equal-sized FFTs of N points to obtain:

$$R[\omega_k] = \mathcal{F}\{r[n]\}, \text{where } \omega_k = \frac{2\pi k}{N}, k = 0, \ldots, N-1 \quad (15)$$

Since $r[n]$ is real, we need only the first $\frac{N}{2} + 1$ values of $R[\omega_k]$, and PCA is then performed on these vectors. The analysis is performed on the complex FFT values in order to preserve the phase information, which is crucial to our noise model. From this, we obtain $\frac{N}{2}$ eigenvectors, which comprise the codebook.

For each noise vector, $\mathbf{r} = r[n]$, we obtain a weighting vector $\mathbf{w}$ corresponding to the contribution of each of the eigenvectors in codebook $\mathbf{C}$:

$$\mathbf{w} = \mathbf{C}\mathbf{r} \quad (16)$$

The n_c highest weighted codebook vectors, corresponding to the highest n_c values of $\mathbf{w}$ are then used to estimate $r[n]$ for each period.

2.2. Phonetic Segmentation

The analysis framework used in the singer identification system operates on individual vowel segments, requiring the input to be phonetically segmented. Using labeled data from the TIMIT speech database, simple templates were built for each of the 42 English phonemes. The templates were created by averaging the MFCCs calculated from short-time frames of the labeled data. The phonetic segmentation system assumes that a phonetic transcript of the singing input is available *a priori*, a restriction currently necessary for accurate segmentation. The system calculates MFCCs for each short-time frame of input singing data and determines the L_2 norm from each MFCC frame to each phoneme template. The phonemes are aligned to the singing data using dynamic programming to find the shortest-distance path over the matrix of MFCC-to-phoneme template distances, given the constraint of phoneme order.

2.3. HMM Model Training

We trained individual HMMs for each vowel ([a], [e], [i], [o], [u]) and each singer in our data set. Our observations consisted of the source-filter features defined in the previous sections. Each HMM state corresponds to one period, and transitions occur at each period boundary. The identification system operates by determining which singer's HMM has the highest likelihood for a given vowel, as established by the phonetic segmentation. The system currently

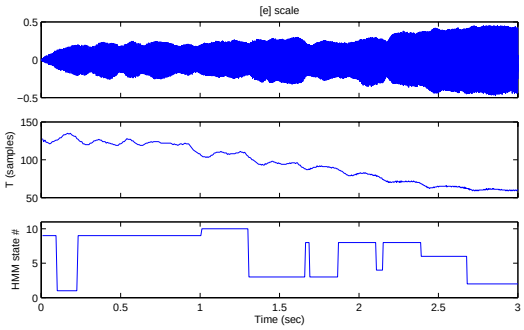


Figure 3: **HMM state path for a scale passage on the vowel [e].**

uses 10 states for each HMM. This value was determined experimentally to cover the wide variability of each vowel's observed features while limiting the computational complexity of the model. An example state path is shown in Figure 3.

3. SINGER IDENTIFICATION

The current data set consists of recordings from 4 conservatory-trained classical singers (two sopranos, one tenor, and one bass-baritone). Each singer performed a variety of vocal exercises (such as scales and arpeggios) emphasizing the 5 major vowels in addition to one entire piece from the classical repertoire. The exercises were segmented by vowel, and each vowel segment was used to train the vowel HMMs specific to each singer.

Vowel segments were extracted from the actual pieces and evaluated against the trained HMMs. The identification system operates by determining which singer's HMM has the highest likelihood for a given vowel, and the singer with the most vowel HMM matches in the excerpt is identified as the source performer. Five 5-7 second excerpts from each piece were used from each singer. On this admittedly small data set, the identification system performed with an overall accuracy of $>$90% when operating over entire excerpts and $\sim$70% over the individual vowel segments. More thorough results are presented in [8].

4. VOICE TRANSFORMATION

The system described in this paper has also been used for analysis/synthesis, where the HMM state paths are used to reconstruct the voice output. Because the various parameters for each pitch period can be represented by a single state value, a great deal of compression can be achieved. The sound quality, however, can be inconsistent, depending on how well the HMM states represent the input signal. Informal listening has shown that the sound quality over the vowel segments is mostly preserved. Higher sound quality is likely to be achieved with larger amounts of training data and a higher number of HMM states per model.

The HMM states provide a common point of reference between differing voice models, allowing us to map parameters from one to another. The states themselves represent clusters of significant concentration in each singer's parameter space. Since there is no consistency in the labeling of states between singers, we reorder each HMM's state labels according to their frequency of occurrence (state 1 becomes the most often occurring state, followed by state 2, etc.), enforcing a semi-statistical relationship between the state numbers (Figure 4). We then used the revised state path from one singer to drive another singer's HMM for re-synthesis. Informal listening does demonstrate a definite transformation in voice quality, though the accuracy of the effect is difficult to quantify. This is a simple mapping and we are currently investigating alternative methods for mapping between the states of different singers.

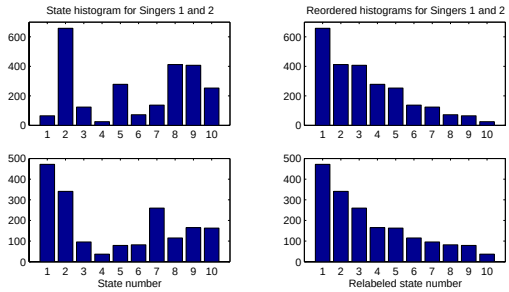


Figure 4: **HMM state histogram for two singers for vowel [e].**

We did not utilize the stochastic codebooks in the these preliminary voice transformation experiments. There is no simple way to map from one codebook to another. An initial attempt at simply applying the code vector weights calculated from one singer's codebook to another resulted in a more distorted sound, similar to phase distortion. A different parameterization may be needed to provide a more suitable mapping of the stochastic component of the excitation.

5. REFERENCES

[1] R. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58–71, 1996.

[2] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, Sept. 1999.

[3] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *JASA*, vol. 82, no. 3, pp. 737–793, 1990.

[4] H.-L. Lu, "Toward a high-quality singing synthesizer with vocal texture control," Ph.D. dissertation, Stanford University, 2002.

[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[6] A. Härmä, "A comparison of warped and conventional linear predictive coding," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 579–588, 2001.

[7] R. Smits and B. Yegnanayarana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 325–333, 1995.

[8] Y. E. Kim, "Singing voice identification via dynamic parameter modeling," *submitted to the 2003 International Symposium on Music Information Retrieval*, October 2003.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

SPECTRAL MODELLING OF THE SINGING VOICE USING ASYMMETRIC GENERALIZED GAUSSIAN FUNCTIONS

Matthew E. Lee

Center for Signal and Image Processing
School of Electrical and Computer Engineering
Georgia Institute of Technology
mattlee@ece.gatech.edu

Mark J. T. Smith

School of Electrical and Computer Engineering
Purdue University
mjts@purdue.edu

ABSTRACT

Analysis and synthesis of the human singing voice has been a focal point of much research within both the music and speech processing communities. In many applications, it is desirable to modify spectral characteristics prior to the synthesis process to enhance pleasing qualities, suppress unwanted ones, or to impart qualities that are specific to a particular style of singing. Existing formant models such as all-pole models offer little flexibility with respect to a formant's shape and form. Furthermore, methods for formant modification can be difficult to control. We propose a new spectral model for formant modification in which formants in the spectral envelope are represented as asymmetric generalized Gaussian functions. These functions do not restrict formants to be symmetric and characterize each formant's shape in addition to its amplitude and bandwidth. An investigation into the ability of this model to characterize spectral differences between classical and belt singing styles is also presented. We show that the new spectral model is capable of identifying certain formant characteristics of each of the singing styles. In addition, by combining these parameters with a sinusoidal synthesis procedure, vocal quality enhancements can be made that can improve style-specific qualities as well as accommodate pitch corrections.

1. INTRODUCTION

For voiced portions of speech, spectral characteristics of the human vocal tract have been shown to have a large impact not only on intelligibility but also on vocal qualities. In singing, this is especially important since overall quality often takes precedence over intelligibility. Studies have identified several spectral characteristics that affect overall vocal quality as well as qualities that are specific to particular styles of singing. The desire for a spectral model that enables modification and insertion of these spectral qualities is apparent in a variety of applications.

Formant models offer an accurate and efficient representation by modelling each of the dominant resonances of the vocal tract with a function based on its characteristics. Formant models typically rely on an all-pole representation such as Linear Predictive Coding (LPC) in which each formant is modelled as a second-order resonant filter. While allowing frequency shifting and amplitude scaling of each formant, this type of model does not enable independent control of other important formant characteristics such as bandwidth. Formant bandwidth has been shown to be correlated with the closed phase of the glottal source [1] which can be an indicator of registration and singing style [2, 3].

In this paper, we propose a model that uses asymmetric generalized Gaussian (AGG) functions to characterize the formant structure of the vocal tract. These functions enable greater control over formant characteristics by independently parameterizing each formant's bandwidth and shape. Additionally, they are not restricted to being symmetric in the frequency domain. By using this model in combination with the Analysis-by-Synthesis/Overlap-Add (AbS/OLA) sinusoidal model for speech, a method is presented which enables control over important vocal characteristics in the singing voice as well as overall quality.

By using these methods to analyze recorded samples, spectral characteristics of *classical* and *belt* singing are identified. While the western classical tradition has been the subject of much research and discussion, the belt voice has received less attention and has not been as well-defined. The belt voice is a vocal genre that is used extensively in both the pop arena and on the musical theatre stage [4]. In this study, the spectral characteristics of these two singing voice styles are parameterized according to our model and used in an attempt to produce vocal quality enhancements that correspond to a desired singing style.

2. SPECTRAL MODELLING AND MODIFICATION

All-pole formant models for speech, such as LPC, represent the formant structure of the vocal tract as a set of cascaded

second-order IIR filters of the form

$$F(z) = \frac{1}{(1 - az^{-1})(1 - a^* z^{-1})}. \quad (1)$$

The conjugate pole pair $z = |a|e^{\pm j\angle a}$ and sampling frequency f_s determine the formant frequency F and 3-dB bandwidth B according to

$$F = \frac{f_s}{2\pi}\angle a \text{ Hz}, \quad (2)$$

$$B = -\frac{f_s}{\pi}\ln|a| \text{ Hz}. \quad (3)$$

Spectral modifications are performed by scaling the angle $\angle a$ and magnitude $|a|$ of each pole. One drawback to this model is that the bandwidth B of each formant is dependent on the magnitude of the corresponding pole. Therefore, the amplitude and bandwidth of each formant cannot be independently modified.

Our approach models formant structure as a weighted sum of asymmetric generalized Gaussian functions which are specified by

$$G[k] = \begin{cases} \exp\left[-\left(\frac{|k-\mu|}{\beta^l}\right)^{\alpha^l}\right], & \text{for } k \le \mu, \\ \exp\left[-\left(\frac{|k-\mu|}{\beta^r}\right)^{\alpha^r}\right], & \text{for } k > \mu, \end{cases} \quad (4)$$

where k is the discrete frequency index. These functions independently parameterize the width and shape of the left and right sides. The spectral width parameter β dictates the bandwidth of each formant while α is a shaping parameter that controls the rate of decay. Figure 1 illustrates the effect of modifying these parameters.

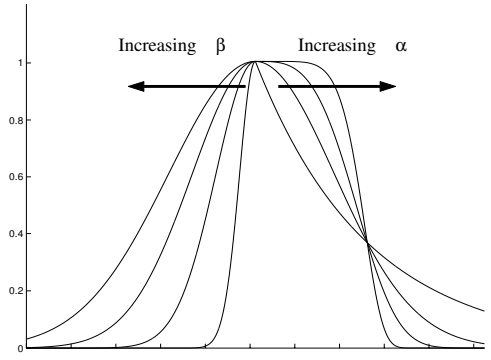


Figure 1: Asymmetric generalized Gaussians with increasing width parameters (β) to the left and increasing shape parameters (α) to the right.

By first calculating the spectral envelope of a given frame, the parameters of the asymmetric generalized Gaussian functions are obtained by considering the spectral envelope as a probability density function and applying a form of the Expectation Maximization (EM) algorithm. The EM algorithm is a broadly applicable algorithm used to maximize the log-likelihood from incomplete data by iteratively maximizing the expectation of the log-likelihood of complete data. The EM algorithm is applied to the spectral envelope of the sinusoids instead of the sinusoidal coefficients themselves in order to ensure that each Gaussian function will not converge to a single harmonic.

The flexibility of the asymmetric generalized Gaussians ensure an accurate fit to the spectral envelope and enable intuitive and independent modification of each formant's frequency, amplitude, bandwidth, and shape. This provides a high level of control over the formant structure of a singer's voice. Figure 2 shows an example of asymmetric generalized Gaussian functions fitted to a spectral envelope.

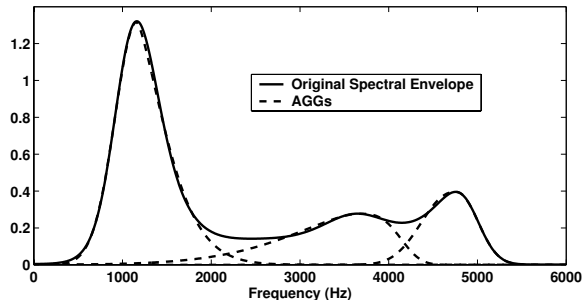


Figure 2: Asymmetric generalized Gaussians fitted to a spectral envelope.

For natural formant modifications to occur, smooth formant trajectories must be obtained based on the generalized Gaussian function parameters. This is accomplished using a process based on McAulay and Quatieri's peak matching algorithm for harmonic tracking [5]. A cost function based on closeness in frequency and amplitude is used in the decision making process. Because formants can merge, split, or disappear altogether, "births" and "deaths" are allowed to account for the possibility of the number of distinguishable formants changing from frame to frame.

3. VOCAL ANALYSIS

In order to test the functionality of our spectral model, an investigation was conducted to identify and characterize spectral differences between classical and belt singing styles. While some research has been done characterizing formant frequencies of each of these styles, little has been done in the way of identifying other formant characteristics associated with classical and belt singing such as bandwidth.

In addition to the analysis procedure, the modification process presented was used in conjunction with a sinusoidal

synthesis procedure in order to make style-specific vocal enhancements to test samples. Figure 3 shows a block-diagram of the analysis, modification, and synthesis processes.

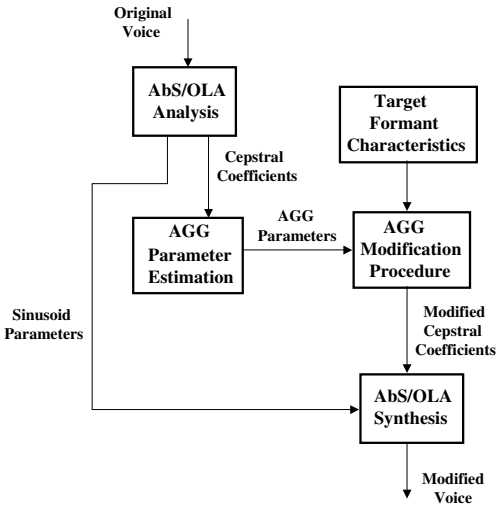


Figure 3: Block diagram of the analysis, modification, and synthesis procedure.

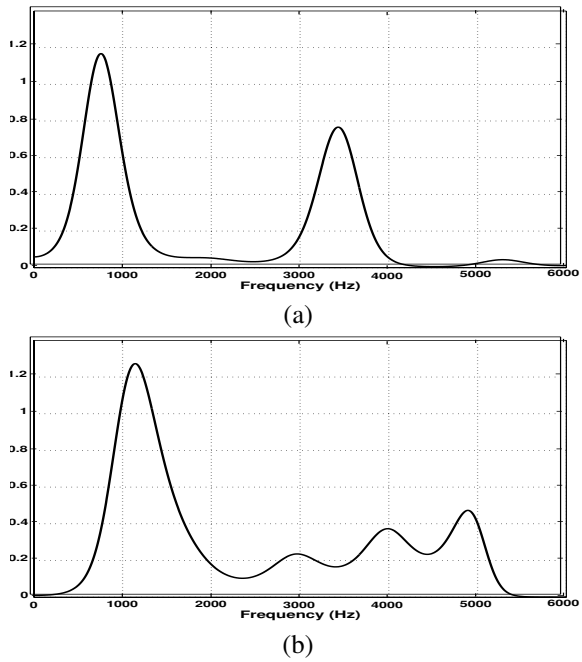


Figure 4: Spectrum of the vowel */a/* sung by (a) a classically trained female singer and (b) a female singer trained in the belt style. Both vowels were sung in C^5.

3.1. Procedure

Samples from four females were recorded for this experiment. Two of the singers had five or more years of classical training and the other two had four or more years of training in the musical theatre singing style known as belting.

Common sustained vowel sounds were isolated and analyzed. Sinusoidal parameters and cepstral coefficients for the spectral envelope were obtained with AbS/OLA for overlapping frames with a duration of 10 ms. A short window length was chosen to accommodate the relatively high fundamental frequency of the singers. AbS/OLA has been shown to be most effective for voiced portions of speech when the window length is set to cover approximately 2.5 pitch periods [6]. The spectral modelling technique described earlier was then used to determine the parameters of the asymmetric generalized Gaussians based on the cepstral coefficients.

3.2. Analysis Results

Although our study was limited to four singers, a number of interesting patterns emerged from the spectral analysis of their voices. Figure 4 will serve to illustrate some of these patterns. The figure shows spectra for the vowel */a/* sung in the note C^5 (one octave above middle C) by a classically trained female and another trained in the belt style. While these are not long-term-average spectra (LTAS), the patterns they illustrate are consistent among the samples recorded.

Comparing the first formant in each of the spectra clearly shows a lower center frequency for the classical singer. This predictable pattern has been shown to contribute to the "covered" or "rounded" sound often attributed to classical singers [7]. The higher frequency of the first formant (F_1) for the belt singer is characteristic of the articulation used by singers of this type. This placement of F_1 is also consistent with Sundberg's data which showed that belters have a weaker fundamental frequency than classical singers for the */i/* vowel [8].

The presence of a well-defined, prominent singer's formant in the classical singer's spectrum around 3400 Hz is unusual given that other studies have shown that sopranos tend not to have a singer's formant or have one with very low amplitude. Sundberg observed little clustering of the 3rd, 4th, and 5th formants in the LTAS of classically trained sopranos [9]. However, he also concludes that the level of the singer's formant is dependent not only on the vowel but also on vocal loudness. No measure of vocal loudness was made in this study. Similar observations regarding the lack of a singer's formant in sopranos were made by Bloothooft and Plomp [10].

The lack of a singer's formant in the belt-style singers

is compensated by the strength of the overtones in the mid and high frequency ranges. This characteristic has also been observed in other studies [8]. LeBorgne speculates that this increase in overtones may contribute to the belt voice's perceived "brightness" [4]. Other research has suggested that the lack of a singer's formant in the belt voice is accompanied by an increase in spectral energy in the lower formants [11]. Within the structure of the asymmetric generalized Gaussian model, this is manifested through an even frequency distribution of formants two through four (2500 Hz to 5000 Hz), each with larger amplitudes and widths than their average classical counterparts. In addition, the right side of the first formant typically exhibited higher width values for the two belt singers.

4. MODIFICATION & SYNTHESIS

The aim of the modification and synthesis processes are to enhance the recorded voices with the desired qualities of a vocal style other than that of the original voice. The AbS/OLA sinusoidal model [12] provides a flexible and accurate framework for implementing spectral modifications through the asymmetric generalized Gaussian model. As shown in Figure 3, the formant parameters are altered based on desired formant specifications.

In order to modify the classical singers' recorded voices to a belt-style of singing, the following modifications were performed:

- Increase the frequency of F_1 by $\sim \frac{3}{4} f_0$.
- Shift F_3 and F_4 so F_2-F_4 are evenly spaced through 5000 Hz.
- Decrease the amplitude of the singer's formant.
- Increase the width of F_2-F_4 and the right side of F_1.

Similarly, the recorded belt voices were modified in the following manner in order to produce a classical sound:

- Decrease the frequency of F_1 by $\sim \frac{3}{4} f_0$.
- Cluster F_3 and F_4 to form the singer's formant at 3300 Hz.
- Decrease the amplitude of F_2.
- Decrease the width of F_2-F_4 and the right side of F_1.

5. REMARKS

Informal listening tests of the modified samples provided encouraging results. The modification procedure was able to introduce vocal qualities of another singing style to a certain degree. However, while the samples did exhibit some of the desired vocal qualities, a certain amount of the original singing style remained. This may be due differences in glottal source characteristics which were not addressed in this paper. Audio results will be presented at the conference.

6. REFERENCES

[1] T. F. Cleveland, J. Sundberg, and R. E. Stone, "Long-term-average spectrum characteristics of country singers during speaking and singing," *Journal of Voice*, vol. 15, no. 1, pp. 54–60, 2001.

[2] J. Sundberg, T. F. Cleveland, R. E. Stone, and J. Iwarsson, "Voice source characteristics in six premeir country singers," *Journal of Voice*, vol. 13, no. 2, pp. 168–183, 1999.

[3] J. Sundberg and C. Högset, "Voice source differences between falsetto and modal registers in counter tenors, tenors and baritones," *Log Phon Vocol*, vol. 26, pp. 26–36, 2001.

[4] W. LeBorgne, *Defining the Belt Voice: Perceptual Judgments and Objective Measures*, Ph.D. thesis, University of Cincinnati, 2001.

[5] T. F. Quatieri and R. J. McAulay, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[6] E. B. George, *An Analysis-by-Synthesis Approach to Sinusoidal Modeling Applied to Speech and Music Processing*, Ph.D. thesis, Georgia Institute of Technology, 1991.

[7] H. K. Schutte and D. G. Miller, "Belting and pop, nonclassical approaches to the female middle voice: some preiminary considerations," *Journal of Voice*, vol. 7, no. 2, pp. 142–150, 1993.

[8] J. Sundberg, P. Gramming, and J. Lovetri, "Comparisons of pharynx, source, formant, and pressure characteristics in operatic and musical theatre singing," *Journal of Voice*, vol. 7, no. 4, pp. 301–310, 1993.

[9] J. Sundberg, "Level and center frequency of the singer's formant," *Journal of Voice*, vol. 15, no. 2, pp. 176–186, 2001.

[10] G. Bloothooft and R. Plomp, "The timbre of sung vowels," *Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 847–860, 1988.

[11] C. Robison, B. Bounous, and R. Bailey, "Vocal beauty: A study proposing its acoustical definition and relevant causes in classical baritones and female belt singers," *The NATS Journal*, pp. 19–30, Sept. 1994.

[12] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Transactions of Speech and Audio Processing*, vol. 5, pp. 389–406, Sept. 1997.

COMPARISON OF DIRECTIONAL SOURCES IN SIMULATING A SOPRANO VOICE

Linda Parati, Felipe Otondo***

*Dipartimento di Ingegneria, Università di Ferrara, Via Saragat 1, 44100 Ferrara, Italy;
**Ørsted-DTU, Acoustic Technology, Technical University of Denmark, build.352,
Ørsteds plads, DK-2800, Kgs.Lyngby, Denmark

lparati@ing.unife.it, fo@oersted.dtu.dk

ABSTRACT

The study of the acoustical balance between the singer and the orchestra by means of room acoustical measurements has shown that the directional characteristics of the source are important. This investigation compares the directional characteristics of two loudspeakers used for room acoustic measurements in an historical opera house when simulating a soprano. Directivity measurements of a soprano singer were done in an anechoic chamber and used as a basis for comparison in room acoustic simulations of the Royal Theatre of Copenhagen. The influence of the directional sources in the distribution of the simulated acoustical parameters in the room was compared and evaluated with the distribution obtained from the soprano.

1. INTRODUCTION

One of the most outstanding acoustical characteristics of an opera house, which still is not sufficiently investigated, is the balance between the sound of the orchestra in the pit and the sound coming from the stage. Previous studies have shown that one of the important factors that affects balance measurements is the directivity of sound sources [1] [2] [3]. Considering the available normative to evaluate the acoustical qualities of an opera house, the ISO 3382, omnidirectional sources have to be placed both in the orchestra pit and on the stage.
In this study the effect of using other types of directivity loudspeakers, besides the dodecahedron omnidirectional source has been investigated. The goal has been determining which source is closer to the directional characteristics of a soprano's singer voice.

2. DIRECTIVITY MEASUREMENT

Directivity measurements of a soprano singer and two loudspeakers were performed in an anechoic chamber to be used for the room simulations in the opera house.

2.1. Directivity measurements of the soprano

Multi-channel anechoic recordings of a soprano singer were made with a setup of 23 microphones. The singer was rotated in four positions and measured with a resolution of 10° in the horizontal and vertical axes. The recordings were made at 24-bit quantization and 44.1 kHz sampling rate and the singer was asked to sing isolated tones in two octaves from A3 (220 Hz) to A5 (879 Hz). The recordings from the four positions were then used to construct the averaged directivity in the octave bands from 125 to 8000 Hz. Figure 1 shows the measured directivity of the soprano in the horizontal plane at 500, 1000, 2000 and 4000 Hz octave bands. Figure 2 shows the measured directivity of the soprano in the vertical plane at 500, 1000, 2000 and 4000 Hz octave bands.

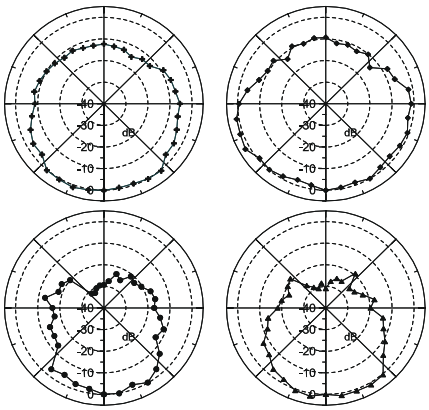


Figure 1: *Measured directivity of the soprano singer in the horizontal plane at 500 Hz (above-left), 1000 Hz (above-right), 2000 Hz (below-left) and 4000 Hz (below-right) octave bands.*

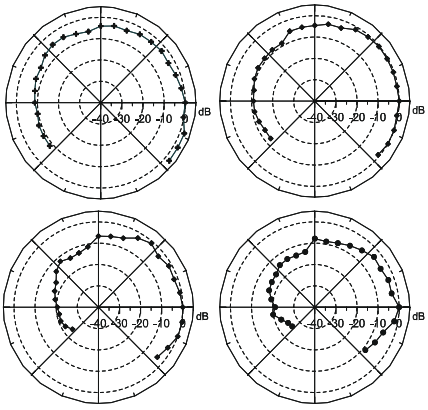


Figure 2: *Measured directivity of the soprano singer in the vertical plane at 500 Hz (above-left), 1000 Hz (above-right), 2000 Hz (below-left) and 4000 Hz (below-right) octave bands.*

2.2. Directivity measurement of the loudspeakers

The directivities of two loudspeakers were measured in an anechoic chamber following the traditional method of rotating the source in steps of 10 degrees in both vertical and horizontal axes with a single microphone. The two full range loudspeakers considered had different shapes in order to evaluate the influence of the loudspeakers cabinet. The first source was a 6 inch unit mounted in a rectangular wooden box of dimensions 20x20x15 cm. The second source was a dodecahedron omnidirectional source with a diameter of 35 cm. Each unit in the dodecahedron omnidirectional source was 5 inches and only one of the vertical units was measured and used for room acoustic simulation. Both sources were selected as representative of conventional loudspeakers that could be used in the context of balance measurements in rooms. The vertical and horizontal directivities of each of the loudspeakers proved to be the same. The directivity of the dodecahedron omnidirectional source was not measured and was assumed to be ideally omnidirectional. Figure 3 and figure 4 show the directivities of the two loudspeakers in the horizontal and vertical planes at 500, 1000, 2000 and 4000 Hz octave bands.

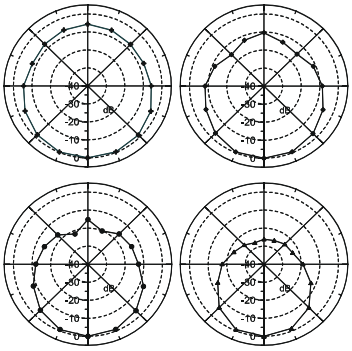


Figure 3: *Measured directivity of the wood box loudspeaker in the horizontal and vertical plane at 500 Hz (above-left), 1000 Hz (above-right), 2000 Hz (below-left) and 4000 Hz (below-right) octave bands.*

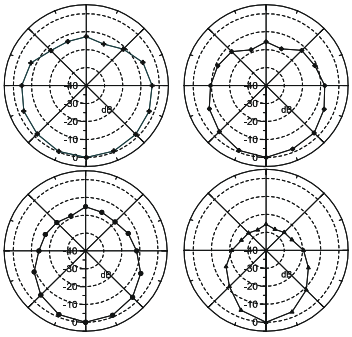


Figure 4*: Measured directivity of the dodecahedron omnidirectional source with just one frontal unit turned on. Horizontal and vertical plane at 500 Hz (above-left), 1000 Hz (above-right), 2000 Hz (below-left) and 4000 Hz (below-right) octave bands.*

3. SIMULATIONS

Room acoustic simulations were carried out on a model of the Royal Theatre of Copenhagen, using the software ODEON [4], in order to compare the directivities of the three acoustical sources with the soprano's one. Two different stage conditions were considered for the simulations: with stage set-up and without stage set-up. This was done in order to consider two realistic situations during the typical acoustical measurements in these kinds of theatres.

3.1. Method

The directivities of four sources were simulated in the theatre. The first source was a dodecahedron omnidirectional source (defined "OMNI"), the second was a rectangular wooden box (defined "Box"), the third was the dodecahedron omnidirectional source with only one unit turned on (defined "D1/12") and the fourth was the soprano's voice. Each source was simulated with the same power level in each octave band. This condition allowed for a comparison between the four different sources.

Three different source positions were chosen on the stage: **S1** at 1 m from the symmetric axis of the hall and 2 m from the fire curtain; **S2** at 1 m from the symmetric axis of the hall and 5 m to the fire curtain, and **S3** at 4 m from the symmetric axis of the hall and 2 m from the fire curtain, all of them at 1.5 m from the stage floor. These positions were considered as typical locations of the singer on stage. All of the four sources were simulated in each of the three positions (S1; S2; S3), one at the time. Due to the symmetry of the hall, positions in only one half of the room were simulated. 21 receiver positions (8 in the stalls and 13 in all the balconies) were selected. The Sound Pressure Level (SPL) simulated by the software in each receiver position was the parameter used to evaluate the similarity between the sources and the soprano's voice.

3.2. Results

An energetic average of the SPL obtained in each receiver's position for each single source, playing one at a time in the three different positions (S1; S2; S3) on the stage, was considered. The results are shown according to the distance between source and receiver in the room.

Two different behaviours were observed for each source in the hall: one in the stalls and another in the balconies. Because of this, the results will be divided between stalls and balconies.

The discussion has focused mostly on the 2 and 4 kHz octave bands, where the formant of the singer's voice is located [5] and where, according to Meyer [6], the voice's emission spectrum can be compared to the one of the orchestra.

Figure 5 shows the simulated energetic average of the SPL at different positions in the stalls, with stage set-up and without stage set-up. In the stalls there is a clear similarity between the behaviour of the two directivity loudspeakers in both the set-ups chosen (with and without stage set-up), while a different behaviour can be observed for the dodecahedron omnidirectional source. None of them proved to be significantly close to the behaviour of the soprano's voice.

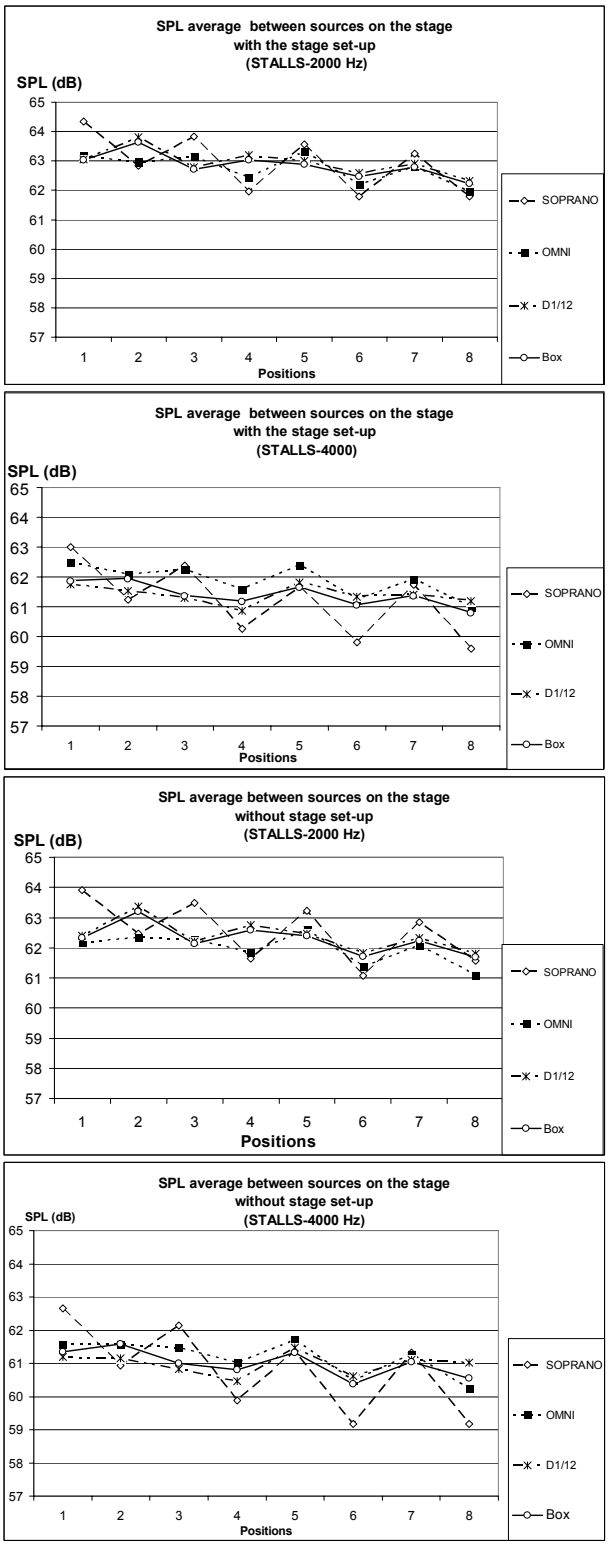


Figure 5: *Energetic average for different sources on the stage with the receivers in the stall. The first and the second are with stages set-up, the third and the fourth without stage set-up.*

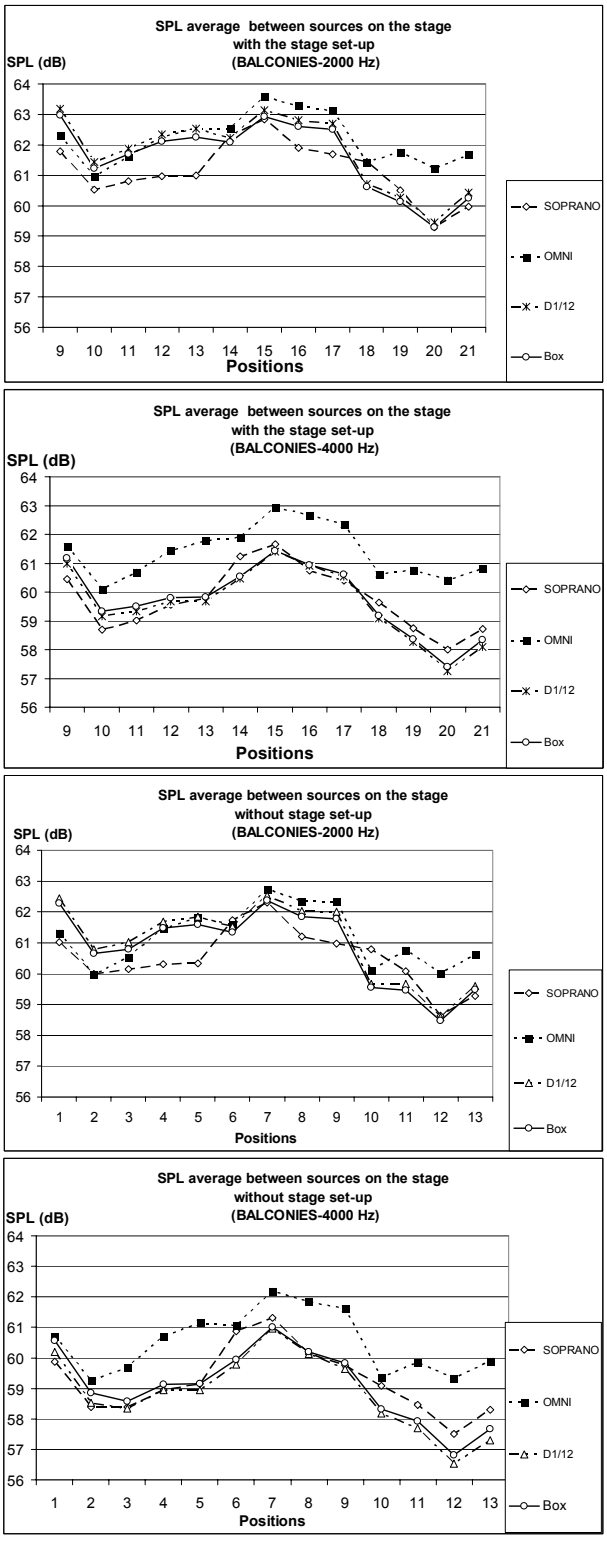


Figure 6: *Energetic average for different sources on the stage with the receivers all over the balconies. The first and the second are with stages set-up, the third and the fourth without stage set-up.*

Figure 6 shows the energetic average of the SPL at different positions in the balconies, with stage set-up and without stage set-up.

The behaviour of the two directivity loudspeakers and the soprano's voice are more similar in the balconies than in the stalls. At 2 kHz, in the receiver positions on the balconies, closer to the symmetric axis (positions 19-20-21), the dodecahedron omnidirectional source is stronger than the other sources. The similarity between the directivity sources and the soprano's voice is more evident at 4 kHz. On the other hand, the dodecahedron omnidirectional source seems to overestimate the soprano's voice by 2-3 dB.

The use of a stage set-up clearly increases the difference between the dodecahedron omnidirectional source and the other sources. This is due to the fact that early reflections are supporting the dodecahedron omnidirectional source more than the other sources, included the soprano's one.

3.3. Discussion

The significant discrepancies observed in the balconies at 2 and 4 kHz for the dodecahedron omnidirectional source seems to be based on two separate aspects. Firstly, this source spreads out more energy towards the balconies due to its uniform directivity distribution. Secondly, the stage tower's characteristics have more influence on the dodecahedron omnidirectional source behaviour than on the directional sources. This effect can be related to the omnidirectional characteristics of the source, an effect which is strongly influenced by the surfaces and the shape of the stage tower. Results obtained in other room models with different stage tower characteristics showed that, based on the same power level in each octave band, the dodecahedron omnidirectional source has a lower SPL over the entire hall than the other directional sources, including the soprano's voice.

4. CONCLUSIONS

The use of a dodecahedron omnidirectional source to emulate a soprano's voice in balance measurements can cause an over or underestimation of the soprano's voice. Simulation showed that the behaviour of the two full range loudspeakers was closer to the one of the soprano's voice compared to the dodecahedron omnidirectional source.

Despite the changes in the directivity and shape of the two full range loudspeakers, their acoustical behaviour in the room proved to be similar. This means that different full range loudspeakers are suitable to emulate a soprano's voice for the balance measurements purposes.

5. ACKNOWLEDGMENTS

This work has been financed by a Marie Curie Fellowship as a part of the programme European Doctorate in Sound and Vibration Studies and the DOREMI European Commission project (IST-2001-39158) [7].

The authors would like to thank the Royal Theatre of Copenhagen, Claus Lynge Christensen for his technical support with the software Odeon, Christina Conway for her English language support and Finn Agerkvist for his support and suggestions regarding loudspeakers.

6. REFERENCES

[1] Barron, M., "Auditorium acoustics and architectural design", E & FN Spon, London, pp. 333, 1993.

[2] O´Keefe, J., "Measurement of the stage to pit balance in four proscenium arch theatres", in Proceeding of IOA – Auditorium design of the millennium, Belfast, 1997.

[3] Pompoli, R. and Prodi, N., "A study on balance inside an historical opera house", in Proceeding of ICA2001, Rome, 2001.

[4] ODEON home page. http://www.odeon.dk

[5] Sundberg, J., "The acoustics of the singing voice", Scient. Amer. 236, 1977, pp 82.

[6] Meyer, J., "Some problems of opera house acoustics", in Proceedings of 12th ICA, Vancouver, 1986.

[7] "The DOREMI project home page." http://www.at.oersted.dtu.dk/~doremi/

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

THE ACTIVATION OF SHOULDER AND NECK MUSCLES IN PROFESSIONAL CLASSICAL SINGING

[1]Pettersen, V. and [2]Westgaard, R.H

[1]Stavanger University College. School of Art Education. N-4068 Stavanger, Norway
[2]Norwegian University of Science and Technology, N-7491 Trondheim, Norway
viggo.pettersen@kf.his.no

ABSTRACT

The study aimed to examine electromyographic (EMG) activity of muscles in the neck and shoulder. Five professional opera singers participated in the study.

Neck EMG activity was recorded from Sternocleido-mastoideus (STM), the scalenus muscles (SC) and muscles on the posterior part of the neck (PN). Shoulder EMG activity was recorded from upper trapezius (TR). STM and SC sites EMG activity was analysed to determine the respiratory phasing of the activity pattern. PN site EMG activity was analysed to quantify the level and pattern of activity, but there was no attempt to quantify any respiratory phasing. We attempted to lower the EMG activity of TR and STM by EMG biofeedback (BF) to examine whether changes in the activity of these muscles have consequences for the activation of the SC and PN muscles. Upper thorax (TX) respiratory movement was traced with a strain gauge sensor.

Three singing tasks were performed: aria, sustained tones and extreme tones. Each task was performed three times with variation in volume or pitch. Following the first performance of the singing tasks, a BF session was carried out and muscle activity recorded in a repeat performance of the same tasks. The EMG activity levels for all muscles were compared before and after BF. We found that STM and SC showed correlated activity patterns when supporting inhalation by classical singers. Substantial muscle activity was observed in posterior neck muscles during inhalation and phonation. Finally, EMG biofeedback performed on TR and STM had a secondary effect of lowering SC and PN site EMG activity after BF.

1. INTRODUCTION

When performing classical singing economical muscle usage is heavily emphasized in voice pedagogy and clinical voice care. Wasted energy and motion as well as excessive muscle tension are incorrect and usually deleterious. Motion economy is a basic principle of all art forms (1). This may be one of the reasons why singers with formal vocal training appear to have less muscle tension (MT) than singers without this training. Classical singers had lower MT scores than non-classical singers when Koufman et al (2) studied patterns of laryngeal tension during normal singing by 100 healthy singers. Nevertheless, tension–fatigue (TF) will occasionally lead to voice disorder, also by trained classical singers, but in particularly by the hard-driving vocal professional. The condition of TF is characterised by excessive muscular tension in the neck musculature, and the STM and the TR muscles being tight (3).

Since it is strongly advised to avoid excessive use of external neck muscles during the phonation process (4), we therefore investigated muscles in the neck of performing classical singers by help of electromyography (EMG).

Earlier studies by us have characterized the use of TR in classical singing (5-7). Other studies have shown that the STM and SC muscles, besides responding to postural demands, show some activity during forced voluntary inhalation (8-12). The PN muscles serve exclusively postural demands. Nevertheless, this area can be exposed to stiffness and pain in some singers. Thus, it would be of interest to compare the activity in the PN muscles to the activity in the lateral part of the neck.

To summarize, this study aims to examine the pattern of use of the STM and SC muscles during singing. EMG activity of the PN and TR muscles is also recorded. Furthermore, we try to lower TR and STM activity by use of EMG biofeedback, and explore the possible consequences on SC and PN sites EMG activity due to reduced TR and STM sites EMG activity. Finally, to compare neck muscles activity to the phasing of breathing, we trace upper thorax (TX) respiratory movement with a strain gauge sensor (RES-117).

2. MATERIAL AND METHODS

Subjects

Five professional opera singers, 2 males and 4 females, age ranging from 30 to 40 years, participated. They were all under contract with a European opera house.

Electromyographic recordings

Muscle activity was recorded from SC, SP, STM, and TR by an ambulatory monitoring system with surface EMG recording (four-channel Physiometers, PHY 400, Premed, Norway), using silver/silver chloride bipolar electrodes with an active diameter of 6 mm. A center-to-center distance of 20 mm was used for all EMG recordings. The EMG activity was band-pass filtered at 20-800 Hz and sampled at 1600 Hz. The EMG signals were thereafter A/D converted, the root-mean-square (RMS) value calculated and transmitted at 10 Hz on a serial interface to two PC Laptops. The processed signals were further analyzed by use of the Physiometer software, using a time resolution of 0.2 s. Electrode placing on TR and STM is described in previous reports (4,5). For SC site the first electrode was placed 2-3 cm caudal to Processus Mastoideus, the second 2 cm further down on a vertical line. PN site had the first electrode 3-4 cm cranial to the spine of the C7 vertebra and second electrode 2 cm further

up the same vertical line, both electrodes 2 cm to the right of the midline.

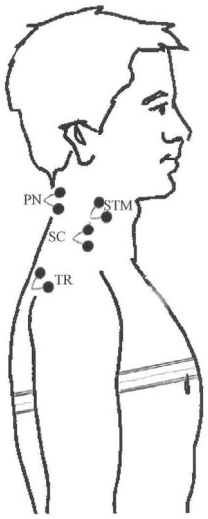


Figure 1. *Electrode placement on the scalenus (SC), posterior neck (PN), sternocleidomastoideus (STM) and the trapezius (TR) muscles.*

Procedure

Three singing tasks were performed; 1) an aria, 2) sustained tones maintained to exhaustion, and 3) extreme tones maintained to exhaustion. Pilot studies had shown that muscle activity varied with pitch and volume when performing. To trace possible variances, the aria was performed three times in three different volumes (normal, forte, piano). The sustained tone was sung in three different pitches (normal, third up, third down). The extreme tones consisted of three tones. The first was the comfort tone (used to test if the voice was functioning well, chosen freely by the singer), the second was the individual singers highest tone used in professional singing, and the third tone was the lowest tone used likewise. The BF procedure was carried out by help of a PC laptop monitor, which was also used to monitor the EMG signals. During the BF session the singer looked at the monitor and observed the recorded muscle activity from TR and STM. By looking at the monitor while varying the activity level, the singer could experience that he/she had an influence on the level of activity by varying the muscular effort. The singer was then instructed in maneuvers used to lower the activity level to a minimum in the two muscles. The singer was not able to observe the monitor during the singing sessions before and after BF. Before the last three sessions after BF they were instructed to do exactly as they did in the nine first sessions. In addition, they were asked to keep a mental focus on keeping the activity level in TR and STM as low as possible, using their experiences from the BF session.

Analysis

All recordings were transferred as ASCII files to a PC for further analyses. The data were statistically tested by the SPSS 11,0 for Windows. The responses before and after BF were tested by a paired, two-tailed t-test. For testing correlations, the Spearman's nonparametric correlation test was used.

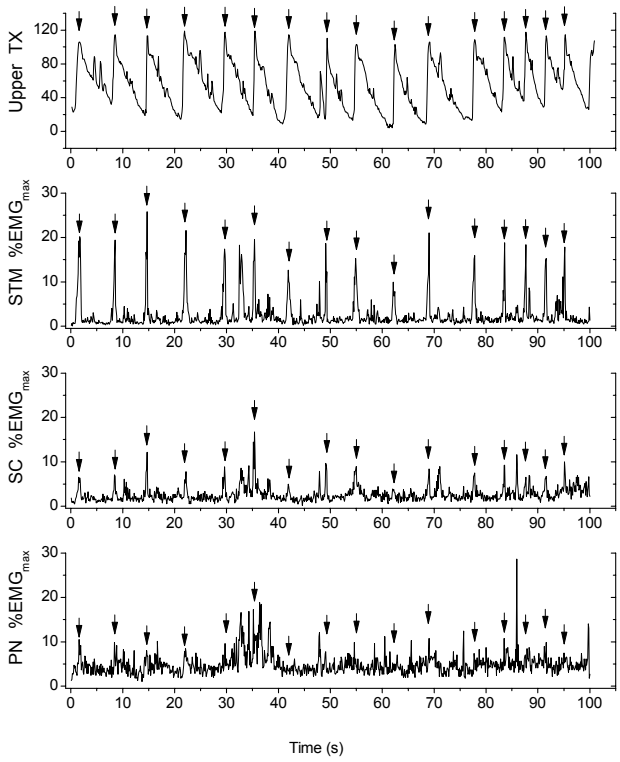


Figure 2. *Individual time plot of the normal version of the aria. Vertical arrows denote end of inhalation.*

3. RESULTS

Figure 2 illustrates results typical of most recordings in this study. The conclusions are supported by the calculations of correlation coefficients (r) between channels, which show high correlations between STM and SC site EMG activity. In addition, positive correlations between STM/SC activity and TX movement pattern are observed. Correlation coefficients for the time plots in Figure 2 are: STM vs. SC: ρ= 0,59, STM vs. TX: ρ = 0,18 and SC vs. TX: ρ = 0,14.

Furthermore, PN site EMG activity is positively correlated to upper TX movement (ρ = 0,17) and positive correlations are also observed for PN vs. TR activity (ρ = 0,18). PN site EMG activity is detected both during the inhalation and the exhalation phase (7). Due to this complex EMG activity pattern, PN vs. TR/TX correlation coefficients are low. PN site EMG activity in Figure 2 shows small peaks in inhalation and some unstable EMG activity during phonation. The primary function of the muscles at the PN site is to maintain the head posture. This indicates to us that PN activity is neither supporting inhalation nor exhalation, but must obviously serve postural demands. In addition the muscles are contracting coincidentally, as random muscular activity, due to the efforts connected to a powerful vocal performance (8).

As expected from an earlier study (5), the analyses of the arias showed significantly lowered EMG activity in both TR and STM after BF. In addition, both SC site and PN site EMG activity were significantly lowered after BF (Figure 3). In the recordings of the three versions of the aria, which should reveal a habitual way of singing, the highest potential for reduction in neck muscle activity was found in the PN site EMG activity (1,2% EMG_{max}). This was expected since activity recorded here purely should serve postural demands, not intended to execute behavior related to breathing for singing. The most predominant change in correlations after BF was found in TR to PN correlations, which were much higher after BF. At the same time SC to PN correlations were reduced after BF.

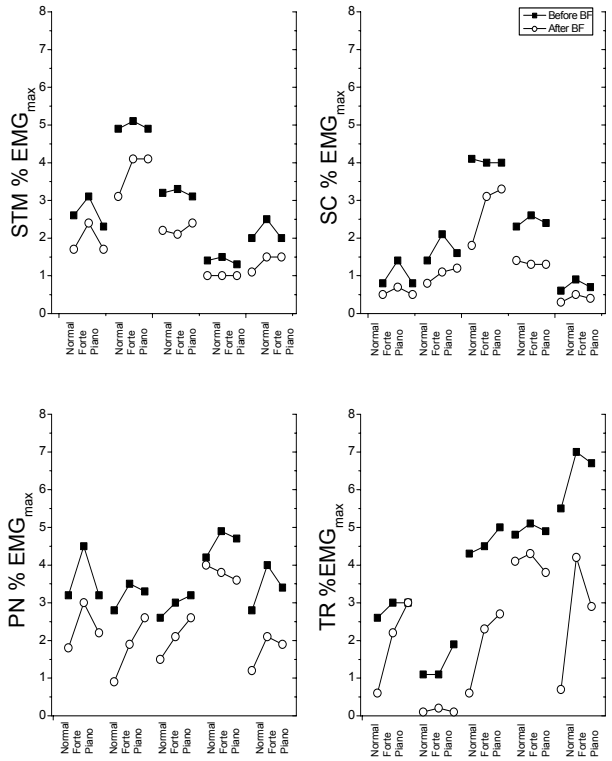


Figure 7. *3- point segment graph (normal, forte, piano) to show median EMG_{max} values for all muscles recorded, in all recordings of the aria. The filled squares denote the results before BF, open circles the results after BF.*

4. DISCUSSION

STM and SC sites EMG activity. The main findings in this study are that STM and SC clearly support inhalation by performing classical singers. The results support earlier findings (8-12), which have reported some EMG activity in these two muscles during voluntary forced inspiration and extend those results by quantifying the muscle activity during the inhalation and the phonation phases. Visual inspection of the time plots, as illustrated in Figure 2, shows that the inward movement of air (i.e., expansion of TX) and onset of STM and SC activity occurred simultaneously. This confirms the findings of Raper et al. (12) concerning very quick inhalations. Caution is however warranted regarding the interpretation of EMG signals from both sites. Potentially, signals from several layers of underlying muscles and crosstalk signals from neighbouring muscles may contaminate the recordings. However, both sites show a clear inhalatory phasing by contributing most of the activity during the inhalation phase and only minor activity during phonation, clearest shown in STM site EMG activity. This indicates that most of the signal recorded at STM site should come from the STM, and a major part of the signal recorded at SC site should be from the SC, also supported by the high correlations between STM and SC sites EMG activity.

PN site EMG activity was higher than anticipated. The TR muscle forms the first layer of muscles. The second and third layers are mainly formed by the splenius capitis.(SP) and the semispinalis capitis (SSP) muscles. It is not possible to determine the contribution from each of these muscles to the recorded EMG signal. However, the low correlations between TR site EMG activity and PN site activity before BF, point to a low TR contribution to the PN EMG signal, at least before BF. Most likely, the dominant part of the signal comes from the SP and the SSP muscles. Takebe et al. (13) reported either inactive or negligible activity in SP and SSP during deep breathing, flexion of the head, free lateral bending and lateral bending against resistance. SP showed activity in the rotation of the head while SSP did not. The main function of SSP was shown to be extension of the head, and the two main functions of SP are extension of the head and rotation of the head to same side. It is therefore not easy to understand the apparent unmotivated EMG activity at the PN site by most singers. (The singers were singing in upright positions, with eye focus on a recording video camera. Thus, the head was kept in a near invariant position during the recording sequence.) Even if high correlations were found especially between PN and SC sites EMG activity, no connections to breathing muscle activity should be made. Campbell (8) pointed out that although a muscle is contracting in a particular phase of breathing does not prove it is a prime mover of respiratory movement in that phase. It may be a synergist or an antagonist, or its contraction may even be coincidental, due to, for example, posture or random activity often found during concentrated efforts.

The EMG biofeedback session (BF) performed on TR and STM gave as secondary effect also reductions in SC and PN sites EMG activity (Figure 3). The largest reduction in muscle activity after BF was observed for TR activity, as expected (5). It is thought provoking that STM activity after BF was not lowered any more than SC and PN sites EMG activity. In fact, PN activity was reduced much more than STM activity after BF. Thus, the reduced STM, SC, and PN sites EMG values after BF are most likely a consequence of the lowering of TR activity. We showed earlier (7) that TR activity during phonation contributed to the compression of upper TX, and that lowered TR activity by EMG BF lead to upper TX being kept in a more expanded position. By the reduced TR activity after BF (Figure 3) and upper TX kept in a higher position (7), it is reason to believe that STM and SC muscles must work less during inhalation.

5. CONCLUSION

We conclude that STM and SC show correlated activity patterns when supporting inhalation by classical singers. Secondly, substantial muscle activity is observed in posterior neck muscles (PN) during inhalation and phonation. Thirdly, EMG biofeedback performed on TR and STM had secondary effect on lowering SC and PN site EMG activity after BF.

6. REFERENCES

[1] Sataloff RT, Spiegel RJ, Hawkshaw MJ. History and Physical Examination of Patients with Voice Disorders. In: Rubin JS, Sataloff RT, Korovin GS ed. Diagnosis and Treatment of Voice Disorders, 2nd ed. TM, Delemar Learning, US, 2003;137-158.

[2] Koufman JA, Radomski TA, Joharji GM, Russel GB, Pillsbury DC. Laryngeal biomechanics of the singing voice. Otolaryngol Head Neck Surg 1996;115:527-37.

[3] Rubin JS, Korovin GS, Epstein R. Special Considerations for the Professional Voice User. In: Rubin JS, Sataloff RT, Korovin GS ed. Diagnosis and Treatment of Voice Disorders, 2nd ed. TM, Delemar Learning, US, 2003;583-595.

[4] Carroll LM. The Role of the Voice Specialist in the Non-medical management of Beginning Voice Disorders. In: Rubin JS, Sataloff RT, Korovin GS ed. Diagnosis and Treatment of Voice Disorders, 2nd ed. TM, Delemar Learning, US, 2003;537-559.

[5] Pettersen V, Westgaard RH. Muscle activity in the classical singer's shoulder and neck region. Log Phon Vocol 2002;27:169-178.

[6] Pettersen V, Westgaard R.H. Muscle activity in professional classical singing. -A study on muscles in the trunk, shoulder and neck regions. Submitted.

[7] Pettersen V, Westgaard R.H. The association between auxiliary breathing-muscle activity and thorax movement pattern in classical singing. Submitted.

[8] Campbell EJM. The Respiratory Muscles. Annals N Y Ac Sciences 1968; 155: 135-139.

[9] Jones FP, Hanson J A, Gray FE. Head balance and sitting posture: II. The role of the sternomastoid muscle. J Psychol 1961; 52: 363-367.

[10] McFarland D, Smith A. Surface recordings of respiratory muscle activity during speech: Some preliminary findings. Journal of Speech and Hearing Research. 1989;32:657-667.

[11] Vitti M, Fujiwara M, Iida M, Basmajian JV: The integrated roles of longus colli and sternocleidomastoid muscles. Anat rec 1973;177;471-484.

[12] Raper AJ, Thompson WT Jr, Shapiro W, Patterson JL Jr: Scalene and sternomastoid muscle function. J Appl Physiol 1966;21:497-502.

[13] Takabe K, Vitti M, Basmajian JV: The function of semispinalis capitis and splenius capitis muscles: an electromyographic study. Anat rec 1974;179:477-480.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

THE LARYNGEAL FLOW MODEL FOR PRESSED-TYPE SINGING VOICES

Ken-Ichi Sakakibara [*1], *Hiroshi Imagawa* [*2], *Seiji Niimi* [*3], *Naotoshi Osaka* [*1]

[*1] NTT Communication Science Laboratories
3-1, Morinosato Wakamiya, Atsugi-shi, Kanagawa, 243-0198, Japan
kis@brl.ntt.co.jp

[*2] Department of Speech Physiology, The University of Tokyo
7-3-1, Hongou, Bunkyo-ku, Tokyo, 113-0033, Japan
imagawa@m.u-tokyo.ac.jp

[*3] Speech and Hearing Center, International University of Health and Welfare
2600-6, Kitakanemaru, Ohtawara, 324-0011, Japan
niimi@iuhw.ac.jp

ABSTRACT

Asian traditional pressed-type singing voices are different from the European traditional singing voice in their timbre and voice production mechanism. In throat singing, the ventricular folds and true vocal folds vibrate, resulting in the generation of the special laryngeal voice. On the other hand, in some other pressed-type singing voices, such as Japanese Min-yoh, the ventricular folds only approximate but do not vibrate.

We propose a new laryngeal flow model incorporating the effect of the ventricular fold vibration and laryngeal ventricle resonance. The model is a combination of the known glottal airflow model (R-model), the laryngeal ventricle resonance (Helmholtz resonator), and the modulation of ventricular fold vibration. We will also demonstrate the relation between model parameters and voice quality. The results show that the proposed model is effective for synthesizing the pressed-type singing voices.

1. INTRODUCTION

Non-interactive parametric glottal models assume that there are no interactions between the glottal source and vocal tract [2]. The glottal source is described by using mathematical equations. Such models are very effective in speech synthesis and coding and therefore have bee used in many studies. The R-model [8] and LF-model [3] have become reference models for this type of model.

All of these models assume that the laryngeal voice source is determined by vocal fold vibratory patterns and intend to control voice quality by changing vocal fold vibratory parameters, such as the open quotient (OQ), speed quotient (SQ), closing quotient (CQ), and amplitude quotient (AQ) [1].

However, in throat singing, the vibration of the ventricular folds (VTFs) (also referred to as the false vocal fold) and strong constriction of the supraglottic structure are observed [11], and in some Asian traditional pressed-type singing, such as Japanese Min-yoh, the constriction of the supraglottic structure is also observed, though the VTFs do not vibrate [5]. Therefore, for synthesis of various styles of singing voices, besides the vocal fold vibration, the effects of the VTF vibration and resonance of the laryngeal ventricle must be considered.

In this paper, we propose a new laryngeal model based on glottal flow, laryngeal ventricle resonance, and the modulation of the VTF vibration. As the laryngeal source for the source-filter synthesis, the proposed model is able to control various timbres of singing voices.

2. THE LARYNGEAL FLOW MODEL WITH VENTRICULAR-FOLD VIBRATORY MODULATION

2.1. VTF-modulation model

VTF vibration is observed in various types of phonation. In throat singing, both drone and kargyraa voice phonations are always accompanied by VTF vibrations, as well as vocal fold (VF) vibrations. In the drone voice, the ventricular folds vibrate in the same period as the VFs, and in the kargyraa voice, the VTF vibrate in an integer multiple (usually double or triple) period of the VFs [4, 6, 11]. The results of a simulation using a 2×2-mass model suggest the possible vibratory patterns of the VFs and VTFs [9, 11].

Here, we use "laryngeal flow (source)" to mean the airflow through the VTF slit, and "glottal airflow (source)" to mean the airflow through the slit of the VFs.

The laryngeal flows of drone and kargyraa for different two singers are shown in Fig. 1. These flows were obtained from recorded sounds using an inverse-filter analysis. We marked five poles on spectrum in the range from 0 to 5 kHz, constructed the inverse-filter, and manually adjusted it to make the result smooth. By combining the results of high-speed images, EGG waveforms, and these inverse-filtered laryngeal sources, we concluded that, in throat singing, the VTF vibration is indispensable for the generation of the laryngeal flow. Therefore, modeling the laryngeal flow in throat singing requires a new laryngeal model that includes the effect of VTF vibration.

The VTF-modulation model $\tilde{u}(t)$ is simply defined as follows:

$$\tilde{u}(t) := M(t)u(t) \tag{1}$$

The block-diagram of the model is depicted as shown in Fig. 2.

In this paper, we choose a simple R-model [8] for the glottal flow. The R-model is described as follows:

$$u(t) = \begin{cases} \alpha\left(3\left(\frac{t}{T_p}\right)^2 - 2\left(\frac{t}{T_p}\right)^3\right), & \text{if} \quad 0 \le t \le T_p \\ \alpha\left(1 - \left(\frac{t-T_p}{T_p}\right)^2\right), & \text{if} \quad T_p \le t \le T_p + T_n \le T_0 \end{cases} \tag{2}$$

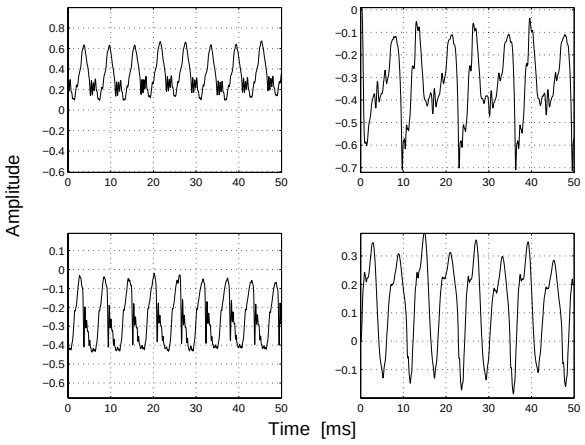


Figure 1: Inverse-filtered laryngeal flow. Left upper: drone by subject A, left lower:drone by subject B, right upper: karygyraa by A, right lower kargyraa by B.

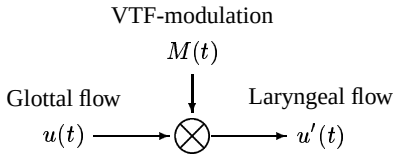


Figure 2: Block diagram of VTF-modulation model

where α is amplitude, T_p opening time, T_n closing time, and T_0 period. All of these variables are in $\mathbf{R}_{>0}$. The open quotient (OQ) is written as $(T_p + T_n)/T_0$.

The vibratory patterns of the VTFs were observed using the high-speed images and seem to be not exactly sine-shape [7, 10, 11]. However, here we define the VTF-modulation function $M(t)$ by multiplication by constant M of the false glottal area function A'_g. We also define A'_g as a sine function:

$$M(t) := MA'_g(t) \tag{3}$$

$$A'_g(t) := \alpha' \sin(\omega' t + \theta') + {A_g}'_0 \tag{4}$$

where α' represents the amplitude of the VTF vibration, ${A_g}'_0$ the area between the VTFs at rest, ω' the frequency of the VTF vibration, and θ' the phase difference of the VTF vibration from VF vibration. All of these are in $\mathbf{R}_{>0}$. Physiological observations and the simulation using 2×2-mass model suggest that the periods of the VF and VTF vibration satisfy $2\pi/\omega' = nT_0$, where $n \in \mathbf{Z}_{>0}$.

2.2. VTF-modulation and LVT-resonance model

The laryngeal ventricle is the space between the VFs and VTFs. When the VTFs are strongly constricted, it seems the effect of this small space on the laryngeal voice can not be ignored. The physical model simulation suggest that some acoustic effects occur around 2000 Hz [9, 11]. The inverse-filtered laryngeal voices of throat singing have some ripples (Fig. 1), which almost agree with the physical model simulation results, [7], hence, some appropriate model with laryngeal ventricle resonance is required. Fig. 3 shows spectra of the drone voices of two different singers.

A block diagram of our proposed model (VTF-modulation and LVT-resonance model) is shown in Fig. 4.

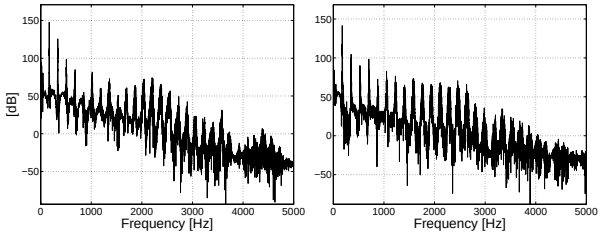


Figure 3: Spectra of inverse-filtered laryngeal flows. Right: drone by A, left drone by B.

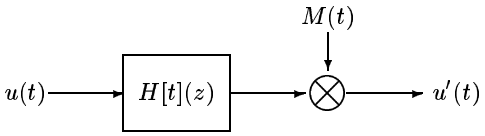


Figure 4: Block diagram of VTF-modulation and LVT-resonance model.

The model was obtained as follows: The glottal airflow is convoluted with the time-variant laryngeal ventricle resonator depending on the VTF vibration , and modulated by the vibration of the VTFs.

We denote the resonator by the laryngeal ventricle by $H[t](z)$. Then, the laryngeal voice with the laryngeal ventricle and VTF modulation is described as:

$$\tilde{u}(t) := M(t)H[t](u(t)) \tag{5}$$

We realize $H[t](z)$ as a time-varying one-pole filter. We calculate the resonance frequency of the laryngeal ventricle, i.e. the frequency of the pole of $H[t]$, by means of a Helmholtz resonator. Let $F_v(t)$ be the resonance frequency, d' be thickness of the VTF, and V_v be volume of the laryngeal ventricle. Then,

$$F_v(t) = \frac{c}{2\pi}\sqrt{\frac{A'_g(t)}{d'V_v}} \tag{6}$$

where c is the sound velocity, 3.53×10^4 cm/s.

To permit control flexibility, we define the bandwidth of the resonance by the multiple of variable K, which changes depending on phonation types, and the bandwidth as a Helmholtz resonator.

$$B_v(t) = K\left(\frac{R_v(t)}{2\pi L_v(t)} + \frac{G}{2\pi C}\right) \tag{7}$$

The resistance $R_v(t)$, inductance $L_v(t)$, conductance G, and capacitance C satisfy the following equations.

$$R_v(t) = d'\frac{2\sqrt{\pi A'_g(t)}}{{A'_g(t)}^2}\sqrt{\frac{1}{2}\rho\mu\omega}, \quad L_v(t) = \frac{\rho d'}{A'_g(t)}$$

$$G = \frac{2d_v\sqrt{\pi V_v}(\eta - 1)}{\rho c^2}\sqrt{\frac{\lambda\omega}{2\xi\rho}}, \quad C = \frac{V_v}{\rho c^2}$$

where $\omega := 2\pi/T_0$ is the frequency of the VF vibration, and d_v the thickness of the laryngeal ventricle. The constants are set as follows: the density of air $\rho = 1.14 \times 10^{-3}$ g/cm^3; the viscosity $\mu = 1.86 \times 10^{-4}$ dyn · s/cm^2; the adiabatic gas constant $\eta = 1.4$; and the specific heat $\xi = 0.24$ cal/gm · degree.

3. ACOUSTICAL CHARACTERISTICS

3.1. VTF-modulation model

We study the effect of the phase difference between the VF and VTF vibrations. In the equation

$$u'(t) = \left(\alpha' \sin(\omega' t + \theta') + A'_{g0}\right) u(t), \tag{8}$$

we fix $A'_{g0} = 0.5 \max u(t)$, $\alpha' = 0.35 \max u(t)$. For $u(t)$, we also set $T_0 = 8$ ms, $T_p/T_0 = 0.42$, $T_n/T_0 = 0.18$, and hence $\mathrm{OQ} = 0.6$. As these settings, the spectral tilt of $u(t)$ is close to 12 dB. We also set $\omega' = \omega$, i.e. study the laryngeal flow, such as the drone voice of throat singing. The laryngeal flows of various θ' are as shown in Fig. 5.

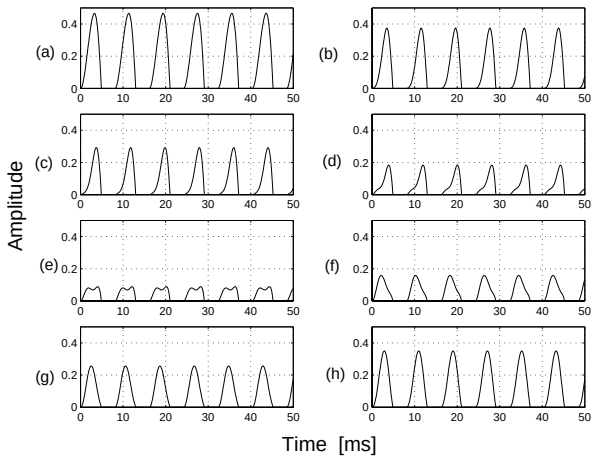


Figure 5: Laryngeal flows in various phases. (a) Original $u(t)$, (b) $\theta' = \pi/2$, (c) $\theta' = 3\pi/4$, (d) $\theta' = \pi$, (e) $\theta' = -3\pi/4$, (f) $\theta' = -\pi/2$, (g) $\theta' = -\pi/4$, (h) $\theta' = 0$.

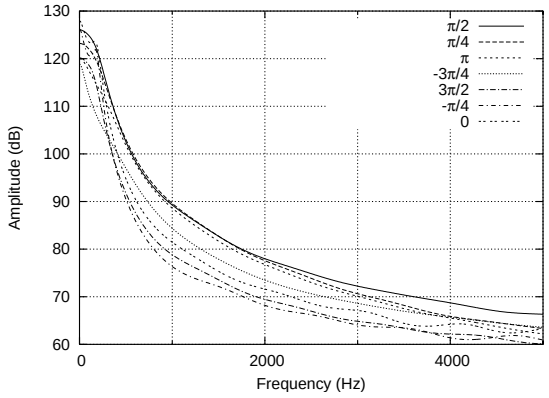


Figure 6: Spectral envelopes of laryngeal flows in various phases.

The EGG waveforms and high-speed images of the same subjects in Fig. 1 suggested that the phase delay of the VTF vibration to the VFs should be around $\pi/4$, i.e. $\theta' = -\pi/4$ [11]. This value is also supported by its frequent appearance in physical model simulation [9]. In Fig. 5, the laryngeal flow for θ's shows the similar characteristics of the drone in Fig. 1 and the opening duration is relatively less than the closing duration.

Fig. 6 show the spectral envelops of the laryngeal flows for different θ'. Among θ's in Fig. 6, the spectral tilt is the largest when $\theta' = -\pi/4$ (-16 dB/octave), smallest when $\theta' = \pi/2$ (-12 dB/octave).

3.2. VTF-modulation and LVT-resonance model

We set $A'_{g0} = 0.10$ cm and $\alpha' = 0.05$ cm. We normalize $u(t)$ by multiplying some real positive value and assume $u(t)$ as the glottal area function. We set the maximal glottal area $\max u(t)$ to 0.2 cm^2.We set the thickness of the VTF d' to 1.0 cm, the cross-sectional area of the laryngeal ventricle to 1.5 cm^2, and the depth of the laryngeal ventricle to 0.5 cm, $K = 20$ in Eq. (7). We used these values for calculation of the $H[t](z)$. The other values are the same as above. The laryngeal flows with LVT resonance of various θ's are as shown in Fig. 7.

In all cases, ripples are observed after the closure of the glottis. Fig. 8 shows spectra of two flows. The effect of the VTF resonance is observed around 2000 Hz. This feature is observed in all the synthesized sources.

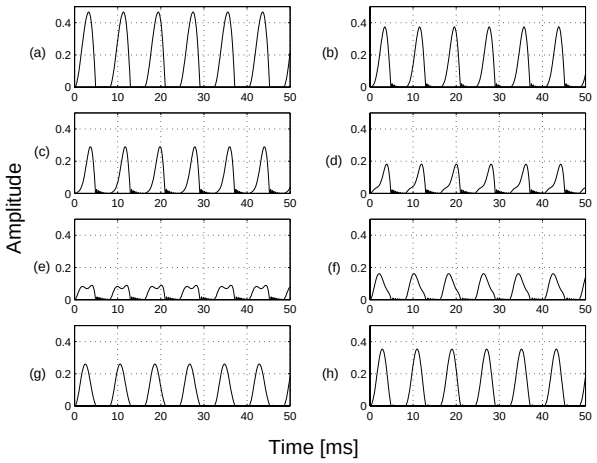


Figure 7: Laryngeal flows in various phases, with laryngeal ventricle resonance. (a) Original $u(t)$ without LVT-resonance, (b) — (h), $\theta' = \pi/2, 3\pi/4, \pi, -3\pi/4, -\pi/2, -\pi/4, 0$ respectively.

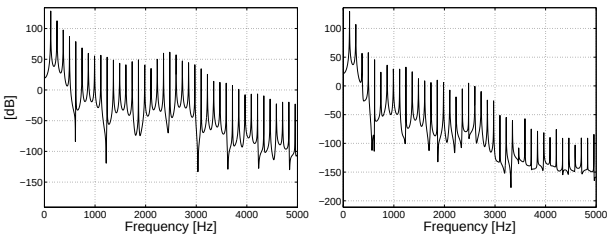


Figure 8: Spectra of the synthesized laryngeal voice. Left: $\theta' = \pi/4$, Right: $\theta' = -\pi/4$.

3.3. False glottal area at rest

When A'_{g0} is decreased, Eq. (6) implies that the resonance frequency is pushed higher. The spectra in Fig. 9 shows the spectra for different A'_{g0}. Other conditions are the same as above.

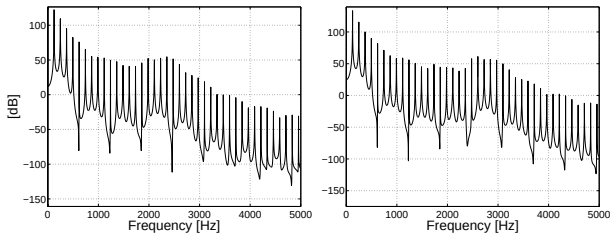


Figure 9: Spectra of the synthesized laryngeal voice with $\theta' = \pi/4$. Left: $A'_{g0} = 0.35 \max u(t)$, Right: $A'_{g0} = 0.65 \max u(t)$.

3.4. Modulation amplitude

We synthesized laryngeal flows by changing the amplitude of VTF vibrations α'. No significant trends are observed in the behaviors of the synthesized flows.

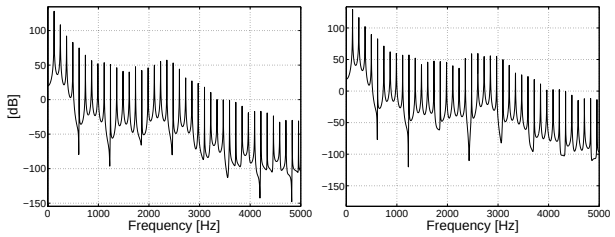


Figure 10: Spectra of the synthesized laryngeal voice with $\theta' = \pi/4$, $A'_{g0} = 0.35 \max u(t)$. Left: $\alpha' = 0.2 \max u(t)$, right: $\alpha' = 0.5 \max u(t)$.

3.5. Laryngeal source for kargyraa

If $\omega' = 2\omega$, then $u'(t)$ has a double-period of $u(t)$ and shows behavior similar to kargyraa phonation. From the characteristics of $u(t)$, in the middle of each period, the laryngeal flow reach to 0. However, the inverse-filtered karygraa voice maintains flow in each period. Uncompleted closure of the VFs are also observed in the physical model simulation [9]. In order to obtain the similar laryngeal flow shape, the second $u(t)$ flow must start before $T_n + T_p$ or $u(t)$ needs sufficiently large OQ.

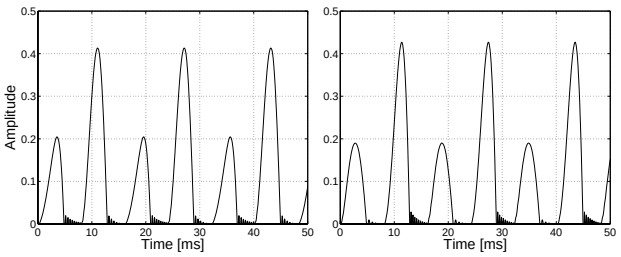


Figure 11: Spectra of the synthesized laryngeal voice with $A'_{g0} = 0.65 \max u(t)$, $\alpha' = 0.35 \max u(t)$. Left $\theta' = \pi/4$, right: $\theta' = -\pi/4$.

4. CONCLUSIONS

A new laryngeal flow model was proposed. We studied the acoustic characteristics of the model by changing its parameters. To obtain the laryngeal voice shape of the drone voice, VTF modulation is indispensable. In addition, to obtain ripples after the closure of the vocal folds, laryngeal ventricle resonance is effective. These results show the proposed model is effective for synthesizing pressed-type singing voices, such as throat singing. Parameter fitting in terms of analysis-by-synthesis and perceptual evaluation will be addressed as future works. In addition, an effective inverse-filtering method in cases that the source has poles and the filter has zeros must be studies.

Acknowledgments

We thank Seiji Adachi, Parham Mokhtari, Yoshinao Shiraki, Niro Tayama, and Masahiko Todoriki for their helpful discussions.

5. REFERENCES

[1] P. Alku, T. Bäckström, and E.Vilkman. Normalized amplitude quatient for parametrization of the glottal flow. *J. Acoust. Soc. Am.*, 112(2):701–710, 2002.

[2] K. E. Cummings and M. A. Clements. Glottal models for digital speech processing: A historical survey and new results. *Digital Signal Processing*, 5:21–42, 1995.

[3] G. Fant, J. Liljencrants, and Q.-A. Lin. A four-parameter model of glottal flow. *KTH STL QPSR*, pages 1–14, 1985.

[4] L. Fuks, B. Hammarberg, and J. Sundberg. A self-sustained vocal-ventricular phonation mode: acoustical, aerodynamic and glottographic evidences. *KTH TMH-QPSR*, 3/1998:49–59, 1998.

[5] N. Kobayashi, Y. Tohkura, S. Tenpaku, and S. Niimi. Acoustic and physiological characteristics of traditional singing in Japan. *Tech. Rep. IEICE*, SP89-147:39–45, 1990.

[6] T. C. Levin and M. E. Edgerton. The throat singers of tuva. *Scientific America*, Sep-1999:80–87, 1999.

[7] P.-Å. Lindestad, M. Sodersten, B. Merker, and S. Granqvist. Voice source characteristics in mongolian "throat singing" studied with high-speed imaging technique, acoustic spectra, and inverse filtering. *J. Voice*, 15(1):78–85, 2001.

[8] A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.*, 49(2):583–590, 1970.

[9] K.-I. Sakakibara, H. Imagawa, S. Niimi, and N. Osaka. Synthesis of the laryngeal source of throat singing using a 2×2-mass model. In *Proc. ICMC 2002*, pages 5–8, 2002.

[10] K.-I. Sakakibara, T. Konishi, H. Imagawa, E. Z. Murano, K. Kondo, M. Kumada, and S. Niimi. Observation of the laryngeal movements for throat singing — vibration of two pairs of the folds in human larynx. *Acoust. Soc. Am. World Wide Press Room*, 144th meeting of the ASA, 2002. `http://www.acoustics.org/press/`.

[11] K.-I. Sakakibara, T. Konishi, K. Kondo, E. Z. Murano, M. Kumada, H. Imagawa, and S. Niimi. Vocal fold and false vocal fold vibrations and synthesis of khöömei. In *Proc. ICMC 2001*, pages 135–138. ICMA, 2001.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

ACOUSTIC AND AERODYNAMIC ANALYSES OF SUPPORT AND RESONANCE DEMONSTRATED BY AN ELITE BARITONE SINGING TEACHER

R.C. Scherer, N. Radhakrishnan*, A. Poulimenos**, and F. Alipour****

*Department of Communications Disorders, Bowling Green State University
**School of Music, Indiana University
***Department of Speech Pathology and Audiology, University of Iowa
ronalds@bgnet.bgsu.edu

ABSTRACT

An elite singer/singing teacher (baritone) used a pedagogical approach to demonstrate levels of support and resonance: (1) the location of support – glottis (worst), chest (better), abdomen (best), and (2) locations of resonance – hard palate/straight tone (worst), mouth (better), sinus/head (best), using a single frequency (196 Hz), mezzoforte loudness, and the smooth /pae pae pae/ technique. The singer used consistently high levels of Ps, glottal adduction, F3 intensity level, and source spectral energy. The primary spectral differentiation was the general reduction of the formant frequencies from worst to best for both sequences, consistent with increased vocal tract length seen with nasal endoscopy. The singer raised Ps slightly, and kept OQ and SQ nearly constant, thus increasing aerodynamic power, acoustic power, AC glottal flow, and maximum flow declination rate. Acoustic efficiency was not used differentially.

1. INTRODUCTION

Typical research on speech and voice is driven by experimental designs where independent variables are well controlled. Here a different approach was taken. The singer/teacher/subject (coauthor AP) proposed two pedagogical concepts that have reality for him and his singing students, namely, "support" and "resonance". These became the "independent variables", and the control of these two variables was the "locations" of where they were sensed by AP when he sings.

For teacher AP, "support" is related to locations where the primary sensation of "work" is perceived to take place to create the balance of respiratory air pressures and laryngeal flow resistance during singing. For AP, support varies in the sense of location from a high position, the glottis, to a low position, the lower abdomen, and anywhere in between. From his pedagogical point of view for classical singing training, the lower in the torso the support is sensed, the better. The other concept, "resonance", is related to where the brightness (or presence) of the sound appears to have primary location to the singer. In the teaching of AP, usually the sense of resonance that is relatively high is better than if it were low, except when women sing in the chest register on their lowest pitches. The concepts and sensations of both support and resonance are taught as primary and interrelated components to singing improvement and the attainment of superior vocal skills.

The purpose of this study was to explore the vocal production differences among variations of support and resonance. Examination of the results may create hypotheses that link different sensations (auditory, kinesthetic) to different values of phonation variables, different levels of vocal technique in singing development, and different levels of aesthetic judgment about sound quality and artistry. In addition, the results may relate to theories and models of phonation and insights into optimal voicing.

2. METHODS

The subject (coauthor AP) was a 59-year-old baritone with a rich history of professional performance in opera (21 roles), oratorio (12 roles), and recital, with 32 years of singing teaching experience (and numerous students with major performance careers). He was in good physical health with no vocal complaints during the research recordings.

Three levels of support were studied. These were labeled "glottal", "chest", and "abdomen", and thus were spatially and kinesthetically varied. The glottal support level ("worst") was sensed as "work" primarily at the larynx level of the body. The second level, chest ("better"), corresponded to a circumferential body emphasis at the level of the lower ribs and the upper rectus abdominal muscles. The third location was "low abdominal" ("best"), with attention to circumferential action below the level of the umbilicus. The three levels of resonance locations were "hard palate" ("worst"), "mouth" ("better"), and "sinus" ("best"). The hard palate resonance location was associated with the perception of vibration felt on the hard palate. For the mouth resonance location, the sound appeared to emanate primarily from the mouth. For the sinus resonance, the auditory and vibratory sense of the sound was perceived by the singer to exist above the hard palate and in front of the face.

The subject stood in a sound-treated single-walled IAC booth. The subject himself placed the pneumotach facemask upon his face. He was trained to sense a complete closure of the cushion on the surface of his face. The comfortable pitch G3 (196 Hz) and a mezzoforte loudness were used throughout the experiment. He produced sung utterances of the syllable sequence without the mask, and then with the mask, for both the support variable and the resonance variable. Following these recordings, the subject was scoped with a flexible nasendoscope while repeating the two sequences.

Wide-band flow, oral pressure, and subsequent inverse filtered flow were obtained by using a Glottal Enterprises aerodynamic system. A Bruel & Kjaer sound level meter transduced the sound pressure level at 0.5 m from the mouth of

the subject. Signals were recorded onto an 8-channel Sony DAT recorder. The three middle /pae/s of a smoothly produced five-/pae/ string were chosen for analysis. LPC formant structures were obtained from two locations within each of the three /pae/ syllables using the Computerized Speech Lab (CSL, Kay Elemetrics Corporation). The locations were centered at a vibrato maximum past the first vibrato cycle and before the last vibrato maximum.

There was unwanted noise on the inverse filtered flow signal. The signal was low-pass filtered at 1500 Hz, then high-pass filtered at 100 Hz, using Matlab linear phase filters. Various measures were obtained: AC flow (from a hand-picked baseline, essentially through the middle of the baseline signal variation, to maximum flow), OQ (open quotient) and SQ (skewing quotient) obtained at the level of the average of the flow signal, and the MFDR (maximum flow declination rate).

The estimate of the subglottal pressure for a particular vowel was obtained as the average of the two adjacent /p/ occlusion pressure peaks. The flow for a /pae/ syllable was obtained as the average flow mid-way between the pressure peaks, which usually was near the middle of the vowel prolongation. Aerodynamic power (subglottal pressure times the average flow) was obtained. Acoustic output power was defined by the expression given in Titze [1], where SPL was corrected by adding 3.26 dB to the SPL values to account for the difference between levels with and without the mask. Acoustic efficiency was defined as the ratio of the output acoustic power and the input aerodynamic power.

3. RESULTS

3.1 Acoustics

The average and standard deviation of the formant frequency and amplitude values for the support sequence are shown in Figure 1. The formant frequency values for F1, F2 and F3 were highest for the glottal support location, next highest for the chest location, and lowest for the abdomen location. These reductions of formant values are consistent with the general finding from the scoping of increasing the vocal tract length from glottal to chest to abdomen support locations.

For the glottal support level, the average F3 level was essentially the same as the F1 level. In contrast, the chest support location had a dominance of F1 intensity (of 7.6 dB) over that of F3. Somewhat intermediate to these first two was the abdomen support location, for which the level of F1 was slightly higher (by 1.6 dB) than for F3. Given that the abdomen was the preferred support location, lower formant frequencies, nearly equal values of F1 and F3 levels, and an F2 level that is about 10 dB lower than F1 and F3 levels may be most preferred.

For the resonance sequence, the average formant frequencies were highest for the hard palate, less for the mouth, and lowest for the sinus location productions, with the exception that the mouth resonance had the lowest average F2 frequency value (Figure 2). These findings are fairly consistent with a lowering of the larynx from the hard palate to the mouth to the sinus location conditions, an observation also supported by scoping views. Similar to the support sequence, the highest intensity level of F3 corresponded to the location considered least desirable (hard palate). The mouth resonance location was distinguished by a low F2 intensity and a relatively low F3 intensity.

The spectral slopes of the glottal flow were relatively small in value, ranging from –6.5 to –8.1 dB/oct for support, and –7.6 to –10.0 dB/oct for resonance across the locations. This suggests strong adduction and relatively high energy levels in the source spectra.

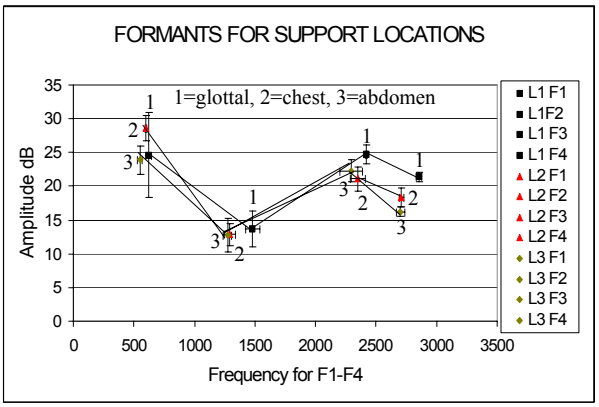


Figure 1: *Formants for support locations. L1: glottal location; L2: chest location; L3: abdomen location.*

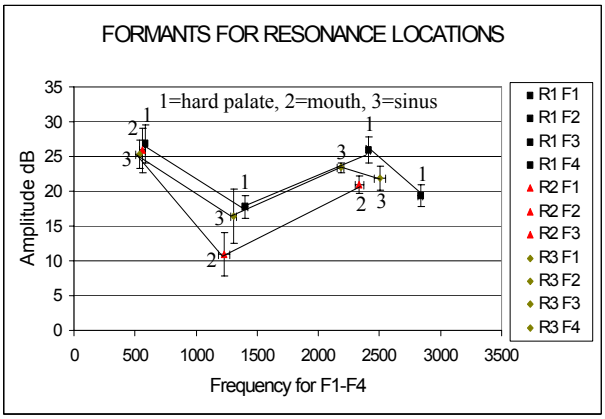


Figure 2: *Formants for resonance locations. R1: hard palate; R2: mouth; R3: sinus.*

3.2 Aerodynamics

Figure 3 indicates that the pressure tended to rise as the support was changed from glottal to chest to abdomen, as was also the case for resonance, from hard palate to mouth to sinus. For both sequences, flow also increased.

The glottal flow resistance was greatest for the glottal support location, and hard palate resonance location (Figure 4), decreasing to approximately half the values for the other two locations within the sequences. The SPL values are also given, and show that the subject increased his vowel intensity from worst to best locations.

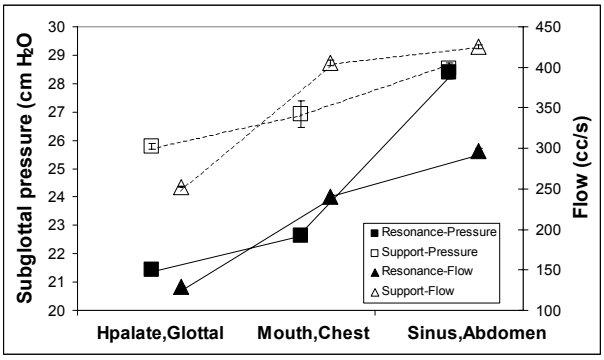


Figure 3: *Estimated subglottal pressure and average flow for both support and resonance production sequences.*

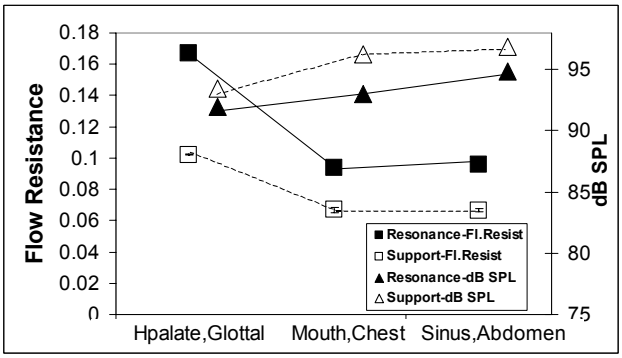


Figure 4: *Glottal flow resistance and SPL for both support and resonance production sequences.*

Figure 5 shows that aerodynamic power and acoustic power also increased from worst to best locations. Acoustic efficiency did not differ among the locations, as shown in Figure 6, with the exception of the hard palate resonance location.

The AC glottal flow and the maximum flow declination rate are shown in Figure 7. The AC glottal flow tended to increase, and the MFDR decrease from worst to best conditions, with one reversal for support. These results suggest that there was more acoustic strength to the better and best locations, and perhaps greater adduction for the worst support and resonance locations.

The open quotient and skewing quotient values were not significantly different across the locations, as can be seen in Figure 8.

4. DISCUSSION

The primary acoustic output factors for distinguishing the different levels of support or resonance appear to be the lowering of the formant frequencies – the lower the more preferred the voice, and maintaining an intensity of the F3 region similar to that of F1. Lower formant frequencies were consistent with the lower larynx positions seen via endoscopy.

The subject used a relatively high value of subglottal pressure, ranging from approximately 25 to 29 cm H_2O for the support location sequence, and 21 to 29 cm H_2O for the resonance location sequence. This suggests the use of a relatively high level of respiratory effort, although the subject did not report any undue exertion. Within these pressure ranges, the subject consistently increased the value from the worst to the best locations for both the support and resonance series. This suggests that the respiratory effort also increased slightly over the sequences. The small rise in SPL values from worst to best voice (approximately 93 to 97 dB for the support sequence, and 92 to 95 dB for the resonance sequence) is consistent with the small rise in subglottal pressure.

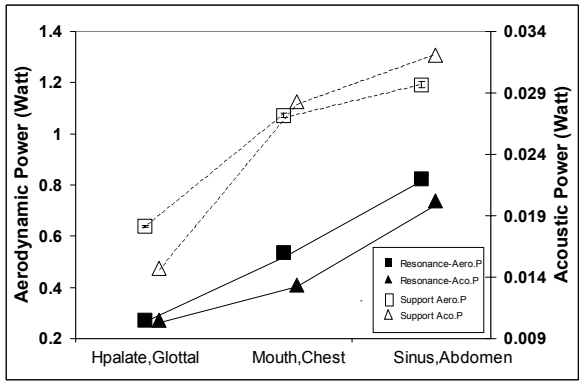


Figure 5: *Aerodynamic and acoustic power for both support and resonance production sequences.*

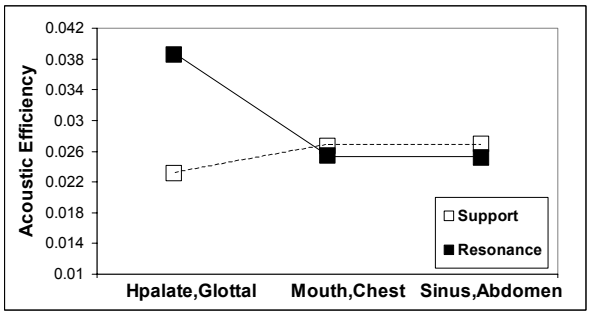


Figure 6: *Acoustic efficiency for both support and resonance production sequences.*

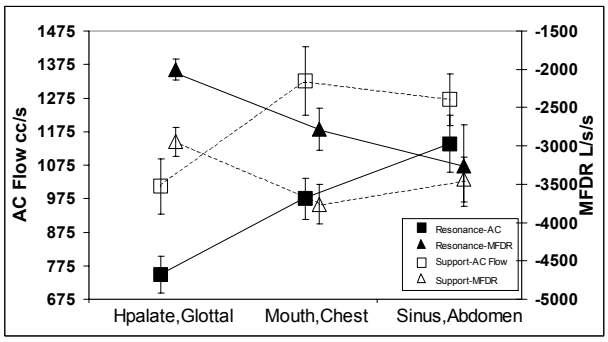


Figure 7: *AC glottal flow and MFDR for both support and resonance production sequences.*

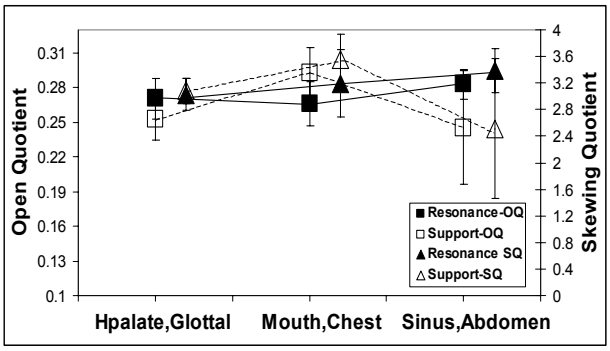


Figure 8: *Open quotient and skewing quotient for both support and resonance production sequences.*

The average airflow also increased from worst to best voice, changing from about 150 to 425 cm^3/s for the support locations, and 125 to 325 cm^3/s for the resonance locations, changes that are greater than a factor of two. The flow resistance values were higher than expected in most speech, and decreased from the worst location to a value that was about the same for the better and best locations for both sequences.

The aerodynamic power (pressure times flow) increased for both the support and resonance sequences. The acoustic power of the output pressure signal also increased across the worst to best sequences. The acoustic efficiency, however, was about the same for all locations of both sequences, with values of approximately 0.02 to 0.03, with the single exception of the hard palate location for resonance, which was approximately 0.04. Thus, better voicing was associated with greater acoustic power but not greater acoustic efficiency.

5. CONCLUSION

This study adopted the singing teacher's pedagogical kinesthetic concepts as the primary control variables and obtained objective acoustic and aerodynamic measures in the attempt to differentiate the levels of the performed control variables. The singer appeared to use the following strategy when demonstrating the change from worst to best voice relative to support and resonance "locations": The worst condition was characterized by compressed adduction, raised larynx, and relatively high subglottal pressure. When moving to improved voice, he then (a) increased his vocal tract length to lower the formant frequencies, and (b) used greater respiratory power and slightly less glottal compression (not discussed) to create greater AC flow and MFDR values to excite the vocal tract, so that the most preferred voice had relatively the same intensity values for the first and third formant regions with an overall higher SPL level. It is noted that acoustic efficiency was not a factor in these changes.

6. ACKNOWLEDGEMENTS

The authors wish to thank Jeff Searl, Sid Khosla, Meena Agarwal, Tracy Sheaks, Jennifer Rettig, Kelly Kaufman, and Stephanie Holcombe for their help in this study.

7. REFERENCE

[1] Titze, I.R. (1994). Principles of Voice Production, Prentice Hall, Englewood Cliffs, 1994.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

PERCEPTUAL RELEVANCE OF THE 5 kHz SPECTRAL REGION TO SEX IDENTIFICATION IN CHILDREN'S SINGING VOICES

Peta J Sjölander

Department of Speech Music and Hearing
Kungl Tekniska Högskolan
peta@speech.kth.se

ABSTRACT

In a previous investigation, the recorded voices of 320 children were evaluated with respect to perceived gender and actual sexual identity by a group of experienced listeners. The audio recordings of a sub-group of those children (30 boys and 29 girls aged between 3 and 12 years) were later subjected to an acoustic evaluation using long-term average spectrum (LTAS) analysis and the results were compared to those of the perceptual evaluation. The results revealed a peak in the average spectrum at 5 kHz for children perceived confidently as boys (whether male or female in actuality), and a flat spectrum at 5 kHz for children perceived as girls. Those findings suggested that an acoustically measurable long-term, and therefore persistent, difference may exist between boys' and girls' voices. However, it was unclear if the peak itself carried any perceptual information. In the present experiment, therefore, the recordings have been re-submitted for perceptual analysis. This time each voice was presented twice to expert listeners, quasi-randomly, the second playing of the sample having been filtered to manipulate the effect of the higher frequencies on perception. The listeners were asked to judge the sex of each subject as presented via audio headphones in order to shed light on the significance of the 5 kHz peak.

1. INTRODUCTION

Long-term average spectrum (LTAS) analysis provides spectral information averaged over time. It is characterised by peaks that relate to the average of formant frequencies produced during longer-term utterances, such as a read text and singing, and thus offers some representative information on voice timbre. The method is particularly useful when persistent spectral features are under investigation, features that might not be apparent in shorter samples or in single speech sounds.

In an experiment performed some years ago, the recorded singing voices of 320 children, aged between 4 and 12 years, were played to a panel of expert listeners (unpublished manuscript, but see White, 2001 and Welch & Howard, 2002 for a description of some of the findings). At the time of recording, all children were enrolled in primary schools in the south-east of England. They all sang the same nursery rhyme "There's a fox in a box". The song, which is around 10-15 seconds in length, had been specially learned by the children for the purpose of the study. The pitch range is one octave and subjects were free to choose their own starting pitch.

The listeners in the experiment were asked to give their opinion as to the sex and age of each voice they heard. They were instructed to rate the perceived sex of the singer on a multi-point scale ranging from "I'm sure it's a boy" to "I'm sure it's a girl", with "could be either" as the mid point of the scale. Averaged across listeners, and normalised for sex, this scaling procedure provided a confidence predictor for each voice yielding a high score (max 7) for voices correctly identified with confidence by all listeners. A low score (min 1) reflected voices that were wrongly identified as being either male or female with confidence by all listeners. In this way, all correctly identified subjects, either male or female, were "high-score" and all incorrectly identified subjects were "low-score".

Using the results from the perceptual test, an investigation of gender-related acoustic differences in children's voices was later carried out (White, 2001). Fifty-nine voices were selected: 20 high-score boys' voices (HM) and 20 high-score girls' voices (HF); plus 10 low-score boys' voices (LM) and 8 low-score girls' voices (LF). There were few children with scores lower than 3 and so this was the criteria chosen for "low score". The investigation showed that the boys' and girls' voices were different acoustically when compared using LTAS analysis. This difference manifested as a peak in the spectrum at around 5 kHz for voices that were clearly identified as male (i.e., high-score boys as well as low-score girls). The same was true regardless of the sex of the singer; where girls had been perceived with confidence as boys, the 5 kHz peak also existed. For voices identified as female (i.e., high-score girls and low-score boys), the spectrum was flat in the 5 kHz region. Confidence in the results was increased as gender had already been identified perceptually in these subjects. The results thus provided some evidence of a relationship between perceived sex and the averaged spectrum. The corroborating factor was that the peak existed also in the spectra of low-score girls' voices (i.e., those voices that were wrongly but confidently identified as belonging to boys). Some questions as to the origin of the 5 kHz peak were raised, however, such as whether harmonic energy was present in the peak, and whether this higher-frequency information was relevant to perception. So, whilst the likelihood existed that LTAS reflected timbral qualities relating to sex, the question of the importance of the 5 kHz region warranted further investigation.

2. METHOD

The voices of a total of 47 subjects (26 boys and 21 girls) were selected from the group of 59 used for the LTAS study (White, 2001)[1]. Table 1 presents the mean scores (taken from the earlier perceptual study) and mean ages for each singer group as used in the present study.

Eight listeners were included in the present perceptual test,

		Mean		Median	
Group	**n**	**Age**	**Score**	**Age**	**Score**
High-score male	16	9	7.0	9	7.0
High-score female	18	8	6.8	8	6.8
Low-score male	7	6	2.4	5	2.5
Low-score female	6	8	2.0	7.5	2.3

Table 1: *Mean and median ages and scores for each group resulting from an earlier perceptual study (see White, 2001; Welch & Howard, 2002. The terms "high-score male" and "high-score female" refer to the subjects who were most often correctly and confidently identified as either male or female, respectively. The terms "low-score male" and "low-score female" refer to the subjects who were least often identified correctly as either male or female, but who were miss-identified with confidence. These facts are reflected in the averaged scores.*

one of whom had also participated in the original perceptual study several years earlier. Four listeners were male and four female. Five were British, two Swedish, and one was from the United States of America. The Swedish nationals were fluent in the English language. Four of the five British listeners, two women (E2 and E5) and two men (E3 and E4), were included in the expert listener group. The fifth British listener described himself as having "very little experience" and so was placed in the naïve listener group (N3). The fifth expert, E1, is a Swedish female. The remaining two naïve listeners (N1, female, and N2, male) had experience of children's singing voices, but only as parents. See Table 2 for a description of each listener.

Identifier	**Sex**	**Nationality**	**Experience**
Expert Group			
E1	F	Swedish	School singing teacher
E2	F	British	Child voice researcher and parent of young child
E3	M	British	Child voice researcher and schoolteacher
E4	M	British	Child voice researcher and parent of young children
E5	F	British	Professional choir director and parent of young child
Naïve Group			
N1	F	American	Parent of young children
N2	M	Swedish	Parent of young children
N3	M	British	No experience

Table 2: *Description of listeners, presented by group, included in the present perceptual test.*

Each of the 47 audio files was manipulated using a so-called *notch* filter provided by a KTH freeware speech editing software system, Wavesurfer™[2]. This notch filter eliminated the part of the audio signal between 4.25 and 5.75 kHz. The procedure produced a total of 94 audio files which were then recorded onto a CD in a quasi-random order, together with 16 repeat files (to test for inter-listener reliability) making a total of 116 files, as follows. First the 116 files were placed randomly into 12 groups

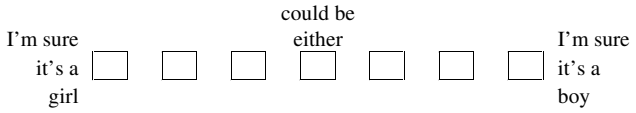


Figure 1: *Multi-point scale used to evaluate each voice. Each listener was asked to place a cross in one box to say he or she believes the voice to be either that of a boy or a girl. The degree of confidence is reflected in where the cross is placed with regard to the extremes "I'm sure it's a girl" versus "I'm sure it's a boy", or the middle point which represents the uncertainty of it "could be either".*

of 10 files (the last group containing just 6 files). Each group contained both filtered and original (non-filtered) files. Next the filenames were checked to ensure no two matching files, i.e., the original and the filtered versions, were presented consecutively on the CD track, in which case one of the files was exchanged for a non-matching file in another group. Lastly, the groups of 10 (or 6) files were placed in order of size, to add an extra degree of randomisation, and were edited into one soundfile using a commercially available speech editing software, Soundswell™[3]. Twelve tracks were thus burned onto the CD. A test sheet was also produced with 116 (10 per page) occurrences of the seven-point scale as seen in Figure 1. The CD and test sheets were given to the listeners with instructions for how to conduct the test. All listeners heard the samples played free-field via a computer-based CD player in a quiet room. They were told they would hear children singing a short song, sometimes more than once, and that there was an uneven number of boys and girls on the CD. They were instructed to indicate a level of confidence that this was a boy's or a girl's voice by placing a cross on the scale for each voice. They were told they should listen only once to each voice in order to give an intuitive evaluation.

The values for each voice as provided by the listeners via the test sheets were entered onto a computer spreadsheet. The results were separated for each listener into filtered and non-filtered files. The results were averaged across singers according to the groups defined in the introduction above (HM, HF, LM and LF). First, the accuracy of each listener's responses with regard to the sex of the singer was calculated. Second, the response for each filtered file was compared to its filtered counterpart, and the difference calculated.

3. RESULTS

Figure 2 shows the individual listener and group means for accuracy in identifying sex correctly. Except for expert listener E2, all listeners, both naïve and expert, were unable to accurately identify the sex of the low-score male (LM) singers from the original recordings (top left-hand corner of the figure). In a similar fashion, none of the listeners were able to discern sex correctly from the filtered versions of those files. The group averages for the original files are a little above 40% and 20%, respectively, for the expert group and naïve group. These values drop by more than a quarter for the filtered files. For the low-score female (LF) group, only two experts were able to tell they were

listening to girls' voices, and then only when presented with the original files (bottom left of the figure). The naïve listeners as a group were slightly more accurate, and the expert listeners slightly less accurate, at identifying the sex of these low-score girls compared to the low-score boys. Two of the experts were markedly more accurate than the other listeners at identifying the sex of the low-score girls from the original files, and all listeners scored 50% or lower for the filtered files.

In general, all listeners were able to identify the high-score boys (HM) and the high-score girls (HF) correctly when listening to both the original recordings and to the filtered files (graphs on the right-hand side of the figure). The only exception was expert listener E3. However, his accuracy increased from 42% to 63% for the filtered HM files. Listener N3 was only accurate to a level of 58% when presented with the original files, but his accuracy also improved to 68% for the filtered files. The naïve group means for the HM files was 77% and 79% for the original and filtered files, respectively. The equivalent mean values for the expert group were 81% and 85%. Accuracy scores for the HF files, compared to all other singer groups, were greatest overall for both expert and naïve listeners. HF identification scores improved in all listeners, except two, when presented with the filtered files. Of the two listeners, both in the expert group, whose scores did not improve, one score remained the same and one was lower for the filtered files.

Figure 3 presents the difference in scores, given by individual listeners, between the original and filtered versions of each file. The top graph shows the results for the naïve listener group and the bottom graph shows the results for the expert group. The results (mean across singer groups for each listener) are displayed as a shift either to the left (negative) or to the right (positive) from zero. Zero represents the score of between 1 and 7 given in the present experiment for the original file, and the shift represents the difference in score between the original file and the filtered file. A negative shift relates to a shift in sex identification towards "I'm sure it's a girl", and a positive shift relates to a shift towards "I'm sure it's a boy". Individual listener results are displayed as black bars. Listener group means are displayed (underneath the individual scores in each case) with white bars. When scores were averaged within singer groups, LM, HM, LF and HF, individual listeners tended to give a shift of less than 0.5 (half of one of the 7 score points), whether in a negative or positive direction. Both positive and negative shifts were observed within each singer group for both naïve and expert listener groups. An exception to this is that, for the LF group, all listeners except one (E5) demonstrated either a negative shift or no shift at all in their scores. Also noteworthy is that for the naïve listeners' group means, all shifts were in a positive direction (a shift of 0.2 or less) except for LF which produced a negative, and greater, shift of 0.7. The equivalent values in the expert group show that

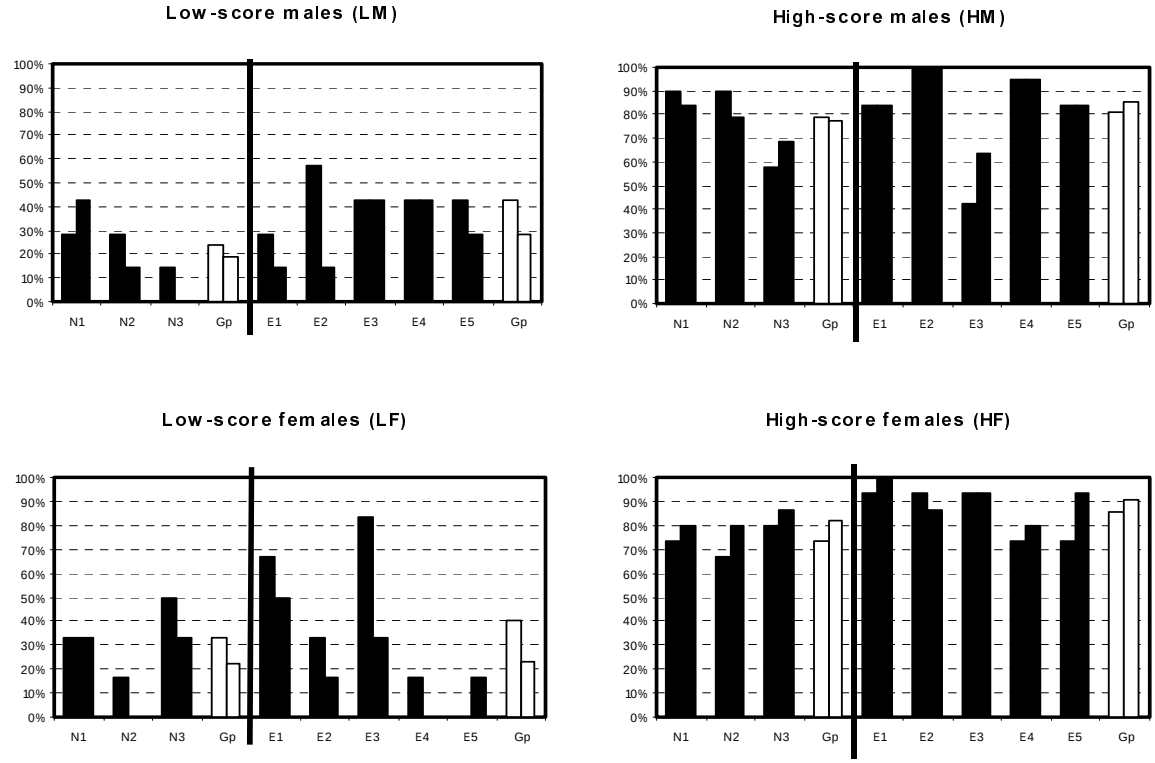


Figure 2: *Accuracy of listener and group means in identifying sex of the singers (max 100%) from the original recordings versus the filtered files. Values are given per singer group, and are presented in pairs for each listener, left column = unfiltered (original) file, right column = filtered file. Naïve listeners' results are to the left of the bold line in each case. Listener's individual scores are in black. Group means are white. "High-score" singers are those who were most often identified correctly for sex in the previous perceptual study. "Low-score" singers are those who were least often identified correctly for sex.*

two singer groups produced negative shifts, HM 0.2 and LF 0.5, and two groups produced positive shifts, LM 0.3 and HF 0.1.

With regard to reliability across listeners, HF was the only singer group that produced consistent results across all expert listeners: they all either showed no shift at all, or they shifted their judgement towards "I'm sure it's a boy". In addition, two of the three naïve listeners, N1 and N2, followed this pattern.

4. DISCUSSION

The present study has attempted to answer the question: How important is the 5 kHz peak to perception? By removing the area of the peak, and by playing the original and filtered files to a panel of expert and naïve listeners, a comparison was made between the listeners' identification of sex both with the 5 kHz area included and with it excluded. It was hypothesised that sex identification would be affected if the 5 kHz region were removed from the signal. This appears to be the case. The results showed that listeners were more likely to be accurate in their identification of sex of the singer if the file was not filtered. The effect was small, however, and not the same within all singer categories. In the case of the high-score girls' voices in particular, the presented filtered files provided greater accuracy in determining the sex of the singer than did the original files. From this, we could hypothesise that some children's voices contain information, such as noise perhaps, in the 5 kHz region that masks vital cues to female voice identification. The greater the amplitude in this area, the more likely we are to identify the voice as being male. The difference in accuracy of sex identification between original and unfiltered files was greater if the voices had been deemed "difficult" to identify in the earlier perceptual study (i.e., low-score male and female voices).

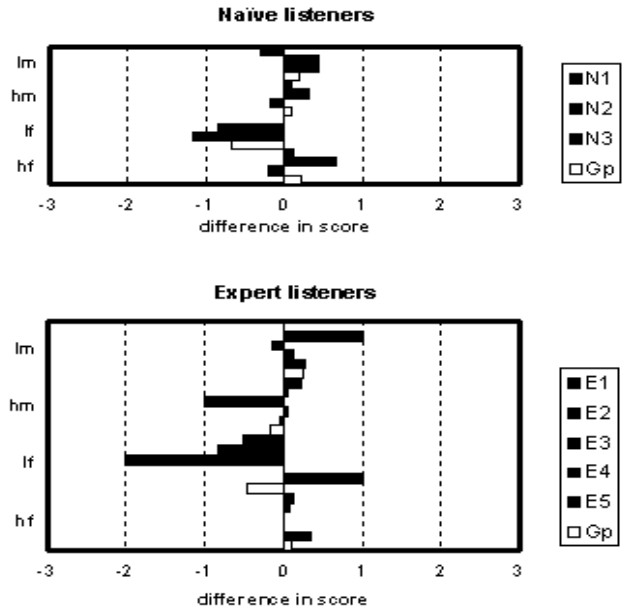


Figure 3: *Difference in sex identification score for each listener (black bars) and group means (white bars) between original and filtered files. A shift to the left indicates a greater confidence that the voices were female, while a shift to the right means greater confidence that the voices were male.*

The effect of the 5 kHz area on perception was not the same for all listeners. For some, boys' voices were more recognisable when listening to the filtered compared to the original files, for others it was the girls' voices that proved easier to identify correctly. With regard to the shift in score between the original and filtered files, it seems there is no outstanding data that can be considered more than a trend. Most of the shifts were small when observing the group data. The singer group LF produced a bigger shift in any direction than the other groups, which would support the theory mentioned above that a peak in the 5 kHz region relates to a boy-like quality.

It should be remembered that there might be some cultural aspects to gender perception. However, there appears to be little evidence of this in the present data. The British listeners performed better than the non-British listeners, but not remarkably so. Indeed, the Swedish expert was at least as accurate in many of her identifications as the British listeners, and the British naïve listener was no more able to identify sex correctly than the Swedish or American naïve listeners.

The present listener group was too small to give much information. However, the results confirm that removing the 5 kHz peak affects listener accuracy in identifying sex of a singer.

5. CONCLUSIONS

Differences were seen both between and within listeners, as well as between listener groups and between singer groups. Some conclusions can be drawn, however, as follows:

- The data support the notion that a peak in the area of 5 kHz produces a greater likelihood that the voice will be identified as male.
- Both the naïve and the expert listeners thought the low-score girls sounded more female when the files were filtered, and the low-score boys sounded more male.
- Conversely, filtered voices of the high-score girls sounded more boy-like, and high-score boys sounded even more like boys to the naïve group but less like boys to the expert group.
- More studies are needed, not least to investigate the phenomenon that the filtered files produced an increase in accuracy of sex identification of the high-score girls' voices.

6. REFERENCES

[1] White, P., "Long-term average spectrum (LTAS) analysis of sex- and gender-related differences in children's voices", Log. Phon. Vocol., 26, 97-1001, 2001

[2] Welch, G. and Howard, D., "Gendered Voice in the Cathedral Choir". Psych. Music, 30, 102-120, 2002.

[1] The voices of 11 subjects, 7 boys and 4 girls, had to be eliminated due to poor quality of the available recordings.

[2] Wavesurfer™ (www.speech.kth.se/wavesurfer)

[3] Soundswell™ (www.hitech.se/development)

MUSIC PERFORMANCE

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

LARGE-SCALE PERFORMANCE STUDIES WITH INTELLIGENT DATA ANALYSIS METHODS

Gerhard Widmer

Department of Medical Cybernetics and Artificial Intelligence,
University of Vienna, and
Austrian Research Institute for Artificial Intelligence, Vienna
gerhard@ai.univie.ac.at

ABSTRACT

The paper reviews current work in the area of AI-based music performance research. The focus is on approaches that study large amounts of empirical data (performance measurements) with advanced data analysis methods from fields like machine learning and data mining. Examples from our recent work demonstrate what is currently possible. In particular, we show how aspects of the individual performance style of concert pianists can be elucidated with novel visualisation and data analysis methods.

1. INTRODUCTION

Expressive music performance as the deliberate use by performers of parameters like tempo, timing, dynamics, articulation, etc., has been the focus of a lot of quantitative and formal modelling research in recent decades (e.g., [6, 7, 9, 14], to name just a few). It has been shown in computer simulations that formal models of expressive performance such as Todd's [10] can produce quite sensible music [14], and even that computers can learn to play new pieces by drawing analogies to known interpretations of other pieces [13].

Several distinct research methodologies have been used for the study of expressive performance, including statistical data analysis [7], analysis by synthesis [9], systematic computer simulation of formal performance models [14], and controlled experiments with real musicians and listeners [1, 6].

It is impossible to give a comprehensive review of the state of the art of performance research in a short paper like this (the bibliography in Gabrielsson's excellent overview article [4] comprises 24 pages!). Instead, we will focus on one particular line of research that our own research group is pursuing — the computer-based analysis of large amounts of musical performances using Artificial Intelligence (AI) methods —, simply in order to advocate the use of AI and intelligent data analysis in performance research, and to indicate where the promise, but also the limits of such an approach are, and where the field of audio analysis may help.

The main characteristic of our approach to the study of expressive performance is that we collect large amounts of 'real-world' data, that is, measurements of expressive aspects such as rubato and dynamics, from recordings of actual performances by highly skilled musicians, and analyse these data with inductive machine learning, data mining, and pattern recognition methods. The emphasis here is on 'large amounts': we are working with hundreds of performances of extended pieces, which makes 'manual' analysis virtually impossible and requires the use of intelligent data analysis methods. In such a data-intensive setting, we will probably lose some detail relative to the more controlled experiments common in performance research, but on the other hand we hope to gain a broader empirical basis. We want to make sure that the patterns discovered are statistically significant and occur in performances of many musical styles, and by many different performers.

Two general types of questions regarding expressive music performance guide our research. First, are there general, fundamental principles of music performance that can be discovered and characterised in formal and quantitative terms? In other words, to what extent can a performer's expressive choices be predicted? And second, is it possible to formally characterise and quantify aspects of individual artistic style?

The first set of questions thus relates to *commonalities* between different performances and different performers, while the second focuses on the *differences*. The following project presentation is structured according to these two types of questions.

2. DISCOVERING COMMON PERFORMANCE PRINCIPLES

In the search for common, fundamental principles of expressive performance, we make use of *inductive machine learning*. The goal is to discover general rules of certain aspects of performance (e.g., expressive timing) that are valid in the sense that they can predict, with a certain level of accuracy, how performers are going to play particular musical passages. Such predictive rules would then point to general principles that performers seem to follow, and might contribute to our understanding of function and workings of expressive performance.

The following presentation is kept very short and only points to the most important results, because most of this work has been extensively published (e.g., [11, 13]).

2.1. Machine Discovery of Performance Rules

In a first attempt at finding general rules of performance, we developed a new inductive rule learning algorithm and applied it to the task of learning note-level rules for timing, dynamics, and articulation, where by 'note-level' we mean rules that predict how a pianist is going to play a particular note in a piece — slower or faster than notated, louder or softer than its predecessor, staccato or legato. This should be contrasted with higher-level expressive strategies like the shaping of an entire musical phrase (e.g., with a gradual ritardando towards the end), which will be addressed in section 2.2 below.

The training data used for the experiments consisted of recordings of 13 complete piano sonatas by W.A. Mozart (K.279–284, 330–333, 457, 475, and 533), performed by a Viennese concert pianist on a Bösendorfer 290SE computer-controlled piano, from which every detail of timing, dynamics, and articulation could be computed. The resulting dataset comprises more than 106,000 performed notes and represents some four hours of music.

The experiments were performed on the melodies (usually the soprano parts) only, which gives an effective training set of 41,116 notes. Each note was described by 29 attributes that represent both intrinsic properties (such as scale degree, duration, metrical position) and some aspects of the local context (e.g., local melodic contour around the note). From these 41,116 examples of played notes, the computer learned a small set of 17 quite simple classification rules that predict a surprisingly large number of the note-level choices of the pianist. For instance, four rules were discovered that together correctly predict almost 23% of all the situations where the pianist lengthened a note relative to how it was notated (which corresponds to a local slowing down of the tempo).

Experiments also showed that most of these rules are highly general and robust: they carry over to other performers and even music of different styles with virtually no loss of coverage and precision. In fact, when the rules were tested on performances of quite different music (Chopin), they exhibited significantly higher coverage and prediction accuracy than on the original (Mozart) data they had been learned from. What the machine has discovered really seem to be fundamental performance principles.

Interestingly, some of the rules discovered by the machine bear a strong resemblance to performance rules postulated by other researchers, based on musical intuition and experience. In particular, some learned rules turned out to be variants of rules specified in the *Director Musices* model developed at KTH [3]. In this way, our experiments produced further empirical evidence for the relevance and validity of the KTH model. A detailed description of the rules, quantitative experimental results, and relations to other rule models of expressive performance can be found in [11].

2.2. Learning Multi-level Phrasing Strategies

Not all of a performer's decisions regarding tempo or dynamics can be predicted on a local, note-to-note basis. Musicians understand the music in terms of a multitude of more abstract patterns and structures (e.g., motifs, groups, phrases), and they use tempo and dynamics to 'shape' these structures, e.g., by applying a gradual crescendo or decrescendo to entire passages. Music performance is a *multi-level* phenomenon, with musical structures and performance patterns at various levels embedded in each other.

Accordingly, the set of note-level performance rules mentioned above is currently being augmented with a multi-level learning strategy where the computer learns to predict elementary tempo and dynamics 'shapes' (like a gradual crescendo-decrescendo) at different levels of the hierarchical musical phrase structure, and combines these predictions with local timing and dynamics predicted by learned note-level models. Preliminary experiments, again with performances of Mozart sonatas, yielded very promising results [13]. In fact, an 'expressive' Mozart performance generated by the computer after learning from real performances of other Mozart pieces won second prize in a 'computer performance rendering contest' in Tokyo in 2002, where computer interpretations of classical music were rated by a panel of listeners. Again, that indicates that what the machine extracts from the performance data seem to be at least 'reasonable' phrase-level performance patterns. A next step will be to build an explicit model of these patterns so that they can be compared to existing models of phrase-level timing and dynamics (e.g., [10]).

3. CURRENT WORK: CHARACTERISING ASPECTS OF INDIVIDUAL STYLE

A second, complementary set of questions guiding our research concerns the *differences* between individual artists. Here, the goal is to characterise and quantify aspects of the individual style of different artists; our current focus is on famous pianists. The work presented in this section is ongoing research, and the following presentation will focus on the general research methodology we are developing, rather than on truly established discovery results.

3.1. Data Extraction and High-level Representation

With famous pianists, the only available source of data are *audio recordings*, which precludes the kind of precise measuring and detailed analysis possible with MIDI (or Bösendorfer) data. At the moment, we analyse recordings at a very crude level. We extract tempo and overall dynamics at the *beat* level. Given an audio recording of a (piano) performance, we measure the time points that correspond to beat locations, and compute the beat-level tempo from the (varying) time intervals between these points. Beat-level dynamics is also computed from the audio signal as the overall loudness/amplitude of the signal at the beat times.

Finding the beat in audio recordings is a research problem in its own right. We have developed a quite effective *beat tracking algorithm* and integrated it into an interactive beat tracking system called BeatRoot [2] that permits a human user to correct mistakes made by the system (which are inevitable, especially in classical music with very expressive timing) and re-track pieces based on new information. Using BeatRoot, we have measured beat-level tempo and dynamics variations in hundreds of commercially available recordings by famous pianists.

The resulting performance data — beat-level tempo and dynamics curves — can be elegantly represented in an integrated way as trajectories in a tempo-loudness space that show the joint development of tempo and dynamics over time [5]. Figure 1 shows a complete trajectory representing a performance of a Chopin Ballade by Artur Rubinstein. The line is produced by interpolating between the measured tempo and dynamics points, and smoothing the result with a Gaussian window to make the general trends visible. The degree of smoothing controls the amount of local variation that becomes visible.

We have developed an interactive visualisation system called the Performance Worm [2] that computes and visualises such performance trajectories, and this tool permits intuitive analysis of high-level strategies in individual performances. But the trajectory representation also provides the basis for more detailed quantitative analysis, again with data analysis (data mining) methods from the field of AI. We are currently exploring various avenues towards the characterisation of individual performance style, and will briefly introduce some of these in the following subsections.

3.2. Performance Alphabets

The performance trajectories must first be converted into a form that is accessible to the data analysis machinery being developed

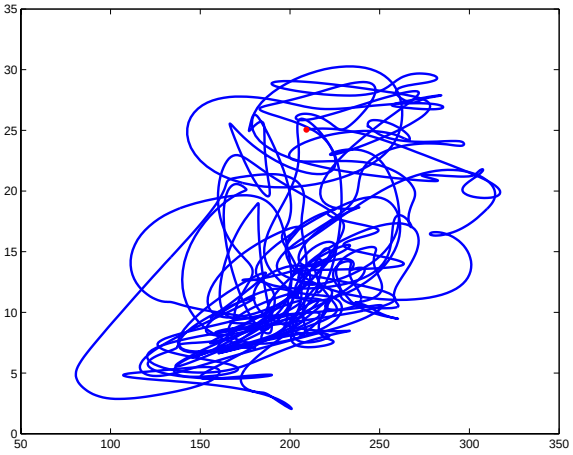


Figure 1: Smoothed tempo-loudness trajectory representing a performance of Frédéric Chopin's Ballade op.47 in A♭ major by Artur Rubinstein. Horizontal axis: tempo in beats per minute (bpm); vertical axis: loudness in *sone*.

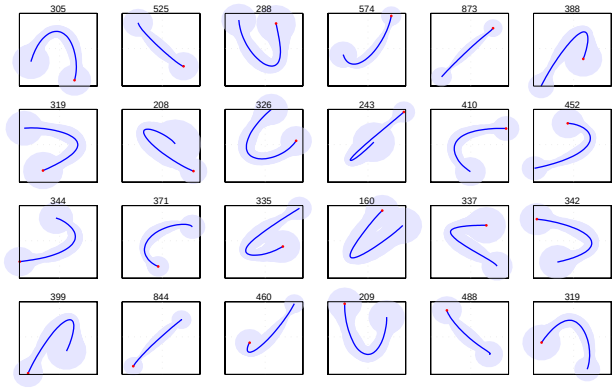


Figure 2: A 'Mozart performance alphabet' (cluster prototypes) computed by segmentation, mean and variance normalization, and clustering, from performances of Mozart piano sonatas by six pianists (Daniel Barenboim, Roland Batik, Vladimir Horowitz, Maria João Pires, András Schiff, Mitsuko Uchida). To indicate directionality, dots mark the end points of segments. Shaded regions indicate the variance within a cluster.

in data mining. To that end, the trajectories are cut into short segments of fixed length, e.g., 2 beats, which are then optionally subjected to various normalisation operations (e.g., mean and/or variance normalisation to abstract away from absolute tempo and loudness and/or absolute pattern size, respectively). The resulting segments are grouped into classes of similar patterns via *clustering*. For each of the resulting clusters, a *prototype* is computed. These prototypes represent a set of typical elementary tempo-loudness patterns that can be used to approximately reconstruct a 'full' trajectory (i.e., a complete performance). In that sense, they can be seen as a simple *alphabet* of performance, restricted to tempo and dynamics. Figure 2 displays a set of prototypical patterns computed from a set of Mozart sonata recordings by different artists.

3.3. Analysis of Global Style Characteristics

On the basis of such performance alphabets, one can perform straightforward statistical analyses regarding the frequency of occurrence of such prototypical patterns in performances by certain pianists. To aid such analysis, we have developed some intuitive visualisation techniques that make the distribution of performance patterns in pianists, and certain differences between them, immediately evident. That provides a very global view of aspects of personal style, such as "pianist A tends to use abrupt tempo turns combined with rather constant dynamics" or "pianist B combines a crescendo with a ritardando much more often than other pianists". An example of such a visualisation can be found in [12]. An extensive study along these lines using large numbers of Mozart and Chopin performances and some 15 pianists is currently under way.

3.4. Discovery of Performance Patterns

A more direct way of quantifying aspects of individual performance style is to search for specific extended patterns in performance trajectories that are somehow typical of a particular pianist. Again, we are currently developing data mining algorithms for this task. Finding distinctive and statistically significant patterns in multivariate time series such as our tempo-loudness performance trajectories is a complex data analysis problem. Our current approach is based on approximating trajectories by sequences of performance alphabet 'letters' and searching for frequent and discriminative substrings in these sequences (see [12] for more details).

Just to illustrate, Figure 3 shows several instances of a particular performance pattern found in various Mozart performances by Mitsuko Uchida, and rarely found in any of the other pianists mentioned in Figure 2. The pattern represents a particular way of combining a crescendo-decrescendo (the vertical movement of the trajectory) with a slowing down during the loudness maximum and afterwards. Assessing the statistical and, above all, musical significance of such discovered patterns is a rather difficult problem that we are currently working on.

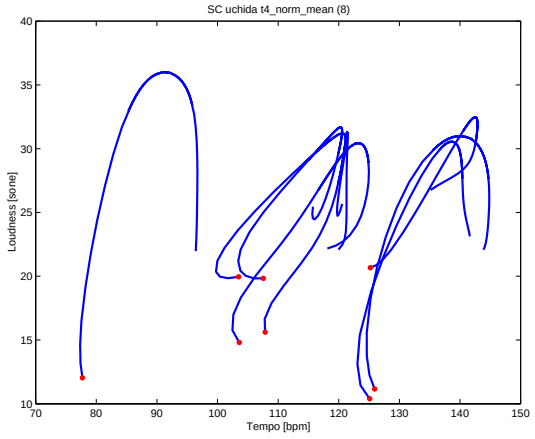


Figure 3: A set of similar performance shapes possibly of Mitsuko Uchida), discovered in tempo-loudness trajectories of Mozart sonata recordings. To indicate directionality, a dot marks the *end point* of a segment.

3.5. Automatic Identification of Performers

Another way of trying to quantify individual performance style is to develop computer programs that attempt to identify artists on the basis of their performance characteristics. In initial experiments with state-of-the-art machine learning algorithms we showed that a computer could learn to partly differentiate several pianists from the Vienna University of Music, given only one test recording by each pianist [8]. The respective performances were made on a Bösendorfer 290SE computer-controlled piano, so that very precise and detailed performance information was available.

Recently, we have managed to show that a computer can also learn, to a certain extent, to distinguish famous pianists based only on aspects of their high-level performance trajectories. The results are still very preliminary, and we have not yet managed to pinpoint precisely which features are the most distinguishing ones, but the current results do indicate that there is still relevant information contained in this high-level representation. More information can be found in another paper in this volume [15].

4. CONCLUSION

As we hope to have shown in this paper, the use of Artificial Intelligence methods (machine learning, data mining, pattern recognition) offers several benefits to empirical performance research. It enables us to analyse large amounts of data, which may make it easier to establish the (at least statistical, if not musical) significance of whatever performance patterns are found. Also, the computer can make autonomous discoveries that may or may not match the biases and pre-conceptions of the investigator. Our computers have already made some quite substantial discoveries [11].

On the other hand, such a data-intensive approach also has its perils. For instance, when relying on given recordings, we have little or no control over the experimental conditions. We depend on recordings with all their (unknown) context (e.g., the specific circumstances of recording and room, artist's particular artistic intention, ...). Also, with very large amounts of data one tends to lose sight of details. Or to put it another way: the kinds of questions that can be studied with such a bottom-up data analysis approach are different from and probably more coarse-grained than the detailed studies possible in controlled experiments. Thus, our approach should be seen as *complementary* to the other strategies used in performance research.

There are innumerable limitations and open problems connected with the current state of the project, and many avenues for further research. The analysis and modeling of performance strategies should be extended to both lower and higher levels. We would like to look at much more detailed levels of expressive performance, such as the timing, dynamics, and articulation of individual voices or individual notes. That will require advanced audio analysis techniques. At the same time, this research must be taken to higher structural levels. It is in the global organization of a performance that great artists express their interpretation and understanding of the music; these high-level 'master plans' do not reveal themselves at the level of local patterns we studied above. To describe such high-level plans, abstract modeling languages will have to be developed, including formal pattern languages for describing performance at different levels of granularity. Our performance alphabets are but a first step in that direction.

5. ACKNOWLEDGMENTS

This research is made possible by the Austrian START project Y99-INF, the EU project HPRN-CT-2000-00115 (MOSART) and the EU COST Action 282. The Austrian Research Institute for Artificial Intelligence acknowledges basic financial support by the Austrian Federal Ministry for Education, Science, and Culture.

6. REFERENCES

[1] Desain, P. and Honing, H. (1994). Does Expressive Timing in Music Performance Scale Proportionally With Tempo? *Psychological Research* 56, 285–292.

[2] Dixon, S. (2003). On the Analysis of Musical Expression in Audio Signals. In *Prodeedings of the Conference on Storage and Retrieval for Media Databases 2003*, SPIE/IS&T Annual Symposium on Electronic Imaging, Santa Clara, CA.

[3] Friberg, A. (1995). *A Quantitative Rule System for Musical Performance.* Doctoral Dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden.

[4] Gabrielsson, A. (1999). The Performance of Music. In D. Deutsch (Ed.), *The Psychology of Music* (2nd ed., pp. 501-602). San Diego: Academic Press.

[5] Langner, J., and Goebl, W. (2003). Visualizing Expressive Performance in Tempo-Loudness Space. *Computer Music Journal*, in press.

[6] Palmer, C. (1989). *Timing in Skilled Music Performance*. Dissertation, Cornell University, Ithaca, N.Y.

[7] Repp, B. (1992). Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's 'Träumerei'. *Journal of the Acoustical Society of America* 92(5), 2546–2568.

[8] Stamatatos, E. and Widmer, G. (2002). Music Performer Recognition Using an Ensemble of Simple Classifiers. In *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI'2002)*, Lyon, France.

[9] Sundberg, J., Askenfelt, A. and Frydén, L. (1983). Musical Performance: A Synthesis-by-Rule Approach. *Computer Music Journal* 7, 37–43.

[10] Todd, N. (1992). The Dynamics of Dynamics: A Model of Musical Expression. *Journal of the Acoustical Society of America* 91, 3540–3550.

[11] Widmer, G. (2002). Machine Discoveries: A Few Simple, Robust Local Expression Principles. *Journal of New Music Research* 31(1), 37–50.

[12] Widmer, G., Dixon, S., Goebl, W., Pampalk, E., and Tobudic, A. (in press). In Search of the Horowitz Factor. *AI Magazine*.

[13] Widmer, G., and Tobudic, A. (2003). Playing Mozart by Analogy: Learning Multi-level Timing and Dynamics Strategies. *Journal of New Music Research*, in press.

[14] Windsor, W.L. and Clarke, E.F. (1997). Expressive Timing and Dynamics in Real and Artificial Musical Performances: Using an Algorithm as an Analytical Tool. *Music Perception* 15(2), 127–152.

[15] Zanon, P. and Widmer, G. (2003). Learning to Recognize Famous Pianists with Machine Learning Techniques. In *Proceedings of the Third Stockholm Music Acoustics Conference (SMAC 03)*, Stockholm, Sweden.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

STUDIES OF MUSIC PERFORMANCE: A THEORETICAL ANALYSIS OF EMPIRICAL FINDINGS

Patrik N. Juslin

Department of Psychology
Uppsala University

Patrik.Juslin@psyk.uu.se

ABSTRACT

The goal of this paper is to outline a psychological approach to expression in music performance that can help to provide a solid foundation for teaching of expression in music education. I will argue that psychological theory is critical to an understanding of performance expression. In my view, a psychological approach to expression should consider how this phenomenon reflects human characteristics that are not necessarily unique to the musical domain. Research reviewed in this paper will provide support for this view. Drawing on previous research, I propose that performance expression is best conceptualized as a multi-dimensional phenomenon consisting of five main components: (a) Generative rules that serve to clarify the musical structure, (b) Emotional expression that function to convey emotions to listeners, (c) Random variations that reflect internal time-keeper variance and motor delays, (d) Motion principles that prescribe that certain aspects of the performance should be shaped in accordance with patterns of biological motion, and (e) Stylistic unexpectedness that involves local deviations from performance conventions. An analysis of expression in terms of these five components - referred to as the GERMS model - has important implications for research and teaching of music performance.

1. MUSICAL EXPRESSION

"He put the bow to his instrument … and then, the first notes, bold and fiery, sang through the hall. At once the spell began to work. Was this really the music of a violin? What grandeur in these slurred notes, what absolute purity! There came roulades of double-stop harmonic notes, and a long run across four octaves, played *staccato* in a single stroke of the bow … Then came a noble, moving theme, which sounded as though a human voice was singing …" (p. 171) [1]

In this description of a performance by Nicolo Paganini in Vienna 1828, something of the powerful impact of an expressive musical performance is captured. Could researchers ever hope to explain this seemingly inexplicable phenomenon? Judging from the comments by some musicians, we would be inclined to think not. Yet, I believe that it is not only possible for researchers to investigate performance expression successfully, it is a matter of responsibility. A genuine understanding of the mechanisms that underlie musical expression has scientific as well as social and educational implications [2].

In this paper, I want to highlight the role of psychological theory in an understanding of performance expression. Looking back, music psychologists have tended to approach performance expression mainly by measuring various acoustic variables of music performances. Often, a purely descriptive approach has been taken, in accordance with the tradition first established by Carl Seashore. As a consequence, the notion of expression has been poorly conceptualized.[1] Matters of expression have largely been reduced to tables or graphs of acoustic data, whereas the question of what these data actually tell us about the origins of musical expression has somehow been lost along the way [2].

But perhaps psychology is uniquely suited to investigating expression? After all, psychology was right from the beginning studying the kinds of psychophysical relationships that seem to underlie musical expression. Still, psychology has arguably not yet fulfilled its potential role in helping to provide a theoretical organization of the field. Possible reasons for this are explored in the following.

2. PERFORMANCE STUDIES

To do music performance justice as a meaningful event in a particular time and place, with certain individuals taking part, we would have to take into account an extremely large number of complicated real-world relationships. To get some sense of the difficulties involved in such an endeavor, consider the list of factors related to the piece, the instrument, the performer, the listener, and the performance context that – in principle – may influence expression. To model all these factors simultaneously may be impossible. To simplify matters, most researchers have focused instead on acoustic measurements of performances with occasional attempts to relate the measures to other aspects (e.g., the notation, the performer's intention, the listener's perception, etc.). Yet, even with such a narrow focus, researchers have been struggling to account for the nature of a typical performance of music. Why is this so?

Clearly, there are a number of problems that performance researchers are confronted with, such as obtaining large samples of music performances from expert musicians, or managing the difficult and time-consuming task of analyzing relevant acoustic variables. (However, see Friberg, Schoonderwaldt, & Juslin, this volume, for a recent attempt to develop *automatic* extraction of performance parameters.) However, the most serious problem is perhaps not to obtain sufficient amounts of data, but rather to find ways of *interpreting* the wealth of data in a meaningful

[1] Meyer argues that "because Seashore advances no theory or attempts no explanation of the relationship between deviation and affective aesthetic experience, his viewpoint lacks substance and plausibility" (p. 203) [3].

way. It may be instructive to look closer at what this endeavor entails. Performance researchers must describe and explain the nature and origin of subtle patterns of variability in acoustic measures shown over the time-course of a performance. The problem is that "using measurements of actual performances alone, it is difficult to separate the effects of combined, though psychologically distinct, expressive actions" (p. 64) [4]. Indeed the same surface pattern could, in principle, be produced by an almost infinite number of different "computational operations", and a pure bottom-up approach is unlikely to allow us to select the correct one. Theoretical guidance is urgently needed. One heuristic tool could be to rely on the advice of professional performers. However, performers are often unable to explain the details of how they apply all expressive features. Psychologists early realized the limitations of introspection in explaining psychological mechanisms, and many of the mechanisms that underlie music performance are *not* available to introspection. It is here that psychological theory could make a contribution.

3. A PSYCHOLOGICAL APPROACH: THE GERMS MODEL

Arguably, the greatest impediment to cumulative progress in explaining expression has been the common tendency to treat expression as a single entity, a homogenous natural category (the expression is "appropriate", "exaggerated", or "lacking"). Considering the controversy that has surrounded the concept of expression throughout history, it is perhaps understandable that many researchers have tended to leave the concept undefined or simply "defined" it in terms of "deviations from the notation". However, as long as expressive features are lumped together as "expressive deviations from the notation", we are not likely to reach a deeper understanding of the origins of these variations. What is at stake here is not the usefulness of the generic word "expression" as such, it is rather whether this is the most fruitful level at which to investigate and teach expression ("Put some expression into it!"). There is now evidence that the expressive variations that are typically found in a human performance are not all of the same kind, and that they do not all share the same origins. Considering that music performance is a form of human behavior, psychology provides a natural point of departure for analyzing performance expression into its constituents.

What would a psychological theory of expression entail? First, a psychological approach would involve an attempt to go beyond the performance and to consider the nature of the person *behind* the performance: "What are the proximate and ultimate causes of the behavior in the performer, and how can we make sense of this behavior from what we know about humans?" Secondly (and partly as a consequence of what we know about the modular organization of human behavior), it would have to be a component approach. Recent studies have indicated that a single theoretical mechanism cannot account for all the variance in expression [4, 5, 6]. A careful review of the literature suggests that performance expression is better conceived of as a multi-dimensional phenomenon, consisting of different components of information. Hence, I have proposed a model of expression – the GERMS model – that *integrates* different kinds of expression [2, 7]. The main assumption is that, in principle (and somewhat akin to a Fourier analysis of wave forms), it is possible to decompose patterns of expression into five distinct components. Moreover, it can be assumed that the different components of expression:

- have different origins
- involve patterns with different characteristics
- are processed by somewhat different brain regions
- have different effects on listeners' perception of music

The last of these general criteria means that each GERMS component also will tend to make a unique contribution to the *emotional* and *aesthetic* impact of a given performance. Further dimensions along which the five components may differ include degree of *dependence on knowledge* and degree of *voluntary control* [2]. The five components are briefly summarized in the following, along with some of their hypothesized characteristics.

3.1. Generative Rules (G)

According to the very influential generative approach [8], expressive variations serve to convey the musical structure to listeners as clearly as possible. By means of variations in such acoustic variables as timing, dynamics, and articulation, a performer is able to clarify group boundaries, metrical accents, and even harmonic structure. The G-component originates in generative transformations of the notated musical structure, and reflects basic characteristics of human information processing, such as categorical perception and gestalt laws. It recruits the human capacity for syntactical processing that is also involved in language. A positron-emission tomography study of sight-reading in piano performance showed activated cortical brain areas distinct from, but adjacent to, those underlying language operations, consistent with the task of translating a generative structure in a musical notation into "expressive markings" of structure [9]. Also, a magnetoencephalography study of musical perception indicated that musical syntax is processed in Broca's area, which is also involved in syntactical processing during auditory language comprehension [10]. Generative rules are mainly bottom-up (i.e., determined by the local structure), but also reflect the hierarchical structure of the music. By conveying the structure and creating coherence and order, generative rules add to the beauty of the music. They may furthermore have an arousing effect, but mainly by enhancing the emotional impact inherent in the structure of the piece [2, 7].

3.2. Emotional Expression (E)

A second function of performance expression might be to render the performance with a particular emotional expression. 83% of the musicians in a recent questionnaire study claimed that they try to express specific emotions in their performances "always" or "often" [11]. A review of 145 studies (101 speech and 41 music studies) strongly indicates that patterns of musical cues used to convey discrete emotions mainly derives from the nonverbal aspects of emotional speech [12], as suggested by Herbert Spencer in the 19th century. The E-component involves a large set of *probabilistic* (uncertain) but partly redundant cues that are additive and compensatory in nature - one cue may to some extent compensate for another cue [6]. These cues involve changes in tempo, sound level, articulation, timbre, timing, tone attack/decay, intonation, vibrato extent and frequency, accents on particular notes, etc. Perception of the E-component can be instantaneous, and involves parallel processing primarily in the

right brain hemisphere, possibly in the basal ganglia, which are involved in the perception of emotions in vocal expression. The E-component is mainly top-down (i.e., a holistic interpretation of the mood of the piece influences local acoustic parameters), and offers a parallel channel of affective information, which may support or contradict the expression of the composed structure. The E-component contributes to the recognition aspect of art, but may also lead to emotional contagion [13].

3.3. Random Variations (R)

It seems generally agreed that human music performance is *not* controlled by a completely deterministic motor-system [7]. On the contrary, performance expression always contains some small random fluctuations, which may contribute to the "living" character of human performance. The R-component originates in a hypothesized *internal timekeeper* and associated *motor delays*, and reflects human limitations with regard to perceptual-motor precision. This seems to involve two kinds of random patterns: *1/f noise* (noise for which the power varies inversely with the frequency) that reflects the timekeeper and changes according to the inter-onset-interval duration, and uncorrelated *white noise* that reflects motor delays that are constant across different inter-onset-interval durations [14]. It may be hypothesized that the internal time-keeper is localized in the lateral and medial parts of the cerebellum. This component involves involuntary patterns of variability that contribute "unevenness" to a performance, and also may enhance the novelty aspect of art.

3.4. Motion Principles (M)

There is a common assumption that music and motion are closely related to one another [15]. Especially, one could assume that performers attempt to achieve an expressive micro-structure that is consistent with the nature of human movement [16]. The M-component is thus believed to originate in subtle patterns of *biological motion* such as arm gestures, which reflect the unique construction of the human body (e.g., the unique composite of anatomical proportions, including both geometrical dimensions and distribution of mass between parts of the body). Perception of the M-component might partly reflect the human ability to distinguish animate objects from inanimate objects on the basis of patterns of sounds, which could have great survival value. This aspect involves complex, dynamic and non-compensatory temporal patterns that are processed by the left hemisphere, possibly by brain regions that are adjacent to those that analyze biological motion in vision, like the superior temporal sulcus. The M-component contributes to the recognition aspect of art – by signaling human intentionality – but also lends balance and smooth human gesturing to a performance [2].

3.5. Stylistic Unexpectedness (S)

The S-component reflects a performer's deliberate attempt to deviate from stylistic expectations concerning performance conventions in order to add tension and unpredictability to the performance [3]. This component originates in a general human tendency to create schematic expectations that may guide future behavior, and the consequent monitoring of such expectations that may give rise to emotional arousal, signaling that something important has happened. This idea is recurrent in psychological theories of emotion since the beginning of the 20th century. The S-component involves locally focused expressive features that contribute to the novelty, originality, and arousal aspects of art. Monitoring of expectations is believed to be processed by parts of the anterior cingulate cortex in the brain.

What is the relationship of the S-component to the others? My estimation is that the first four components are sufficient to achieve an *acceptable* performance, but that the S-component is what makes a performance *special*. Hence, much of the artistic process aims at turning GERM performances into GEMS. That is, reducing random fluctuations to a minimum, and increasing the originality of the musical interpretation [2].

4. PRELIMINARY FINDINGS

Can this kind of psychological meta-theory be turned into something useful in empirical terms? In a recent study, we made our first attempt to implement a multi-component approach in a computational model, the GERM model, that simulates different aspects of expression [7]. The model comprised only four main sources of variability (Generative Rules, Emotional Expression, Random Variations, Motion principles) - we had not figured out how to implement the fifth component (S) at the time.

In a preliminary experiment, some of the rules featured in the GERM model were used to synthesize performances of a brief piece of music. The aim of this experiment was to examine (a) whether the four components would yield predicted effects on listeners' judgments of the performances, and (b) whether the components would yield at least partly independent effects on these judgments (for further information about this experiment, see 7). The basic idea was to manipulate the four components of the GERM model in a so-called *factorial design*; this was done by generating all possible combinations (16) of the presence and absence of each GERM component. For instance, a condition with only the G-component present and all other components absent would include only those rules that serve to convey the structure to listeners. We asked musically trained listeners to rate all 16 performances on scales believed to reflect various aspects of expression: *clear, sad, human, gestural, musical, expressive*.

The results of this experiment suggested, first of all, that the different components yielded predicted effects on listeners' ratings. For instance, the G-component yielded high ratings on the "clear" scale, and the E-component yielded high ratings on the "sad" scale. Multiple regression analyses indicated that listeners' ratings of overall expressivity could be successfully predicted on the basis of a linear combination of the GERM components, which all contributed significantly but in different degrees to the judged expressivity; $R = .77$, $F_{4,184} = 66.143$, $p < .0001$, with beta weights (β) in order of predictive strength: Emotional Expression, $\beta = .64$, $p < .0001$; Motion Principles, $\beta = .32$, $p < .0001$; Generative Rules, $\beta = .23$, $p < .0001$; Random Variations, $\beta = .14$, $p < .01$).

There were, of course, some interactions between the four components in how they affected listeners' judgments. However, only about 20% of the total number of possible interactions were actually significant, and the interaction effects were smaller than the main effects. Finally, the components produced different and at least partly independent effects on listeners' judgments. This tendency is seen in examples of *dissociations* between the four

components. For instance, the performance with only G present yielded a high rating on the "clear" scale, but low mean ratings on the "sad", "human", and "gestural" scales. These findings lend support to a multi-component approach to expression.

5. IMPLICATIONS

What can we conclude from these preliminary data? First, the GERM model suggests that different aspects of performance expression *can* be integrated into a common model. I suspect that researchers have found performance data confusing, partly because they have treated them as if they reflected a single factor. If we are unaware of the different aspects of performance expression, we are likely to be confused by data that reflect an unknown mixture of different modes of behavior [7].

Distinguishing between different components of expression could help us to better explain individual differences between performers. Different performers may be characterized in terms of the "relative weights" they give to different components of expression. Further, different musical styles may put different emphasis on different components. Finally, a multi-component approach could help to resolve some inconsistent findings in the previous literature. For example, the fact that musically trained listeners have shown refined ability to detect expressive features in some studies but not others might be explained by the fact that these studies have investigated different components (e.g., generative structure vs. emotional expression).

How can we be sure that the particular components I have proposed are the correct ones? This is, ultimately, an empirical question, which is far from settled. I remain quite open to the possibility that a different set of components might provide a more cohesive account. However, I believe that psychological theory in combination with findings from listening experiments, performance measurements, and brain imaging could help us to resolve this problem. In particular, a major implication of the GERMS model is that psychological theory concerning various aspects of human behavior, such as language, emotion, motor-coordination, movement, and expectancy monitoring is relevant to an understanding of performance expression [2].

The GERMS model also offers an example of how research could benefit teaching by helping to "demystify" expression. As demonstrated by the GERM experiment, acoustic correlates of perceived expression can be manipulated and reliably related to listeners' judgments of expressivity. This paves the way for a more theoretically informed approach to teaching expressivity. One implication is that, at certain stages of learning, different components of expression might need to be taught separately, since they have different characteristics [17]. Different teaching strategies may be required to help a performer to (a) convey the structure, (b) express emotions, (c) exhibit motor precision, (d) be suggestive of human motion and gesture, and (e) deviate from stylistic expectations in pleasing ways, respectively. Thus, a psychological approach to performance expression could help to provide a solid foundation for teaching of expression, a goal as laudible as any, in view of the reported neglect of expressive skills in music education [17].

6. REFERENCES

[1] Farga, F. (1969). *Violins and violinists* (2nd ed.). London: The Cresset Press.

[2] Juslin, P. N. (2003). Five facets of musical expression: A psychologist's perspective on music performance. *Psychology of Music, 31 (3),* 273-302.

[3] Meyer, L. B. (1956). *Emotion and meaning in music.* Chicago: University of Chicago Press.

[4] Thompson, W. F., Sundberg, J., Friberg, A., & Frydén, L. (1989). The use of rules for expression in the performance of melodies. *Psychology of Music, 17,* 63-82.

[5] Clarke, E. F., & Windsor, W. L. (2000). Real and simulated expression: A listening study. *Music Perception, 17,* 277-313.

[6] Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1797-1813.

[7] Juslin, P. N., Friberg, A., & Bresin, R. (2002). Toward a computational model of expression in music performance: The GERM model. *Musicae Scientiae, Special Issue 2001-2002,* 63-122.

[8] Clarke, E. F. (1988). Generative principles in music performance. In J. A. Sloboda (Ed.), *Generative processes in music* (pp. 1-26). Oxford: Clarendon Press.

[9] Sergent, J., Zuck, E., Tenial, S., & MacDonall, B. (1992). Distributed neural network underlying musical sight-reading and keyboard performance. *Science, 257,* 106-109.

[10] Maess, B., Koelsch, S., Gunter, T. C., & Friederici, A. D. (2001). Musical syntax is processed in Broca's area: A MEG study. *Nature Neuroscience, 4,* 540-545.

[11] Lindström, E., Juslin, P. N., Bresin, R., & Williamon, A. (2003). "Expressivity comes from within your soul": A questionnaire study of music students' perspectives on expressivity. *Research Studies in Music Education, 20,* 30-55.

[12] Juslin, P. N., & Laukka, P. (in press). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*.

[13] Juslin, P. N., & Sloboda, J. A. (Eds.). (2001). *Music and emotion: Theory and research.* Oxford University Press.

[14] Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). 1/f noise in human cognition. *Science, 267,* 1837-1839.

[15] Todd, N. (1999). Motion in music: A neurobiological perspective. *Music Perception, 17,* 115-126.

[16] Friberg, A., & Sundberg, J. (1999). Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America, 105,* 1469-1484.

[17] Juslin, P. N., Friberg, A., Schoonderwaldt, E., & Karlsson, J. (in press). Feedback learning of musical expressivity. In A. Williamon (Ed.), *Enhancing musical performance.* New York: Oxford University press.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

ESTIMATION OF THE NOTE DURATIONS IN A POLYPHONIC PIANO TRANSCRIPTION SYSTEM

A.M. Barbancho, I. Barbancho, L.J. Tardón

Departamento de Ingeniería de Comunicaciones
E.T.S. Ingeniería de Telecomunicación. Malaga (Spain)
abp@ic.uma.es

ABSTRACT

A system for determining the note durations and the measure division of a polyphonic piano piece is presented. The inputs for this system will be the instants when each onset takes place, its intensity and the estimated duration, in seconds, of each note. The beat and the time signature can be either inputs for the system or calculated by the system. In order to perform the mapping into note durations, a modified of the K-means clustering algorithm has been developed.

1. INTRODUCTION

The aim of a transcription system is to convert an acoustic music signal into its symbolic representation. This process involves a number of different tasks, which are necessary to extract all the necessary elements to write down the score from the music signal [1], [2]: pitches and durations of the notes, attack times, time signature and key signature.

Among all the different tasks, in this paper we will deal with the determination of the note durations and the measure division. The inputs to the system are shown in Figure 1 (beat and time signature can be inputs to the systems or they can also be estimated) and they are extracted from the polyphonic piano transcription system presented in [2].

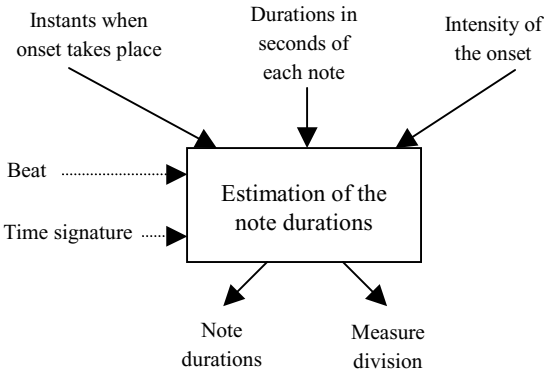


Figure 1: *Overview of the note duration estimation system.*

The problem at hand can be formulated as follows: when a piano piece is performed, the same note durations last a different while, depending on the passage. This is due to non-voluntary effects, like random timing error caused by the limited accuracy of the human motor system [3] and errors in mental time-keeping processes [4] and also due to expressiveness reasons (retardando, accelerando...). So, the note durations must be estimated in a rather flexible way and that implies that also the estimation of the tempo must be flexible [5]. In this paper, we present a system based on the K-means algorithm [6] that analyzes the acoustic music signal as a whole to quantify the note durations. The designed system has no trouble with small variations of the tempo.

The outline of the paper is as follows: section 2 describes the K-means clustering algorithm and how it has been adapted to properly estimate the note durations. In section 3, the procedures to estimate beat ant time signature are presented. Also, the technique to perform measure division is described. Section 4 presents the results obtained by the system described and, finally, section 5 draws some conclusions about the problem considered and the techniques described.

2. NOTE DURATIONS ESTIMATION SYSTEM

In this section, the technique to estimate the note durations is described. Initially, we recall the K-means algorithm, which is a robust technique for data clustering [6], that has been successfully used by the image processing community for image segmentation [7]. Afterwards, the restrictions about the clustering of the note durations are introduced and then, the variations introduced to the algorithm and the final clustering scheme are presented.

2.1. K-means clustering algorithm

The K-means algorithm groups the data items that are similar, according to a measure of similarity and creates clusters of data using a distance function defined upon that measure of similarity. The algorithm is informed about the number of clusters and then, the process is as follows:

1. Begin with an arbitrary assignment of samples to clusters or begin with an arbitrary set of cluster centers and assign samples to nearest centers.
2. Compute the sample mean of each cluster (centroids).
3. Reassign each sample to the cluster with the nearest mean.
4. If the classification of all samples has not change, stop; else go to step 2.

This algorithm cannot be directly employed to estimate the note durations of a music work since in this specific case:

a) The number of centroids is not known.

b) The centroids should represent the note durations that follow the relations shown in Table 1.

	𝅝	𝅗𝅥.	𝅗𝅥	♩.	♩	♪.	♪	𝅘𝅥𝅯.	𝅘𝅥𝅯
rela. value	4	3	2	1.5	1	0.75	0.5	0.375	0.25
$\log_2$	2	1.585	1	0.585	0	-0.415	-1	-1.415	-2

Table 1. *Relationships among the expected note durations (natural and logarithmic scales).*

c) Not all the existing note durations have to appear in a particular piano piece.

To use the K-means algorithm, we will use the following ideas:

a) We will work with the logarithm ($\log_2$) of the note durations, so that the relationships among them become nearly linear.

b) We impose restrictions to the centroids in the logarithm scale according to table 1. To this end, we will add new artificial data as described in the following subsections to drive the centroids to accomplish the desired relationships. Also, the number of dotted notes will be forced to be smaller than the rest of notes.

2.2. Algorithm for the estimation of the note durations

Figure 2 shows a diagram of the algorithm developed. Let

$$\mathbf{c} = \{c_1, c_2, \ldots, c_9\}^t \quad (1)$$

the vector of centroids obtained by the K-means algorithm, where c_n reprents the centroid n. In this vector of centroids, the index $n = 1$ corresponds to the whole note and the index $n = 9$ to the sixteenth note. From $\mathbf{c}$, a new set of constrained centroids $\mathbf{c}r$ is found according to:

$$\min\left[(\mathbf{c} - \mathbf{c}r)^t \cdot \mathbf{N} \cdot (\mathbf{c} - \mathbf{c}r) \right] \quad (2)$$

subject to the constrain :

$$\mathbf{b} = \mathbf{R} \cdot \mathbf{c}r \quad (3)$$

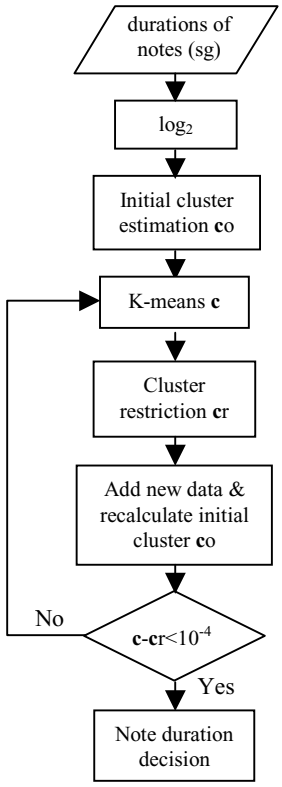


Figure 2: *Algorithm for the estimation of the note durations.*

where:

- $\mathbf{N} = diag(N_1, N_2, \ldots, N_9)$, and N_n is the number of data assigned to cluster n.
- $\mathbf{R}$ is a (8×9) matrix defined as follows:

$$\mathbf{R} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & & -1 & & \vdots \\ \vdots & & & \ddots & \\ 1 & & \cdots & 0 & -1 \end{bmatrix} \quad (4)$$

- $\mathbf{b} = \{0.4, 1, 1.4, 2, 2.4, 3, 3.4, 4\}$ contains the relationship between note durations as related by $\mathbf{R}$

Equation (2) with hard constraint (3), can be solved using Lagrange multipliers to obtain:

$$\mathbf{c}r = \mathbf{c} - \mathbf{N}^{-1} \cdot \mathbf{R}^t \left[\mathbf{R} \cdot \mathbf{N}^{-1} \cdot \mathbf{R}^t \right]^{-1} (\mathbf{R} \cdot \mathbf{c} - \mathbf{b}) \quad (5)$$

However, instead of using this solution, we use $\mathbf{c}r$ to add new data to the input of the K-means algorithm. After the k-th iteration, 'k' durations of each centroid are added to the original

data input to the algorithm. Using this procedure, the constraint is imposed in a soft manner.

The procedure to minimize the number of dotted notes involves the proper selection of the initial clusters of the K-means algorithm. The procedure employed is outlined in Figure 3 and it can be summarized as follows: if the number of dotted notes is too large, then, all the centroids are moved so that the new assignment gives a lower number of dotted notes.

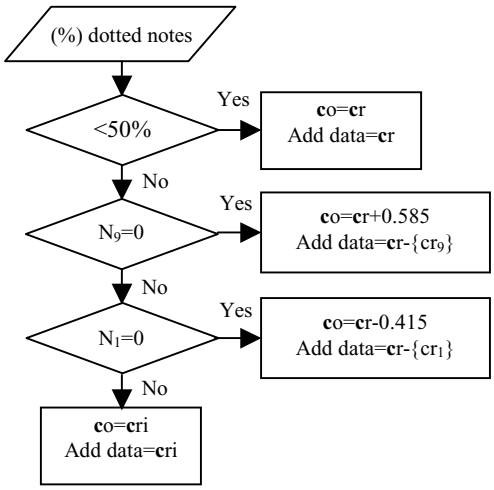


Figure 3. *Add new data and recalculate initial centroids.*

The proposed algorithm ends deciding what note duration corresponds to each centroids. This is implemented assigning the note duration corresponding to the denominator or the time signature to the centroid that has the duration closer to the beat of the piece and not corresponds to a dotted note.

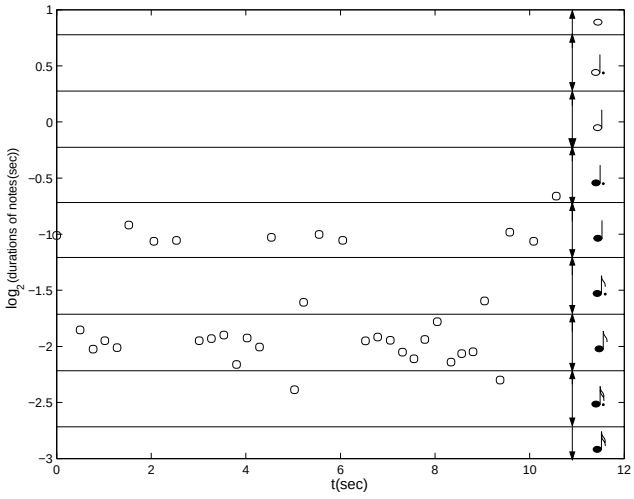


Figure 4. *Application of the algorithm for the estimation of the note durations to Minuet nº 20 (bar 1-8), Notebook 20 pieces, Ana Magdalena Bach.*

In the Figure 4, it is represented the result of applying the note durations estimation algorithm to Minuet nº 20 (bar 1-8), Notebook 20 pieces, Ana Magdalena Bach. If this result is compared with the original scores, four note durations are found to be wrong, sixteenth notes and dotted half notes. The reasons for this result are problems in the interpretation of the piece and in the time resolution of the transcription system [2].

3. ESTIMATION OF BEAT, TIME SIGNATURE AND MEASURE DIVISION

In this section, we describe how beat and time signature are estimated. Also, the measure division algorithm will be presented.

3.1. Estimation of beat

The beat is estimated on the basis of the average duration, in seconds, of the notes. First, the average value is calculated using all the notes, then, to obtain more accurate results, a new average is calculated using only the notes whose duration is ±25% the first calculated average value. Afterwards, the denominator of the time signature and the beat is determined using the table 2, where τ stands for the average time of notes.

	τ>0.75s	0.75s<τ<0.4s	τ<0.4s
Beat	𝅗𝅥=60/τ	♩=60/τ	♪=60/τ
Time signature denominator	2	4	8

Table 2. *Determination of beat and time signature denominator.*

3.2. Estimation of time signature

The denominator of the time signature is obtained from the beat estimation (Table 2). The estimation of the numerator of the time signature is based on the knowledge that the first beat of the measure, the one just after the bar, will have more energy than the rest of beats of the measure. The procedure implemented works as follows: taking into account that each measure lasts numerator*beat, the bar division of the piano piece is done for each possible value of the numerator. Then, the average intensity of the beginning of each measure is calculated. The bar division procedure also takes into account that the piano piece can start with an upbeat. The program chooses the numerator with which the maximum average intensity of the beginning of measures is achieved.

3.3. Measure division algorithm

After the estimation of the note durations and the time signature, we face the problem of finding the bars of a piano work. Using the note durations and the time signature, the bars are selected to minimize the number of syncopes between bars. A schematic of the algorithm is shown in Figure 5.

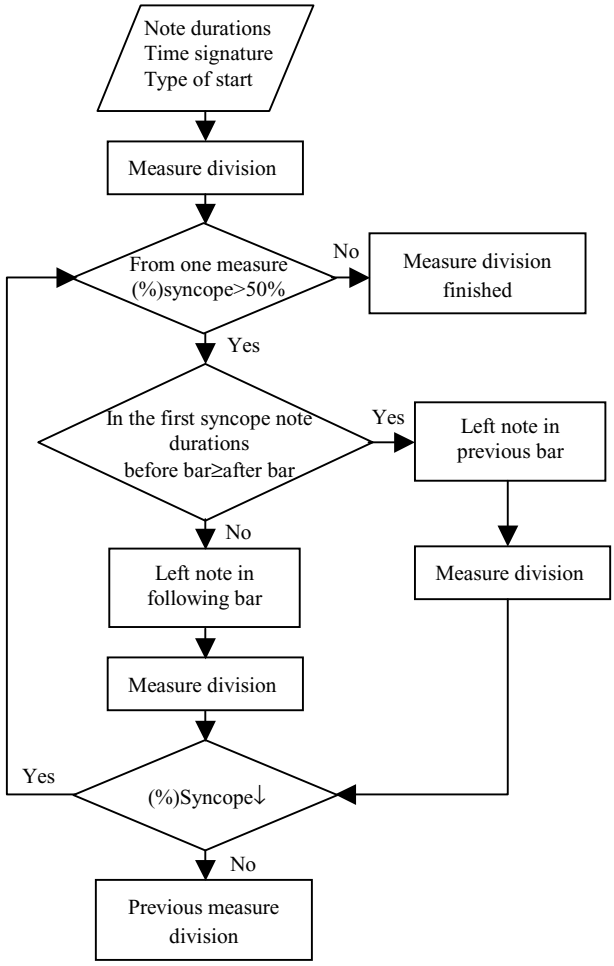


Figure 5. *Measure division algorithm.*

4. RESULTS

In this section we present some of the results obtained by the system described in this contribution (Table 3). In Table 3:

- *Score* is the percentage of correctly detected notes over the total played notes plus false notes reported. This result is obtained from the transcription system [2].
- *Note duration* is the percentage of note durations correctly estimated over the total detected notes.
- *Measure division* is the percentage of correctly placed bars.

Some particular comments about the results must be done. The note duration error is directly related to the score of the piece. In "The Sting" the performance of the program is worse than in the other pieces. This piece has been interpreted with a very irregular tempo and the transcription system makes several mistakes in the onset detection. In general, it can be observed that the performance of the measure division is quite good, despite errors in the estimation of the note duration. The error rate in the bar division of "The happy farmer" is due to an error in the detection of the duration of a single note that produces a displacement of all the bars after it.

Piano pieces	*Score*	*(%)Note duration*	*Measure*	*Measure division*
Ana Magdalena Bach. Notebook 20 pieces. Minuet nº 20	84.81%	82%	3/4	100%
Schumann. Album for the Young. "The happy farmer"	78.24%	92%	4/8 (real 4/4)	50% (*)
Hal Leonard. Adaptation of the "Himmlische, dein Heiligtum" (bars 1-3)	100%	90%	4/4	100%
Scott Joplin. "The Sting", based on "The Entertainer"	82.26%	62%	8/8 (real 4/4)	100%

Table 3. *Results obtained by the system.*

5. CONCLUSIONS

A system for determining the note durations and the measure division of a polyphonic piano piece has been presented. The inputs to the system, extracted from the polyphonic piano transcription system [2], are the instants when each onset takes place, its intensity and the estimation duration, in seconds, of each note. The technique to estimate the note durations is based on a variation of the K-means clustering algorithm that works with the logarithm ($\log_2$) of the note durations. The variations of the algorithm are an application of the following ideas: impose restrictions to the relations between centroids and minimize the number of dotted notes.

6. REFERENCES

[1] Klapuri, A. "Automatic transcription of music", Master of Science Thesis, Department of Information Technology, Tampere University of Technology, 1998.

[2] Barbancho, A.M., Jurado, A., Barbancho, I., "Identification of rhythm and sound in polyphonic piano recordings", Forum Acousticum, Sevilla, Spain, 2002.

[3] Shaffer, L.H., "Performances of Chopin, Bach, Bartok: Studies in Motor Programming", Cognitive Psychology, No. 13, pp. 326-376, 1981.

[4] Vorberg, D.J., Hambuch, R., "On the temporal control of rhythmic performance", J. Requin (ed.), Attention and Performance VII, Lawrence Erlbaum Associates, New Jersey, 1978.

[5] Desain, P., Honing, H., "The quantization of Musical Time: A Connectionist Approach", Computer Music Journal, Vol. 13, No. 3, pp. 150-167, 1989.

[6] Therrien, C.W., "Decision, Estimation and Classification. An Introduction to Pattern Recognition and Related Topics", John Wiley & Sons, New York, 1989.

[7] Coleman, G.B., "Image segmentation by clustering", Proceedings of the IEEE, Vol. 67, No. 5, pp. 773-785, 1979.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

EXPRESSIVE DIRECTOR: A SYSTEM FOR THE REAL-TIME CONTROL OF MUSIC PERFORMANCE SYNTHESIS

Sergio Canazza, Antonio Rodà, Patrick Zanon

CSC – Center of Computational Sonology
DEI – Dep. of Information Engineering
University of Padua, Italy
ar@csc.unipd.it, {canazza,patrick}@dei.unipd.it
http://www.dei.unipd.it/ricerca/csc/

Anders Friberg

KTH – Royal Institute of Technology
Speech, Music and Hearing
Stockholm, Sweden
andersf@speech.kth.se
http://www.speech.kth.se/music/

ABSTRACT

The Expressive Director is a system allowing real-time control of music performance synthesis, in particular regarding expressive and emotional aspects. It allows a user to interact in real time, for example, changing the emotional intent from happy to sad or from a romantic expressive style to a neutral while it is playing. The Expressive Director was designed in order to merge the expressiveness model developed at CSC and at KTH. The control of the synthesis can be obtained using a two-dimensional space (called "Control Space") in which the mouse pointer can be moved by the user from an expressive intention to another continuously. Depending on the position, the system applies suitable expressive deviations profiles. The Control Space can be made so as to represent the Valence-Arousal space from music psychology research.

0.1. Definitions

Acoustic parameters: they specify the low level characteristics of a performance; in the case of MIDI performances they are: the onset time O, the inter onset interval IOI, the duration D, the intensity I (expressed in dB), and the note number (which specify the pitch).

Tick: it is a temporal measure, which refers to the nominal duration (specified in the score); equal durations notes in the score have the same duration expressed in ticks.

Metronome (MM): it is a measure of the instantaneous tempo with which a given note is played; it is measured in beats per minute.

Legato degree (leg): it is the ratio between the duration of a given note and the inter onset interval which occurs between its subsequent note: $leg = D/IOI$.

Key Velocity (KV): it is an intensity measure, which is used in the MIDI representation of a musical performance, and it is approximately linearly related to the intensity reproduced by the sound card; its values ranges from 0 to 127.

Sonologic parameters: they specify the mid level characteristics of a performance; in the case of MIDI performances they are: the metronome MM, the legato degree leg, and the Key Velocity KV.

1. INTRODUCTION

The shaping of the musical expression is a natural part of musicians task when performing music. To become a master musician is a time-consuming endeavor to say the least. A great deal of the time is spent mastering the motoric skills in producing the tones. What happens if the note production is the left to the computer but the expressiveness is still controlled by the musician? In order to explore this we developed the Expressive Director, which combines two different approaches for computer modelling of expressiveness in music performance. The starting point was the KTH rule system that contains an extensive set of context-dependent rules going from score to performance processed in non-real-time, and the CSC expressiveness model that transforms a natural performance (rather than the score) and works in real-time. A score, preprocessed in Director Musices, can be played by the Expressive Director with a real-time control of all the rule parameters.

2. GENERAL OVERVIEW OF DIFFERENT REPRESENTATIONS OF THE EXPRESSIVENESS

2.1. The KTH rule system

The KTH rule system is a set of about 30 rules that transform a score to a musical performance [3, 5]. The rule system models different principles used by musicians when performing a given piece. It is intended to model general principles found to be used by many musicians and not restricted to any particular style or instrument, which however, is not always feasible. Another goal has been that a rule should work on any musical context, implying that the context constraints within the rule itself finds the appropriate musical condition where the rule should trigger and filters out the rest. The rules cover such aspects as phrasing, articulation, timing of small groups, intonation, tonal tension, and accents [7].

Each rule has one main parameter k, which is used to control the overall amount of variation. Most rules also have extra parameters to fine adjust the behavior. By combining k values and extra parameters, different performance styles can be obtained. In such a way, it was possible to model variations among musicians playing the same piece [4], or changing the emotional/motional character [1, 8].

The rules have been implemented in a number of different computer programs. The development platform at KTH is the program Director Musices[1] [6] in which most of the rules are implemented.

[1]The software can be downloaded at http://www.speech.kth.se/music/performance

2.2. The CSC expressiveness model

The model is based on the hypothesis, that different expressive intentions can be obtained by suitable modification of a neutral performance. The transformations realized by the model should satisfy some conditions: 1) they have to maintain the relation between structure and expressive patterns found into the neutral performance, 2) they should introduce as few parameters as possible to keep the model simple. In order to represent the main characteristics of the performances, we used only two transformations: shift and range expansion/compression. Different strategies were tested. Good results were obtained [9] by a linear instantaneous mapping that – for each *sonologic parameter* S and for a given expressive intention e – is formally represented by the equation:

$$S_e(n) = k_e \cdot \overline{S}_0 + m_e \cdot \big(S_0(n) - \overline{S}_0\big) \tag{1}$$

where $S_e(n)$ is the estimated profile of the performance related to expressive intention e for the n-th note, $S_0(n)$ is the value of the S-parameter of the neutral performance, $\overline{S}_0$ is the mean of the profile $S_0(n)$, k_e and m_e are respectively the coefficients of shift and expansion/compression related to expressive intention. We verified that these parameters are very robust in the modification of expressive intentions [10]. Thus, the equation (1) can be generalized to obtain, for each S-parameter, a morphing among different expressive intentions as:

$$S(n) = k(u) \cdot \overline{S}_0 + m(u) \cdot \big(S_0(n) - \overline{S}_0\big) \tag{2}$$

in which the expressive parameters $k(u)$ and $m(u)$ are not fixed anymore, but depend on the user input and can change from an expressive intention to another continuously. In this way, the problem of the morphing among expressive intention has been translated into the problem of the morphing among the parameters k and m. Section 4.1 will show one of the possible choices (the simplest) in order to solve this task.

2.2.1. The expressive profiles

The arithmetic mean $\overline{S}_0$, used in equations (1) and (2), is calculated over a sliding window whose size can be defined by the user (the context size). It was not calculated over the entire piece, since we found that different phrases requires different strategies for the same expressive rendering. Thus, in the implementation of the expressiveness model we used the following formula:

$$S(n) = k^{(S)}(u) \cdot P^{(S)}_{ave}(n) + m^{(S)}(u) \cdot P^{(S)}_{dev}(n) \tag{3}$$

where $P^{(S)}_{ave}(n) = \overline{S}_0(w_n)$ is the *expressive profile* of the sonologic parameter average, which is calculated over the window w_n centered around the n-th note; $P^{(S)}_{dev}(n) = S_0(n) - \overline{S}_0(w_n)$ is the expressive profile of the sonologic parameter deviations around the average.

The expressive profiles contain the information related to the neutral performance, and represent all the timing and intensity nuances that the performer used in his interpretation. These nuances are codified into the profiles $P^{(S)}_{ave}$ and $P^{(S)}_{dev}$ which can be used by the model in real time. The sonologic parameter can be $S \in \{MM, leg, KV\}$, thus we have six profiles representing the neutral performance: $P^{(MM)}_{ave}$ $P^{(MM)}_{dev}$ $P^{(leg)}_{ave}$ $P^{(leg)}_{dev}$ $P^{(KV)}_{ave}$ and $P^{(KV)}_{dev}$.

2.2.2. The Expressive Sequencer

The implementation of the CSC's expressiveness model has been made using the EyesWeb[2] platform, a graphical environment for developing multimedia oriented application developed by the Music and Informatics Lab of the University of Genoa [2]. The core of the model has been written into the `ExpressiveSeq`[3] block which is used to synthesize an expressive performance.

The block inputs are: the information of the score, the expressive profiles, and the control parameters (see figure 1). At the first step, the block reads the score and the profiles storing them into memory. Then, if the button Play is pressed, it starts to sequence the MIDI messages.

Four *acoustical parameters* completely specify these MIDI messages: the note number, the onset time O (expressed in ms), the duration D (expressed in ms) and the intensity I which is expressed through the the Key Velocity KV. The note number is provided by the score, while the other parameter values are calculated using the expressive profiles and the control parameters. More precisely, the calculation of the timing parameters can be made with the subsequent formulas:

$$\begin{aligned} O(n+1) &= O(n) + IOI_{[tick]}(n) \cdot C/MM(n) \\ D(n) &= IOI_{[tick]}(n) \cdot C \cdot leg(n)/MM(n) \end{aligned} \tag{4}$$

where the initial onset time $O(1)$ is fixed to 0 ms, $IOI_{[tick]}$ is the inter onset interval expressed in ticks as stored into the score, and C is a suitable conversion constant; the expressive metronome $MM(n)$, the expressive legato degree $leg(n)$, and the expressive intensity $KV(n)$ are calculated in real time through equation (3) and the values of the six controlling parameters: $k^{(MM)}$ $m^{(MM)}$ $k^{(leg)}$ $m^{(leg)}$ $k^{(KV)}$ and $m^{(KV)}$.

3. INTEGRATION

3.1. From rules to profiles

The integration of the KTH rule model into the CSC's expressiveness system requires that the information regarding the KTH rules is translated into suitable profiles. The KTH rule system adds several deviations to the score. Since the average profiles $P^{(S)}_{ave,N}$ of a nominal performance are mostly constant, then we translated the rules into deviation profiles $P^{(S)}_{dev}$.

Each of the KTH rules can be described in terms of deviations that have to be applied to the score. The deviations are: timing deviations ΔIOI, duration deviations ΔD, and intensity deviations ΔI. The first two quantities are expressed in ms, while the third is expressed in dB. When all the deviations for each rule have been computed, then they are linearly combined, weighted by the respective rule quantities, the so called k_{KTH} parameters.

The timing information can be translated into a profile of metronome deviations. An expression of such a transformation can be derived by considering how the CSC model and the KTH one calculate the IOI, whose expression should produce the same value; for m rules we have:

$$\frac{IOI_{[tick]} \cdot C}{MM_N + \sum_i^m \Delta MM_i} = IOI_N + \sum_i^m \Delta IOI_i \tag{5}$$

[2]The software can be downloaded at `http://www.eyesweb.org`

[3]The libraries and the patches can be downloaded at `http://www.dei.unipd.it/ricerca/csc/research_groups/mega/mega.html`

where the subscript N indicates values which are related to the the score (nominal values): the symbols $IOI_{[tick]}$, IOI_N, and MM_N refer to the nominal inter onset interval expressed in tick, in ms, and to the nominal metronome respectively. The symbol ΔIOI_i indicates the inter onset interval deviation introduced by the i-th rule, and ΔMM_i is the respective metronome deviation.

In this formula we can immediately see that the property of additivity which holds for the deviations introduced by the KTH rule system, is not directly translated into additivity for the metronome profile. Moreover, the effect of different rules on the sum of metronome deviations cannot be decoupled from each other. However, for small deviations of ΔIOI_i, the equation (5) can be simplified into the following expression for the metronome deviations:

$$\Delta MM_i \simeq -\frac{\Delta IOI_i \cdot MM_N}{IOI_N + \Delta IOI_i} \tag{6}$$

in which the effects of the rules are decoupled, and the additivity property is maintained.

The duration information has to be managed similarly, in order to obtain a profile of legato deviations. More precisely, if we have m rules, the durations computed by the two models are combined in the following expression:

$$\frac{IOI_{[tick]} \cdot C \cdot \left(leg_N + \sum_i^m \Delta leg_i\right)}{MM_N + \sum_i^m \Delta MM_i} = D_N + \sum_i^m \Delta D_i$$

Also in this case, a simplification is needed in order to achieve linearity and decoupling of the effects of different rules: for small values of ΔIOI and ΔD, we can approximate the deviation of legato introduced by the i-th rule with:

$$\Delta leg_i \simeq MM_N \frac{IOI_N \Delta D_i - D_N \Delta IOI_i}{IOI_{[tick],N} \cdot C \cdot (IOI_N + \Delta IOI_i)}. \tag{7}$$

Finally, all the intensities I expressed in dB are translated into a profile of Key Velocities KV. As usual, if we have m rules, then:

$$KV_N + \sum_i^m \Delta KV_i = f\left(I_N + \sum_i^m \Delta I_i\right)$$

where f is a suitable conversion function, which depends on the sound card used. Decoupling and additivity can be achieved by considering that f is approximatively linear, so that:

$$\Delta KV_i \simeq f(\Delta I_i). \tag{8}$$

Thus, for each rule, three profiles are obtained: $P_1 = \Delta MM$, $P_2 = \Delta leg$, and $P_3 = \Delta KV$.

3.2. Relations between parameters

In the KTH rule system, different rules can be weighted by the so called k_{KTH} parameters, allowing them to model performances more closely and to adapt the rules to different situations.

When we converted the KTH rules into profiles, we had to face with the conversion of the controlling parameters too. In fact, we want that the effect of a rule – when its weighting parameter is equal to a certain value – should be the same of the effect of the rule when it is codified into the CSC expressiveness model. To do so, a relation between the controlling parameters has to be found.

This relation can be obtained by considering how generally the controlling parameters acts in the KTH rule system and into the CSC one; to do so we take as references the deviations calculated by the rules and by the formulas (6, 7, 8) when $k_{KTH} = 1$. In the case of the inter onset deviations and the metronome deviations we have:

$$\Delta IOI = k_{KTH} \cdot \Delta IOI\,|_{k_{KTH}=1}$$
$$\Delta MM = k_{MM} \cdot \Delta MM\,|_{k_{KTH}=1}$$

which can be substituted into (6) obtaining:

$$k_{MM} \Delta MM\,|_{k_{KTH}=1} \simeq -\frac{k_{KTH} \Delta IOI\,|_{k_{KTH}=1}\ MM_N}{IOI_N + k_{KTH} \Delta IOI\,|_{k_{KTH}=1}}$$

For small deviations of intr onset intervals, we can neglect their contribution into the denominator, leading to the final formula for k_{MM} (and the other controlling parameters in a similar way):

$$k_{MM} \simeq k_{KTH},\ \ k_{leg} \simeq k_{KTH},\ \ k_{KV} \simeq k_{KTH}.$$

Thus, for each rule three profiles are defined which are controlled by three controlling parameters: $k_1 = k_{MM}$, $k_2 = k_{leg}$, and $k_3 = k_{KV}$.

3.3. The Expressive Director

The `ExpressiveSeq` block uses six profiles for the generation of the expressive performances. The number of profiles (and the relative k and m parameters) is strictly what is required by the CSC expressive model, so that no other redundant profile is allowed. More precisely, each sonologic parameter (tempo, legato and intensity) is affected exactly by 2 profiles, the average profile P_{ave} and the deviation profile P_{dev}, see (3). On the other hand each profile P can affect only one sonologic parameter.

When we generalized the model to include the KTH rules, we had to rewrite the block in order to allow more than 6 profiles. This allowed each sonologic parameter to be affected by more than 2 parameters. Thus, the `ExpressiveDirector` was designed to accept more than 6 profiles, each of which can affect any of the 3 sonologic parameters, i.e. each profile P_l can affect more than one sonologic parameter. To do so, the equation (3) has to be rewritten into the subsequent:

$$S = \sum_l^p k_l(u) \cdot \delta_l^{(S)} \cdot P_l$$

where p is the number of profiles, $k_l(u)$ is the controlling parameter for the l-th profile, and the symbol δ_l^S specify if the sonologic parameter S is affected (the value is 1) or not (the value is 0) by the l-th profile. This last information has been included at the beginning of the profiles file.

4. INTERFACES

The control of the Expressive Director is made through the $k_l(u)$ parameters, which depends on the user input. However, the number of parameters is quite large, making difficult handling them. Thus, we developed an interface able to provide a user friendly control of the expressive rendering, allowing the user to change the expressive intention in real time continuously.

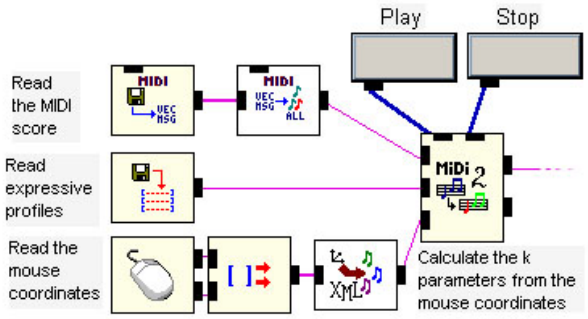


Figure 1: Snapshot of the `ExpressiveDirector` (or `ExpressiveSeq`) at work: the blocks on the left side provide the information needed to synthesize an expressive performance (*score*, *profiles* and k parameters), the central block is the `ExpressiveDirector` (or `ExpressiveSeq`) which control the synthesis, and on the right side, there are the blocks that redirect the output to a MIDI port.

4.1. The control space

The control space controls the expressive content and the interaction between the user and the final expressive performance. In order to realize a morphing among different expressive intentions we developed an abstract control space, called perceptual parametric space (PPS), that is a two-dimensional space derived by multidimensional analysis of perceptual tests on various professionally performed pieces ranging from western classical to popular music [11, 12]. This space reflects how the musical performances might be organized in the listener's mind. It was found that the axes of PPS are correlated to acoustical and musical values perceived by the listeners themselves [13]. We make the hypothesis that a linear relation exists between the PPS axes and each expressive controlling parameter k_l; thus, if x and y are the coordinates of the PPS, then:

$$k_l(x, y) = a_{l,0} + a_{l,1}x + a_{l,2}y.$$

5. CONCLUSIONS

A technical overview of a real time application for expressive rendering of musical performances has been presented, with a brief presentation of the two main contributions that have been merged into the Expressive Director design. Some issues have been examined more deeply, i.e. the translation between the two expressiveness models, both at expressive profiles level, and at the control parameters level.

Some technical refinements have to be carried out:

- implementation into the Expressive Director block of the translation function $KV = f(I)$ in order to be independent from the sound card used and to reduce the approximation of the intensity formulas;
- refinements to the configuration library, in order to manage a large number rules (actually limited at 15).
- assessment of the system has to be made, in order to verify both to which extent the approximated formulas are still valid, and the reliability of the expressive communication channel between the artist using this system and the listeners.

6. ACKNOWLEDGMENTS

This research was supported by the EC project IST-1999-20410 MEGA.

7. REFERENCES

[1] Bresin, R. and Friberg, A. (2000). "Emotional colouring of computer controlled music performance". *Computer Music Journal*, 24(4): 44-62.

[2] Camurri, A., Coletta, P., Peri, M., Ricchetti, M., Ricci, A., Trocca, R., Volpe, G. (2000). "A real-time platform for interactive performance", *Proc. of the ICMC-2000*, Berlin, 374-379.

[3] Sundberg, J., Askenfelt, A. and Frydén, L. (1983). "Musical performance: A synthesis-by-rule approach", *Computer Music Journal*, 7, 37-43.

[4] Friberg, A. (1995). "Matching the rule parameters of Phrase arch to performances of 'Trumerei': A preliminary study", in A. Friberg and J. Sundberg (eds.), *Proceedings of the KTH symposium on Grammars for music performance*, May 27, 1995, pp. 37-44.

[5] Friberg, A. (1995). "A Quantitative Rule System for Musical Expression", *Doctoral dissertation*, Royal Institute of Technology, Sweden.

[6] Friberg, A, Colombo, V, Frydn, L and Sundberg, J (2000). "Generating Musical Performances with Director Musices". *Computer Music Journal*, 24:3, 23-29

[7] Friberg, A. and Battel, G., U. (2002). "Structural Communication". In (R. Parncutt and G. E. McPherson, Eds.) *The Science and Psychology of Music Performance: Creative Strategies for Teaching and Learning*. New York: Oxford University Press, 199-218.

[8] Juslin, P. N., Friberg, A., and Bresin, R. (2002). "Toward a computational model of expression in performance: The GERM model". *Musicae Scientiae*, Special issue 2001-2002, 63-122.

[9] Canazza, S., Rodà, A. (1999). "A parametric model of expressiveness in musical performance based on perceptual and acoustic analyses", *Proc. of the ICMC99 Conf.*, November, 1-4.

[10] Canazza, S., De Poli, G., Drioli, C., Rodà, A., Vidolin, A. (2000). "Audio Morphing Different Expressive Intentions for Multimedia Systems", *IEEE Multimedia*, 7(3), 79-84.

[11] Canazza, S., De Poli, G., A., Vidolin, A. (1997). "Perceptual Analysis of the Musical Expressive Intention in a Clarinet Performance". In (M. Leman ed.) *Music, Gestalt, and Computing*, Springer Verlag, 441–450.

[12] Canazza, S., Orio, N., (1999). "The Communication of Emotions in Jazz Music: a Study on Piano and Saxophone Performances", In (Marta Olivetti Belardinelli ed.) *Musical Behaviour and Cognition*, 263–278.

[13] Canazza, S., De Poli, G., A., Vidolin, A. (1996). "Perceptual analysis of the musical expressive intention in a clarinet performance", *IV International Symposium on Systematic and Comparative Musicology*, September, Brugge, 31–37.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

EXPRESSIVENESS ANALYSIS OF VIRTUAL SOUND MOVEMENTS: AUDIO-VIDEO CROSS CORRELATION

Amalia de Götzen

CSC-DEI, Università di Padova
corvo@dei.unipd.it

ABSTRACT

This paper describes a work which investigates sound motion expressiveness, with particular focus on cross-correlation between audio and video. This work is divided in two parts: an overview of some results obtained in sound motion expressiveness with his consequent musical application and [6]a design of a perceptive audio-video test. The main idea is to find a way to use sound motion like a musical parameter and in a coherent manner with video to convey a specific expressive content. The purposes of this work can be stated as follow:

- To show how multi-modality can be conveyed by acoustic means by changing the position of sound objects into a virtual space;
- To open a new multi-modal channel to convey expressiveness;
- To establish an explicit connection between multi-sensory applications and expressiveness;

These purposes are related to the idea that much contemporary art performances and a lot of applications from Human Computer Interaction can take advantage of a good use of audio-video cross-correlation.

1. INTRODUCTION

The need for this kind of investigation comes from contemporary music works in which sound motion is a musical parameter like intensity, timbre, pitch etc.. While the connection between music and emotion of these afore mentioned parameters has been deeply investigated [1] [2], spatialization and its relations with visual aspects still remains a new research path. On the other hand, interferences between senses has been studied very well from a perceptive point of view [3] [4] but not from an expressive one. This work has been conducted within the MEGA (Multisensory Expressive Gesture Application, cf. `http://www.megaproject.org`) project which concentrates on expressive and emotional content modeling and communication in non-verbal interaction through the use of multi-sensory interfaces in shared interactive Mixed Reality environments.

2. EXPRESSIVENESS OF VIRTUAL SOUND MOVEMENTS

In order to evaluate whether sound movement can be considered a musical parameter and whether it has some expressive content for the listener, a psychoacoustic test was performed. Its design was aimed at establishing a perceptive paradigm to construct a model that can be used in a musical context. This preliminary study focused on three parameters which are considered to be the basic components of sound movement: speed, legato/staccato and path. Variations of these parameters were designed to obtain eight stimuli/gestures. These variables in these parameters were designed to have only two opposed values in each example:

- speed: 0 = 200ms, 1 = 2000ms (where this sustain time represents the sound permanence on every loudspeaker)
- path: 0 = continuous circular, 1 = discontinuous random
- legato/staccato: 0 = 0.2 overlap, 1 = 2 overlap (where the overlap shows the degree of crossed overlap between the loudspeakers amplitude envelopes:0.2 means staccato, while 2 means legato)

With these kinds of movements, 8 audio examples were produced featuring white noise sound, while another 8 examples featured an harmonic continuous sound (a looped trombone sample) in order to evaluate if timbre has an influence on movement perception too. Listeners where placed in the center of a circular 6 loudspeakers setup. 20 subjects (10 musicians, 10 non-musicians) were selected and a dimensional approach was used: the subjects were asked to mark in a bi-dimensional space an increasing order alphanumeric character at each stimulus. The bi-dimensional space was organized to establish valence on the vertical axis and arousal on the horizontal one. This is an approach of representing emotional states, drawn from the psychological tradition, which is alternative to the categorical one. The clearest common element of emotional states is that the person is materially influenced by feelings that possess a valence: they are centrally concerned with positive or negative evaluations of people or things or events; furthermore, emotional states involve a disposition to (re)act in certain ways. Arousal states are simply rated in terms of the person's disposition to take some action rather than no action at all [5].

2.1. Data analysis and preliminary results

The collected data and its statistical analysis allow some conclusions to be drawn:

- The values obtained from each timbre identify a similar space location for the same movements; there are some movements that seem to be further correlated for both tests. These movements are correlated by featuring the same speed and the same legato/staccato characteristics, while path is not overly important; the the T-test further outlines this evaluation and the correlation test shows a better correlation of musician answers;

- Legato/staccato and speed are stronger parameters than path, even though the Russel model features a specific emotion related to path;
- Timbre influences subject answers: it was observed that very often the valence component was also dependent on the pleasantness of the sound heard; for example: in the white noise case, someone referred to it as a reminiscence of sea wave, while others were just plainly annoyed by it; consequently, the standard deviation over this collected data set was found to be bigger than that related to harmonic sound.
- Speed is strongly related to the arousal axis and it is also the most coherent parameter in subjective appreciation in the examples with the two different timbres.
- Different kinds of path (continuous or discontinuous) appear to be a more subtle parameter that can lead to the distinction among emotions which belong to the same area, while speed and legato/staccato appear to be stronger parameters in direct relation to emotions. In fact, these latter parameters have been widely studied in psychoacoustics and they are strongly connotated from the expressive point of view [1]: a smooth amplitude envelope, for example, is associated to emotions like sadness, fear or tenderness, while a steep slope envelope is related to happiness or anger like a staccato articulation is.

2.2. A music application: *Medea* by Adriano Guarnieri

This framework has been used in the opera *Medea* by Italian composer Adriano Guarnieri which was premiered at the PalaFenice in Venezia in October 2002 [7]. *Medea* is an innovative work which stems out of an in-depth research in the multimedia domain; the importance and the amount of research dedicated to sound movement in space is considerable in this work. Guarnieri insisted explicitly on wanting to achieve an expressive matching between instrumental and sound movement gestures so that both gestures would reinforce each other producing a more powerful and complete message. Sound spatialization is a new way of conveying expressive content, and in this perspective it can be used to open new expressive channels that can interact with other audio transformations and visual analyses, broadening the spectrum of multi-modality capabilities. *Medea* has offered the possibility to experiment with spatialization controls driven by the very same instrumental gesture (coming in this case from one of the solo trombone players) while allowing the connection between advanced technological research and a real-world complex musical event. To be specific, the four solo trombone players are located in the hall among the public and follow very distinct individual movements. It is interesting to notice that some of the pages of the score suggest specific gestures to the trombone players (standing up, stirring up, etc.) That is precisely when spatialization is driven by their gestures. A webcam picks up the movements of the players and digital image analysis performed on this signal allows to extract some parameters that control the speed (and therefore part of the expressive features) of spatialization.

3. AUDIO-VIDEO CROSS-CORRELATION: FROM THE EXPRESSIVENESS POINT OF VIEW

Several studies of multi-sensory perceptions have demonstrated that information received from the various sensory systems are not processed independently [8]. On the contrary, multi sensory couplings are believed to be inherent to our perceptual model of the world. In fact, the quality of realism in virtual environments is typically considered to be a function of visual and audio fidelity mutually exclusive of each other. However, the virtual environment participant, being human, is multi-modal by nature. Therefore, in order to validate more accurately the levels of auditory and visual fidelity required in a virtual environment, a better understanding is needed of the inter-sensory or cross-modal effects between the auditory and visual sense modalities. In particular, in this work we strive to investigate the expressiveness possibilities of images and sounds representing movements, starting by the assumption that the video content can be reinforced or weakened by a sound source which is moving around the audience. In other words we have tried to assess if the spatial auditory perception is dependent on the visual one from the expressiveness point of view.

3.1. The perceptive test

In order to evaluate whether sound movement can be considered a musical parameter which can help or contrast the video expressiveness and whether this match or mismatch has some expressive content for the listener a perceptive test was performed using the process and the approach already seen in par. 2. Its design was aimed at establishing a perceptive paradigm to construct a model that can be used in multi-modal contexts. Three kinds of video with 2 opposite speed values (slow and fast) were used:

- an airplane: to perform a straight path
- a merry-go-round: to perform a circular path
- a virtual animation: a ball moving on the screen in circular paths with different articulations

In order to perform some sound movements matching or contrasting the videos a white noise sound was used performing the 28 stimuli shown in tab. 1. For each kind of video several sound stimuli were produced: two kinds of spatialization movements were used in order to match or to contrast the visual perception. The circular movements on the video were matched with circular movements of the virtual sound source around the subject with different speeds and rotation spins. In the airplane example a straight motion of sound parallel to the video was created as shown in fig. 1; in this case several speed and direction variations of the sound source were applied. The test setup was the same used for the

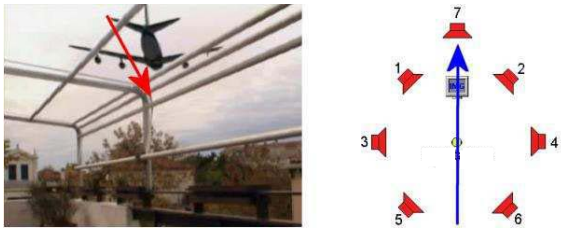


Figure 1: *Video and audio setup for the airplane example*

investigation mentioned in par.2.

3.2. Data analysis

The data collected has been analyzed statistically in order to obtain the desired information.

	Video			Audio	
N	Subject	Path	Speed	Path	Speed
1	airplane	straight	slow	MUTE	
2	airplane	straight	slow	matched	slow
3	airplane	straight	slow	mismatched	fast
4	airplane	straight	fast	MUTE	
5	airplane	straight	fast	matched	slow
6	airplane	straight	fast	mismatched	fast
7	merry-go-round	circular	slow	MUTE	
8	merry-go-round	circular	slow	circular matched	slow
9	merry-go-round	circular	slow	circular mismatched	slow
10	merry-go-round	circular	slow	circular matched	fast
11	merry-go-round	circular	slow	circular mismatched	fast
12	merry-go-round	circular	fast	MUTE	
13	merry-go-round	circular	fast	circular matched	fast
14	merry-go-round	circular	fast	circular mismatched	fast
15	merry-go-round	circular	fast	circular matched	slow
16	merry-go-round	circular	fast	circular mismatched	slow
17	animation	circular legato	slow	MUTE	
18	animation	circular legato	slow	circ. legato matched	slow
19	animation	circular legato	slow	circ. staccato mismatched	fast
20	animation	circular legato	fast	MUTE	
21	animation	circular legato	fast	circ. legato matched	fast
22	animation	circular legato	fast	circ. staccato mismatched	slow
23	animation	circular staccato	slow	MUTE	
24	animation	circular staccato	slow	circ. staccato matched	slow
25	animation	circular staccato	slow	circ. legato mismatched	fast
26	animation	circular staccato	fast	MUTE	
27	animation	circular staccato	fast	circ. staccato matched	fast
28	animation	circular staccato	fast	circ. legato mismatched	slow

Table 1: Stimuli

A one factor ANOVA test for each dimension and for each kind of video was performed in order to evaluate whether the mean value of the collected data was significant. The results of the ANOVA test show that all mean values are significant both for the valence and the arousal axis in every video. The results are better than with just audio where the significance values were not so good. Drawing mean values it can be observed that for each video the mute stimuli have low arousal values, while stimuli with sound matched with video have high arousal values. According to the results of the previous work it becomes clear that the rotation sense is not perceived, while the sound articulation used in the last video is a strong parameter. The cluster analysis has been performed using the Complete Linkage method: three groups have been found, as shown in fig. 2 Using the cluster analysis some peculiar re-

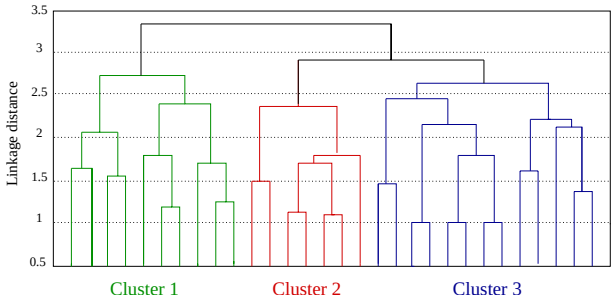


Figure 2: *Tree diagram for 30 variables using the complete linkage method.*

actions may be observed: cluster 2, composed by 7 subjects, did not recognize the different sounds matched or mismatched with video: it separated plainly the different kinds of video providing very similar values for different sound movements. Clusters 1 and 3, composed by 13 and 10 subjects respectively, have been able to recognize very well the different stimuli, in particular the mute one from the others. While cluster 1 choose a high value of valence and arousal for stimuli with sound in agreement with video, cluster 3 seemed to have a bigger attention to sound, choosing high values for stimuli with fast and staccato audio, without any correlation with video speed or path. In any case, clusters 1 and 3 have separated very clearly the different sound-track of each video, using the expressive space in a different way but keeping the same relative distance among the different stimuli. In Figs. 3, 4, 5 is shown the mean values for each video and for each matched or mismatched sound used. The airplane example seems to be the best result because each stimulus is very well distinguished from the other: we can suppose that a sound movement which involve the body of the listener (as a sound which is moving far away from the listener) can be perceived as a stronger stimulus. The suggestion is that the movement far-away is the best expressive way in order to do expressive sound movements. We used two arrows in order to underline the mute-mismatched-matched 'andamento'. In the merry-go-round video the sound, when it was present, could move with path and speed matched or mismatched with the video: the video rotation speed determine the arousal value while the matched or mismatched path modified the valence: fast speed implies high valence and arousal values, the matched audio increase this result. In the animation video sound movement have also the articulation parameter: here the results are somewhat diffiult to read in particular for the fast video, where the stimuli are too much close. Anyway we can see that speed and articulation modified the arousal value according to previous works.

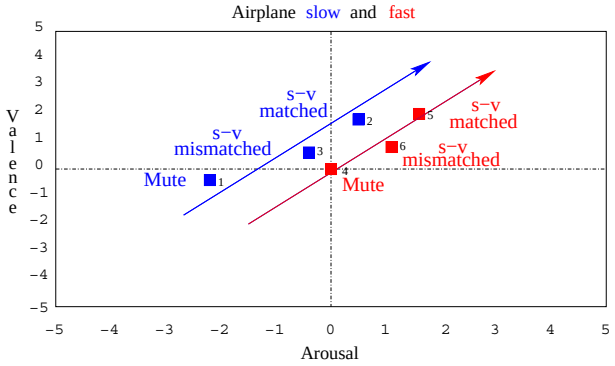


Figure 3: *Mean values for airplane stimuli*

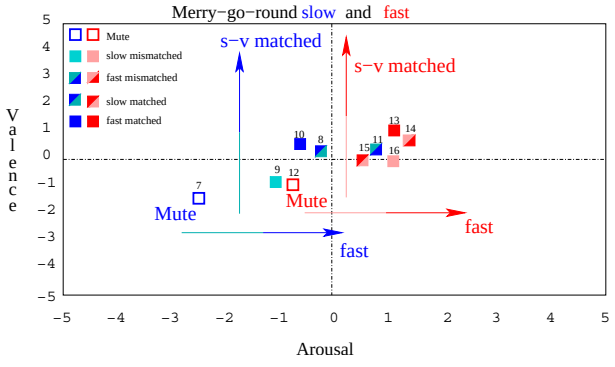


Figure 4: *Mean values for the animation stimuli*

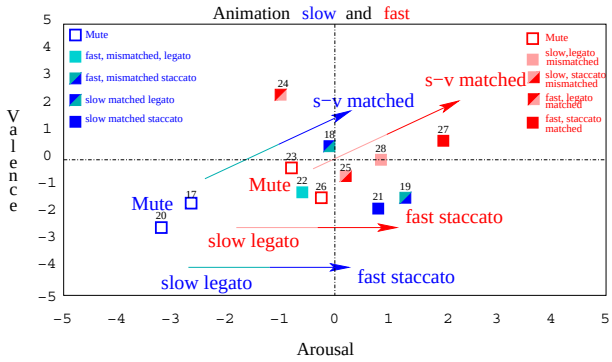


Figure 5: *Mean values for the merry-go-round stimuli*

3.3. Conclusions

The data obtained in the perceptive tests have underlined a correlation between the proposed sound-track and the position over the expressive space. This is a preliminary study about audio-video cross-correlation; however, interesting results with further investigation on spatialization can be expected. Comparing these results with former ones the following observations can be made:

- Timbre is very important for the valence axis; in particular, using sound matched with video leads to a rise of "perceivable" sound motion in the valence axis
- In the tests carried out, speed is strongly related to the arousal axis and fast speed leads to a rise of the arousal in each example
- While the subjects did not separate very well circular and random path in the former tests, they were able to perceive different kinds of movements very accurately in this one (probably suggested by the accompanying video)

In general a stronger involvement of the audience and a stronger perceived expressive intention can be noticed using matched sound-video sequences. This kind of relationship is bound to be used very carefully in multi-modal contexts.

4. ACKNOWLEDGEMENT

The author wishes to thank all the brave souls who have accepted to patiently sit through the test sessions, Prof.Giovanni De Poli for allowing the work to be carried out at the CSC, Alvise Vidolin, Nicola Bernardini and Adriano Guarnieri for allowing her to participate to the making of a master-work. Ing. Antonio Rodà for his constant help and useful suggestions and Gabriele Zulini who carried out the largest part of the cross-correlation work.

5. REFERENCES

[1] Juslin, Sloboda, Psychological perspectives on music and emotion, Music and Emotion - Theory and research, Oxford University Press, Oxford, 2001

[2] Gabrielsson and Lindstrom, The influence of musical structure on emotional expression, Music and Emotion - Theory and research, Oxford University Press, Oxford, 2001.

[3] Shams et al., What you see is what you hear, Nature, 2000

[4] McGurck, MacDonalds, Hearing lips and seeing voices, Nature, 746-748, 1976

[5] Cowie, R., Douglas-Cowie, E., Tsuputsoulis, G., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., Emotion Recognition in Human-Computer Interaction, IEEE Signal Processing Magazine, January 2001.

[6] Amalia de Götzen, "Expressiveness analysis of virtual sound movements and its musical applications", in *Proceedings of XIV Colloquium in Musical Informatics*, Firenze, 2003

[7] Nicola Bernardini, Alvise Vidolin, "*Medea* by Adriano Guarnieri: a report on extreme Live-Electronics", in *Proceedings of XIV Colloquium in Musical Informatics*, Firenze, 2003

[8] A. Kohlrausch and S. van de Par, "Audio-Visual Interaction: From Fundamental Research in Cognitive Psychology to (possible) Applications", *Human Vision and Electronic Imaging IV*, vol 3644, pp 34-44, 1993

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

A STATISTICAL APPROACH TO EXPRESSIVE INTENTION RECOGNITION IN VIOLIN PERFORMANCES

Roberto Dillon

Laboratory of Musical Informatics – University of Genoa

rdillon@dist.unige.it

ABSTRACT

This paper presents a possible approach to the automatic recognition of expressive intentions in actual music performance. In particular we will analyze three performances (played with different expressive characteristics such as "brillante", "agitato" and "cupo") of the same excerpt played by a professional violinist and propose a classification system based on the analysis of a set of simple audio parameters, extracted in real time by a set of tools implemented in the EyesWeb open platform.

The cue extraction process is carried out by looking at a "note" and a "phrase" profile (obtained by squaring and low pass filtering the incoming signal using different cut off frequencies). These are compared to obtain, for each note, parameters regarding note duration, articulation and dynamics (needed to point out how the performance is evolving). Statistical analysis is then performed on the extracted data to see if it is possible to find remarkable differences between the original performances.

Once this is confirmed, a Hidden Markov Model (HMM), developed in MatLab and able to recognize the player's intentions, having the audio cues as inputs, is explained.

1. INTRODUCTION

This paper concerns the study of sonological parameters involved in expressive performance [1], [2] and tries to find out if a basic set of cues, related to metronomical and dynamics aspects such as tempo, sound level, articulation and attack velocity, can be useful for classifying performances characterized by different expressive intentions.

The application framework is the EyesWeb open real-time platform [3] and the developed audio cues extraction libraries are integrated within the EyesWeb environment.

The paper is divided into four sections covering all the steps involved:

- Choice and extraction of audio cues
- Statistical analysis
- HMM based recognition system
- Results and conclusions

2. CHOICE AND EXTRACTION OF AUDIO CUES

For this experiment we selected three performances of the first part of the second movement from Arcangelo Corelli's Sonata Op.5 n.8 performed by a professional violinist following different expressive intentions: *agitato* (restless, hectic), *brillante* (glittering, sparkling) and *cupo* (dark).

Each intention was performed more than once and recorded as wave file in mono with a sampling rate of 22050 Hz and 16 bit resolution.

To analyse the input audio stream in the EyesWeb environment, we followed the approach used in [4] and used an improved version of the tools first proposed in [5].

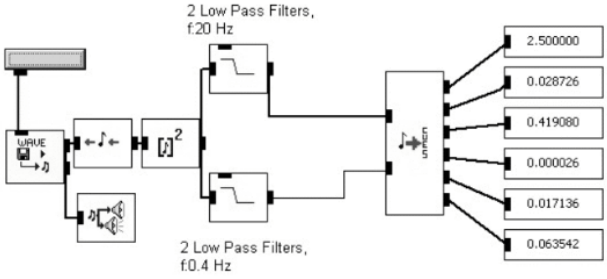


Figure 1: *the EyesWeb patch: from sound input to cues*

The patch in fig.1 takes an input and stores the last 4 seconds of music. This time interval seemed a good compromise between the short time needed by the system to react in real time and the longer time required to understand what is going on from a musical point of view. The 4 seconds time frame is updated 5 times per second (by the first block after the input wave reader in fig.1. This overlapping between two following buffers allows a very smooth analysis of the incoming signal, useful to appreciate any variations in the characteristics of the performance under analysis (such as following the development of a crescendo, etc.).

The data in the current frame are then elaborated following the suggestions firstly presented in [6] to extract event information from an actual audio input (the music stream is simply considered like a standard PAM signal): the buffers are first squared and then low pass filtered using 2 first order IIR filters in cascade (to get zero ripple) with very low cut off frequencies (in this case around 20 Hz and 0.4 Hz).

In this way we extract two profiles: the first, extracted with a cut off frequency of 20Hz, is the amplitude envelope of the signal and can be seen as related to fast events, or "notes". The other one is the energy of the signal and can be interpreted as an average "phrase" profile.

These two profiles are overlapped and compared by the "Cues" block in the patch (fig.1) so as to get a set of sonological parameters defined as follows:

• ***Tempo***, defined as the average Note Duration (DR), in seconds, of the notes in the buffer
• ***Articulation***, defined as Actual Note Duration / Inter Onset Interval averaged across the notes contained in the buffer
• ***Standard Deviation of Articulation***, computed over the notes contained in the buffer
• ***Mean Sound Level*** of the notes contained in the buffer where *Sound Level* is the amplitude measured at the beginning of the note, i.e. at the intersection between the two profiles
• ***Mean Sound Level Difference***, where *Sound Level Difference* is defined as the difference between the current and the preceding note. This cue offers useful information for identifying local crescendo/decrescendo patterns.
• ***Mean Attack Velocity*** of the notes contained in the buffer where *Attack Velocity* is defined as the derivative of the note profile at the intersection with the phrase profile

3. STATISTICAL ANALYSIS

In agreement with the violinst, a particular performance was selected for analysis purposes.

By analyzing all the cues over the whole piece as performed in the various expressive intentions, the following overall Mean (M) and Standard Deviation (SD) values were obtained (Table 1):

Mean	Cupo	Brillante	Agitato
Tempo	0.20455	0.11263	0.14558
Articulation	0.59644	0.50132	0.55864
SD of Articulation	0.08214	0.05669	0.06693
Sound Level	0.05284	0.01156	0.03844
Sound Level Difference	-2.1226	-0.6427	-0.8611
Attack Velocity	0.08992	0.03093	0.07226

Standard Deviation	Cupo	Brillante	Agitato
Tempo	0.01402	0.00071	0.00203
Articulation	0.02154	0.01309	0.01556
SD of Articulation	0.00260	0.00092	0.00148
Sound Level	0.00165	0.00003	0.00021
Sound Level Difference	193.694	12.2506	16.1832
Attack Velocity	0.00775	0.00027	0.00100

Table 1: *overall mean and standard deviation*

To check whether the extracted cues are actually able to show statistically meaningful differences among the three performances, a single factor ANOVA analysis was carried out. Thereby we assumed that the cues' distributions were Gaussian. This assuption is appropriate since the values refer to an average across numbers extracted in the current buffer, and therefore satisfy the hypothesis of the Central Limit Theorem regardless of the distribution of the original samples (for a comprehensive explanation of this fundamental theorem of probability theory see, for instance, [7]).

The computed F and p values ($\alpha = 0.05$. F critical value is 3.0191) are shown in Table 2:

Expressive Cue	**Computed F**	**P-value**
Tempo	44.0845	5.6921E-18
Articulation	17.8136	3.9822E-08
SD of Articulation	12.8306	4.0286E-06
Sound Level	73.187	< E-20
Sound Level Difference	0.97481	0.3781891
Attack Velocity	31.6994	1.7922E-13

Table 2: *ANOVA results*

As we see, the results are very good with the only exception of the sound level difference. Actually the performer followed the same kind of crescendo decrescendo patterns, so this bad result is understandable.

To have a better understanding of the results, also F and T tests were computed for every possible couple of performances for each expressive cue. In this way we can determine if means and/or standard deviations of the Gaussian distributions were different enough to guarantee a good identification between different performances.

The results are shown in Table 3. The value shown is the probability of incurring in a Type-I error ($\alpha = 0.05$) i.e. rejecting the null hypotesis (samples are coming from the same distribution) while it is actually true.

Tempo	**F-Test**	**T- Test**
Agitato/Brillante	1.1 E-07	1.1 E-8
Agitato/Cupo	< E-10	< E-10
Brillante / Cupo	< E-10	< E-10
Mean of Articulation	**F-Test**	**T- Test**
Agitato/Brillante	0.3741	0.0017
Agitato/Cupo	0.1599	0.0039
Brillante / Cupo	0.0181	< E-10
Standard Deviation of Articulation	**F-Test**	**T- Test**
Agitato/Brillante	0.0158	0.0376
Agitato/Cupo	0.0243	0.0052
Brillante / Cupo	9.09 E-7	2.29 E-7
Mean of Sound Level	**F-Test**	**T- Test**
Agitato/Brillante	< E-10	< E-10
Agitato/Cupo	< E-10	2.96 E-5
Brillante / Cupo	< E-10	< E-10
Sound Level Difference	**F-Test**	**T- Test**
Agitato/Brillante	0.1518	0.6756
Agitato/Cupo	< E-10	0.2623
Brillante / Cupo	< E-10	0.1801
Mean of Attack Velocity	**F-Test**	**T- Test**
Agitato/Brillante	< E-10	< E-10
Agitato/Cupo	< E-10	0.0161
Brillante / Cupo	< E-10	< E-10

Table 3: *F and T Tests for every pair of performances*

As shown in Table 3, the statistical analysis produced interesting results, again with relevant exception of the Sound Level Difference. There are problems just in a few cases but when one test fails, the other test shows quite good results (like in the Agitato/Brillante for the Mean of Articulation cue: the F test is very bad but the T test is good).

Hence the probability of wrongly classifying a performance is low, provided both mean and standard deviation of the distributions are known.

4. HMM BASED RECOGNITION SYSTEM

Having proved that the chosen cues show statistically meaningful differences between the various performances, we can now proceed to propose a system that attempts to recognize the particular expressive intention on the basis of the overall mean and standard deviation for each cue. The proposed system is based on a Hidden Markov Model (HMM) that was implemented in MatLab. This yields the probability that the cues extracted by the EyesWeb patch at a certain time belong to a given intention. The system is tested on a different set of performances from the one used for extracting the global parameters, but still played by the same violinist, so as to find out if we can correctly classify a cue vector from a particular time frame on the sole basis of knowing a set of global parameters, i.e. mean and standard deviation.

The Observable States of the HMM are the cue values while the Hidden States, i.e. those we want to guess by looking at the obervable states, are the different expressive intentions.

The HMM is defined by a *Transition Matrix* (TM), a *Confusion Matrix* (CM) and by a Π-vector representing the initial state, which is the hidden state we choose the system to start from. It should be stressed that, in our case, the TM is not constant in time and is a 6x3 matrix defined so as each row yields, for each cue, the conditional probability (Pr.) that the current value belongs to one of the possible hidden states. Note that the sum of the elements in each row must be 1 since the incoming value must be from one of the possible hidden states.

$$\begin{pmatrix} \Pr.Tempo|Agitato & \Pr.Tempo|Brillante & \Pr.Tempo|Cupo \\ \Pr.Art.M|Agitato & \Pr.Art.M|Brillante & \Pr.Art.M|Cupo \\ \Pr.ArtSD|Agitato & \Pr.Art.SD|Brillante & \Pr.Art.SD|Cupo \\ \Pr.SoundM|Agitato & \Pr.SoundM|Brillante & \Pr.SoundM|Cupo \\ \Pr.SoundDiff|Agitato & \Pr.SoundDiff|Brillante & \Pr.SoundDiff|Cupo \\ \Pr.AttackVel|Agitato & \Pr.AttackVel|Brillante & \Pr.AttackVel|Cupo \end{pmatrix}$$

Where, for example

$$\Pr.Tempo \mid Agitato = \int_{A}^{B} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx \quad (1)$$

σ and μ are the overall standard deviation and mean previously calculated for the cue Tempo, respectively, in the agitato performance. The integration interval [A,B] is centered around the current cue value and its amplitude D is equal to

$$D = (Vmax - Vmin) / N \quad (2)$$

Here Vmax and Vmin are the maximum and minimum values assumed by the cue, as calculated previously, and N is equal to 1+1.43ln(n) where n is the overall number of samples (i.e. time frames).

The TM instead, which defines the probability of being in state X at time *t*, having been in state Y at time *t-1*, is constant in time and has been defined by optimizing the probability of correctly classyfing the performances having the CM already defined and the original set of performances as input.

The initial state is assigned to the Agitato (Π-vector = [1 0 0]). The reason for this choice was that the results in Table 1 showed this performance to be often half way between the two others.

5. RESULTS AND CONCLUSIONS

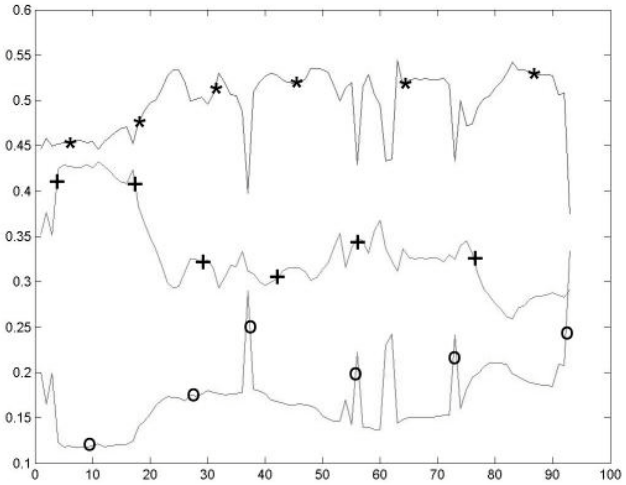


Figure 2: *Listening to the 'Agitato' performance. The abscissa shows the sample number (there are 5 samples per second) while the ordinate is the probability assigned by the system to a particular expressive intention. The line marked with '*' refers to Agitato, '+' to Brillante, 'o' to Cupo.*

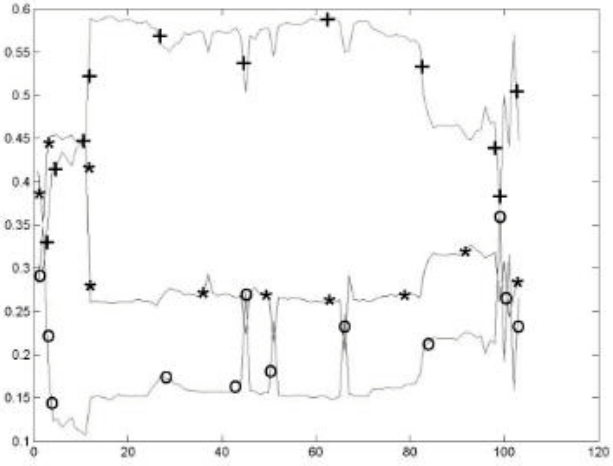


Figure 3: *Listening to the 'Brillante' performance. Symbols as in fig.2*

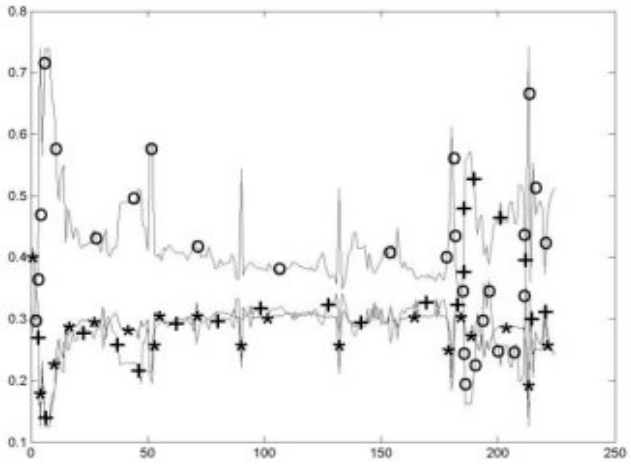


Figure 4: *Listening to the 'Cupo' performance. Symbols as in fig.2*

As we see from fig.2-4, the system shows very good results: the 'Agitato' (fig.2) is correctly recognized during the whole performance without any problem, in fact its probability is the highest from the beginning to the end of the piece. The 'Brillante' is recognized very well too: the system switches on the correct state just in a couple of seconds. The 'Cupo' is recognized almost istantaneously and correctly identified over the whole performance with the exception of a frame between samples [180-210] where it's misunderstood for being 'Brillante'. In those couple of bars actually we have some repeated 8th notes that were performed with the same kind of bowing between the different performances and this succeeded in "confusing" the system for a short while (it's interesting to note also that in fig.3, at that point, the 'Brillante' and 'Cupo' probabilities get close to each other, although in that case the correct 'Brillante' state is mantained throughout the whole passage), anyway, during the cadenza ending on the dominant, which ends the played excerpt, the performance is again recognized correctly.

In conclusion, the system displays good results and seems to be able to recognize the particular expressive intention, in this particular piece, despite the rather simplistic tools used.

Surely further research is needed to refine the overall detection and recognition approach so as to make the system more precise and general. Once this goal is achieved, a system like this could be useful in developing new computer teaching tools addressed to amateur players or conservatory level students.

For example, once the system is trained on particular expressive intentions, it may provide valuable feedback on the student general understanding of the piece and whether he/she is able to correctly convey a desired expressive intention or not.

ACKNOWLEDGEMENTS

This research was funded by the EU-IST Project MEGA (Multisensory Expressive Gesture Applications) no. IST-1999-20410 (www.megaproject.org).

I'd like to thank Prof. Johan Sundberg and Dr. Anders Friberg for providing many valuable suggestions during all the stages of the development of this experiment, Dr.Paolo Coletta for debugging the EyesWeb libraries, the InfoMus Lab staff for several suggestions and testing and maestro Fabrizio Ferrari for his kindness and help.

REFERENCES

[1] Juslin P.: "Cue utilization in communication of emotion in music performance: relating performance to perception", Journal of Experimental Psychology: Human perception and performance, 26 (6),2000, pp.1797-1813

[2] Sundberg J., Friberg A., Fryden L.: "Common secrets of musicians and listeners: an analysis-by-synthesis study of musical performance", in Representing Musical structure, Academic Press., 1991

[3] Camurri A., Hashimoto S., Ricchetti M., Suzuki K., Trocca R., Volpe G.: "EyesWeb – Toward gesture and affect recognition in dance/music interactive systems" Computer Music Journal, 24 (1), MIT, 2000

[4] Friberg A., Schoonderwaldt E., Juslin P. and Bresin R.: "Automatic Real-Time Extraction of Musical Expression", Proceedings ICMC 2002, pp.365-367, Goteborg 2002

[5] Dillon R.: "Extracting audio cues in real time to understand musical expressiveness", Proceedings "Current research directions in computer music", MOSART Workshop, pp.41-44, Barcelona 2001

[6] Todd N.: "Music and Motion: a personal view", Proceedings 4th Workshop on Rhythm Perception, Bourges, 1992

[7] Papoulis A.: "Probability, Random Variables and Stochastic Processes", McGrawHill, 3rd Edition, 1991, pp.214-221

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

MELODIC CHARACTERIZATION OF MONOPHONIC RECORDINGS FOR EXPRESSIVE TEMPO TRANSFORMATIONS

Emilia Gómez[1], *Maarten Grachten*[2], *Xavier Amatriain*[1], *Josep Lluís Arcos*[2]

[1] Music Technology Group - Universitat Pompeu Fabra
Ocata, 1. 08003 Barcelona SPAIN
http://www.iua.upf.es/mtg
{emilia.gomez,xavier.amatriain}@iua.upf.es
[2] IIIA-CSIC-Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research
Campus UAB. 08193 Bellaterra, Barcelona SPAIN
http://www.iiia.csic.es
{maarten,arcos}@iiia.csic.es

ABSTRACT

The work described in this paper aims at developing a system that could perform expressive tempo transformations in monophonic instrument phrases. We have first developed a melodic description subsystem that extracts a set of acoustic features from monophonic recordings. This set of features is structured and stored following a description scheme that is derived from the current MPEG-7 standard. These performance descriptions are then compared with their corresponding scores, using *edit distance* techniques, for automatically annotating the expressive transformations performed by the musician. Then, these annotated performance descriptions are incorporated in a case-based reasoning (CBR) subsystem in order to build an expressive tempo transformations case base. The transformation subsystem will use this CBR system to perform tempo transformations in an expressive manner. Saxophone performances of jazz standards played by a professional performer have been recorded for this study.

In this paper, we first describe which are the melodic features that have been extracted and how they are structured and stored. Then, we explain the analysis methods that have been implemented to extract this set of features from audio signals and how they are processed by the CBR subsystem.

1. INTRODUCTION

As shown by Desain and Honing [1], the way a performer changes tempo in an expressive manner is not equivalent to performing a time-scale modification of the piece recorded in its nominal tempo. This is the fact that we address in this work. We want to study the deviations and changes that the performer introduces when changing the tempo of a piece. The study of these variations will be used as guidelines to perform tempo transformations in a expressive manner in the context of a content-based transformation system. Figure 1 represents the basic block diagram of the system. The audio processing part of the system consists of two different subsystems: the melodic description subsystem, that it is explained in section 2, and the transformation block, that is based in Spectral Modeling Analysis and Synthesis. The CBR part of the system basically consists of two subsystems: first there is the pre-processing part that consists of performance annotation, and musical processing (see subsection 3.1). In this part the cases are constructed. The second subsystem is the core reasoning procedure. It is explained in more detail insections 3.2, 3.3, and 3.4.

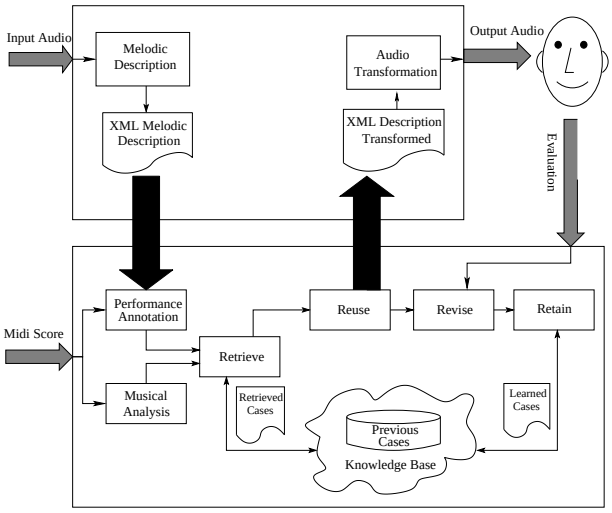


Figure 1: System architecture

2. THE MELODIC DESCRIPTION SUBSYSTEM

The melodic description block extracts melodic descriptions from monophonic recordings and stores them in a structured format. In this section we explain how the acoustic features are computed from audio and stored.

2.1. Description scheme

We define three structural levels of description: the analysis frame, the note and the global level. According to these structural levels,

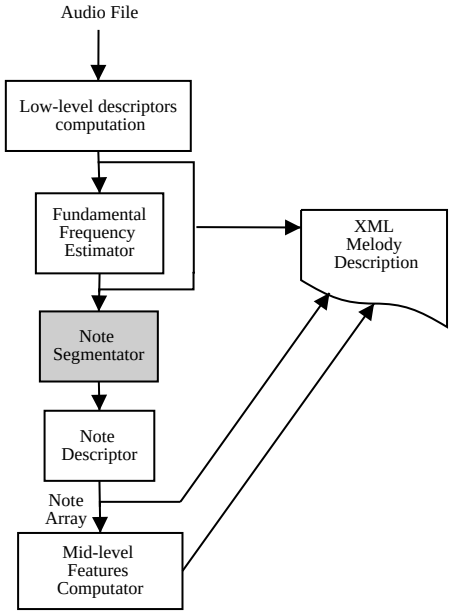


Figure 2: Block diagram of the melody descriptor

we define instantaneous descriptors, i.e. associated to an analysis frame, note descriptors attached to a note segment and global descriptors attached to the whole monophonic excerpt. All these descriptors are stored into a XML document. A detailed explanation about the description scheme and the MPEG-7 standard can be found in [2, 3].

2.2. Methods to extract the descriptors

Figure 2 represents the steps that are performed to obtain a melodic description from audio. First, the audio signal is divided into analysis frames, and the set of low-level descriptors are computed for each analysis frame. These low-level descriptors are used by the note segmentation algorithm, as well as in a preprocessing step of the fundamental frequency algorithm.

The fundamental frequency detector outputs an estimates for each analysis frame. Using these values and the low-level descriptors, the note segmentation block detects the note boundaries. Once the note boundaries are known, the note descriptors are computed from the low-level and the fundamental frequency values. Finally, note and low-level descriptors are combined to compute the global descriptors, associated to the whole audio segment.

2.2.1. Fundamental frequency estimation

For the estimation of the instantaneous fundamental frequency we use a harmonic matching model derived from the one proposed by Maher and Beauchamp [6], the Two-Way Mismatch procedure (TWM).

First of all, we perform a spectral analysis of a portion of sound, called analysis frame. Secondly, the prominent spectral peaks of the spectrum are detected from the spectrum magnitude. These spectral peaks of the spectrum are defined as the local maxima of the spectrum which magnitude is greater than a threshold. These spectral peaks are compared to a harmonic series and an TWM error is computed for each fundamental frequency candidates. The candidate with the minimum error is chosen to be the fundamental frequency estimate. After a first test of this implementation, some improvements to the original algorithm where implemented to deal with some errors of the algorithm:

- Peak selection: a peak selection routine has been added in order to eliminate spectral peaks corresponding to noise. The peak selection is done according to a masking threshold around each of the maximum magnitude peaks. The form of the masking threshold depends on the peak amplitude, and uses there different slopes depending on the frequency distance to the peak frequency.
- Context awareness: we take into account previous values of the fundamental frequency estimation and instrument dependencies to obtain a more adapted result.
- Noise gate: a noise gate based on some low-level signal descriptor is applied to detect silences, so that the estimation is only performed in non-silences segments of the sound.

2.2.2. Note segmentation

Note segmentation is performed using a set of frame descriptors, which are energy computation in different frequency bands and fundamental frequency. Energy onsets are first detected following a band-wise algorithm that uses some psycho-acoustical knowledge [5]. In a second step, fundamental frequency transitions are also detected. Finally, both results are merged to find the note boundaries (onset and offset information).

2.2.3. Note descriptors

We compute note descriptors using the note boundaries and the low-level descriptors values. The low-level descriptors associated to a note segment (as e.g. energy, centroid, spectral flatness, etc) are computed by averaging the frame values within this note segment. Pitch histograms have been used to compute the pitch note and the fundamental frequency that represents each note segment, as found in [7]. This is done to avoid taking into account mistaken frames in the fundamental frequency mean computation.

3. THE CASE BASED REASONING SUBSYSTEM

To apply musically meaningful transformations to the melodic descriptions obtained by the melodic analysis, it is of importance to know what musical material was actually performed and how the performance relates to this material. This information can be extracted from the score and the melodic descriptions of the performance. We constructed a CBR subsystem that uses this information to construct cases. These cases serve as ‘experience’ in the system. When a tempo transformation must be applied to a given musical performance, relevant cases are retrieved from the case base and using the retrieved cases a new performance is constructed at the desired tempo.

3.1. Construction of the cases

In the current system, a case consists of basically two kinds of information:

- A musical analysis of the score.
- An annotation of the performance that relates the performance to the score.

The musical analysis of the score is provided as a criterion for assessing similarities between scores and to split the phrases in smaller segments (see subsection 3.2). As an analysis-scheme, we use Narmour's Implication/Realization (I/R) model [9]. This is a model of melodic structure, based on principles akin to Gestalt Theory. An I/R analysis consists of a grouping of notes and categorizing these groups into a set of predefined categories. We have developed a parser for melodies that automatically generates I/R analyses. It implements most of the basic ideas from the I/R model.

The annotation of the performance is a reconstruction of how the musician actually performed the score. This annotation contains information such as which timing deviations were played, which score notes were not performed and which notes in the performance were not present in the score. Such an annotation can be conveniently obtained by mapping a melody description to the corresponding score. We use edit-distance techniques as described in [8] to generate mappings between the score and the melody description.

3.2. Retrieval of stored cases

The retrieval mechanism is organized in three phases:

3.2.1. Phrase retrieval

In a first phase the input melody (the score) is compared with the melodies of the case base using melodic similarity measures for retrieving only those case melodies really similar—For instance, given a slow ballad as input, we are not interested in comparing it with be-bop themes. To this end, similarities are computed between the input melody and each of the melodies in the case base. The similarities can be computed either using the note representations of the melodies, or using other representations such as melodic contour representations, or the musical models derived from the melodies. Combinations of these different measures can also be used to retrieve a subset of melodies. In [4], a comparison of various similarity measures is reported. It turns out that a similarity measure based on note level representations is not very discriminative when applied to melodies that are very different (i.e. all phrases that are not virtually identical, are assessed more or less equally dissimilar). Therefore, it seems more promising to use similarities between musical models or contour representations for retrieving similar melodies from the case base. The final output of this phase is a subset of a melodies of the case base close to the input melody. Only performances from these retrieved melodies will be taken into account in the following phases.

3.2.2. Motif retrieval

In a second phase, we try to find similar melodic fragments for segments of the input melody. The input melody is segmented based on the musical model that was constructed for the melody (including musical information such as the metrical strengths of the notes). In particular, I/R structures (or sequences of two or three structures) usually coincide with melodic motifs. For each of these segments, the most similar parts of the retrieved melodies can be selected (again using a similarity measure on either of the melodic representations). The result is that for each segment/motif of the input melody a set of melodic fragments is available, each fragment with one or more performance annotations.

3.2.3. Performance ranking

Finally, in the third phase the performances that were retrieved for each segment of the input melody are ranked using a similarity measure for the performance annotations. The idea behind this step is to use the input performance as a guide for how the input melody should be performed at the desired tempo. For illustration, say that an input performance P_{in} of a melodic segment M at tempo T_{in} was given and a new performance P_{out} of M at the desired tempo T_{out} must be generated. Suppose that there is a retrieved melodic fragment that has a performance P_1 at tempo T_i close to T_{in} and in addition, it has a performance P_2 at T_j close to the desired tempo T_{out}. Then, if performance P_{in} is similar to P_1, we assume that P_2 is a good basis for constructing performance P_{out} at tempo T_{out}. In this way, performances at tempos close to the desired tempo can be selected and ordered according to their expected relevance for constructing the solution.

In conclusion, the output of the retrieval step is an ordered collection of candidate annotations for each segment/motif in the input melody.

3.3. Reuse of retrieved cases

The reuse (or adaptation) mechanism is being implemented using *constructive adaptation* [10], a generative technique for reuse in CBR systems. The reuse mechanism deals with two kinds of criteria: local criteria and coherence criteria. Local criteria deal with the transformations to be performed to each note—i.e. how retrieved candidate annotations can be reused in each input note. Coherence criteria try to balance smoothness and hardness. Smoothness and hardness are basically contradictory: the first tends to iron out strong deviations with respect to the score, while the second tends to favor strong deviations. The resulting expressive performance is a compromise between the smoothness and hardness criteria, with the aim of keeping an overall balance pleasant to the ear.

3.4. Revising and Retaining cases

In the revise step, the user has the opportunity to either confirm the solution proposed by the system, discard it as being a bad solution, or manipulate it to correct local imperfections. When the result is satisfying to the user it is stored in the case base for solving future problems (the retain step).

4. AUDIO MATERIAL

For this experiment, we set up an audio database consisting of 5 jazz standards played at 11 different tempos around the nominal one, played by a professional musician. Most of the phrases where repeated to test consistency between performances. The jazz standards recorded where chosen to be of different moods: *Body and Soul, Once I Loved, Donna Lee, Like Someone In Love* and *Up Jumped Spring*.

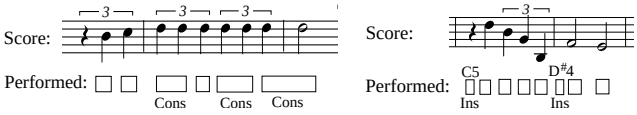


Figure 3: Two examples of the performer's interpretation of the score. *Cons* refers to a consolidation of notes, and *Ins* to a note insertion.

5. CASE-STUDY

We are currently in the process of analyzing the audio material, and constructing the CBR system. Nevertheless, in this section we will present some of the data we obtained by analyzing the audio recordings. First, we show two cases where the performer deviated from the score by inserting or consolidating (i.e. playing two or more notes as one) notes. On the left, figure 3 shows a fragment of *Once I Loved*, performed at 80 bpm. The blocks underneath the notes show what the performer played. It can be seen that the repetitions of the *D* pitch were not played literally, but instead some notes were played as one (this interpretation is preferred over the interpretation that some notes were *deleted*, since the rests caused by the omitted notes were filled up by an adjacent note). The example on the right shows a fragment of *Body and Soul*, also performed at 80 bpm. Here, the performer introduced two extra, ornamental notes. The inserted notes were of very short duration and the pitches of the notes were a semitone and whole tone below their successor notes, respectively.

Secondly, figure 4 shows the note onset deviations for performances of *Body and Soul* at different tempos. The y-axis shows the difference between performance onset and score onset, where negative values correspond to early performance onsets. It can be seen that the deviations are not constant under tempo changes. Nevertheless, there appear to be regularities. For instance, for the faster tempos, the beginning of the phrase was played with rather late timing, whereas for slower tempos the timing was neutral or slightly early.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have explained the set up of a system to automatically construct a knowledge base for expressive tempo transformations and use this knowledge base to perform expressive tempo changes to audio material. We also presented the results of the analysis and annotation of some audio recordings. This revealed that performances contained occasional deletions and insertions of notes, and that timing deviations tend to vary with tempo. Given the fact that the performances under consideration were rather neutrally played, this implies that a tempo transformation system for arbitrary jazz performances (which are often rather liberal interpretations of the score) cannot do without an analysis framework to interpret the performances. We intend to continue working on the analysis of the audio recordings to construct an initial case base. Moreover, the CBR system itself needs more elaboration and specification, in particular the Reuse step. Once the transformation system is fully operational, a user listening test could be set up in order to validate the transformation results.

Acknowledgments This work has been partially funded by the European IST Project CUIDADO (IST-1999-20194) and the Spanish TIC project TABASCO (TIC 2000-1094-C02).

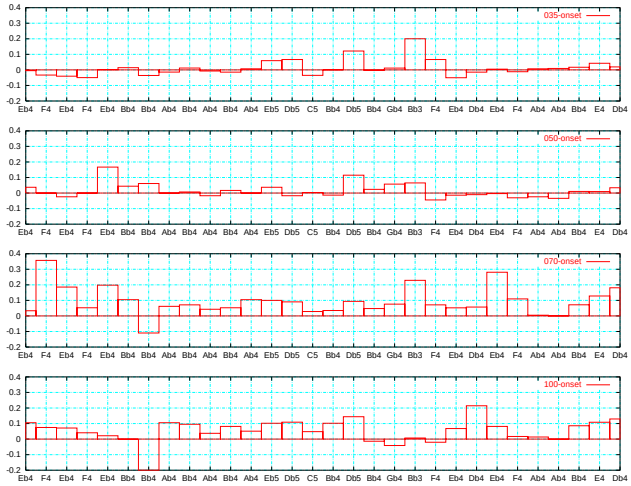


Figure 4: Onset deviations of *Body and Soul* for tempos 35, 50, 70 and 100 bpm.

7. REFERENCES

[1] Peter Desain and Henkjan Honing, *Tempo curves considered harmful. a critical review of the representation of timing in computer music*. Proceedings of the 1991 International Computer Music Conference, San Francisco, 1991.

[2] Emilia Gómez, Fabien Gouyon, Perfecto Herrera and Xavier Amatriain, *Using and enhancing the current MPEG-7 standard for a music content processing tool*, Proceedings of the 114th Audio Engineering Society Convention, Amsterdam, The Netherlands, March 2003.

[3] Emilia Gómez, Fabien Gouyon, Perfecto Herrera and Xavier Amatriain, *MPEG-7 for content-based music processing*, Proceedings of the 4th WIAMIS Special Session on Audio Segmentation and Digital Music, London UK, April 2003.

[4] Maarten Grachten, Josep Lluís Arcos, and Ramon López de Mántaras, *A comparison of different approaches to melodic similarity*. IInd International Conference on Music and Artificial Intelligence, 2002.

[5] Anssi Klapuri, *Sound Onset Detection by Applying Psychoacoustic Knowledge*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 1999.

[6] Robert C. Maher and James W. Beauchamp, *Fundamental frequency estimation of musical signals using a two-way mismatch procedure*, Journal of the Acoustic Society of America, vol. 95 pp. 2254-2263, 1994.

[7] Roger J. McNab, Lliyd A. Smith and Ian H. Witten, *Signal Processing for Melody Transcription*,SIG working paper, vol. 95-22, 1996.

[8] Marcel Mongeau and David Sankoff, *Comparison of musical sequences*, Computers and the Humanities, 24:161–175, 1990.

[9] Eugene Narmour, *The Analysis and cognition of basic melodic structures : the implication-realization model*, University of Chicago Press, 1990.

[10] Enric Plaza and Josep Lluís Arcos, *Constructive adaptation*, In Susan Craw and Alun Preece, editors, Advances in Case-Based Reasoning, number 2416 in Lecture Notes in Artificial Intelligence, pages 306–320. Springer-Verlag, 2002.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

MEASUREMENTS AND MODELS OF MUSICAL ARTICULATION

Jesper Jerkert

Department of Speech Music and Hearing
Kungl Tekniska Högskolan, Stockholm
jesperj@speech.kth.se

ABSTRACT

In four Bach organ fugue openings, each of which was played by three or four organists, the mean articulation was found to be 0.84, with an average standard deviation of 0.18. Articulation for tone n was taken to be $(\text{offset}_n - \text{onset}_n)/\text{IOI}_n$, where the inter-onset interval $\text{IOI}_n = (\text{onset}_{n+1} - \text{onset}_n)$. Articulation consistency, as measured by the mean correlation coefficient between performers, was almost as high as for the relative IOI (0.46 vs. 0.52). The mean correlation between articulation and relative IOI was –0.32. This means that in the studied material lengthening of relative IOI typically coincides with shorter (more *staccato*-like) articulation. In one more carefully studied fugue three out of four performers used comprehensible articulation strategies which were distinguishable from mere relative IOI adjustments.

1. INTRODUCTION

"Articulation" is a word used for many purposes. In musicology the term is sometimes attributed a very wide scope. In a much-read book on style analysis [1], for example, every change, from one symphony movement to the next, as well as from one tone to the next, is said to constitute "articulation". However, articulation is normally associated with small dimensions, whereas phrasing acts on a larger scale (though admittedly not as large as to include whole symphony movements). In this paper we shall consider strategies for making adjacent tones belong together or for making them seem separated. If the tones are tied together, so that there is no silence between the notes, we speak of a *legato*. If the tones are clearly separated, we speak of *staccato*. But this description is not the final word. Tones might be joined or separated by other means than shortening or lengthening of the intervening rest. For example, tones can be played longer or shorter than the nominal duration. Furthermore, the adjacent tones can be played with different sound levels, or with different vibrato, or with different tone attacks, etc.

Here I will only deal with methods involving the timing of the tones. There are two such methods. The first is to alter the relative length of the rest, i.e. what is normally called *legato* vs. *staccato*. In this paper I will call this articulation. If one note ends exactly when the following note starts, the articulation (or articulation ratio) is 1 (one). If one note lasts for only half the time between its onset and the onset of the next tone, the articulation is 0.5. In other words the articulation for tone n is $(\text{offset}_n - \text{onset}_n)/\text{IOI}_n$, where IOI is the *interonset interval* (see e.g. [2]), defined as the time interval between the onset of a given tone and the onset of the immediately following tone, i.e. $\text{IOI}_n = \text{onset}_{n+1} - \text{onset}_n$.

The second method is to play the tones longer or shorter than their nominal duration, i.e. the tempo is altered locally. This can be expressed in terms of relative IOI. A relative IOI of 1 (one) means that the duration of the tone is exactly what is to be expected from the nominal (notated) length given the mean tempo of the piece. For example, if a piece is performed at the mean tempo 120 quarter-notes per minute, we would expect an eighth-note to last for 0.25 s. A certain eighth-note with a duration of 0.30 s then has the relative IOI 0.30/0.25 = 1.2.

I will report some basic findings on articulation, as defined above. In particular I will investigate the following questions: What is "normal" articulation in one-part music lacking notated articulation marks, i.e. what is its mean and standard deviation? How large differences are there between performers in the same piece? Is there any detectable relationship between articulation and relative IOI? If so, how could it be explained? It seems that no one has investigated these questions before, at least not in a quantitative manner. There are some studies on how to play a predetermined articulation type. For example, [3] reports that a typical staccato in the pianoforte has an articulation ratio of approximately 0.4. However, there are no reports on the articulation of music without notated articulation marks. Finally I will comment on the presence or absence of specific articulation strategies in one of the pieces analyzed.

BWV	Performer	Recording
537	Hans Fagius	BIS 397/398
537	Peter Hurford	Decca 443485-2
537	Wolfgang Rübsam	Naxos 8553150
542	Helmuth Rilling	Denon 38C37-7039
542	Peter Hurford	Decca 443485-2
542	Ton Koopman	Hänssler 98182
542	Hans Fagius	BIS 308/309
564	Peter Hurford	Decca 443485-2
564	Ton Koopman	Archiv 410999-2
564	Wolfgang Rübsam	Naxos 8553859
564	Hans Fagius	BIS 343/344
578	Helmuth Rilling	Denon 38C37-7039
578	Wolfgang Rübsam	Naxos 8553859
578	Hans Fagius	BIS 439/440

Table 1: *The Recordings used for analysis.*

2. METHOD

The material used for analysis was commercially available recordings of Bach organ fugue theme openings. By "opening" I mean from the beginning of the fugue to the entrance of the second voice (*comes*) or even shorter; the material is thus monophonic. Some of Bach's fugue themes are rhythmically very monotonous (e.g. the fugue of the famous BWV 565) and so have not been considered. The analyzed recordings comprise

four different fugue openings, played by three or four performers each, see Table 1 for details. Notes followed by pauses have been excluded from analysis (this occurs only in BWV 564). Musical notes of the fugue themes are presented in Figure 1. These notes were redrawn from [4].
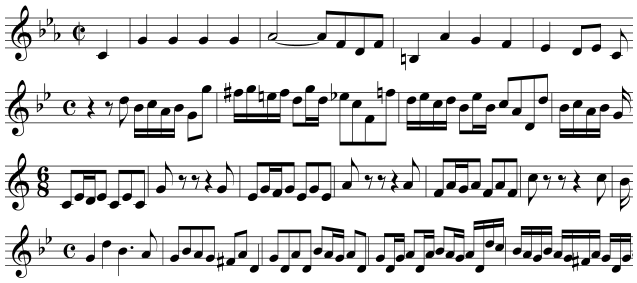


Figure 1: *The four analyzed fugue openings. From the top and down: BWV 537, BWV 542, BWV 564, BWV 578.*

BWV	Performer	Mean art. ± SD
537	Fagius	0.93 ± 0.13
537	Hurford	0.89 ± 0.18
537	Rübsam	0.90 ± 0.13
542	Rilling	0.93 ± 0.23
542	Hurford	0.96 ± 0.15
542	Koopman	0.70 ± 0.21
542	Fagius	0.83 ± 0.14
564	Hurford	0.66 ± 0.27
564	Koopman	0.65 ± 0.23
564	Rübsam	0.86 ± 0.18
564	Fagius	0.73 ± 0.23
578	Rilling	0.98 ± 0.13
578	Rübsam	0.88 ± 0.16
578	Fagius	0.91 ± 0.11
	Overall mean	0.84 ± 0.18

Table 2: *Mean articulations and standard deviations.*

The reason for analyzing organ recordings rather than, say, violin recordings is that the articulation is straight-forward in the organ; it is only a matter of measuring the onset and offset of each tone. Furthermore, the organ tone is steady, being (very nearly) equally loud from onset to offset. There are some possibly complicating factors in that there is no standard organ. For example, organs might be mechanical or electro-pneumatic. Different organs do not respond in the same way to a given touch. These things have been ignored in this paper. The rationale for this is the assumption that a skilful organist can produce a desired articulation with almost any organ or at least with the organs of the analyzed recordings.

The onsets and offsets of each tone were estimated through spectrogram inspections. The programs WaveSurfer 1.4.5 [5] and Soundswell 4.0 [6] were used to supply spectrograms.

3. RESULTS

3.1. Overall articulation

The mean articulation, for all pieces and all performances, was 0.84, with an average standard deviation of about 0.18. The results for each performance are presented in Table 2. As can be seen, there are some differences between pieces, which is quite expected because different pieces seem to invite the performers to articulate differently. A rough measure of the articulation consistency is the mean value of all pairwise correlations between performers. For example, there are three performer combinations for BWV 537: Fagius-Hurford, Fagius-Rübsam, and Hurford-Rübsam. This mean value, for all four fugues (18 performer pairs), was 0.46 (appr. 95% CI: 0.32-0.60).

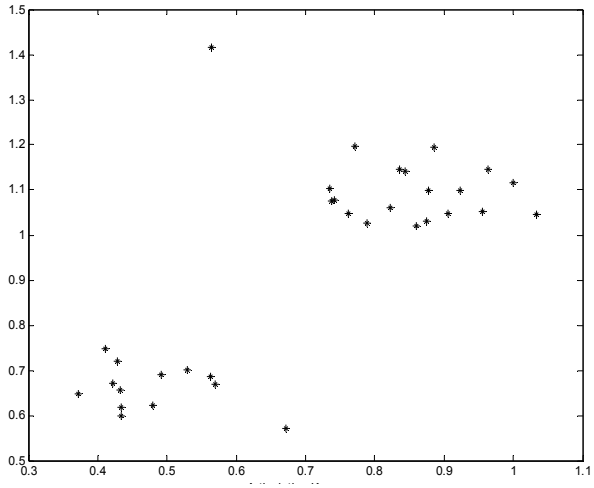


Figure 2: *Articulation in BWV 542 for Koopman vs. Rilling. Koopman is on the x-axis, Rilling on the y-axis. The correlation coefficient r is 0.760. If the most obvious outlier is removed, r will be 0.873.*

3.2. Differences between performers

As can be seen In Table 2, there are some substantial differences in the mean articulation between performers in the same piece. For example, in BWV 542 Koopman has a mean articulation of 0.70, whereas Rilling has 0.93. Still, their articulation profiles are in reality very similar, as shown in Figure 2. They have both adopted a "binary" type of articulation; either short (for Koopman 0.37–0.57, for Rilling 0.57–0.75) or long (for Koopman 0.73–1.04, for Rilling 1.02–1.20). Since they have chosen the same type of articulation for all tones, with only one or two exceptions, the points in Figure 2 are gathered in two distinct groups.

Plottings of other performer combinations in BWV 542 (Rilling-Hurford, Hurford-Fagius, etc.) are not as clearcut as Figure 2. The combination Koopman-Rilling is thus an exceptional case for BWV 542, but there are some other striking articulation similarities in the other fugues, so Figure 2 is not unique in the total set of data. However, I will return to all BWV 542 performances under "Discussion".

3.3. Articulation and relative IOI

Articulation and adjustment of relative IOI are the two temporal strategies for joining or separating adjacent tones. The relative IOI consistency, measured as the mean correlation between performers, was 0.52 (appr. 95% CI: 0.42-0.63). No correlation was negative (out of 18 combinations). The value is somewhat higher than the mean correlation for articulation between performers. The difference is too small to be statistically significant, but if it is real I believe it can be explained by the relative IOI being not only an articulation strategy, but most of all a phrasing strategy. Relative IOI adjustments act on a larger scale, making them more robust.

What relationship can be expected between articulation and relative IOI? A negative correlation might seem obvious: if a tone is given an absolute duration, then prolonging its IOI would mean adding time to the following rest, thereby reducing its articulation. This would mean that longer IOI is more likely to be associated with shorter (more *staccato*-like) articulation. But *a priori*, one could as well argue that the opposite could be true: perhaps shorter tones must be articulated more clearly (i.e. more *staccato*) to be heard in their own right, whereas longer tones do not need such clarification. In this study, the mean correlation between articulation and relative IOI for all performances was negative, –0.32 (appr. 95% CI –0.47 - –0.17). All correlations were negative except one (out of 14).

4. DISCUSSION: ARTICULATION STRATEGIES IN BWV 542

Why are there agreements between some of the interpretations (as in Figure 2)? What is the code behind the similar articulation strategies? Let me underline that not all performers articulated similarly in the same piece. Still, if there are similarities, there is hope of finding underlying principles. Since the space is limited I will here comment only on the articulation of the BWV 542 performances.
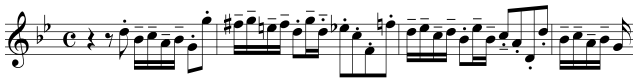


Figure 3: *Articulation in BWV 542 for Koopman and Rilling. Staccato dots indicate articulations of 0.37–0.57 for Koopman, 0.57–0.75 for Rilling. Portato strokes indicate articulations of 0.73–1.04 for Koopman, 1.02–1.20 for Rilling.*

When checking which tones correspond to which points for Koopman and Rilling (Figure 2), the result is simple. All long notes (eighth-notes) are short, and all short notes (sixteenth-notes) are long, i.e. all long notes are found in the lower left group of Figure 2, and all short notes are found in the upper right group. In Figure 3 the notes have been marked accordingly, with *staccato* dots or *portato* strokes. Two exceptional cases (the outliers in Figure 2) have been marked with both dots and strokes.

Turning to Hurford, it is not as easy to find distinct articulation types. Almost all articulation ratios are between 0.86 and 1.15. Four notes have articulation values below 0.76. These four notes are marked with staccato dots in Figure 4. They are not randomly scattered but are the last two notes in a recurrent group of four. All other notes have articulation values of 0.86 or more.
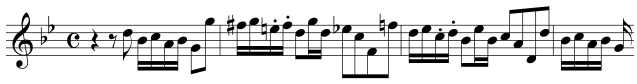


Figure 4. *Articulation in BWV 542 for Hurford. Staccato dots indicate articulations of less than 0.76. All other notes have articulations of 0.86 or more.*

In Fagius, it is even harder to distinguish clear articulation types. In Figure 5 three different articulation types have been indicated, but the boundaries between them are quite arbitrary. As can be seen, there is a tendency of shortening notes immediately preceding the beats, but this seems to be fully explained by corresponding lengthenings of the relative IOI's. In this way Fagius has emphasized the metrical structure, but he has not used any specific articulation strategy separable from simply adjusting the relative IOI's.
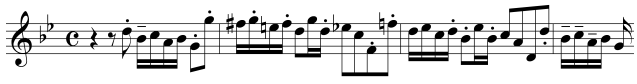


Figure 5. *Articulation in BWV 542 for Fagius. Staccato dots indicate articulations of 0.56–0.79, portato strokes indicate articulations of 1.03–1.05, neither dots nor strokes indicate articulations of 0.81–0.98.*

In conclusion, Koopman and Rilling adopted a simple and almost perfectly consistent strategy: long notes were played short and short notes were played long. Hurford played most notes quite long, but showed a tendency of shortening the last two sixteenth-notes in the groups of four sixteenth-notes (though he did it in only two out of four groups). This pattern – binding the first two notes and separating the last two when encountering groups of four short notes – is well-known in classical Western music, not only in the organ (see e.g. [7], pp. 95f, 107). Fagius, finally, exhibited no obvious articulation pattern. Figure 6 shows the complete series of relative IOI and articulation in BWV 542.

One should remember that the search for articulation strategies in BWV 542 was done in a typical post hoc fashion; i.e. there were no prior hypotheses on what strategies to look for. Nevertheless, the strategies found in Koopman, Rilling, and Hurford were so simple and familiar that I consider them to be real.

The point here in analyzing the four performances of the BWV 542 fugue opening is not to establish exactly what articulation strategies were utilized, but to find out whether such strategies were used at all. In a forthcoming, more extensive study I will detail the articulation strategies in all performances, not only in BWV 542. As for now, suffice it to say that three out of four performances of the BWV 542 opening showed some comprehensible articulation strategy. The presence of articulation strategies is consistent with the regularities often found in other interpretation aspects, such as phrasing.

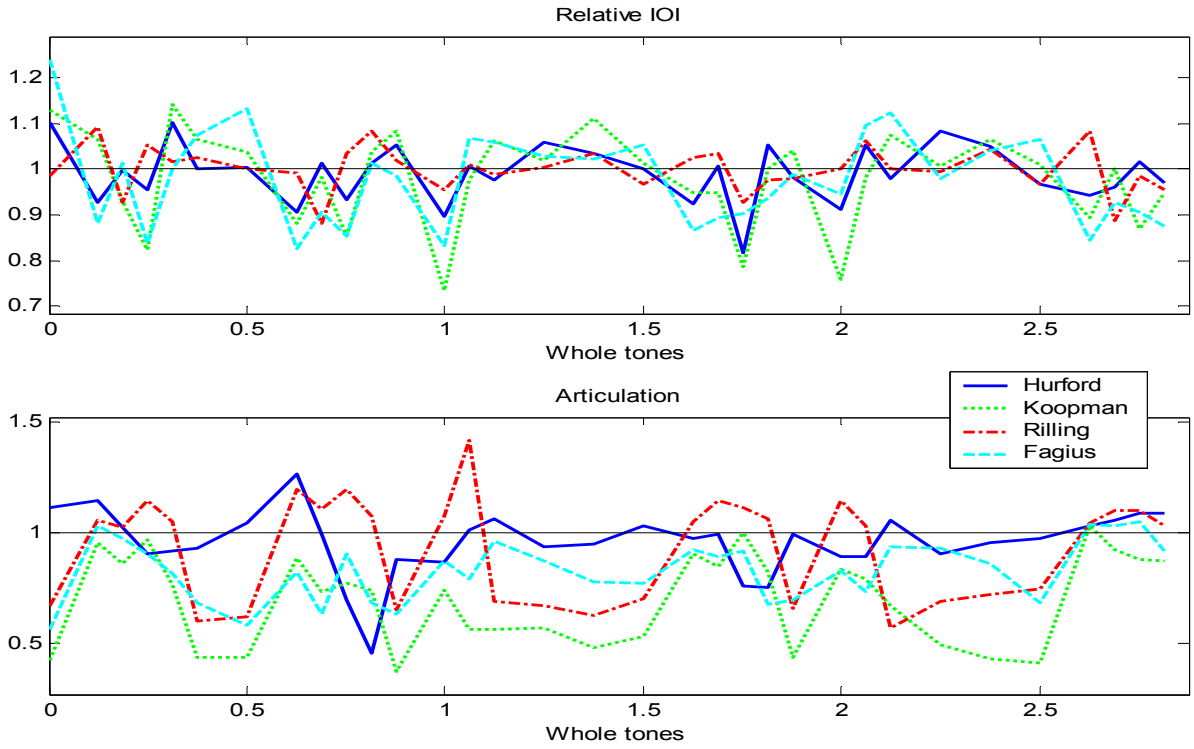


Figure 6: *Relative IOI (top) and articulation (bottom) for all four performances of BWV 542.*

5. CONCLUSIONS

In four Bach organ fugue openings, each of which was played by three or four organists, the mean articulation was found to be 0.84, with an average standard deviation of 0.18. Note that this value needs not be valid for all sorts of instruments or other kinds of music.

Articulation consistency, as measured by the mean correlation coefficient between performers, was almost as high as for the relative IOI (0.46 vs. 0.52). The mean correlation between articulation and relative IOI was –0.32. Thus lengthening of relative IOI typically corresponded to shorter (more *staccato*-like) articulation. In one of the fugues, BWV 542, three out of four performers used an easily comprehensible articulation strategy.

6. REFERENCES

[1] LaRue, J., Guidelines for Style Analysis, 2nd ed., Harmonie Park Press, Pinewood, Michigan, 1992.

[2] Friberg, A. and Battel, G.U. "Structural Communication", in: R. Parncutt & G. E. McPherson (eds.), The Science and Psychology of Music Performance: Creative Strategies for Teaching and Learning, Oxford University Press, New York, pp. 199–218, 2002.

[3] Bresin, R. and Battel, G.U. "Articulation strategies in expressive piano performance", J. New Mus. Res. 29(3), 211–224, 2000.

[4] Keller, H., Die musikalische Artikulation insbesondere bei Joh. Seb. Bach., C. L. Schultheiß Musikverlag, Stuttgart, 1925.

[5] Sjölander, K. and Beskow, J. WaveSurfer. Available at http://www.speech.kth.se/wavesurfer/. 2002.

[6] Hitech Development. Soundswell. Information available at http://www.hitech.se/development/. 2002.

[7] Keller, H., Phrasing and Articulation, translated by L. Gerdine, Barrie and Rockliff, London, 1966.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

THE PERCUSSION INSTRUMENT OPERATION SYNCHRONIZED WITH THE VISUAL AND/OR AUDITORY STIMULI

Tatsuo Kamitani[*1], *Kagemasa Kouzuki*[*2], *Minoru Matsuda*[*3]

Osaka Electro-Communication University Joint Institute for Advanced Multimedia Studies[*1]
Konami Corporation[*2]
Graduate school of Engineering, Osaka Electro-Communication University[*3]

kami@dmic.org

ABSTRACT

In recent years, the video game machines, imitated the actual musical instrument, have been developed. By using this kind of machine, those who are inexperienced in a musical instrument can also enjoy a percussion instrument by synchronizing the video images and the sounds. However the human recognition mechanism induced by the visual and/or auditory stimulus is not clear. Then the accuracy of human response time induces by visual and auditory stimuli is measured and discussed in this paper.

From the result of the experiment, it became clear that the accuracy of the percussion instrument tapping is improved when the specific visual stimuli are given. And the auditory stimuli did not influence on the accuracy of the percussion instrument tapping under the kind of visual stimulus.

1. INTRODUCTION

Advances in computer technology in recent years have resulted in the development of various forms of video game machines. In particular, sophistication of the man-machine interface has made it possible to develop game machines that simulate actual devices so as to give even novice players the feeling of operating that equipment. Although video game machines that simulate airplanes and other transportation vehicles are well known, game machines that simulate musical instruments and give the player the feeling of playing the instrument even if they cannot actually play it have also been developed in recent years.

The most well-known types of video game machines that simulate a musical instrument are those that simulate percussion instruments. A game machine that simulates a drum set (Fig. 1) and one that imitates Japanese taiko (Fig. 2) are well known in Japan. With either of these game machines, the player can listen to music and see instructions on the screen that are synchronized with the music to guide the player's decisions on playing the instruments. That makes it possible for even players that have not mastered percussion instruments to play the drums along with the music, at least to a certain degree, and thus enjoy the game. Because the points earned during the game depend on how accurately the player can hit the drums, it does not meant that following the instructions presented on the screen will make it possible to play with perfect accuracy. Nevertheless, even players that have no experience with percussion instruments at all can use this game machine fully, and many players become able to play the drums in synchronization with the musical performance in far less time than it takes to master the actual instruments.



Figure 1: Game machine 1.
The game center video game machine known as "Drum Mania" has a user interface that simulates a drum set. Rubber pads arranged like tom-toms, snare drum, cymbal and high hat and are struck by the player to accompany the music produced by the game machine. There is also a pedal to simulate the kick drum.



Figure 2: Game machine 2.
The game center video game called "Taiko Master" has a user interface that simulates the traditional Japanese drums called taiko. The player beats the taiko according to the music produced by the game machine and instructions that are displayed on the screen. The game can also be played by two persons.

When designing such instrument-simulating game machines, it is necessary to design the screens so that the percussion instruments can be easily played together with the music produced by the game machine. However, how people react when visual stimuli such as video and auditory stimuli such as music are presented at the same time is not fully understood. Current machine design therefore involves empirical adjustments in the design of the game parts that involve the effects of multimodal stimulation. Such design is a matter of expertise possessed by the game machine manufacturers. Therefore, the objective of this research was to present test subjects with visual and auditory stimuli and measure their response times so as to obtain fundamental data for use in the design of game machines and other such purposes.

2. EXPERIMENTS

2.1. Experimental arrangement

The experiments employed equipment that is a simplification of game machine 1. The user interface for game machine 1 uses drum pads made for commercial electronic instruments and are readily available. We were thus able to approximate the environment of game machine 1 with the experimental equipment.

The experimental equipment consists of a digital VTR (VTR), cathode-ray tube display (CRT), headphones (HP), a data recorder (DRD) and a drum pad (Fig. 3). The visual and auditory stimuli were pre-recorded on videotape and presented to the subjects by playing the tape on a 28-in CRT.

The visual stimulus was the video signal output from the VTR displayed on the CRT for presentation to the subject. The auditory stimulus was the audio signal output from the VTR converted to sound by the HP for presentation to the subject.

The subjects are presented with either a visual stimulus only, an auditory stimulus only, or both visual and auditory stimuli. They are instructed to strike the drum pad when indicated. When the drum pad is struck, it generates an audio signal. That signal and the auditory stimulus signal are recorded simultaneously by the DRD.

The test subjects were 18 healthy individuals (average age of 22.4±1.09 years old); 16 were male and 2 were female.

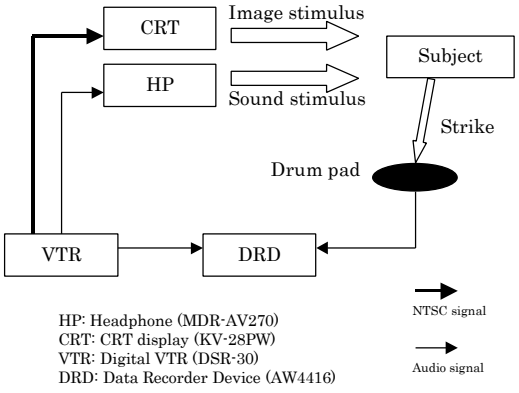


Figure 3: Experimental arrangement

2.2. Visual stimuli

The visual stimuli used in the experiments were video images created using computer graphic techniques (Fig. 4). The visual stimuli were presented to the subjects as video media on a CRT.

Denoting the visual stimulus presented to the subject as **G**, **G** can be considered as an ordered set of n still images, g_i.

$$\mathbf{G} = (g_i) \equiv (g_1, g_2, \ldots, g_n) \quad \text{where} \quad i \in \{1, 2, \ldots, n\} \tag{1}$$

Various kinds of visual stimuli can be obtained by changing the still images, g_i.

The visual stimulus, g_i, can be separated into a background image that does not change over time, h_0, and a series of images that does change over time, h_i. That is to say, denoting the image compositing operation by the symbol ' ∘ ', g_i can be represented as Equation (2).

$$g_i = h_0 \circ h_i \tag{2}$$

In the experiment, visual stimuli $\mathbf{G_0}$ to $\mathbf{G_2}$ were used. $\mathbf{G_0}$ is a blank visual stimulus in which nothing at all was shown on the CRT.

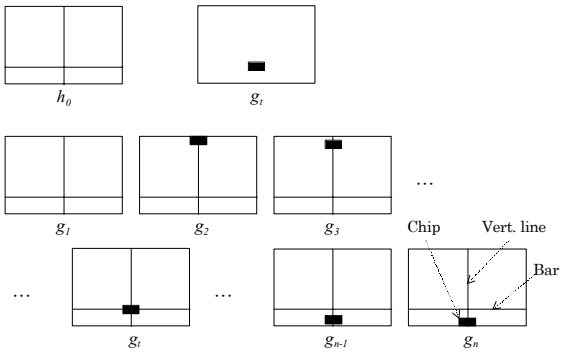


Figure 4: Visual stimulus

$\mathbf{G_1}$ is an image that suddenly appears to indicate that the drum is to be hit. It is presented in a rectangle (referred to as a chip in the rest of this paper) that is displayed within $h_i (2<i \leqq n)$ in the background image, h_0. At times other than when the drum is to be struck ($i = t$), only the background image is shown. $\mathbf{G_1}$ is expressed by Equation (3). Here, n is the number of images to be used for one indication to strike the drum.

$$\mathbf{G_1} = (g_i) \equiv (g_1, g_2, \ldots, g_n) \quad \text{where} \quad i \in \{1, 2, \ldots, n\} \tag{3}$$
$$g_i = \begin{cases} h_0 \circ h_i & \text{where } i = t \\ h_0 & \text{elsewhere} \end{cases}$$

In the experiments, each image of g_i is displayed every 1/29.97 [s]. Thus, $\mathbf{G_2}$ appears as a video image that moves smoothly from the top of the chip to the bottom of the chip above the background image. $\mathbf{G_2}$ is expressed by Equation (4).

$$\mathbf{G_2} = (g_i) \equiv (g_1, g_2, \ldots, g_n) \quad \text{where} \quad i \in \{1, 2, \ldots, n\} \tag{4}$$
$$g_i = (h_0 \circ h_i)$$

When the horizontal line of image h_0 (referred to as the bar in the rest of this paper) changes to a chip, it is an indication for the

test subject to hit the drum pad. Denoting the image for the time at which the subject is to hit the drum as g_t, g_t is expressed by Equation (4), because g_t is a composite of h_0 and h_t.

$$g_t = h_0 \circ h_t \tag{5}$$

That is to say, the subject is supposed to hit the drum pad when image g_t is seen. The values $n=56$ and $t=47$ were used in the experiments reported here.

2.3. Auditory stimuli

The auditory stimuli used in the experiments is sound generated by a decaying sine wave signal. The auditory stimulus is converted from the decaying sine wave signal into sound and presented to the subject via the headphones. When the drum pad used in the experiment is struck, it generates a sound that has a frequency near 400 [Hz]. Therefore, the frequency of the auditory stimulus signal was set to 400[Hz] to provide an auditory stimulus to the subject that is close to the sound that can be heard when the pad is struck.

The waveform observed when the pad is struck has a decay of approximately 32 [ms]. Furthermore, the display time for a single image is 1/29.97 [s], which is close to the duration of the drum pad waveform. Therefore, the duration of the auditory stimulus was set to the output time of one image, 1/29.97 [s]. The signal waveform is set to decay exponentially, and a time constant that produces a final waveform output level of -40 [dB] was used. For time constant τ, sampling frequency f_s, and auditory stimulus waveform sample count ν Equations (6) and (7) hold. The signal frequency, f, is 400 [Hz] and the sampling frequency f_s is 48 [kHz].

$$\nu = \mathrm{round}(\frac{f_s}{29.97}) = 1602 \tag{6}$$

$$\tau = \frac{\nu - 1}{\log(0.01)} \tag{7}$$

The authors denote silence as η_0 and the decaying sine wave as η_1. The auditory stimulus is a function in the time domain so these values are expressed by Equations (8) and (9).

$$\eta_0(x) = 0 \quad x \in \{0,1,2,\cdots,\nu-1\} \tag{8}$$

$$\eta_1(x) = A \cdot \exp(-\frac{x}{\tau}) \cdot \sin 2\pi \frac{f}{f_s} x \quad x \in \{0,1,2,\cdots,\nu-1\} \tag{9}$$

On the other hand, denoting the sound that is output every 1/29.97[s] in synchronization with the images g_i as s_i, the auditory stimulus **S** can be considered as an ordered set of n sounds, s_i.

$$\mathbf{S} = (s_i) \equiv (s_1, s_2, \ldots, s_n) \quad \text{where } i \in \{1,2,\ldots,n\} \tag{10}$$

Here, it is considering as the synchronization that sound s_i is emitted simultaneously with the output start of image g_i. The auditory stimuli for the experiments are generated by assigning either η_0 or η_1 to each s_i. The following auditory stimulus $\mathbf{S_0}$ and $\mathbf{S_2}$ were used in the experiments.

Auditory stimulus $\mathbf{S_0}$ (No generation of sound - silence)

$$\mathbf{S_0} = (s_i) \equiv (s_1, s_2, \ldots, s_n) \quad \text{where } i \in \{1,2,\ldots,n\} \quad s_i = \eta_0 \tag{11}$$

Auditory stimulus $\mathbf{S_1}$ (The time of the strike indication and the time of presentation of the auditory stimulus are the same.)

$$\mathbf{S_1} = (s_i) \equiv (s_1, s_2, \ldots, s_n) \quad \text{where } i \in \{1,2,\ldots,n\} \quad s_i = \begin{cases} \eta_1 & \text{where } i = t \\ \eta_0 & \text{elsewhere} \end{cases} \tag{12}$$

In the case that only the auditory stimulus is presented to the subject, auditory stimulus $\mathbf{S_1}$ is presented. Because the respective s_i is output every 1/29.97 [s], the duration of the auditory stimulus is set to be the same as when the visual stimulus is also displayed.

2.4. Visual and/or auditory stimuli

In the work described here, experiments were conducted in which multimodal stimuli consisting of combinations of the visual and auditory stimuli described above were presented to the subjects. The multimodal stimuli included five combinations of three types of visual stimuli and two types of auditory stimuli, $\{\mathbf{G_0,S_1}\}$, $\{\mathbf{G_1,S_0}\}$, $\{\mathbf{G_1,S_1}\}$, $\{\mathbf{G_2,S_0}\}$ and $\{\mathbf{G_2,S_1}\}$. Each multimodal stimulus was presented to each subject 21 times. The presentation order was random, so that the subject was not able to predict the appearance of the striking indications from the intervals between the stimuli.

3. RESULTS AND DISCUSSION

3.1. Accuracy of operation

In the experiments, the mean reaction time for when the multimodal stimuli with only the auditory stimulus $\{\mathbf{G_0,S_1}\}$ were presented was 295.0 [ms]. When the multimodal stimuli with only the visual stimulus $\{\mathbf{G_1,S_0}\}$ were presented, the mean reaction time was 324.6 [ms]. The mean reaction time for when the multimodal stimuli consisting of both visual and auditory stimuli $\{\mathbf{G_1,S_1}\}$ were presented, the mean reaction time was 264.1 [ms]. (See Table 1.) When both visual and auditory stimuli were presented, the reaction time was shorter by 30.72 [ms] than when the auditory stimulus was presented alone and 60.4 [ms] shorter than when the visual stimulus was presented alone. This shows that presenting both types of stimuli can result in more accurate drumming.

The phenomenon of a faster human reaction time in response to both visual stimuli and auditory stimuli is known as the intersensory interaction and is being investigated further in various forms [1-6]. In this experiment reported here, the authors think that the reason for the shortened reaction time that resulted from the presentation of both visual and auditory stimuli is depended on the intersensory interaction. The experiment result at the time of using simple sensory stimulus apparatus and a simple switch input is reported to Nickerson (1970) [7], and is in good agreement with the result of this experiment.

The multimodal stimulus $\{\mathbf{G_2,S_1}\}$ consists of the video image of the chip approaching the bar and the decaying sine wave

auditory stimulus. The mean reaction time for that multimodal stimulus was -9.363 [ms] (Table 1). The negative value indicates that the drum pad was struck before the strike indication appeared. Because the mean reaction time for when the multimodal stimulus **{G_1,S_1}** was used was 264.1[ms], the striking was done 304.4 [ms] earlier when **{G_2,S_1}** was used. That result confirms the shorter reaction time. Also, using the absolute value of the difference between the striking indication time and the mean value as a measure of striking accuracy to compare drumming accuracy for the multimodal stimuli **{G_2,S_1}** and **{G_1,S_1}** reveals a mean improvement in of 254.9 [ms] with **{G_2,S_1}**. The authors have thus demonstrated that the use of **{G_2,S_1}** resulted in a ten-fold or greater improvement in accuracy.

Stimulus type	Mean [ms]	Std. dev. [ms]	95% confidence interval [ms]	
{G_0,S_1}	295.0	72.32	287.3	302.7
{G_1,S_0}	324.6	53.59	319.1	330.1
{G_1,S_1}	264.1	57.68	258.8	269.8
{G_2,S_0}	-16.17	51.17	-21.33	-11.01
{G_2,S_1}	-9.363	49.26	-14.33	-4.397

Table 1: Response time statistics

3.2. Influence of an auditory stimulus together with G_2

When visual stimulus **G_2** is used, the approach of the chip to the bar can be perceived on the screen. That allows the subject to prepare for the reaction motion before the indication appears by predicting the time of the strike indication. This is considered to be the reason for the improvement in striking accuracy when **G_2** is used. The authors therefore investigated the effect of the presence or absence of the auditory stimulus when a visual stimulus that allows prediction of the time at which the pad should be struck, such as **G_2**, is used.

The response times for when multimodal stimuli **{G_2,S_0}** and **{G_2,S_1}** were presented were -16.17 [ms] and -9.363 [ms] (Table 1). The difference between these values is a few milliseconds, which is small relative to the size of the standard deviation. The authors therefore performed a one-way analysis of variance on the experimental results for **{G_2,S_0}** and **{G_2,S_1}**. The result showed no significant difference at the 5% level of significance (Table 2).

	Sum of squares	DF	Mean square	F_0
Between Groups	1.296×10^7	1	1.296×10^7	2.6513
Within Groups	3.685×10^9	754	4.887×10^6	
Total	3.698×10^9	755		

F(1,755;0.05)=3.854
P-value: 0.1039

Table 2. One-Way ANOVA (**{G_2,S_0}**, **{G_2,S_1}**)

Because the P-value is 0.1039, the significant difference is detected at over 10% level of significance. This indicates that when **G_2** is used, the effect of the presence or absence of an auditory stimulus on response time is slight. From this fact the authors conclude that **G_2** contributes greatly to the improvement of drumming accuracy, and that the presentation of a video image like **G_2** to indicate the timing with which the player should strike the drum is a valid man-machine interface for such game machines.

4. CONCLUSIONS

The authors used equipment that represents a simplification of an actual game machine to present multimodal stimuli consisting of video and sound to test subjects and measured the subjects' response times.

The results of the experiments confirmed the occurrence of the intersensory interaction when sound and image stimuli were presented together. The authors also showed that presentation of a visual stimulus that allowed prediction of the time to strike the drum pad improved drumming accuracy by a factor of ten or more. Furthermore, when this kind of visual stimulus is presented, the improvement in accuracy is not affected by the presence or absence of an auditory stimulus.

The equipment used in this research simulated actual musical instruments, and the experimental results are considered to approximate the reactions in the operation of the actual instruments. Precise timing is important when playing musical instruments, and the practice needed to acquire that precision requires much time. However, the results of this research show that the addition of equipment that guides the player with appropriate visual stimuli to a musical instrument allows even novice players to play with high accuracy and with relative ease. The results of this research are therefore expected to be applicable to the playing of musical instruments as well as to the design of game machines.

5. REFERENCES

[1] M. Hershenson, "Reaction time as a measure of intersensory facilitation," J. Experimental Psychol. 63, pp. 289-293, 1962.

[2] David H. Raab, "Statistical facilitation of simple reaction times", Trans. N.Y. Acad. Sci. 24,pp. 574-590, 1962.

[3] Donald Reynolods, "Effects of double stimulation: Temporary inhibition of response", Psychol. Bul. 62, pp. 333-347, 1964.

[4] Francis B. Colavita, "Human sensory dominance", Percept. Psychophys. 16, pp. 409-412, 1974.

[5] Francis B. Colavita, Daniel Weisberg, "A further investigation of visual dominance", Percept. Psychophys. 25, pp.345-347, 1979.

[6] R.A. Schmidt, S.C. Gielen and P.J. van den Heuvel, "the locus of intersensory facilitation of reaction time", Acta Psychol. 57, pp. 145-164, 1984.

[7] Raymond S. Nickerson, "The effect of preceding and following auditory stimuli on response times to visual stimuli", Acta Psycho. 33, pp. 5-20, 1970.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

ASPECTS ON TIME MODIFICATION IN SCORE-BASED PERFORMANCE CONTROL

Mikael Laurson[1] and Mika Kuuskankare[2]

[1]Center for Music and Technology, [2]Department of Doctoral Studies in Musical Performance and Research, Sibelius Academy, P.O.Box 86, 00251 Helsinki, Finland
laurson@siba.fi, mkuuskan@siba.fi

ABSTRACT

This paper presents our recent developments in synthesis control of physical models of acoustical instruments. We will focus on the timing control of musical scores with the help of a notation package. Our system allows to control several aspects of timing using tempo functions, performance rules, visual editors and special expressions dealing with time intervals and durations.

1. INTRODUCTION

There has already been several attempts to fine-tune timing information when translating a musical score into a computer generated performance (see for instance [1]). Lately this approach has become even more important. First, recent advances within the research dealing with modeling of acoustical instruments have lead to a situation where we must find new efficient control strategies to play these instruments. Second, while various real-time controllers are being actively developed, this technology is not yet mature enough to mimic the complex interaction between a human performer and an acoustical instrument. Our approach - using a musical score as a starting point to generate performance control information - has several attractive features: high-level musician-oriented user-interface, flexibility (almost immediate feedback of the system after editing the score) and precision.

Musical scores in our system are situated within a larger environment called PWGL. PWGL [2] is a visual programming language based on Lisp, CLOS and OpenGL. PWGL is a tool for computer assisted composition and sound synthesis. PWGL is based on similar concepts than PatchWork (PW, [3]) such as direct relation to its base languages Lisp and CLOS, musical objects and visual editors. The main difference between PW and PWGL is that the graphics part of the implementation is based now on OpenGL. Scores are of primary importance in our system and they can be used in many compositional and analytical applications such as to produce musical material for instrumental music. Scores have been also used successfully to produce control information for physical models of acoustical instruments [4], [5].

The Score-editor of PWGL utilizes internally the ENP2.0 music notation package [6]. ENP2.0 is based on the former ENP user library of PW [7]. The purpose of ENP2.0 is to provide professional notational capabilities, a graphical user interface and powerful object structures for compositional use. As PWGL, ENP2.0 is programmed with LISP, CLOS and OpenGL.

In this paper we describe how our notation package can be used to modify timing of an input-score containing basic musical information such as pitches and rhythms.

2. BACKGROUND

Timing information is probably one of the most difficult parameters to modify when translating a score to a performance. In order to meet this challenge we have considered the following general design issues when implementing our system.

First, the input process is interactive. After listening to the result the user can modify the score and recalculate the score until satisfied with the outcome. The user can select and edit any range from the score, polish it and hear the refinements in real-time, without re-synthesizing the whole piece. The ability to work with only a small amount of musical material at a time has proven to be very useful. This is especially important when working with musical pieces of considerable length.

Second, our system allows to use performance rules that generate timing information automatically in a similar fashion than in [1]. The user can, however, also work by hand using the graphical front-end of the notation package. In this case special expression markings can be inserted directly in the score. We have found that this kind of mixed approach - using automated rules and hand-given timing information - is very practical and allows to define time modifications in a more flexible way than using automatic rules only.

Third, the system supports both local and global time modifications. The importance of this kind of approach has also been discussed in [8]. Local modifications involve only one note or chord (such as an expression that changes the time interval between notes). A global modification, in turn, handles a group of notes or chords (a typical example of this is a tempo function).

Fourth, ENP2.0 includes several specialized editors that allow editing while displaying the musical score in the background. This visual approach helps the user to localize the exact time point where the change should occur. It allows to edit with great precision both local and global time modifications.

Finally, time modifications can be either absolute or relative. Absolute modifications - given in seconds - alter the start-time or duration of the notes without taking into consideration the current tempo of the piece. A typical

example is for instance an arpeggio or trill gesture that should retain its speed even if the user modifies the tempo. Relative modifications, in turn, work typically by scaling the original values thus reflecting the tempo of the input score. One example could be a staccato expression that would change the duration of a note depending on the current tempo. Furthermore, it is often necessary to use various representations for timing information, such as start-time (an absolute time-point in the score) vs. delta-time (a time interval between two adjacent notes). The former case is used when making local changes dealing only with the modified notes (i.e. the rest of the score will not be affected by this change, see for instance Section 5). The latter case, in turn, is useful when a change in one note will affect all notes after the current one (Section 4 will discuss such an example).

In the next sections we describe the various tools provided by our system in more detail. We start with tempo functions and performance rules. After this we introduce a visual editor that allows to adjust start-times of notes. Then we go over to special expressions that are used to modify delta-times and durations. We end with a section describing how the final control information is calculated in our system.

3. TEMPO FUNCTIONS

ENP2.0 supports a rich set of standard and non-standard expressions. An expression can contain other objects such as break-point functions. This type of expression is called a score-BPF. Break-point functions have many applications in our system and can be used to define tempo functions, envelopes, pitch contours, velocity curves, etc.

ENP2.0 allows fine-tuning of timing with the help of graphical tempo functions. Tempo functions are defined by first selecting a range in the score where the tempo change should occur. After this a score-BPF is applied to the range. The score-BPF can be opened and edited with the mouse. In order to facilitate the editing process, the score is shown in the background (Figure 1):

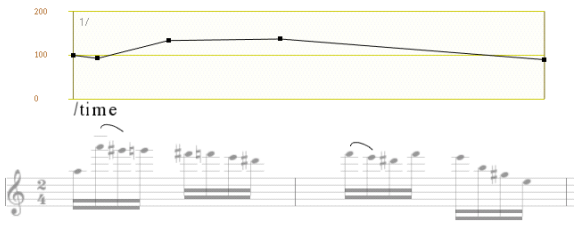


Figure 1. Editing a tempo function. The example by Nicolò Paganini shows a complex accelerando/ ritardando gesture.

Tempo function values are given in percentage (i.e. 100 means 'a tempo', 200 twice as fast, 50 half tempo). For instance an accelerando from 'a tempo' to 'twice as fast' is achieved with a ramp that starts from 100 and that ends at 200.

In addition to conventional accelerandi and ritardandi, the user can apply special rubato effects ("give and take"). As in the previous case the user starts with a selection in the score and applies a tempo function to it. In this case, however, the duration of the range is not affected by the time modification. Time modifications are only effective inside the selected range.

4. PERFORMANCE RULES

ENP2.0 supports user definable performance rules which allow to modify score information in a similar way as in [1]. In ENP2.0, however, rules use a syntax which was originally designed for PWConstraints [3]. PWConstraints has been used to solve difficult compositional search problems dealing with melodic lines, counterpoint, voice-leading and harmony. PWConstraints provides a rich set of tools that allow to access various structural information from a musical score. Furthermore, it allows to formulate rules in a compact manner using a pattern-matching language. Thus this mechanism provides us an ideal tool to write performance rules where one has to access musical information from several points of view.

Rules are written in two parts: (1) a pattern-matching part which is followed by (2) a Lisp-code part. The pattern-matching part checks when the rule should be applied and also extracts relevant note information which is used later by the Lisp-code part. We use a fairly typical pattern-matching syntax that can contain a wild-card ('*'), variables ('?1', '?2', '?3', etc.), anonymous-variables ('?') and index-variables ('i1', 'i2', 'i3', etc.). The Lisp expression, in turn, executes the actual modification.

ENP2.0 performance rules that deal with timing work typically using delta-time representation. Thus a change in one note's delta-time will shift all notes that occur after the current one.

For instance a variant of the well known "notes inégales" rule ("in a beat of two eighth-notes the first note is made longer than the second one") can be translated into PWConstraints syntax as follows:

```
(* ?1 ?2 (rtm-match? (1 ((?1 1) (?2 1)))          ;; (1)
   (?if (write-key ?1 :dt (+ (read-key ?1 :dt) 0.1));; (2)
        (write-key ?2 :dt (- (read-key ?2 :dt) 0.1)))))
```

The pattern-matching part - line (1) - states that if two notes (?1 and ?2) are adjacent and form a two eighth-note beat, then the time interval (given by the keyword ':dt') of the first note (?1) is lengthened by 0.1 s and the second one (?2) shortened by 0.1 s. This change is absolute, i.e. the modification is independent of the tempo. Sometimes, however, it makes more sense to make this change relative. This can be accomplished by multiplying the original time interval value by some appropriate scaler.

This rule, when applied to a score, is global: it will affect all note groups forming a two eighth-note beat. Although global rules can be useful they can result in a mechanical interpretation if they are applied very often in the input score.

To avoid this one can use more local rules, i.e. rules that can be turned on or off in the score depending on the musical situation and the taste of the user. One way to localize a rule is to add an extra check that has to be true before the rule is to be applied:

```
(* ?1 ?2 (and (not (belong-to-group? ?1 "Even 8th's"))
           (rtm-match? (1 ((?1 1) (?2 1))))
  (?if (write-key ?1 :dt (+ (read-key ?1 :dt) 0.1))
       (write-key ?2 :dt (- (read-key ?2 :dt) 0.1)))))
```

In this case the rule is applied only if the first note (?1) does not belong to a group-expression called "Even 8th's". Figure 2 gives an example score where the first two eighth-note group is active and second one inactive:
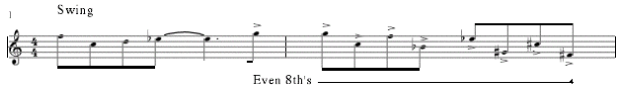


Figure 2: Applying local performance rules to modify timing. An excerpt from a jazz standard 'Harlem Nocturne' by Earle Hagen

5. START-TIME EDITOR

ENP2.0 contains a specialized editor that allows to edit directly the start-times of notes. The start-time editor belongs to a family of editors that allow to inspect and modify various note properties such as velocity, start-time and duration. All these editors share the same basic idea where the editable data is visually synchronized with the underlying musical structures. For instance Figure 3 shows a start-time editor in an editable state. The score is shown in the background drawn in gray color. The editor consists of a coordinate system where the x-dimension denotes time and y-dimension pitch. The individual notes of the chords of the background score are represented as points. The y value of a point is given by the midi value of the respective note and the x value by the start-time. While editing, the start-time of a note is displayed relative to the start-time of the chord that owns the note. Thus when it equals to 0 then the note will be played at the start-time of the chord, if it is negative then it will be played before the chord and if positive after the chord.

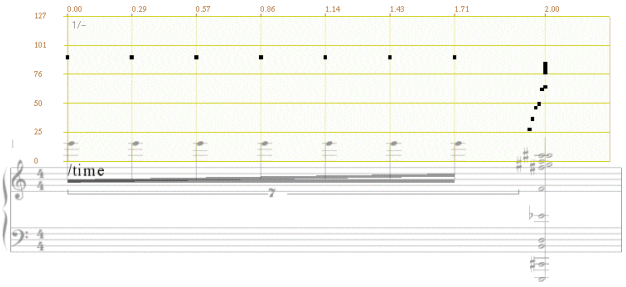


Figure 3. An editor modifying start-times of a keyboard texture.

6. EXPRESSIONS CONTROLLING TIME INTERVALS AND DURATIONS

ENP2.0 expressions can be used to modify time intervals and durations. For instance, a tenuto expression can be interpreted so that it will lengthen the time interval between the current note and the next one. These kinds of local time modifications are especially useful when used with global modifications such as tempo functions. ENP expressions are often used in conjunction with performance rules. For instance, in the following rule delta-times of tenuto notes will be lengthened by 10%:

```
(* ?1 (belong-to-group ?1 :tenuto)
    (?if (write-key ?1 :dt (* (read-key ?1 :dt) 1.1)))
```

The user can define similar kind of behavior for other types of expressions such as trills, accents, etc.

Our system contains two types of expressions that control the duration of notes. The first one is used to realize staccatos while the second one is used to define overlaps. If no duration expressions are found in the score then all note durations will be 100%. Staccato expressions are named using 'st' with a suffix (an integer from 1 to 9) indicating the amount of staccato. For example, an expression 'st5' would denote a staccato where the resulting note duration is 50% of the original.

Overlaps are useful as they allow to mimic polyphony in a homophonic texture and/or to realize various pedal effects. Overlaps can be defined using an expression with the name 's' ('s' stands for 'sustain'). In this case the note in question will be sustained to the end of the expression. Partial overlaps can be realized by adding an integer (ranging from 1-9) to the name. Thus 's1' would denote a 10% overlap. Finally, our system provides yet another way to define overlaps using an expression with the name 'gs' (for 'general sustain'). In this case all notes within the expression are held until the end of the expression.

Figure 4 gives an example that contains both tenuto expressions and overlaps (notes marked with 's'):

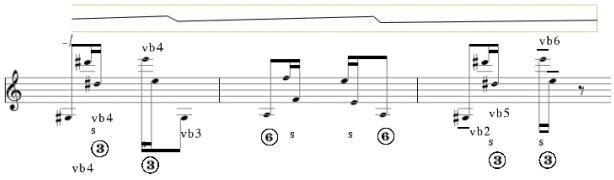


Figure 4. A classical guitar example (by Nicolò Paganini) with instrument specific expressions, a tempo function, tenuto expressions and overlaps.

7. CALCULATION OF CONTROL INFORMATION

Previous sections have described various timing tools as separate entities. We end with a discussion of how these

modifications are combined and how the final calculation of the control information is executed. First, the note information provided by the input score is modified by the tempo functions and performance rules. In order to assure synchronization of polyphonic scores, all tempo functions are merged and translated internally into a global time-map [9]. This phase is responsible for all tempo changes and local time shifts given in the score. After this, local timing data that is not dependant of the tempo is added to the timing information. These local modifications were discussed in Sections 5 and 6 and include note start-time, staccato and overlaps.

In the second step, all the notes of the input score are scheduled. While the scheduler is running, each note sends a special method to its instrument which in turn starts other scheduled methods which produce the final control data. These methods are responsible for creating discrete control data (such as excitation information) or continuous data (filter gains, filter coefficients, or other low-level data). This part of the system is described in more detail in [10]. If a note contains a special macro expression (trill, portamento, tremolo) then new note instances are created on the fly by instrument specific methods and inserted in the scheduler queue.

8. CONCLUSIONS

This paper has presented a notation-based environment to generate performances for synthesis control. Our system is highly interactive as it allows to use a graphical front-end to define time modifications. Several approaches (tempo functions, local and global performance rules, expression markings, visual editors) can be used to gain the desired result. The system has been used to realize large scale musical examples for various instrument models. Sound examples can be found in our homepage:

http://www.siba.fi/soundingscore.

This research has been conducted within the project "Sounding Score—Modelling of Musical Instruments, Virtual Musical Instruments and their Control" financed by the Academy of Finland.

9. REFERENCES

[1] Friberg, A., "Generative Rules for Music Performance: A Formal Description of a Rule System", Computer Music Journal, Vol. 15, no. 2, pp. 49–55, 1991.

[2] Laurson M., and Kuuskankare M., "PWGL: A Novel Visual Language based on Common Lisp, CLOS and OpenGL", in Proc. ICMC'02, pp. 142-145, 2002.

[3] Laurson M., PATCHWORK: A Visual Programming Language and Some Musical Applications. Doctoral dissertation, Sibelius Academy, Helsinki, Finland, 1996.

[4] Laurson M., Erkut C., Välimäki V., and Kuuskankare M., "Methods for Modeling Realistic Playing in Acoustic Guitar Synthesis", Computer Music Journal, Vol. 25, no. 3, pp. 38-49, 2001.

[5] Välimäki V., Laurson M., and Erkut C., "Commuted Waveguide Synthesis of the Clavichord", To be published in Computer Music Journal, 2003.

[6] Kuuskankare M., and Laurson M., "ENP2.0, A Music Notation Program Implemented in Common Lisp and OpenGL", in Proc. ICMC'02, pp. 463-466, 2002.

[7] Kuuskankare M., and Laurson M., "ENP, Musical Notation Library based on Common Lisp and CLOS", In Proc. ICMC'01, pp. 131-133, 2001.

[8] Honing H., "From Time to Time: The Representation of Timing and Tempo", Computer Music Journal, Vol. 25, no. 3, pp. 50-61, 2001.

[9] Jaffe, D. A., "Ensemble Timing in Computer Music", Computer Music Journal, Vol. 9, no. 4, pp. 38–48, 1985.

[10] Laurson M., and Kuuskankare M., "Instrument Concept in ENP and Sound Synthesis Control", in Journées d'Informatique Musicale, Marseille, 2002.

INTERACTION AMONG PHYSICAL ACOUSTIC KNOWLEDGE AND PERFORMANCE IN THE ACCORDION AND ITS EDUCATIONAL SIDE

Llanos-Vázquez, R.[1]; Alonso-Moral, J.[2]; Elejalde-García, M.J.[1]; Macho-Stadler, E.[1]

[1]Dpto. Física Aplicada 1. Escuela Superior de Ingenieros. UPV/EHU.
Address: Alameda de Urquijo s.n., 48013 Bilbao, Spain.
wupelgam@bi.ehu.es
[2]Conservatorio Superior de Música Juan Crisóstomo de Arriaga
Address: General Concha 20, 48010 Bilbao, Spain.
JESALONSO1@terra.es

ABSTRACT

The vibration function of the free tongue, as well as the mechanisms of buttons and bellows, has a fundamental importance for the sound production of the accordion. The objective of this work is to show how to organize a corpus of physical acoustic knowledge useful for the playing and for the teaching of this instrument. Concretely, we focus on the analysis of hard attacks (of finger) versus soft attacks (of bellows), different button lowering velocities in the attacks, spectral analysis of the same tones in different registers, bellows direction change, differentiations among different "musettes", as well as a connection among sound envelopes and expressivity's elements. The interactive development among the physical side and the musical side seems to provide some clear and precise possibilities for the educational [1] and playing activity, and it contributes knowledge so that those activities are more effective. Likewise, with the physical acoustic references, in this work there are generated coherent fundamentals on different ways of the accordion playing.

1. VIBRATIONAL FUNCTION OF THE FREE REED

1.1. Technical data of the experimental set up

We will work with a Pigini's concert accordion, Sirius model, with buttons in its right manual and free bass (besides the traditional system of standard bass) in its left manual. When studying the musette effect, we will also deal with a Guerrini's accordion designed for the interpretation of popular music.

The analyzer is a Larson Davis LD 2900-B and the microphone a Brüel & Kjær type 4176, placed to 40 cm from the center of the front plane of the accordion.

1.2. Description

The sound generators of the accordion are the free reeds stimulated by the airflow supplied by the bellows. As indicated in the Figure 1, the basic constituent elements of the reed are a plate, the own reed (or tongue) and the rivet. The not riveted end of the reed is free to vibrate from one side and the other of the slot.

The reed will only vibrate when the air comes from the side of the plate on which the reed is fixed.

But on the other side of the plate there is another tongue that will vibrate when the air comes from the opposite direction to the one indicated in Figure 1.

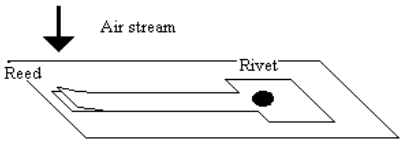


Figure 1: *A free reed on its reed plate.*

1.3. Pitch

The pitch of any note is determined by the natural frequency of the corresponding reed. It is accepted that the reed is a thin free-fixed bar.

When exposed to the flow of air supplied by the bellows, the reed, having little resistance, will vibrate with a frequency just below the frequency of lower resonance of the fixed-free bar. Apart from this, due to the non linearity of the problem, a great number of harmonic of that frequency is found, instead of the non-harmonic overtones expected for a free-fixed bar vibrating freely after being disturbed from its equilibrium position [2], [3], [4], [5].

2. VOICES AND REGISTERS

The normal concert accordion [6] consists of four groups of reeds called voices (a lower one, two central ones and a higher one) for the right hand and three voices (according to the accordion type, either a lower, a central and a sharp one or two lowers and a higher one) for the free bass of the left hand. These groups of reeds can be combined by means of the registers: a mechanism whose action makes possible to put in action several reeds corresponding to the same note (but to different voices) in the same or different octave, by means of the pulsation of a single button or key.

2.1. Loaded reeds

The lowest reeds in the right hand are smaller than the corresponding ones of the same pitch in the left manual. The reason is to maintain the accordion weight under reasonable limits. So, in order to achieve that the reeds of the right hand,

being shorter than the left ones, sound so low as those of the left manual, the right hand ones are loaded with more mass in the tip than those of the left hand (which also have some load) [7]. But, in similar conditions of playing, this additional load makes their sonority smaller and its timbre not so full (fundamental and first harmonics less intense) compared with the left manual reeds.

Pedagogic application: when writing a transcription for accordion, if one has the opportunity to choose the manual with which a very low passage will be played, it should be keep in mind the biggest sonority and richer timbre of the left one.

2.2. Cassotto

Each register has its own timbre, not because the reeds are different, but because of their location inside the accordion (apart from what was commented in the previous section). In the right hand of the concert accordions, two of the four voices (the lowest and a one of the central ones) are placed inside a resonance chamber called cassotto, which gives the instrument a warmer and more velvet sound eliminating or softening part of the higher harmonics. The cassotto is only present in high quality instruments and in voices of the right hand, not in the left one and never in accordions designed for popular music [6]. In the Figure 2a the harmonic amplitude distribution is represented for a La_4 sound with a one central voice register in cassotto and then, in the Figure 2b the distribution of amplitudes for a central voice out of cassotto. The attenuation of the higher harmonic by the cassotto is reflected in a sharper fall of their amplitudes as we move toward higher overtones.

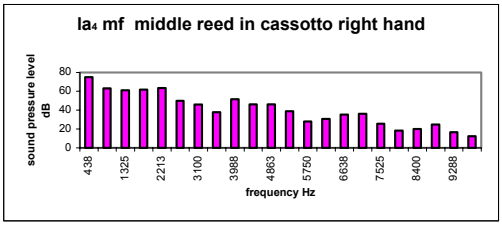
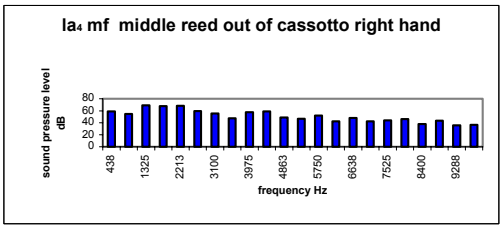


Figure 2a: *Harmonic amplitude distribution for a La_4 sound*



with a one central cassotto voice register.

Figure 2b: *Harmonic amplitude distribution for a La_4 sound with a one central out of cassotto voice register.*

Pedagogic application: depending on the desired timbre (we deal with an original music but no registration is indicated, we tackle with a transcription, with chamber or orchestral music, acoustics of the room or concert hall) we will use cassotto or no cassotto voices. For example, the fact that registers without cassotto have more intense high harmonics can be used to make the accordion be heard over an orchestral background (analogous to what happens with the singers formant [8]).

2.3. Musette

Sometimes, one of the complete sets of reeds corresponding to one of the central voices of the right hand is tuned somewhat different than the other one so to create the "musette" effect: the beats generated in this way give a characteristic color of the popular accordion (in fact, some accordions designed for folk music are provided with up to four central voices, with several of them slightly out of tune).

In our experiments to study the musette effect, we have worked on two different accordions. Besides the Pigini Sirius concert accordion on which we have developed most of the measures; we have also work with a Guerrini accordion, designed to interpret popular music. In the case of the concert accordion, when measuring the spectrum corresponding to the La_4 with the two central voices register, only one fundamental of 442Hz, with a 0,2Hz resolution, has been found, that is, both reeds were tuned to the unison. This is the usual when dealing with concert accordions, where it is tried to avoid all kind of beats in the tuning.

However, in the case of the folk Guerrini accordion, when studying the spectrum of the same note, but with a three central voices register, we founded three peaks centered at 440Hz, 442Hz and 447Hz (plus those associated to each harmonic series); this implies a considerable confusion due to the beats, which is a characteristic effect of the interpretation with folk accordions.

3. TRANSIENTS: FINGERS AND BELLOWS ACTION

The growth and decay of the sounds are relatively short, of 30 at 120ms approximately, according to the reed and the type of attack. The kind of attack depends on what is activated before the button or the bellows, and on the speed with which the button is pressed down.

In Table 1 we have schematized the four basic attacks. The data correspond to a La_4 played mezzo forte using a cassotto central voice register. All these effects are better noticed using single voice registers. Evidently, between these extreme cases there are many intermediate possibilities.

Something similar could be said with regard to the use of the different decays. If the button is suddenly released, maintaining the bellows tight, the decay will be hard (determined by the speed of recuperation of the button), while if the decay of the sound is carried out either stopping the bellows without releasing the button or maintaining tight the bellows but releasing slowly the button, the decay will be softer.

3.1. Dependence on the pitch.

In the measures we have made it has been found that the sharper the fundamental sound of the studied reed, the shorter the transients, supposed that the rest of parameters maintains constant, and that we don't work with very strong sonorities (see paragraph 3.2). This relation is due to a bigger inertia of the reeds corresponding to low-pitched fundamentals (bigger

tongues). To be precise, using a cassotto central voice register, Hard-Hard attacks and a dynamics in forte, the transients times go from about 100ms for the La_2, 35ms for the La_4, until arriving at about 30ms for the La_6.

Attack type	Description	Hardness of the attack (Observations)	Transient time
Hard-Hard	1st we pull (push) of the bellows 2nd the button is pressed down quickly	the faster the button is pressed down, the harder the attack is	30ms
Hard-Soft	1st we pull (push) of the bellows 2nd the button is pressed down slowly	the slower the button is pressed down, the softer the attack is	42ms (depending strongly on how slowly is pressed down the button)
Soft-Hard	1st the button is pressed and hold down 2nd the bellows is quickly activated	the faster the bellows is activated, the harder the attack is	37ms
Soft-Soft	1st the button is pressed and hold down 2nd the bellows is slowly activated	the slower the bellows is activated, the softer the attack is	44ms (depending strongly on how slowly is activated the bellows and the button is pressed down)

Table 1: *Basic attacks and their transients.*

3.2. Dependence on the sonority.

In the same way, maintaining constant the rest of parameters, the stronger the sonority, the shorter the transients are [9]. Using again a cassotto central voice register, the La_4 and Hard-Hard attacks, the transients last from about 40ms for the pianissimo to about 34ms for the fortissimo. In the case of carrying out this experiment with a larger reed, corresponding to a La_2, it is found that the transients are made longer again when using sonorities in fortissimo, since the reed tends to drown. Pedagogic application: the students should learn that when dealing with very low reeds, and if they don't want to drown the sound, it is necessary not to attack abruptly in fortissimo.

4. HARMONIC AMPLITUDE DISTRIBUTION

To study the spectra we have differentiated between right and left manual, having centered us more in the right one. We have distinguished between low (La_2), middle (La_4) and high sounds (La_6), and between pianissimo, mezzoforte and fortissimo sonorities. We have also worked separately the cassotto and the non-cassotto.

4.1. Dependence on the register.

If we analyze the spectrum of a sound with a central voice register, it is observed that, independently of the intensity and of the pitch (La_2, La_4 or La_6) of the fundamental sound, the spectra corresponding to a one central voice cassotto register, present much more attenuated high harmonics that those corresponding to one central voice non cassotto register.

4.2. Dependence on the intensity.

In the same way, we see that, maintaining fixed the rest of parameters, the greater the intensity of the analyzed sound is, the higher the number of noticeable harmonics in the spectrum [9]. But it seems to be that if we increase too much the sonority, the amplitude of the high-pitched harmonics doesn't already grow more, and it even will diminish, what is consistent with which the fact that a reed can get drown if the pressure goes beyond some high limit. Pedagogic application: as we said before, the students should learn how to use the bellows cleverly, without exercising a greater force than the one that the reeds can support without beginning to drown and to diminish their answer.

4.3. Dependence on the frequency of the fundamental sound.

If we maintain fixed the rest of parameters, we see that the higher is the fundamental frequency, the more intensive are the high-pitched harmonics (between 5000 and 10000 Hz).

4.4. Right hand-left hand differences.

The differences of sonority and timbre between the right and left manuals are due to two causes. On one hand, the right and left chambers of resonance are different in shape and have different elements placed inside them and surrounding the reeds under study. On the other hand, we have already said that the lowest reeds of the left manual are less loaded and they are bigger than those of the right one. Pedagogic application: because of these differences both manuals never give the same sound, what has to be kept in mind by educators and players when writing a transcription of a work for another keyboard instrument, or by the composer in the moment to write for accordion.

4.5. Cassotto effect in the left hand.

There is no cassotto corresponding to the left manual, that is, there is no special chamber of resonance such as the one in the right side. However, if we compare the three one voice registers of the left hand (low, central and high), we observe that the low one voice register presents its high harmonics more attenuated than the central and high one voice registers, a somewhat similar effect to the cassotto of the right hand. This is due to the disposition of the reeds inside the left harmonic box

4.6. Free bass manual peculiarities.

We describe now another effect that we find in the left harmonic box. We have worked on a concert accordion with three independent voices in the left manual. When choosing an instrument of these characteristics, the players can order to the manufacturer either the disposition of the three voices in three

heights (low, central and high) or the so called "double bass" lay-out (two low voices and a high voice, without any central voice) [6]. When comparing the lowest octave (mi_1 to $fa\#_2$) with the one low voice register of the left hand of both instruments, it is observed that in the accordion "without double bass" this octave is more attenuated (its spectra show weaker high harmonics) with regard to the rest of notes of the registration. This effect is not found in the accordion of "double bass", that presents a more homogeneous sonority and color. It is because the reeds of that lowest octave are all in the same block of reeds and the mechanisms of registration of the accordion "without double bass" impede the radiation of the sound from that block.

Pedagogic application: the accordionist will have to choose between the homogeneity and sonority of accordion of double bass or the greater flexibility allowed in the election of registrations by the accordion without double bass.

5. EXPRESSIVE POSSIBILITIES

When focusing on the interpretation, apart from the basic use of the bellows (to vary the dynamics) and of the diverse articulation types (legato, non legato, staccato...), the accordion student should learn how to use other expressive resources: these are based in the use of the different types of attacks achieved by the combination bellows and fingers and in the small dynamic control achieved with the depth of the pulsation without necessity of involving the bellows.

5.1. Applications of the diverse types of attack.

In general, soft attacks are used with soft dynamics and slow movements, in "expressive" passages. On the other hand, the hard attacks achieved with the bellows are usually used to get very marked accents that involve the group of all the notes that sound in that moment. These are also the attacks used to carry out a bellows change with the minimum breakdown in the continuity of the sound. One can take advantage of the hard attacks achieved with the fingers to highlight a given passage or a note or melodic line without necessity of varying the dynamics. They are also used to achieve the maximum clarity in quick passages. As far the decays is concerned, the hard extinctions achieved loosing abruptly the button are usually applied in quick passages, sharp accents, staccati with the fingers... while the softest extinctions are usually used for more slow passages, delicate finals...

5.2. Applications of the dynamic control by the pulsation.

The bellows is not the only source of the variation in the dynamics of the accordion. Indeed, small (but perceptible) variations of the sonority can be gotten modifying the depth of the pulsation. The applications are varied:

- To highlight a high-pitched melody over a slow and low background: the highest-pitched reeds usually provide less sonority, so that we can balance their sonority (compared with the lowest ones) by means of this control of the pulsation, pressing down slightly the lowest-pitched reeds but completely the highest-pitched ones.
- To obtain a very sonorous and marked bass, when using the standard bass in the left hand, without necessity of using the bellows, what would imply an unwanted disturbing in the sonority of the right manual.
- To highlight somewhat a voice in a polyphonic texture regulating the depth of their pulsation.
- To soften dynamically the end of a sound played in a manual without varying the dynamics of the other sounds, in the same or in the other manual: this effect is carried out by releasing more or less slowly the corresponding button. And, in the same way, to enter smoothly a voice the button must be pressed down slowly maintaining constant the tight of the bellows. These effects are more easily noticeable and more convenient to carry out in slow movements, with soft dynamics and with single voice registrations.

6. CONCLUSIONS

In this approach to the acoustics of the accordion it has been shown a foundation of some basic elements of the accordion playing. More concretely, it is appreciated a strong coherence between the structure of acoustic phenomenon and the corpus of the accordion technique. A deeper study in this direction will suppose a greater foundation of the interpretation and teaching, determining with accuracy what can and cannot be demanded to the accordion in each musical context.

7. REFERENCES

[1] Llanos-Vázquez, R.; Elejalde-García, Mª J.; Macho-Stadler, E.; Alonso-Moral, J., "A pedagogical utilisation of the accordion to study the vibration behaviour of free reeds", Revista de Acústica, Vol. XXXIII, 2002; Sociedad Española de Acústica, Madrid.

[2] Fletcher, N. H.; Rossing T. D., "The Physics of Musical Instruments", 2nd ed. Springer, New York, 1998.

[3] Fletcher, N. H., "Autonomous vibration of simple pressure-controlled valves in gas flows", J. Acoust. Soc. Am., Vol. 93, 1993, pp 2172-2180.

[4] Cottingham, J. P.; Reed, C. H.; Busha, M.; "Variation of frequency with blowing pressure for an air-driven free reed", Forum Acusticum, Berlin, March 14-19 (1999).

[5] St Hilarie, A. O.; Wilson T. A.; Beavers, G. S.; "Aerodynamic excitation of the harmonium", J. Fluid Mech., vol. 49, 1971, pp. 803-816.

[6] Llanos-Vázquez, R.; Alberdi I., "Accordion for composers", Self Published, ISBN 84-607-6778-7, Lasarte-Oria, Spain, 2003.

[7] Benetoux, T., "The Ins and Outs of the Accordion", Self Published, ISBN 2-9517184-1-1, Le Thor, France, 2002.

[8] Campbell, M.; Greated, C.; "The Musician's Guide to Acoustics", Oxford University Press, New York, 1987.

[9] Meyer, J., "Acoustics and the performance of music", Verlag Das Musikinstrument, Frankfurt/Main, 1978.

8. ACKNOWLEDGEMENTS

This work has been partially supported by the project 1/UPV 00057.345-E-14783/2002 of the Universidad del País Vasco/ Euskal Herriko Unibertsitatea.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

TIME DOMAIN NOTE AVERAGE ENERGY BASED MUSIC ONSET DETECTION

Ruolun Liu[1], *Niall Griffth*[1], *Jacqueline Walker*[2] *and Peter Murphy*[2]

1: Department of Computer Science and Information System
2: Department of Electronic and Computer Engineering
University of Limerick, Limerick, Ireland

Ruolun.Liu@ul.ie

ABSTRACT

A novel time domain strategy is proposed for the detection of the onset of musical notes based on the changing energy level. By calculating the note average energy (NAE), the proposed method is insensitive to both the dynamic range of the energy levels in a piece and whether or not the piece is monophonic or polyphonic. More importantly, the new strategy tackles the thorny 'threshold' problem that is always being avoided unsuccessfully. The detection performance of the new method is illustrated by its performance over a range of music pieces played on different instruments.

1. INTRODUCTION

Onset detection is one of the most important components of computational auditory scene analysis (CASA). Subsequent analysis of rhythm, melody, and harmony depend on the accuracy of onset detection. It's also of importance for transcription, music editing and recording.

Conventional onset detection methods are usually based on a peak-searching technique where a threshold is needed to detect onsets from a measured curve. The problem is how to determine the threshold. It is either superficially "*properly chosen*", or a fixed proportion of a specified value. From simulations, we found that the appropriate threshold varies with the music in question. A good threshold for piano doesn't guarantee a good result for violin. Even worse, for the same piece of music, the threshold suitable in one part is not suitable for some other parts. None of the existing methods provides a generalized method for determining the threshold for music of different styles and instruments to avoid missing notes and spurious notes being detected.

The key issue for onset detection is how to determine the *noticeable* change in the sound energy. Since the noticeable level varies from song to song and from style to style, no one fixed threshold is suitable for all kinds of music. However, is it possible to find the onsets without any threshold?

2. EXISTING METHODS

To try and answer this question, we investigate existing methods of measuring the prominence of the note energy level change. Among the existing methods, curves that can be used in the peak searching procedure are the different formulations of the evolution of sound energy. They can be roughly classified into time domain methods and frequency domain methods.

2.1. Time Domain Methods

To begin with, a power envelope must be extracted. Assume $s(t)$ is the sound wave signal and $h(t)$ is the smoothing window. The instant power envelope is the convolution of the instantaneous power, $s^2(t)$, and the smoothing window, that is:

$$p(t) = \int_0^t s^2(\tau)h(t-\tau)d\tau \quad (1)$$

While listening to music, people can relatively easily tell where the onset is, from the falls and rises of the instant power envelope $p(t)$. To make a computer perform this task, a better curve is needed in which the peaks can be easily found.

First Order Difference & Evolution Ratio

Among the different developed formulations of $p(t)$, the most traditional ones include First Order Difference and the Evolution Ratio,

$$FOD(t) = p(t) - p(t-T) \quad (2)$$

$$R(t) = p(t) / p(t-T) \quad (3)$$

where T is the sampling interval.

Both the first order difference and the evolution ratio measure the degree of change in the sound power level in a quite natural way. While $FOD(t)$ observes the absolute change, $R(t)$ measures the relative change. When the sound power increases, $FOD(t)$ returns a positive value, and a negative value for decreasing power. On the other hand, the faster the power changes, the greater the degree by which $R(t)$ differs from 1: greater than 1 for increase and less than 1 for decrease.

Since $FOD(t)$ peaks at the highest local change in the power envelope, it doesn't peak at the very beginning of the note. In addition, small oscillations in the decline of a note from a higher power level of $p(t)$ may cause peaks that are larger than peaks corresponding to notes with relatively lower power levels. In Figure 1(b), the 4th note has 2 by-peaks that are bigger than the main peak for the 7th note.

Leaving aside this problem, $R(t)$ is better than $FOD(t)$. It peaks quite close to the beginning of the note, but if there are small sharp surges at lower power level in $p(t)$, it returns peaks competitive to the peaks corresponding to notes at a relatively higher level. As shown in Figure 1(c), after the 7th note, there are 2 extra peaks that are bigger than the main peak for the 5th note.

First Order Relative Differential

To overcome the problems of the first order difference, Klapuri[1] proposed a first order relative differential which is the first order differential of $p(t)$ divided by $p(t)$ itself.

$$FORD(t) = p'(t) / p(t) \tag{4}$$

This is equivalent to the first order differential of the logarithm of $p(t)$. $FORD(t)$ is said to be psychologically relevant because people perceive the sound energy level in a logarithmical scale. Equal increases are more prominent in a quiet signal than in a loud signal. However,, we found little difference between the first order relative differential and the evolution ratio, and it is hard to find any improvement, see Figure 1(d).

2.2. Frequency Domain Methods

The power envelope can also be obtained in the frequency domain from the short time Fourier transform (STFT) of the sound signal.

$$q(t) = \int_0^{+\infty} |STFT_s(f;t)|^2 \, df \tag{5}$$

where $STFT_s(f\,;\,t)$ is the STFT coefficient, at frequency f, of the sound frame of $s(t)$ beginning at time t. Two consecutive frames may also be overlapped.

Theoretically, the time domain methods mentioned above can also use the power envelope obtained by (5), which results in time-frequency hybrid methods; but practically, $q(t)$ given by (5) is not smooth enough to give good results. Researchers have developed frequency domain methods based on measurements from the frequency domain.

High Frequency Content (*HFC*)

$$HFC(t) = \int_0^{+\infty} f \, |STFT_s(f;t)|^2 \, df \tag{6}$$

By weighting the power spectral intensity linearly toward the high frequency, HFC emphasises the high frequency component of the power spectral intensity. This strengthens the measurement of a sharp attack in a sound signal because a sharp attack involves more higher frequency components. However, $HFC(t)$ doesn't peak at the start of the rise in $q(t)$. In Figure 2(b), it is easy to see $HFC(t)$ peaking at the middle of the rise of $q(t)$.

Masri's Detection Function

Masri[2] proposed a detection function based on $HFC(t)$ which peaks at the instant of the start of a new note.

$$M(t) = \frac{HFC(t)}{q(t)} \times \frac{HFC(t)}{HFC(t-T)} \tag{7}$$

$M(t)$ observes the relative power changes in both the higher frequency components and $HFC(t)$ for all frequency components.

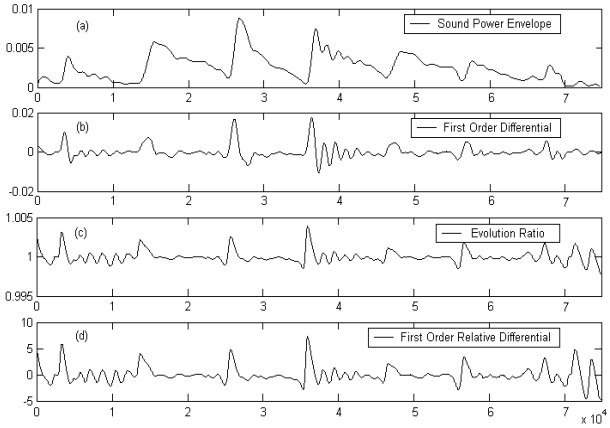


Figure 1*: Example of FOD* (b), *revolution ratio* (c) *and FORD* (d) *from power envelope* (a).

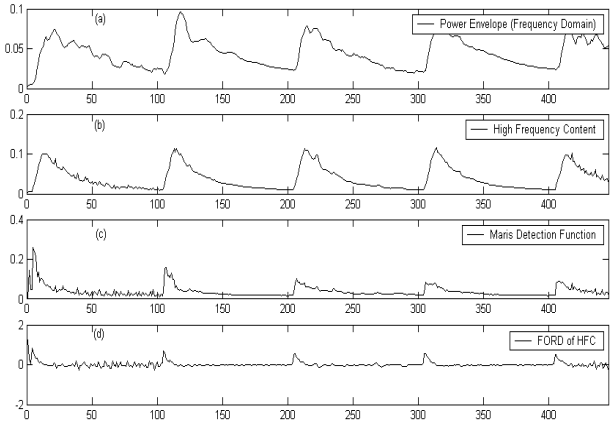


Figure 2: *Example of HFC* (b), *Masri's detection function* (c) *and FORD of HFC* (d) *from power envelope* (a).

One big drawback of $M(t)$ is the imprecision of the peaks for notes with a slow rise. See Figure 2(c).

FORD of HFC

Giuliano[3] proposed a combination of first order relative differential and high frequency content.

$$D(t) = HFC'(t) / HFC(t) \tag{8}$$

As shown in Figure 2(d), $D(t)$ performs better than $HFC(t)$ and Masri's method to some degree, but the inherent shortcomings of the two original methods don't automatically compensate for each other and no additional techniques are used to overcome them.

2.3. Band Wise Processing

Another possibility is to detect onsets $o_i(t)$, within frequency bands selected by a Gammatone filter. By setting a weight to each band according to the contribution of the components in its band, we can expect a better detection of the onset.

$$o_w(t) = \sum_{i=1}^{I} w_i o_i(t) \tag{9}$$

This idea was first proposed by Scheirer[4]. It may be a good complement if no further improvement can be found without band wise processing. But it's also possible to create spurious onsets by frequency band splitting. We will now consider a solution that removes the threshold problem and is applicable to both monophonic and polyphonic music played on a wide range of instruments.

3. NOTE AVERAGE ENERGY ONSET DETECTION

As addressed in section one, the key issue for onset detection is how to determine 'noticeable' changes in sound energy. We noted that none of the methods outlined above really investigate the energy of the music signal. True, power reflects the energy quite precisely, but it's not real energy after all.

Energy is the integration of power with respect to time *t*.

$$e(t) = \int_0^t p(t)dt \tag{10}$$

In Figure 3(b), we can see a bending corresponding to each note, even some small sharp oscillations in its decline. However, the bending is not very clear because $e(t)$ is an increasing function. As it is not convenient to find the start or end of the bending, we decided to investigate the average energy.

$$a(t) = \frac{1}{t}\int_0^t p(t)dt \tag{11}$$

In Fig.3(c), the bending is much easier to detect, but not for the later notes. Since the energy difference between consequent notes is bounded, but time *t* keeps increasing, when *t* is big enough, no energy difference could be found between two consequent notes from $a(t)$.

Enlightened by the features of the averaged energy, $a(t)$, we investigated $a(t)$ within the duration of each note, namely Note Averaged Energy (NAE). We found as expected that NAE declines when no new note follows; while NAE rises when a following note adds additional energy. This suggested that NAE could be a better measurement of the 'noticeability' of energy level changes.

Suppose we know the onset instant, t_n, of every note where n= 1, 2, … N, and N is the number of notes. NAE is defined as:

$$NAE(t) = \frac{1}{t - t_n}\int_{t_n}^{t} p(t)dt, \quad (t_n < t < t_{n+1}) \tag{12}$$

When $t = t_n$, $NAE(t_n) = p(t_n)$. From Figure 3(d), $NAE(t)$ has only one increase and decrease for each note. So onsets can be determined simply by checking the local minima.

Since $\{t_n ; n = 1, 2, \ldots N\}$ is the set of onsets to be determined, $NAE(t)$ can not be calculated directly. However, this doesn't mean we cannot decide where an onset is. Let n=0 and t_n=0, if we calculate $NAE(t)$ in a range of *t* with a limit open to the right. If the following relation:

$$NAE(t - T) > NAE(t) < NAE(t + T) \tag{13}$$

is met, we can set $n = n+1$ and $t_n = t + T$, and move on to the next detection also in a range of *t* with a limit open to the right. In this way, we remove the threshold problem.

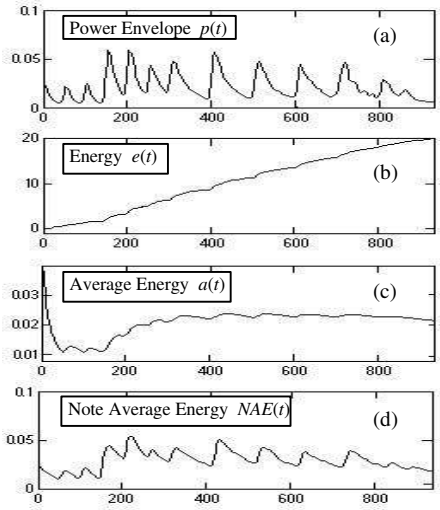


Figure 3 *Energy* (b), *averaged energy* (c) *and note average energy* (d) *of power envelope* (a).

Other strengths of NAE include insensitivity to the overall sound energy variation, and applicability to both monophonic and polyphonic music. This comes from the inherent property of $NAE(t)$. Since $NAE(t)$ investigates one note only, the overall power evolution has little effect on onset detection. $NAE(t)$ examines the onset from the view of note energy changing in the time domain, it has nothing to do with whether the music is monophonic or polyphonic. Since no frequency domain measurement is taken into account in the detection, the proposed method has no trade off between resolutions of time and frequency, and less complexity.

4. IMPLEMENTATION

To implement onset detection based on NAE, several additional techniques were used.

Trimming: keeps only instant power samples between the first and last two samples that are stronger than a certain proportion of the maximum. This procedure removes the recording noise that causes extra spurious notes at the start and end.

Time Warping: compacts instant power samples of one frame by the maximum of this frame with a certain overlap. It can be treated as a special over-sampling that significantly reduces the size of data.

The trimmed, warped instant power signal is not smooth enough to be forwarded to detection. Considering that different music at different sampling rates has different smoothness, we introduce an *Adaptive Smoothing* technique, where the length of the smoothing window is decided according to the smoothness of curve to be smoothed. Smoothing is repeated until the number of peaks in the curve maintains the same for two consequent applications.

Though the overall power evolution has little affect on onset detection based on NAE, since $NAE(t)$ is calculated with a right

open range of t, a very strong note can still overshadow a weak following note. It will be better if we can lower the strong notes and raise the weak ones to the same level. *Evening:* the power envelope achieves this to some degree by dividing the smoothed power envelope by its baseline.

Sharpening: sets the decay of the evened, smoothed power envelope to be exponential. This gives a clearer change in $NAE(t)$ between two consequent notes.

Refining: the detected onset time, $\{t_n ; n = 1, 2, \ldots, N\}$, can be refined by looking back into the smoothed power envelope to find a set of local minima. It can be refined still further by looking back into the instant power and repeatedly smoothing the corresponding frame of the instant power with the Hanning window of half of the frame length till the number of peaks in the curve is the same for two consecutive applications. With this baseline, the minimum, indicates time instant of the onset, $\{\hat{t}_n ; n = 1, 2, \ldots N\}$.

5. EXPERIMENTS

With a range of music pieces, we evaluated the proposed method using a measure of Correct Detection Rate (CDR), which is defined as:

$$CDR = \frac{total - missing - spurious}{total} \times 100\% \quad (14)$$

"*total*" takes the bigger of the number of true notes and number of notes detected. The music covers different instruments and genres and dynamic ranges. The first six songs are generated from midi files. Most of them are simple pieces. The rest are taken from the recordings of symphonies and pop song. *Ava Maria* of Bach is adagio of harp and violins. Pachelbel's *Canon* is the mixture of harp, violins and clarinet. Steely's song consists of drum, electric guitar, piano and voice. Cut-off for trimming, warping frame length and frame overlap are 1/20, 1ms and 50% respectively for all the songs. From the results we obtain, the CDR is very satisfactory.

6. CONCLUSIONS

Based on the observation that existing note onset detection methods detect the onset from the real energy of sound, we developed a new detection scheme using the NAE.

NAE focuses on the energy within one note, and thus it relatively insensitive to both the dynamic range of the song and whether the song in question is monophonic or polyphonic. The new strategy removes the threshold problem with a better detection performance for a large range of music pieces of different instruments. The proposed method has no trade off between time resolution and frequency resolution.

7. REFERENCES

[1] Klapuri, A., "Sound Onset Detection by Applying Psychoacoustic Knowledge", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3089-3092, 1999.

[2] Masri, P., Bateman, A., "Improved Modeling of Attack Transients in Music Analysis-Resynthesis", International Computer Music Conference, pp.100-103. Hong Kong; 1996.

[3] Monti, G., "Signal Processing and Music Analysis", First Year Report, Department of Electronic Engineering, King's College London, 2000

[4] Scheirer, E., "Tempo and Beat Analysis of Acoustic Music Signal", Machine Listening Group, MIT Media Lab., 1996.

Song	True	Detected	Missing	Spurious	CDR
Bass	15	16	0	1	93.75%
Piano	32	32	0	0	100%
Trombone	22	22	0	0	100%
Ocarina	52	52	0	0	100%
Banjo	52	52	0	0	100%
Clarinet	52	52	0	0	100%
Violin (vivacissimo)	809	832	12	35	94.35%
Ava Maria (Bach)	813	907	23	71	89.64%
Canon (Pachedel)	556	825	21	248	67.39%
Rikki Don't Lose That Number (Steely Dan)	68	71	3	6	87.32%

Table 1: *Summary of detection result based on NAE.*

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

APPLICATION OF BAYESIAN NETWORKS TO AUTOMATIC RECOGNITION OF EXPRESSIVE CONTENT OF PIANO IMPROVISATIONS

Luca Mion

CSC - DEI, University of Padua
Via Gradenigo 6/a, 35131 Padova, Italy
randy@dei.unipd.it

ABSTRACT

Expressive content detection in human communications can be approached from various perspectives. We focus on the expressiveness conveyed by subjects while performing piano improvisations. A model based on Bayesian Networks for the automatic recognition of a performer's expressive intentions will be presented. Performances were inspired by a set of eight sensorial adjectives, and a Bayesian Network for each of the suggested intentions was designed. Improvisations of six performers were recorded and analyzed yielding several relations between expressive intentions and acoustic parameters. Factor analysis on perceptual evaluation of performances determined in how many dimensions listeners discriminate the performances. This analysis was used to set the conditional probabilities among the nodes of each network. The purpose is to learn Bayesian Networks according to listeners' judgement, in order to make networks' behaviour as similar as possible to human perception. The whole set of analysis and methodologies will be described, explaining how results are taken into account to learn the networks. Results show that better automatic recognition is gained when intentions can be explained in terms of variations of acoustic parameters within a short time window. When performers used very different strategies for rendering a given expressive intention, the corresponding network presents uncertain decision rules, thus making the recognition harder.

1. INTRODUCTION

The necessity of algorithms for the automatic recognition of contents in multimedia objects is pointed out, from a side, by the spreading of systems for the document indexing used by search engines or standards for the fruition of multimedia contents such as mpeg7, and by the other, by the development of human-machine interfaces, based on expressive paradigms. As regards musical media, several works aim to recognize musical event such as notes or motifs ([1], [2]), other works aim to identify expressive performance rules [3], but very few works aim to recognize how (i.d. with what expressive intention) a musical gesture is performed. Dannenberg [4] developed a system for the automatic recognition of some characteristics in trumpet improvisations. The system was based on an empiric approach and works with a limited number of adjectives. In this system, a different approach has been used, starting with some works concerning the transmission of the expressive intents in music, such as [5], [6]. These works analyze the main musical parameters and their relations with the musician's expressive intentions, defining a multi-level model of abstraction, but without defining a procedure for the automatic analysis of the expressiveness. This research is part of a project founded by the European Community, that aims to design a conceptual framework and a software/hardware environment for the developing of multimodal interactive applications. The project is based on the concept of expressive gesture (movement or musical gesture) and is focused on understanding the way and with which expressiveness a gesture is performed [7]. One problem to face is that expressiveness is a weakly defined concept, as it is dependent on subjective choices and personal bents. For the above reason, the design of the algorithm for the automatic recognition should take into account aspects related to human perception and all the interactive systems should be validated by human judgments. This work follows a previous experiment on piano improvisations [8], in which a perceptual test has been carried out; following factor analysis and all the other used analysis on performances will be described, understanding how they fix the topology and the decision rules for the Bayesian networks.

2. EXPERIMENT AND METODOLOGIES

Six performers (three pianists and three non-musicians) have been invited to improvise, in separate sessions, on a digital piano according to eight expressive intentions suggested by means of as many groups of adjectives. Three adjectives have been chosen for each group in order to better specify each expressive intention. The eight groups are: 1. (SLA) Slashing, Impetuous, Resolute; 2. (HEA) Heavy, Hard, Rigid; 3. (HOP) Hopping, Galloping, Springing; 4. (VAC) Vacuous, Hesitant, Tired; 5. (BOL) Bold, Torrential, Unbridled; 6. (HOL) Hollow, Solemn, Obscure; 7. (FLU) Fluid, Fluent, Fleetly; 8. (TEN) Tender, Sweet, Simple. Performers could choose a note but they could not play any melody (i.e. they had to improvise without changing the pitch). The experiment had been planned in four phases, in order to progressively limit the musical means available to the performer. In this way, it was possible to evidence the relations between expressive intentions and each of the sonologic parameters pitch, intensity, articulation and tempo, that were one by one excluded.

Phase 1. Performers could choose a note but could not develop any melody (they had to improvise without changing the pitch). The sonologic parameters they could use, in order to communicate the eight expressive intentions, were: pitch of the chosen note, intensity, articulation and tempo.

Phase 2. The note to play has been imposed to middle C. Now, performers should improvise acting on intensity, articulation and tempo.

Phase 3. Besides imposing the same note (middle C) the intensity (volume) has been fixed to a key velocity value = 100. Performers, then, should express all eight expressive intentions by acting just

on articulation and tempo.

Phase 4. Finally only the tempo parameter was available. In fact, a middle C with a fixed duration (260 ms) has been pre-recorded: the duration of the sound uttered did not change if the piano key was longer pressed. As for phase 3, pitch and intensity have been imposed too. Performers were left free to repeat at will their improvisations, in the desired order concerning the eight expressive intentions, but respecting the given sequence of the four phases, until they were pleased with the recorded performances. The length of each performance was indicated as no shorter than six seconds. Data have been analyzed by a perceptual test (2.1), a mean values' analysis (2.2), a factor analysis on expressive profiles (2.3), and note-by-note analysis (2.4).

2.1. Perceptual test

Listeners' perception of performers' expressive intentions had been analyzed; fifteen subjects without any specific musical education participated to the perceptual experiment. The task consisted in qualitatively describing the expressive content of the 32 improvisations (8 performances*4 phases), realized by a professional pianist (R.Basilicati), by means of the same semantic groups previously proposed to the performer. The order of the four phases was maintained in the listening session, but listeners were not informed about the nature of the experiment. A factor analysis on listeners' responses followed, taking the expressive stimuli as variables. The analysis had the purpose of understanding how similar or different the expressive performances were perceived by listeners and determined in how many dimensions listeners discriminate the performances. In our case, only the factors with eigenvalues greater than 1 were taken into account, and then associated to the sonologic parameters that better explain each dimension. We have never found more than three factors per phase, and in most of the cases we have found two significant factors. In the case of the phase 1, we obtained 2 significant factors (see figure 1). The linear combination of these two factors yields a bi-dimensional space where each axis represents a factor. The factor analysis results in performances scores, which are expressed in terms of the just obtained factors; this fact allows us to plot the position of each performance in the space. In figure 1 these factors are depicted, and the analysis of the mean values can reveal relations among acoustic parameters and factors, which, consequently, can assume a meaning. For example, along the horizontal axis, the Fluid, Hopping, and Bold intentions show the lowest scores, while the Heavy and Hollow intentions show the highest values. Considering that the Fluid intention was played with a fast tempo, and that the Hollow was played with a slow tempo, we can reasonably associate this axis to tempo parameter. We found intermediate values of tempo for Tender, Vacuous and Slashing, confirming the consistence of the meaning we gave to this axis. On the other axis, intentions Slashing and Bold have the highest scores, while Vacuous and Tender intentions have the lowest scores; these intentions show very different intensity values, thus we related this axis to the Intensity. Intermediate values were found for Hollow, Fluid, again confirming the consistence of the interpretation of this factor.

2.2. Mean values'analysis

For each performance, a mean values' analysis and the ANOVA test have been done, quantitatively evaluating the sonologic parameters as it follows: Intensity directly evaluated by means of the Key Velocity value (KV) of the MIDI files. It ranges between 0

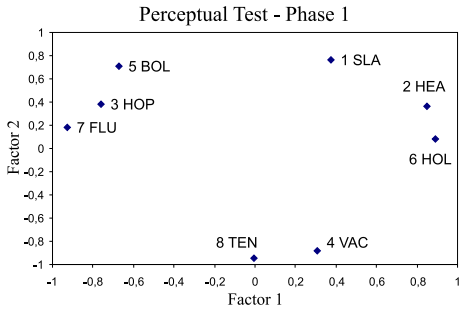


Figure 1: *Factor analysis on perceptual test on phase 1. Factor 1 opposes performances 3, 5 and 7, on one extremity, with performances 2 and 6 on the other, clearly indicating a style characterized by the Tempo parameter (fast performances 3, 5 and 7, vs. slow ones 2 and 6). Factor 2 opposes performances 1 and 5 with performances 4 and 8, defining a style characterized by the Intensity parameter (high volume of performances 1 and 5, vs. low volume of per. 4 and 8.)*

(no volume) and 127 (maximum volume). Articulation evaluated by the legato (LEG) parameter, defined as the ratio between the duration of the first note (D) and the Inter-Onset-Interval (IOI) of the two notes (IOI=duration between the beginning of the first note and the beginning of the second note). LEG values range between 0 (minimum legato, or staccato, → high articulation) and 1 (maximum legato → low articulation). Tempo evaluated as the (mean value) number of Notes Per Second (NPS). We assumed that those intentions which show highest and lowest mean values of KV, LEG and NPS, have a more direct and stronger relation with intensity, articulation and tempo respectively; while those intentions which present medium values have a weaker relation. An important fact comes from ANOVA test: pitch and intensity values were often chosen similar by performers, whereas articulation was used with more personality. The tempo parameter is the one that showed the maximal concordance with the perceptual test.

2.3. Factor analysis on expressive profiles

On the basis of a previous work [9], we made a factor analysis on expressive profiles for each of the parameters. This factor analysis requests as variables the eight performances, and take the notes as statistical units; in this way it is possible a statistical survey for each parameter, once more with the purpose of understanding how the expressive performances are similarly or differently played by Basilicati, and determines in how many directions and dimensions the musician discriminates the performances, according to the particular sonologic parameter. Principal factors for tempo parameter are extracted in phase 4, for articulation in phase 3 and for intensity in phase 2.

2.4. Note-by-note analysis

Finally, a note-by-note analysis showed the presence of any relationships among the parameters profiles. This analysis yielded some interesting result: for example, in the case of the hopping performance, intensity and the articulation profiles revealed a close proportionality.

3. GRAPH STRUCTURE

The general structure of the Bayesian networks is shown in figure 2. There were six-performers' performances available, but we decided to make the analysis only on the Basilicati's ones: we had to made this restriction because the analyzes involved were highly time consuming, thus we choose Basilicati supposing his higher bent to interpretation.

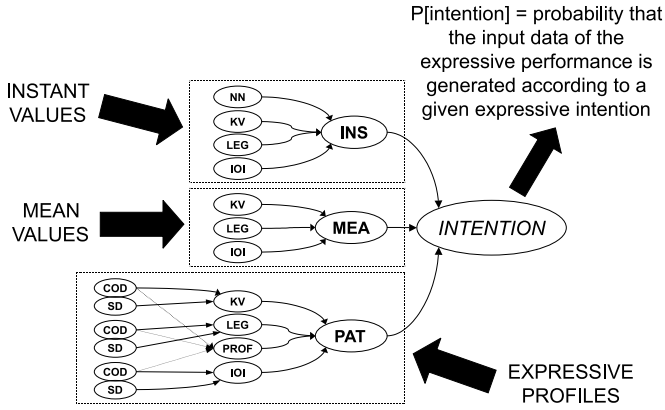


Figure 2: *The Bayesian network structure. It's possible to see the three main categories of cues used (in the rectangles): the instant values, the mean values, and the pattern. Marginal nodes (on the left) are inferred firstly. Each node conditions its descendant nodes until node intention, which value is the probability that input performance is played with a given expressive intention.*

On the basis of the previous analysis, we organized the topology of the Bayesian Networks (see figure 2), and each of them is intended to recognize a given expressive intention: each net is build using 21 nodes and organized in three main parts. In the upper side of the network, the sonologic parameters are evaluated by their instant values; the four upper nodes manage pitch, key velocity, legato and tempo respectively. In the second part, mean values (KV, LEG and IOI) are taken into account while the third part is devoted to evaluating if the input data contains some relevant patterns which can be used to recognize the intention. We assume that each node is discrete and binary, thus the number of possible values it can take on is two. Marginal nodes are inferred firstly, and they lie at the left of the network; once they are observed they have only one value. Regarding instant and mean values, the marginal nodes are properly set according to the comparison between input data and Basilicati's values; for each parameter we computed the 90% confidence intervals on the mean. Thus, if the input parameter aims to a value within its confidence interval, the output of the node is set to 1, otherwise to 0. Regarding marginal nodes of expressive profiles' section (the lower on figure 2), we set them on the basis of a factor analysis inspired by a previous work [9]. The factor analysis we made on expressive profiles yielded some relevant pattern for each parameter. Actually, not all the adjectives have pointed out remarkable patterns; this fact is evidenced by the arcs between nodes COD and PROF. If present, the patterns we found for a given parameter become the reference patterns; if a pattern sequence is not present, in the related network we disconnected that arc using only Standard Deviation features. Once a pattern has been found, input parameters KV, LEG and IOI are coded by a three-level-discretization (low, medium, and high) within an 8-note window, and their Standard Deviations are computed. For example, the pattern discretization for the timing profile when the IOI among the notes is low-high-low-high-etc. becomes 0-2-0-2-etc. Then, the value of nodes COD (Codification) and SD (Standard Deviation) are inferred to 1 or 0 according to the similarity among the just computed discretizations and reference ones: if two discretized profiles show 5 equal values, then the node is set to 1, otherwise to 0. Standard deviations were evaluated in the same manner as the upper instant values using 90% confidence intervals. Finally, any relationship among the parameters is observed in terms of proportionality of their expressive profiles by a note-by-note analysis; any relation found is compared to reference relationships allowing node PROF to be set. Regarding inner nodes (INS, MEA, PAT and INTENTION), each of them has a Conditional Probability Table (CPT) which defines the probability distribution of the node given its parents; CPTs are thought of as multidimensional arrays, whose cells are filled with the whole set of combinations of parent values. Each combination gives a fixed value of probability to the descendant node. To fill these tables, we detected the rules among nodes parents/descendants; founding our work basically on the perceptive test, we hand-filled the CPTs after qualitative observations of perceptive test results and considerations about the different phases. As we already mentioned, the experiment has been planned in four phases; each of the sonologic parameters (pitch, intensity, articulation and tempo) were one by one excluded. In order to evidence which are the relevant parameters used to play with a given expressive intention we used the differences in the listening results between phases. For example, let us consider the Hollow adjective: in the phase number 2 the pitch choice was eliminated, by fixing it to middle C; this resulted in the perceptual test being at a lower rate of recognition of this intention. Thus, we argued that pitch was a significant parameter to convey the Hollow intention. In the same manner we understood that the tempo is another important parameter. Articulation and intensity were found to be less important than the other parameters. We decided to express this fact setting the weights of the parameters respectively to 0.4 for pitch, 0.4 for tempo, 0.1 for articulation, and 0.1 for intensity (the values have been chosen qualitatively). The same kind of reasoning has been made to assign the values of the CPTs of the other nodes (MEA and PAT). Regarding the weight to assign to nodes INS, MEA and PAT in order to fill the CPT of the node INTENTION, we used ANOVA test results. The higher the significance of the mean, the higher weight we assigned to the node, i.e., for Hollow intention we found high significance for instant and mean values, and the pattern had low significance; we express this fact giving to nodes INS, MEA and PAT weights 0.4, 0.4, 0.2 respectively. The three analyzes are joined together, to provide as output the probability that the performance in the input is played with the given expressive intention. According to independence assumption among the nodes, one can compute the final probability as follows:

$$P[\textit{intention}] = \sum_{cue} P[\textit{intention}|cue]P[cue] \qquad (1)$$

Where cue are the three main categories of cues used: instant values, mean values, and pattern. The computation has been made using a suitable Toolbox for Matlab [10].

4. VALIDATION

Each network was tested with the whole set of Basilicati's performances. Thus, we performed 64 simulations (8 networks subjected to 8 performances), then for each simulation we computed the mean output probability; the table 1 shows the results.

Input Data

Networks	SLA	HEA	HOP	VAC	BOL	HOL	FLU	TEN
SLA	0,70	0,46	0,64	0,49	0,54	0,35	0,52	0,48
HEA	0,36	0,77	0,21	0,25	0,34	0,13	0,21	0,28
HOP	0,68	0,76	0,77	0,48	0,67	0,50	0,72	0,67
VAC	0,34	0,20	0,27	0,41	0,21	0,48	0,40	0,54
BOL	0,63	0,52	0,58	0,08	0,69	0,39	0,61	0,31
HOL	0,35	0,48	0,25	0,69	0,36	0,89	0,29	0,25
FLU	0,59	0,39	0,58	0,42	0,57	0,49	0,83	0,59
TEN	0,44	0,49	0,37	0,26	0,40	0,29	0,56	0,52

Table 1: *Mean probabilities given by Bayesian networks using the eight performances as input: the cell (x, y) contains the mean P value given by network x subjected to performance y; the values on the diagonal represents the mean values given by networks subjected to the performances they were trained with.*

As shown, most of the networks give higher probability (gray cells) when they are subjected to the data used for their training. Hollow and Fluid networks give the highest value of probability to the corresponding performances; Vacuous and Tender networks confuse the input data, giving higher probability to other intentions (circled cells). Anyway, Tender network recognize the correspondent performance as second (circled beside). We noticed that intentions confused in perceptual test (as mean and factor analysis revealed) caused difficulties in assigning decision rules in their networks; thus, those intentions are recognized with difficulty. For example, we can see that Vacuous and Tender networks give low reliability to their recognitions indeed. This is an interesting result: it underlines that these Bayesian networks' behavior is similar to human perception. As the ANOVA test revealed, articulation was used with more personality and the listeners in perceptual test confused performers' intentions: adjectives characterized the most by the articulation (Vacuous and Tender) are difficult to be recognized. As we noticed instead, performers' choices in terms of pitch, tempo and intensity were recognized by listeners and decision rules in correspondent networks were easily found. For example, Hollow (low pitch, low tempo) and Fluid (low intensity, high tempo) intentions are recognized by correspondent networks.

5. CONCLUSIONS

A model for expressive intentions' recognition on musical improvisation has been presented. Slashing and hollow intentions give high values of probability for their performances; vacuous and tender intentions were confused both in perception and in factor analysis, in fact networks 4 and 8 give low reliability values for their recognitions. Anyway the Bayesian network approach need further works to find other intentions' features in terms of pattern and relations between the parameters' profiles, which showed to be a very important aspect. For example, an integrated approach using also the HMMs could give us more information regarding the expressive profiles' timing development. Another imperfection is due to the adjectives' choice: we had just sensorial adjectives in this task, while in next works will consider affective adjective discovering further relationships among the parameters.

6. ACKNOWLEDGEMENTS

This research was supported by the EC project IST-1999-20410 MEGA and by the EC project HPRN-CT-2000-00115 MOSART.

7. REFERENCES

[1] Melucci, M., Orio, N. and Gambalunga, N., "An Evaluation Study on Music Perception for Content-based Information Retrieval", Proc. of International Computer Music Conference, Berlin, Germany, pp. 162-165, 2000.

[2] Cambouropoulos, E., "The Local Boundary Detection Model (LBDM) and its Application in the Study of Expressive Timing", Proceedings of the International Computer Music Conference (ICMC 2001), 17-22 September, Havana, Cuba, 2001.

[3] Bresin, R. and Friberg, A., "Espressive musical icons", J.Hiipakka, N.Zakarov and T.Takala (Eds.), Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29 - August 1, pp. 141-143, 2001.

[4] Dannenberg, R., Thom, B. and Watson, D., "A Machine Learning Approach to Musical Style Recognition", 1997 International Computer Music Conference, International Computer Music Association (September 1997), pp. 344-347, 1997.

[5] Canazza, S., De Poli, G., Drioli, C., Rodà, A. and Vidolin, A., "Audio morphing different expressive intentions for Multimedia Systems", IEEE Multimedia, July-September, (7)3:79–83, 2000.

[6] Gabrielsson, Alf, Lindstrom and Erik, "Emotional expression in synthesizer and sentograph performance. Psychomusicology, 14(1-2):94–116, 1995.

[7] Cirotteau, D., De Poli, G., Mion, L., Vidolin, A. and Zanon, P., "Recognition of musical gestures in known pieces and in improvisations" V International Gesture Workshop, Genoa, Italy, 2003.

[8] Bonini, F. and Rodà, A., " Expressive content analysis of musical gesture: an experiment on piano improvisation. Workshop on Current Research Directions in Computer Music, Barcelona, Spain, 2001.

[9] Repp B. H., "Patterns of expressive timing in performances of a Beethoven minuet by nineteen pianists", Journal of Acoustical Society of America, Volume 2, pp. 622-641, August 1990.

[10] Murphy, K., " The Bayes Net Toolbox for Matlab, in Computing Science and Staistics , vol. 33, 2001.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

EPS - AN INTERACTIVE COLLABORATIVE GAME USING NON-VERBAL COMMUNICATION

Marie-Louise Rinman[1], Anders Friberg[2], Ivar Kjellmo[3], Antonio Camurri[4], Damien Cirotteau[5], Sofia Dahl[2], Barbara Mazzarino[4], Bendik Bendikesen[3], Hugh McCarthy[5]

[1] Centre of User Oriented IT-design, KTH, Stockholm, rinman@nada.kth.se

[2]Speech Music and Hearing, KTH, Stockholm, [andersf, sofia.dahl]@speech.kth.se

[3]Octaga / Telenor, Oslo, [ivar.kjellmo, bendik.beniksen]@octaga.com

[4]InfoMus Lab, DIST – University of Genoa, Genoa, [music, bunny]@dist.unige.it

[5] CSC – Center of Computational Sonology, DEI – Dept of Information Engineering, University of Padua, [Cirotteau, maccarthyhugh]@csc.unipd.it

ABSTRACT

The interactive game environment EPS (expressive performance space), presented in this short paper, is a work still in progress. EPS involves participants in an activity using non-verbal emotional expressions. Two teams use expressive gestures in either voice or body movements to compete. Each team has an avatar controlled either by singing into a microphone or by moving in front of a video camera. Participants/players control their avatars by using acoustical or motion cues. The avatar is navigated /moving around in a 3D distributed virtual environment using the Octagon server and player system. The voice input is processed using a musical cue analysis module yielding performance variables such as tempo, sound level and articulation as well as an emotional prediction. Similarly, movements captured from the video camera are analyzed in terms of different movement cues. The target group is children aged 13-16 and the purpose is to elaborate new forms of collaboration.

1. INTRODUCTION

Participants involved in the EPS game use (emotional) non-verbal communication in music and dance to control and interact with an avatar. Expressive Gesture, as applied in EPS, is using recent models of expressive communication in music and dance developed by Camurri and others [3,] [1], [4] and [8] in the on-going EU project MEGA. These models are based upon a large body of research in expressive communication [9], [5].

The aim of the EPS game, and presented in this paper, is to elaborate new forms of collaboration for children as well as to test the applicability and intelligibility of expressive gesture in a new setting. First, could expressive gesture, e.g. music and dance, be used as a basis for (audience) collaboration and participation? Second, could it be used as an input control/means of communication in an interactive game environment? And third, the benefits of using a game format when young people are involved.

2. THE EPS GAME

Main features and aims

A multi-user multi-relationship real-time activity including multisensory interaction and multimodality.

Non-verbal communication:

Both in the interaction with the 3D world projected on a screen and between audience and players. Non-verbal expression in voice and body movements.

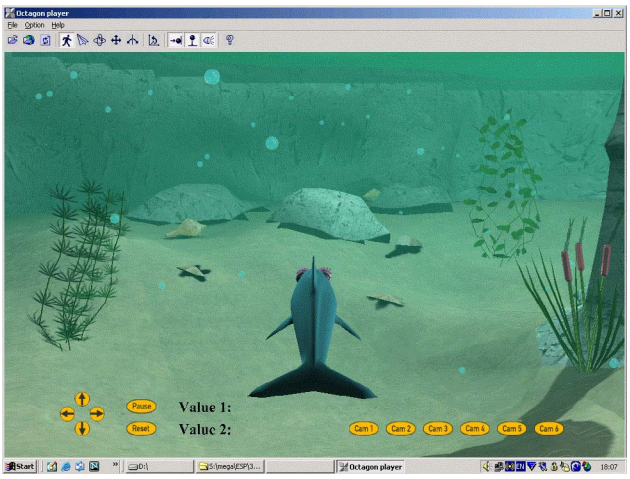


Fig. 1. *Screenshot from the players' point-of-view of one of the teams.*

Collaboration:
Between team members as well as between audience and teams. The audience influences the speed of the avatar and creates sound through their movements.
Mixed-reality features:
Face-to-face meetings between players in the teams as well as between teams and public.
Interaction with a virtual world. Each team has an avatar controlled by one of its players though expressive sounds and movements.
Task:
The main task is to find three caves in the virtual world and solve the riddles/tasks presented in each cave assisted by the audience. To get into the caves teams have to collaborate by solving the problem how to open the gates. The team that finds to the third and last cave first may enter without waiting for the competing team – and wins.

EPS allows participants to interact with each other using verbal communication as well as interacting with a 3D environment using non-verbal communication. The participants are divided into two competing teams (Team 1 and Team 2. There are three players in each team) and members of each team use expressive gestures in voice and in body movements. One team member in each controls an avatar by singing into a microphone or by moving in front of a video camera. Meanwhile the other two team-members may discuss alternative solutions helping the navigator executing his or her task. The avatar changes position and (in a later version of the game) colour, shape and size depending on the acoustical and motion cues conveyed by the participants/players. The first challenge for each team is to figure out how to use the gestures to control the position (and shape) of the avatar in order to play. An example of the controllers viewpoint is shown in Figure 1.

Members of the audience are divided into two groups and invited to take part in the game (Audience group 1 and Audience group 2). Their task is to serve as extra helpers for each team. The audience groups generate a music background by moving and dancing. They may follow the game via the two screens onto which the representations of each team are projected. They also have access to a screen that shows the full overview of the 3D environment. Figure 2 gives an overview of the game setup. The two screens on stage show a third-person view of the virtual world. The audience screen is placed above the stage showing an overview of the complete virtual world.

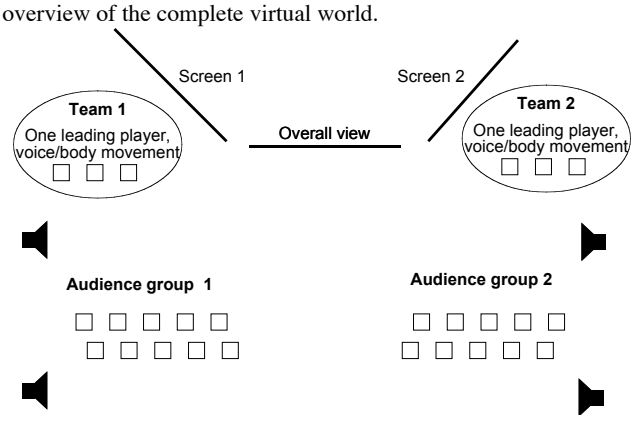


Fig. 2. Overview of game setup.

3. EXPRESSIVENESS IN MUSIC AND DANCE

Music performances can be described by their performance variables such as tone interonsets (IOI), tone sound level, articulation, and intonation. Overall measures of these performance variables over a section, henceforth as cues, have a strong coupling to the performed and perceived emotional expressive content in the performance [7, 8]. For example, happiness is communicated by fast tempo, staccato articulation, high sound level, and fast tone attacks, while sadness is communicated by slow tempo, low sound level, legato articulation, and slow tone attacks. These rather general descriptions have been collected from a large number of investigations and are surprisingly consistent for different music examples, performers or listeners.

Dance performance may be analysed in a similar way using overall motion cues such as quantity of motion, velocity, amount of motion/rest, and fluency [2]. Here too, there is a coupling between these cues and the expressive character of the dance according to Camurri, Lagerlof and Volpe [3].

Recently, real-time computer tools for analysing expressive gestures in music and dance have been developed within the MEGA project (www.megaproject.org). These tools have been developed within the software platforms EyesWeb and pd and are used as a starting point in the non-verbal input modules in the game. The body gesture analysis package in EyesWeb has a large number of different gesture cues and is using imaging algorithms applied to video input. The musical gesture analysis is using signal processing algorithms applied on audio input and outputs both the basic musical cues such as tempo and sound level, but also the emotional intent [2, 6].

4. TECHNICAL IMPLEMENTATION

The virtual 3D environment is run in the Octagon server and player system. This system is a distributed virtual reality system, that uses MPEG4 and VRML file for-mats. Multiple users are represented in the virtual world by their own figure or avatar. Here they are present in the same shared virtual environment and can interact with each other. The distribution goes through a network using the MPEG4 protocol. (More on Octaga and the Octagon system at: www.octaga.com)

There are two types of expressive input: voice or movements. An avatar can be controlled by either voice or movement input using expressive features in each do-main. The voice of the player is captured by a microphone connected to a computer.

The player is supposed to sing or make sounds and different music performance variables are analysed using the musical cue analysis software [6]. The movement of the player is captured by a video camera connected to a computer. The player is supposed to move/dance and different overall motion cues are analysed using a cue analysis patch implemented in the EyesWeb software.

EyesWeb is connected to each Octagon client so the clients can receive expressive parameters. In this case the avatars receive the parameters and change according to them.

5. USING CONVENTIONS IN GAMES AND NARRATIVES TO ENGAGE PARTICIPANTS AND ORGANISE PARTICIPATION

The aim of the activity is to encourage collaboration between young people coming from different cultural and social backgrounds taking specifically language and communication problems into consideration. The reason why a game format is used is that it is familiar to the target group and may thus serve the purpose of establishing a common ground [10].

How could participation in complex settings involving young users be facilitated? Conventions used in games (rules for instance), narratives (a protagonist – an avatar; a story line; organizing the story and participants roles and actions) and drama (three caves, three acts; limitations in time and space; beginning, middle and end) could be used: a) to design participants' actions and responses; b) to create engaging tasks for the target group and allow a certain amount of free space for them to interpret and solve them in a meaningful way; c) to encourage collaboration. The properties of networked computer games are familiar to most 13 – 16 year olds. A large number consists of a mix of teamwork, competition and action (such as *Capture The Flag,* a networked multi-player version of the computer game Quake).

Children having problems with expressing themselves verbally due to language problems may interact with the 3D world in EPS as well as with co-players using other ways of expression such as singing, dancing, and/or rhythmic sounds (clapping their hands etc.). They may solve riddles and overcome obstacles individually as well as together with other players. The use of non-verbal communication in expressive gesture (movements and sound) will hopefully encourage participants to use other forms of communication such as the written and spoken language.

Finally the use of a theatre and game metaphors influences the designer's approach to the user. In an interactive game environment the user becomes a member of an audience, a participant or a player. They are made participants and become co-writers, co-directors and co-actors with a possibility to influence the outcome of the game [10]. An ideal situation would be to involve young participants at an early stage in the design process on order to find out what parts of the game/narrative they find engaging and attractive. The task and challenge for the designers then would be how to put all the bits and pieces together. As a suggestion a game setting may be constructed as a conventional play or story having a beginning, middle and end, containing obstacles, riddles and tasks to be solved, all put together into a meaningful whole.

6. ACKNOWLEDGEMENTS

This work was partially supported by the EU project MEGA IST-20410.

7. REFERENCES

[1] Camurri A (2002). *Interactive Systems Design: a KANSEI-based Approach*, in NIME-02 Intl. Conference on New Interfaces for Musical Expression, Dublin, Ireland

[2] Camurri A, Coletta P, Mazzarino B, Trocca R, &Volpe G (2002). *Improving the man-machine interface through the analysis of expressiveness in human movement*. In Proc. Intl. Conf. IEEE ROMAN-2002, Sept.2002, Berlin. IEEE CS Press.

[3] Camurri A, Lagerlof I & Volpe G (in press). *Recognizing Emotion from Dance Movement: Comparison of Spectator Recognition and Automated Techniques*. International Journal of Human Computer Studies.

[4] Canazza S (in press). *An abstract control space for communication of sensory expressive intentions in music performance*. Computer Music Journal.

[5] Friberg A & Battel G U (2002). Structural Communication. In (R. Parncutt & G. E. McPherson, Eds.) *The Science and Psychology of Music Performance: Creative Strategies for Teaching and Learning*. New York: Oxford University Press, 199-218

[6] Friberg A, Schoonderwaldt E, Juslin, PN & Bresin, R (2002). *Automatic Real-Time Extraction of Musical Expression*. In Proceedings of the International Computer Music Conference 2002, San Francisco: International Computer Music Association, 365-367.

[7] Juslin P N (2001). *Communication of emotion in music performance: A review and a theoretical framework*. In P. N. Juslin & J. A. Sloboda (Eds.), Music and emotion: Theory and research (pp. 309-337). New York: Oxford University Press.

[8] Juslin, P N, Friberg A & Bresin, R (2002). *Toward a computational model of expression in performance: The GERM model.* Musicae Scientiae specal issue 2001-2002, 63-122.

[9] Juslin P N & Sloboda, J A, eds. (2002). *Music and emotion: Theory and research.* New York: Oxford University Press.

[10] Rinman M-L (2002). *Forms of Interaction in Mixed Reality Performance – A study of the artistic event Desert Rain.* Licentiate thesis, Royal Institute of Technology (KTH), Stockholm.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

MAPPING A PHYSICAL CORRELATE OF LOUDNESS INTO THE VELOCITY SPACE OF MIDI-CONTROLLED PIANO TONES

Tomoyasu Taguti

Department of Information Science and Systems Engineering
Konan University
taguti@konan-u.ac.jp

ABSTRACT

This paper presents a method of mapping a physical correlate of loudness into the velocity space for piano tones of MIDI-controlled musical instruments. The proposed method is based on (1) a listening experiment to determine the contour of equal loudness, and (2) a measurement of the values of the concerned physical correlate to the loudness. To fix the method, the equivalent continuous A-weighted sound pressure level (L_{Aeq}) was taken as the physical correlate. The experiment was done with a particular sound synthesis module. The paired comparison was conducted on the notes in the range C3 $\sim$ C7 by setting C4 of different velocities as the standard stimuli, where the durations of the standard and comparison stimuli were 500 ms each, being separated by a 500 ms silence. The subject was a female who was expert in piano playing. By using the L_{Aeq} values in dB of note C4 as the interval scale, an interpolation formula that maps dB to velocity was established. This formula was implemented as part of a computer-aided piano performance system so that the dynamic expression is specified in the units of dB in the above sense.

1. INTRODUCTION

The sound intensity (or sound volume) of MIDI-controlled musical instruments (MIDI instruments, for short) are specified with a 7-bit integer code called the velocity. But the MIDI standard document does not define quantitatively the acoustical characteristic of the velocity, only saying that the larger the velocity the larger the volume of the generated sound, with a suggestion of use of a logarithmic scale for its dependence relationship.

This vague functional definition of velocity creates a problem in dynamic expression reproduced on different MIDI instruments, that is, the same sequence of MIDI codes results in different responses in dynamics.

Further, It seems rather common in the present-day MIDI instruments so that the same velocity number given to different notes, even to the notes of simitone intervals, on a particular MIDI instrument does produce a different sound intensity (or volume, loudness, or what may be termed for similar auditory impression.) Manually adjusting such an uneven response of the velocity leads to almost an endless task in creating music performance.

This paper proposes an experimental method to regulate the response of sound "volume" on the piano tones of MIDI instruments by software. The method is based on (1) determining the contour of equal loudness by the method of paired comparison, and (2) representing the velocity numbers in terms of a physical correlate of loudness.

On point worth to mention is that we have to admit a practical limit in the precision of such an experiment for the following reasons: first, since we are dealing with the waveforms of real piano tones in a sense, as contrasted to those of test tones prepared for a psychoacoustic listening experiment in the laboratory, the temporal development of their amplitude, as well as their harmonic construction, varies note by note. Second, since the MIDI velocity system takes on integral numbers at 127 steps, the degree of resolution of the loudness is limited; even worse is that there seems to be cases where some MIDI instruments assign a group of several consecutive velocity numbers to the same waveform (i.e., to exactly the same tone).

Our experiment was made on the piano tones, whose timbre is called "stretched tuning grand piano stereo", of a MIDI instrument made by a Japanese manufacturer. Let us call this instrument Module M. The instrumental parameters were set to the initial, factory setting that turns off all the supplemental sound effects and equalizing functions.

2. LOUDNESS MATCHING BY THE METHOD OF PAIRED COMPARISON

2.1. Method

Note C4 at several velocities were taken as the standard stimuli, and each of C3 $\sim$ C7 (the comparison stimuli) was matched to C4 by adjusting the velocity of the comparison stimulus. Because of ease of the experimental task, however, for matching of a comparison note whose pitch is higher by more than one octave from C4, the closest C from below is chosen as the substitute for C4. (The detail of this substitution will be given below.)

Subject: A female student majoring piano playing in a graduate school of music education.

Listening environment: The experiment was conducted in a low reverberant listening room (volume 61 m^3, reverberation time 0.3 s at 500 Hz, background noise level 26 dBA). See Fig.1 for the setup of the experiment. The output level of the stereo loudpspeakers was set so that the sound of C4 at velocity 127 becomes 90 dB SPL (fast) at the listening point of the subject. Then the subject was listening to the stimulus pair by controlling the loudness by herself with velocity numbers step by step according to the experimental procedure which follows.

Stimuli: **A** and **B**, respectively, are the standard and comparison stimuli with velocities V_A and V_B, each taking a 500 ms duration separated by a 500 ms silence. The duration 500 ms was chosen by taking into consideration (1) the effect of temporal summation of loudness [1], and (2) the fact that it is the duration of a quarter

note played at 120 beats per minute. The silent interval 500 ms was chosen by following [2].

A and **B** were arranged in six ways for the experiment as:

A A A A
B B B B
A B A B
B A B A
A B A B A B
B A B A B A

where these stimuli (a 500 ms duration each) were separated by a 500 ms silence. The subject was allowed to listen to any of these sequences as many times as she became confident.

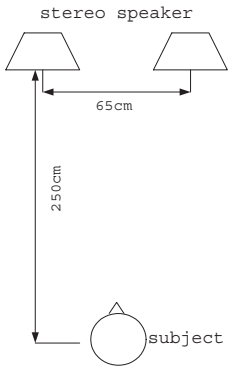


Figure 1: *Setup of the listening experiment.*

Procedure: The subject was asked to choose a best-fitted velocity of the comparison stimuli, $\bar{V}_B$, by adjusting V_B's so that it becomes to be subjectively equal to that of **A**, by following the method of limits.

The stimulus pairs: The comparison stimuli were notes C3 $\sim$ C7, and V_A (the velocity of C4) were 20, 40, 60, 80, and 100, where pairs **A** and **B** were taken from (1), (2), (3), and (4) in this order from the table below.

	A	**B**
(1)	C4 with V_A	C4#, $\cdots$, C5
(2)	C5 with $\bar{V}_B$ †	C5#, $\cdots$, C6
(3)	C6 with $\bar{V}_B$ ‡	C6#, $\cdots$, C7
(4)	C4	B3, $\cdots$, C3

†) The velocity number chosen at (1)
‡) The velocity number chosen at (2).

2.2. Result

There were 240 pairs for comparison in total, among which 216 cases yielded a single number of $\bar{V}_B$ while 24 cases yielded two consecutive integral numbers of $\bar{V}_B$. For the latter cases, the average of the two consecutive integral numbers was taken as the effective $\bar{V}_B$ for further proceesing.

Let $v = \bar{V}_B$ be written in the form of

$$v = f(n, V_A) \tag{1}$$

where n is the note of comparison stimulus. Here, we assume the trivial relation that $V_A = f(\mathrm{C4}, V_A)$.

Figure 2 summarizes the result of the listening experiment. This figure will be called the equal loudness contour of velocity.

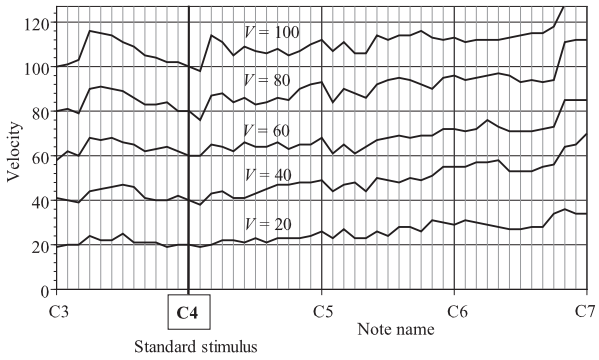


Figure 2: *Equal loudness contour of velocity for $C_3 \sim C_7$.*

3. DEPENDENCE OF L_{AEQ} ON THE VELOCITY

Several physical correlates to the loudness of unsteady sounds are known. We have chosen an equivalent continuous A-weighted sound pressure, L_{Aeq}, as the physical correlate to the present problem. But its absolute value could not be obtained because the sound source was the analog output singnal of Module M. So, the energy of their A-weighted waveforms were used in reference to that of C4 at velocity 40.

Measurement: 88 notes, i.e, n = A0, A0#, $\cdots\cdots$, and C8; 14 steps of velocity as V = 5, 10, 20, 30, $\cdots\cdots$, 110, 120, and 127. These velocity numbers were labeled V_i such that V_1 = 5, V_2 = 10, $\cdots\cdots$, and V_{14} = 127.

Energy computation: The analog, stereo waveforms were sampled digitally at 48 kHz sampling, and were mixed down to a monoral singal, $p(t)$. Then $p(t)$ was processed to yield an A-weighted signal, $p_A(t)$.

Compute the following quantity of each waveform with T = 500 ms

$$I_{n,V} = \max_t \int_t^{t+T} |p_A(s)|^2 \, ds \tag{2}$$

and get ℓ_{n,V_i}

$$\ell_{n,V} = 10 \log(I_{n,V} / I_{\mathrm{C4},40}) \tag{3}$$

for all pairs (n, V_i). Figure 3 shows the quantity $\ell_{n,V}$ as a function of n, n = A0, A0#, $\cdots\cdots$, and C8.

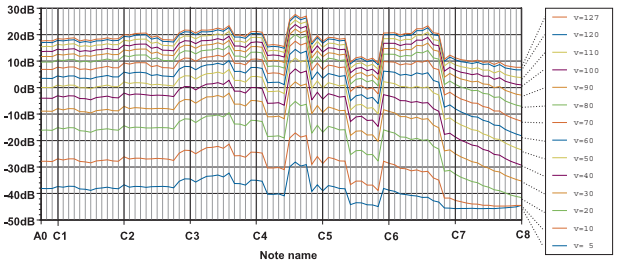


Figure 3: *Equal velocity contour of L_{Aeq} (dB).*

The figure reveals that the dynamic range of Module M is about 50 dB in the low to middle registers, and is about 40 dB in high registers, in terms of L_{Aeq}.

4. VELOCITY OF EQUAL LOUDNESS

Let us take $\ell = \ell_{C4,V}$ for the interval scale variable representing the loudness, where ℓ = 0 dB corresponds to the loudness of note C4 at velocity 40 by the present definition.

The function $v = g(n, \ell)$ determined by the following interpolation procedure gives the velocity v of note n whose loudness is equal to that of C4 at ℓ dB.

1) Get i that satisfies $\ell_i \leq \ell < \ell_{i+1}$ where $\ell_i = \ell_{\mathrm{C4},V_i}$
2) Compute $a = (\ell_{i+1} - \ell)/(\ell_{i+1} - \ell_i)$
3) Compute $v = af(n, V_i) + (1 - a)f(n, V_{i+1})$

In order to utilize this interpolation formula effectively, we made a supplementary listening experiment for the velocity numbers V = 5, 10, 30, 50, 90, 110, 120, and 127. By merging these data with those of the previous experiment, we obtained a denser set of equal loudness contours.

The formula $v = g(n, \ell)$ is valid for $-20\,\mathrm{dB} \leq \ell \leq 20\,\mathrm{dB}$ approximately. The thick lines in Fig. 4 show the contours of equal loudness at $\ell = 0, \pm 10, \pm 20$ dB for notes C4 $\sim$ C6.

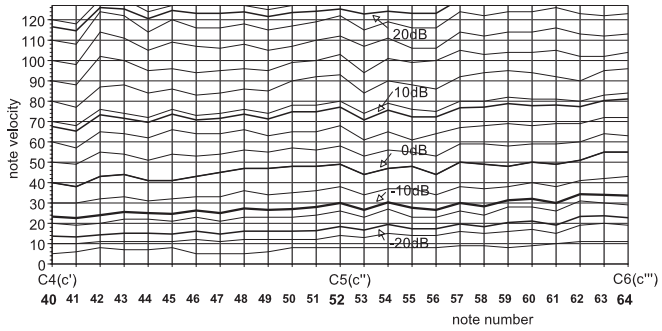


Figure 4: *Equal loudness contours in terms of $\ell = \ell_{C4,V}$ in the note–velocity space.*

5. CONCLUSIONS

We presented a method for describing the loudness characteristic of piano tone generated by MIDI instruments by use of the method of paired comparison on the loudness and its physical correlate.

The method cannot give a very accurate result in the sense of laboratory experiment on well-controlled testing tones, because the concerned sounds are pre-fixed real piano tones whose temporal development of waveform and harmonic construction differ note by note, and because the loudness is controlled with the velocity that takes on only 127 discrete steps. Nevertheless, it gives practically a uniform way of representation of loundess of piano tones not only on a particular MIDI instrument but also across over different MIDI instruments, provided that the necessary data are prepared for each of those MIDI instruments.

The given interpolation formula was implemented in a software of piano music performance for Module M by the author. The use of the dB value enabled the author to create his desired dynamic expression, e.g., a smooth change of crescendo and diminuendo, very easily without struggling with the magic numbers of velocity[3].

We claim that the proposed way of representing the loudness of piano tones will give us a means for quantitative study of "performance rules" on the dynamic expression.

6. REFERENCES

[1] Zwisloski, J.J., "Temporal summation of loudness: An analysis", J. Acoust. Soc. Am., 46(2), 431–441, 1969.

[2] Reichardt, W. and Nisse, H., "Choice of sound duration and silent intervals for test and comparison signals in the subjective measurement of loudness level", J. Acout. Soc. Am., 47(4), 1083–1090, 1970.

[3] Taguti, T., "Expression of motion in piano performance: A quantitative study with Mozart's piano sonata K. 311", Proceedings of IAEA2002 (17th Congress of the International Association of Empirical Aesthetics, 4-8 August 2002, Takarazuka, Japan), pp. 469–472, 2002.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

PERFORMANCE CUES FOR LISTENERS' EMOTIONAL ENGAGEMENT

Renee Timmers (1, 2), Antonio Camurri (2), Gualtiero Volpe (2)

(1) Austrian Institute for Artificial Intelligence
(2) Infomus, DIST, University of Genova
renee74@xs4all.nl, music@dist.unige.it, volpe@infomus.dist.unige.it

ABSTRACT

A variety of factors may influence listeners' emotional engagement with a musical performance. It may depend on the sight of the pianist, the performers' interpretation of the music, or the musical experience of the listener. In an explorative study, we investigated what aspects of a performance are most strongly related to listeners' emotional engagement and considered cues present in sound as well as in the performer's movement. Twelve participants were presented audio recordings and twelve participants were presented audio and video recordings of three performances of an Etude of Skriabin. Their task was to indicate their emotional engagement with the music by moving a slider up and down. The results can to a large extent be described as listeners' synchronization with performance aspects: higher energy levels in sound and movement of the performance corresponded to higher levels of emotional engagement, while lower levels had lower engagement. Considerable overlap existed between the cues of the movement and the audio suggesting an unambiguous interpretation of the music by the performer. This interpretation was however most clearly revealed by the dynamics and tempo of the performances, where tempo communicated local phrase structure and dynamics the overall pattern of tension and release.

1. INTRODUCTION

With the perfection of audio playback systems and with the rise of computer-generated music, one could expect the role of the musical performer and the need for live concert attendance to decay. Fortunately this is not the case and people still wholeheartedly enjoy concert visits. One of the reasons for this might be, besides social and cultural reasons, the communication that goes on between performer and listener and the excitement of witnessing the moment-to-moment realization of the performance.

Communication from performer to listener has been demonstrated in several studies. Performers communicate a personal interpretation of the music, which concerns not only the structure of the music [1] [2] [3], but also the expressive or emotional content [4] [5] [6] [7]. When asked to choose between different possible interpretations, listeners are rather accurate at choosing the one expressed by the pianist.

What has been less investigated and what is probably harder to investigate is the communication of emotional engagement from performer to public, but also from public to performer. The performer may show emotional engagement in his movements, facial expressions and in his sound production, and also may explicitly express a pattern of tension – release or variations in emotional intensity that guide and influence listener's emotional engagement. The responses of listeners may influence and inspire the performance.

We do not expect to be able to investigate this process quite accurately, but we do think it is possible to investigate at least one aspect of it: the communication of emotional engagement from performer to listener. To investigate this, we adopted a conceptual framework previously defined for the analyses of expressive communication in dance that consists of four layers [8]. (1) a performance data preprocessing layer, (2) a performance data feature extraction layer, (3) a performance data segmentation, mapping and interpretation layer, and (4) a relation to listener's responses layer. In other words, we used a bottom-up approach that assumes that measurable features from video and audio recordings of a performance (that is measurable aspects of the movement and the sound of a performance) are able to account for the observed communication. Layers 1 and 2 are technical processing layers, but layers 3 and 4 are aimed to have psychological relevance as well. Layers 3 and 4 could be compared with the distinction that Juslin [7] makes between the expression of emotion by performers and the understanding of emotion by listeners: the performer may or may not be clear and distinct in his emotional expression and listeners may or may not be accurate and sensitive to this expression.

2. METHOD

2.1. Musical performance

A professional pianist performed Etude Op. 8 no. 11 by Alexander Skriabin at a concert that was organised for the experiment's purpose. He first performed the piece without public in a normal manner (to be referred to as p1) and an exaggerated manner (to be referred to as p2) and then performed it with public in a normal, concert manner (to be referred to as p3). He performed on a Yamaha Disklavier, which made it possible to register MIDI information of the performance. In addition, audio recordings were made and presented to the participants of the experiment.

Etude Op. 8 no. 11 is a slow and lyrical piece (Andante cantabile) in a late Romantic style. According to the pianist, the piece can be played with a lot of freedom. Theoretically, the piece has a simple A B A with coda structure, but the pianist interpreted the line of the music differently: The first main target of the music is a release of tension halfway the B

section (left downward pointing arrow in Figure 1). Everything preceding this target point is a preparation for this tension release. The A section is anyway preparatory; it leads towards the start of the B section, which is the real beginning of the piece. After this release of tension, the music builds up towards the dramatic return of the theme of the A section. This prepares for the second possible point of tension release halfway the coda at a general pause (right downward pointing arrow in Figure 1). The release is however not continued and the piece ends most sad.

2.2. Participants and procedure

24 people participated in the experiment among them were eight musicians. The participants were divided in two groups. One group saw the performances at a computer screen and heard the performances over speakers (referred to as the video condition). The other group only heard the performances over speakers (referred to as the audio condition). The participants sat behind a desk with a slider and a joystick before them. They saw and/or heard the performances two times. The first time they heard the music, they indicated the phrase boundaries in the music by pressing the button of the joystick. The second time they heard the music, participants indicated to what extent they were emotionally engaged with it by moving a MIDI-slider up and down while listening to the music. The whole procedure was explained in a written instruction and a practice trial.

3. RESULTS

3.1. Layers 1 and 2: preprocessing and feature extraction of performance data

Auditory data. The note-onset times and the key-velocity per note were taken from the MIDI recordings from the Yamaha Disklavier. The note-offset information was not used, because the pianist used too much pedal for articulation measures to be reliable.

Two features – local duration and local dynamics – were extracted from this information. Local duration was measured as the inter-onset-interval (IOI) between successive quarter note onsets of a middle voice that had a rhythm of repeated 8th notes. Local dynamics was measured as the average key-velocity of all note onsets within a time-span of a quarter note.

Visual data. For the analysis of the movement of the pianist, we concentrated on the movement of the head, which closely relates to the movement of the back of the pianist. Techniques for motion detection and tracking were applied to the recorded images for segmenting the silhouette of the head, individuating its bounding box, and tracking its movement.

Three features were extracted. First of all, the position of the head was measured in two directions: x-direction showing left-right position and y-direction showing backward-forward position. We used the Lucas and Kanade feature-tracking algorithm [9] that assigns and tracks a specified number (in our case 40) of randomly assigned moving points within a region.

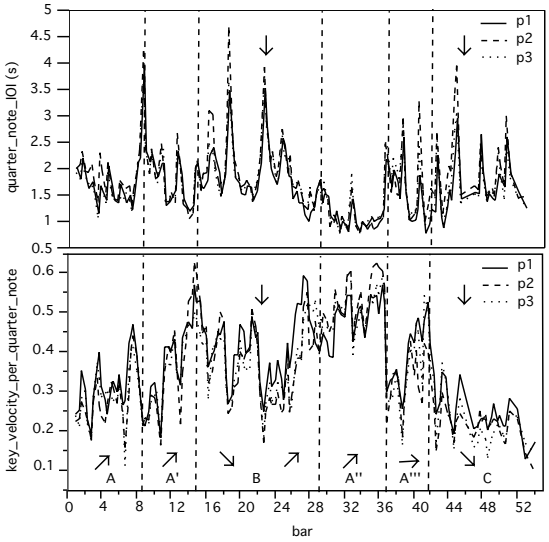


Figure 1:Quarter note IOI and key-velocity per quarter note for p1, p2 and p3. Dotted vertical lines indicate section boundaries. Arrows are explained in the text.

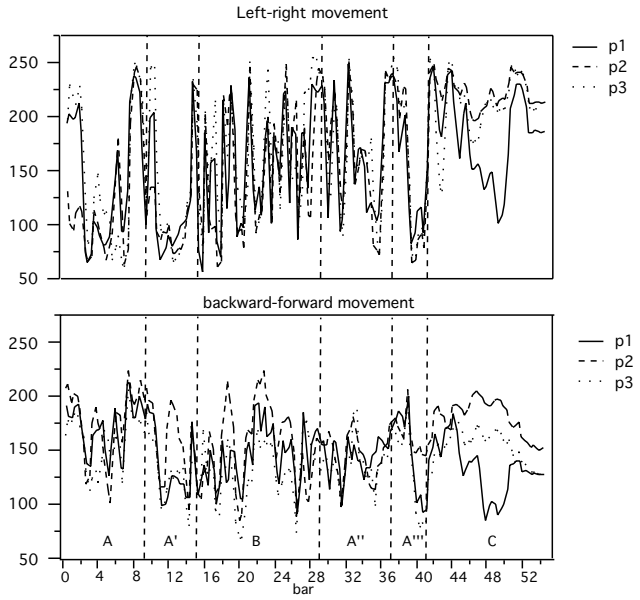


Figure 2: Left-right movement and Backward-forward movement of the head of the pianist. High values correspond to a right and forward position, while low values correspond to left and backward position. Separate plots for the three performances show especially consistent left-right movement.

Thirdly, velocity was calculated for each trajectory using the symmetric backward technique for the numeric derivative. Average values of position and velocity among the forty trajectories were calculated for both the x and y component. In addition, the velocity values were integrated for the x and y direction to get a general measure of movement velocity.

Redundancy in the number of points (i.e., forty points instead, for example, of just the barycentre of the blob) allowed us to get more robust and reliable values for velocity. A low-pass filter was applied to smooth the obtained data. Measures were summarized per quarter note in order to be comparable to the other measures.

3.2. Layer 3: Mapping and first interpretation of features

Correlations between performance measures were calculated to check the coherence between them. Key-velocity and IOI are negatively correlated (r = -.51 on average). Velocity of head movement is positively correlated with key-velocity (r = .45 on average) and negatively with IOI (r = -.25 on average). The low correlation between values is partly due to the asynchrony between the periodicity of the measures. If peak-values (maximum for key- and movement velocity and minimum for IOI) per two bars are correlated, agreement between movement and sound measures becomes higher. Especially the two velocity measures turn out to be highly correlated (r = .79 on average for key and movement velocity, versus r = -.38 on average for movement velocity and IOI).

Backward-forward position is negatively correlated with key-velocity (r = -.39 on average) and positively with IOI (r = .26 on average), which means that the performer moves forward in soft and slow passages and backwards in louder and faster and therefore more intense passages. Left-right position is also negatively correlated with key-velocity and positively with IOI, but correlations are fairly low (r = -.16 and r = .15 on average, respectively). Notice however that left-right movement was highly consistent over the three performances, which suggests that these movements were systematic and intended. The negative correlation with key-velocity and positive correlation with IOI suggests a tendency for a right position to coincide relatively often with longer durations and softer notes as at the end of phrases and a left position to coincide relatively often with shorter durations and louder notes as in the middle of phrases.

Notably, the contour of the key-velocity pattern closely relates to the pianist's interpretation of the musical piece. In the interview, he described the structure of the piece as starting with a moderate introduction that builds up to the B section, which he considered as the real beginning of the piece. This beginning is again a preparation for the first target of the piece: the release of tension at the middle of the B section. Hereafter tension builds up towards the dramatic return of the theme, which leads via a repeat of the theme in contrasting dynamics to the second important target of the theme: the second possible release of tension at the general pause. Key-velocity reflects this interpretation by showing higher values for tense passages and diminuendi at points of tension release (see arrows at the bottom of Figure 1). Only the soft ending does not correlate with the description of tension-release given by the pianist: this soft ending is not a release, but is loaded with emotion. In addition, this global pattern in key-velocity could be seen as communicating the overall form of the piece that subdivides the entire piece at measures 22 and 36. The return of the theme is the culminating point of the piece where after tension can release.

All performance measures show a periodic increase and decrease. To check the relation between these periodicities and the musical structure, the location of mimima in key-velocity, and maxima in IOI, left-right and backward-forward position were compared to the location of phrase-boundaries. Generally, the Skriabin etude has a local structure of two-bar phrases. The forward and the right position of the performer were taken as start/end point for periodicity. IOI was most systematically related to the two-bar phrasing of the Skriabin piece, followed by key-velocity. 55% of the phrase-boundaries were joined by a slowing down in tempo. The other phrase-boundaries were directly followed by a slowing down in tempo (a delay of 1 quarter note). For the key-velocity, 42% of the phrase-boundaries coincided with a minimum in key-velocity, 15% was anticipated by a minimum and 28% followed by a minimum. The period-boundaries of the movement of the pianist hardly synchronized with the score-phrasing, instead the location of these boundaries varied greatly with respect to the two-bar score-phrasing.

These results indicate a differentiation in function between the performance measures, where local duration especially communicates local phrasing and key-velocity and movement velocity communicate the pattern of tension-release of the music. Minima in key-velocity reinforce the segmentation indicated by local duration now and then. And local duration joins key-velocity to some extent in communicating the musical tension.

3.3. Layer 4: Conceptual layer and relation to responses of listeners

The participants of the experiment indicated the end of phrases by pressing a button and the degree of emotional engagement by moving a MIDI slider. To get a measure of subjective segmentation of the piece (SM), the number of people indicating a phrase boundary for each quarter note was calculated and expressed as a multiple of chance level. The average indicated emotional engagement per quarter note was taken as measure of emotional engagement (EM).

To check the relation between extracted performance features and listeners' responses, a series of multiple regression analyses were done. In the first analysis, quarter note IOI, key-velocity, and movement velocity were used to predict SM. In the second analysis, the same variables were used to predict EM. In the third analysis, peak values per hyper-bar were used to predict EM per hyper-bar. All analyses were done directly and with a time-delay of one, two and three quarter notes of the performance data with respect to the listeners' data. The best R^2 obtained are reported in Tables 1 and 2. These were obtained with a delay of either zero or one for SM, and either two or three for EM.

SM was significantly predicted by the model. From this model, IOI was the only significant variable. In other words, duration was a predictor of the variation in number of

participants indicating a section-boundary. More participants indicated a phrase-boundary for longer durations.

EM was well predicted by the quarter note model, but even better by the second model that took the peak-values per hyper-bar to predict EM per hyper-bar. Key-velocity was in the majority of cases the only significant variable, but was sometimes joined by movement velocity. IOI reached significance only for p2, for which variations in location duration are relatively large. Notably, all explained variances were higher for the non-musicians than for the musicians. The non-musicians seem especially to have paid attention to the performance variables, while the musicians either paid less attention to them or provided a more abstract response.

Measure	Audio Mus	Audio Non-mus	Video Mus.	Video Non-mus
SM	.18	.23	.24	.26
EM				
Quarter	.32	.57	.29	.56
Peak	.48	.71	.40	.71

Table 1: Average R^2 for each condition obtained for the regression analyses explaining the variance in SM and EM on the basis of quarter note or peak per two bars IOI, key-velocity, and, for the video conditions, movement velocity.

As a comparison, the analyses were repeated with x-position and y-position as independent movement variables instead of the more general movement velocity variable. The results did not improve or change from this alteration, instead x and y-position did not contribute significantly to any of the regressions.

These results confirm a differentiation between expressive means: tempo primarily communicates segmentation, while dynamics communicates musical tension, which is interpreted especially by the non-musicians as variations in emotional intensity. Velocity of the movement is correlated with dynamics and may therefore also reflect musical tension, but the sounding parameters seem to be the main communicative factors.

4. SUMMARY AND CONCLUSION

In an explorative study, the communication of emotional engagement from performer to listener was investigated. A analysis framework was used where extracted cues from recordings of performances were analyzed for their communicational function.

Features were extracted from the video and MIDI data of three performances of a Skriabin etude that characterized the position and movement velocity of the head of the pianist and the tempo and dynamics of the performance. Correlation between the features and their relation to the score-phrasing suggested a differentiation in function between the features. Movement measures were most strongly related to the dynamics of the performance and less strongly to the tempo of the performance and especially peak-values correlated highly. The periodicity of rise and fall in tempo was systematically related to the score phrasing. This was less so for the dynamics of the performance and hardly the case for the movement of the head of the pianist. The close relation of dynamics to the description of the pianist of tension and release finally suggested that the dynamics of the performance communicated the pianist's interpretation of musical tension, while tempo communicated local phrasing.

The differentiation in function between tempo on the one hand and dynamics and movement velocity on the other hand was substantiated by the responses of the listeners. Tempo was the only significant factor that could explain the subjective segmentation of the participants, while dynamics was the main factor that explained emotional engagement. These results can to a large extent be described as listeners' synchronization with performance aspects: higher energy levels in sound and movement of the performance corresponded to higher levels of emotional engagement, while lower levels had lower engagement. The non-musicians showed these tendencies most strongly, while the musicians often deviated from this pattern. An aspect that seems worthwhile to be investigated further in future research.

5. REFERENCES

[1] Palmer, C., "Mapping Musical thought to musical performance", Journ. Exp. Psych.: Human Perception and Performance, 15(12):331--346, 1989.

[2] Palmer, C., "On the Assignment of Structure in Music Performance" Music Perception, 14(1):23--56, 1996.

[3] Sloboda, J. A., "The communication of musical metre in piano performance", Quarterly Journal of Experimental Psychology, 35(A):377--396, 1983.

[4] Gabrielsson, A., and Juslin, P. N., "Emotional expression in music performance: Between the performer's intention and the listener's experience", Psychology of Music, 24(1):68-91.

[5] Juslin, P. N., "Communicating emotion in music performance: a review and a theoretical framework". In Juslin, P., and Sloboda, J. A. (eds.), Music and Emotion: theory and research, Oxford University Press, Oxford, pp. 309-337, 2001.

[6] Davidson, J. W., "Visual perception of performance manner in the movements of solo musicians" Psychology of Music, 21:103-113, 1993.

[7] Kendall, R. A., and Carterette, E. C., "The Communication of Musical Expression", Music Perception, 8(2):129-164, 1990.

[8] Camurri, A., De Poli G., Leman M., "MEGASE - A Multisensory Expressive Gesture Applications System Environment for Artistic Performances", Proc. Intl Conf CAST01, GMD, St Augustin-Bonn, pp.59 – 62, 2001.

[9] Lucas B., Kanade T., "An iterative image registration technique with an application to stereo vision", in Proceedings of the International Joint Conference on Artificial Intelligence, 1981.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

ANALYSIS OF JAZZ DRUMMERS' MOVEMENTS IN PERFORMANCE OF SWING GROOVES – A PRELIMINARY REPORT

Carl Haakon Waadeland

Department of Music
Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway
carl.haakon.waadeland@hf.ntnu.no

ABSTRACT

'Swing groove' is often used to denote a typical rhythmic cymbal pattern played by jazz drummers. Various ways of swing performances may be highly individual for different drummers. Moreover, various performances of the swing groove may be more or less musical appropriate, and may also in varying degree be "swinging".

In this investigation we pose the following question: What can empirical studies of a jazz drummer's movements tell us about the way the drummer is performing swing grooves? In our search for some answers to this question we have constructed a series of experiments where kinematic as well as dynamic aspects of jazz drummers' performances of swing grooves are measured and analysed.

To give an idea of our approach, we outline the basic constituents of the experiments and suggest some possible consequences of our investigation.

1. INTRODUCTION

A swing groove, as played on a cymbal by a jazz drummer, is often written in the following way:



A jazz drummer may perform a swing groove in a number of different ways. Typically, the drummer will perform the rhythm by making various deviations from the exact note values in the notation above. E.g., some notes are played with shorter duration, other notes with longer duration compared to a strict "metronomic" ("deadpan") performance where each note is played according to the note values defined by musical notation. These various ways of swing performance may be highly individual for different drummers. (In fact, the drummer may sometimes be identified by the way the swing groove is performed.) Moreover, various performances of the swing groove may in varying degree be musical appropriate (more or less "correct" related to various styles or traditions of performance), and may also in varying degree be swinging (i.e. make you want to "swing along with the music").

Interesting studies of performances of swing grooves in jazz are presented in [1], [2], [3] and [4]. The main strategy in these investigations has been to measure to what extent live performances show deviations from exact note values, and to relate these deviations to various musical/contextual parameters, such as individual preferences, musical style, tempo, and inter-ensemble relations (i.e. rhythmic deviations between the different musicians in an ensemble). Several empirical investigations detecting various deviations in performances of music belonging to other traditions than the jazz tradition have also been carried out (for an overview, see [5], [6]).

1.1. Gesture vs. attack-point rhythm

The majority of research on rhythm performance has been concerned with investigating attack-point rhythm, i.e. attack points and durations, and is thus based on information of a finite number of discrete points along the one-dimensional axis of time. The very object of study, the musical performance itself, however, is unfolded as a continuous movement in time and space, created through an interaction between the musician, a musical instrument, and different physical/social environments related to which the performance takes place.

In addition to investigating attack points and durations, it is therefore important to view musical performance as a continuous phenomenon, and take gestural aspects of the performance into consideration. Based on this epistemological point of view we now pose the following general questions:

(A) How do the musicians' different movements relate to different rhythmic performances? (This is also a major concern in [7].)
(B) Is it possible to identify different movement categories that are typical of different rhythmic expressions in music?

And, to be somewhat more specific:

(C) What can studies of a jazz drummer's movements tell us about the way the drummer is performing swing grooves?

This last question is the one we address here. Trying to find some answers to this question we have constructed a series of experiments where jazz drummers' performances of swing grooves are measured and analyzed. These experiments are, at present, part of an ongoing investigation, and careful analyses have still to be done in order to give valid results. However, to give an idea of our approach, we now briefly outline the basic constituents of the experiments. Moreover, we present some speculative thoughts and suggest some possible consequences of our investigation.- Parts of this report were also presented at 6th NOFOD Conference, Trondheim, 2002 (see [8]).

2. METHOD

2.1. Subjects

The subjects participating in these experiments are semi-pro/professional drummers acquainted with jazz drumming. All of the drummers were, at the time of the experiments, students or teachers at Section of Jazz Education, Department of Music, NTNU.

2.2. Task

The subjects were asked to use one drumstick to play swing groove restricted to various performance conditions that were defined according to the main focus of this investigation, which is: To study possible relations between the drummers' movemenst and the following performance parameters:

(1) *Tempo:* The subjects were asked to play swing groove at 7 different tempi, ranging from 60 bpm (beats pr. minute) to 300 bpm. Among the possible relations that will be analyzed on the basis of this experiment are: movement vs. tempo, dynamics vs. tempo, rhythmic phrasing vs. tempo.

(2) *Accents:* The subjects were asked to play three different versions of swing groove with accent on the first, second and third note, repectively, in the swing pattern. Dahl, [9], [10], has carried out interesting research showing how the playing of accents have influence on temporal and kinematic aspects of drumming. It will be investigated to what extent Dahl's findings apply to performances of swing grooves.

(3) *"Swing ratio" (rhythmic subdivision/rhythmic phrasing):* The subjects were asked to play swing grooves applying three different rhythmic subdivisions of the quarter note: subdivision in eighth notes, subdivision in eighth note triplets, subdivision in sixteenth notes. (This corresponds to the swing ratios 1, 2, 3, respectively. For more information of swing ratios, see [3], [4].) Among the possible relations to be examined on the basis of this experiment are: movement vs. swing ratios, accents vs. swing ratios.

2.3. Equipment

The measurements were carried out assisted by Geir Oterhals, Section of Movement Science, NTNU. Figure 1 illustrates how the experimental situation was constructed. The figure shows the drummer playing on a "force plate" using one drumstick. Markers were placed on the drummer's arm and on the drumstick.

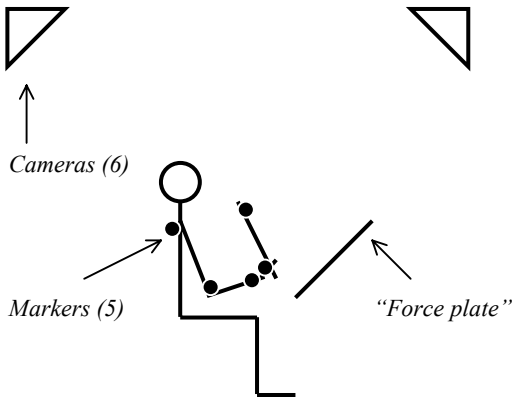


Figure 1: *Illustration of experimental set up for measuring kinematic and dynamic aspects of performance of swing grooves.*

The equipment used for this investigation is well known within empirical movement science. For our set up we applied the following components:

- 6 cameras (Proreflex camera system) transmitting infra red light were used to measure movements of the arm and the drumstick (kinematics).
- 5 markers reflecting the light were placed on the drummer's shoulder, elbow, wrist and hand, as well as on the drumstick.
- A force plate (Kistler) measuring force from the drumstick was used to give accurate measurements of attack points and musical dynamics.
- A minidisk was used for audio recording of the experiment.

3. WHAT DO THE MEASUREMENTS LOOK LIKE?

Using the equipment described above, we obtain graphic representations of relations between movement, dynamics and timing in performance of swing grooves. It is expected that analysis of such representations should give interesting information about formation and stability of movement patterns in drumming.

Let us give an example to indicate what the measurements might look like. This example also suggest what kind of information we may be able to extract from these experiments.

3.1. Example: Movement categories and related movements

The graphic representation in Figure 2 below shows the vertical movement (height vs. time) of drumstick, hand and wrist, respectively, in a performance of swing groove.

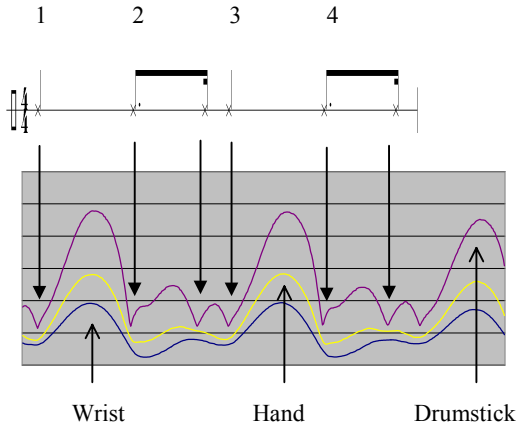


Figure 2: *Illustration of vertical movement of drumstick, hand and wrist in a performance of swing groove.*

Looking at the illustration in Figure 2 we observe that the movements of drumstick, hand and wrist all seem to show a rather cyclic/periodic movement pattern, indicating "a certain way of moving", i.e. a movement category, in this performance of swing groove. It seems plausible to hypothesize that this movement category is dependent on both the performing musician's individual preferences and the tempo in which the performance is played.- Given this hypothesis it is interesting to know that already in 1929 Bernstein and Popova carried out empirical investigations of movements in piano playing, and they stated: "One can say that, with the slow, middle and fastest paces, we are dealing with three totally different dynamical constructions, with 3 dissimilar movement mechanisms" (see [11]). On the basis of the information given from our experiments, we will investigate to what extent we may identify relations between movement categories and tempo in performances of swing grooves, and thereby support the findings of Bernstein and Popova.

Moreover, it is interesting in this case to note how the three movements in Figure 2 are related. For example, we observe that the downbeats of the movements are not synchronized: On the 2 and 4 in the pattern we see that the downbeats of hand and wrist are both delayed compared to the downbeat of the drumstick, whereas on the sixteenth note preceeding the 1 and 3 the movements of hand/wrist have no downbeat at all. This might raise the following interesting question: From the performer's point of view, which movements define the feeling/perception of the beat?

4. SOME POSSIBLE CONSEQUENCES AND RESULTS

The large number of data and information given from these experiments must be carefully examined and analyzed. At this point we merely suggest some possible consequences and results:

(I) With reference to the findings of Bernstein and Popova, it will be investigated if we can identify categorical relations between tempo regions and movement patterns in performance of swing grooves (cf. 3.1. above).

(II) From an ethnomusicological, anthropological point of view Keil [12] discusses motion and feeling through music. In a jazz rhythm section he makes a distinction between drummers who perform the swing groove "on top" of the pulse, and those who "lay back" behind it. Observing (not measuring) various well known jazz drummers, Keil asserts: "Although it is difficult to generalize about the attacking motion, on-top drummers tend to keep the stick close to the cymbal and the arm fairly stationary, with the stroke moving perpendicular to the cymbal, such that each beat lands on the cymbal in the vicinity of its predecessor. Conversely, drummers of the "lay-back" school (…..) seem to attack horizontally, so to speak, placing each beat on a different part of the cymbal as the arm moves back and forth slightly." ([12], pp. 61-62) This we take as a very interesting observation, and it will be studied to what degree our experiments support Keil's statement.

(III) On a more general level, this investigation may contribute to develop a new methodology in the study of music performance: Instead of measuring only attack points and durations, a description of music performance as a continuous process taking gestural aspects into account, is offered. (This is also supported by the research of Dahl [7].)

(IV) Moreover, analysis of gestural data may provide new insight into basic features of rhythmic performance, which is likely to have interesting pedagogical applications, and might thus be of interest to musicians and music teachers.

(V) The findings of this empirical investigation are also important for the construction of continuous (computer) models generating synthetic performances that approximate reality. One such theoretical model simulating rhythmic expression, is developed by the author (see [13]). It seems likely that results derived from these experiments will be valuable in the further development of this model.

5. ACKNOWLEDGEMENTS

The author is very grateful to Geir Oterhals, Section of Movement Science, NTNU, for fruitful assistance and cooperation in the measurements of rhythmic movements. Special thanks are also due to the drummers participating in the experiments. Economical support to this project was provided by The Faculty of Arts, NTNU.

6. REFERENCES

[1] Reinholdsson, P., "Approaching Jazz Performances Empirically. Some Reflections on Methods and Problems". In A. Gabrielsson (Ed.), Action and Perception in Rhythm and Music. The Royal Swedish Academy of Music, No. 55, pp. 105-125, 1987.

[2] Prögler, J.A., "Searching for Swing: Participatory Discrepancies in the Jazz Rhythm Section", Ethnomusicology, 39(1), pp. 21-54, 1995.

[3] Friberg, A. & Sundström, A., "Preferred swing ration in jazz as a function of tempo", Speech Music and Hearing Quarterly Progress and Status Report, 4/1997, Stockholm, pp. 19-28, 1997.

[4] Friberg,A. & Sundström,A., "Swing Ratios and Ensemble Timing in Jazz Performance: Evidence for a Common Rhythmic Pattern", Music Perception, 19(3), pp. 333-349, 2002.

[5] Clarke, E.F., "Rhythm and Timing in Music". In D. Deutsch (Ed.), The Psychology of Music, Second Edition, San Diego: Academic Press, pp. 473-500, 1999.

[6] Gabrielsson, A., "The Performance of Music". In D. Deutsch (Ed.), The Psychology of Music, Second Edition, San Diego: Academic Press, pp. 501-602, 1999.

[7] Dahl, S., "Striking movements: Movement strategies and expression in percussive playing", Licentiate Thesis, Royal Institute of Technology, Dept. of Speech, Music and Hearing, Stockholm, 2003.

[8] Waadeland, C.H., "Movements in Rhythmic Performance of Swing in Jazz – Report from an Empirical Investigation", In Fiskvik & Bakka (Eds.) Proceedings 6th NOFOD Conference, Trondheim, pp. 193-199, 2002.

[9] Dahl, S., "The Playing of an Accent – Preliminary Observations from Temporal and Kinematic Analysis of Percussionists", Journal of New Music Research, 29(3), pp. 225-233, 2000.

[10] Dahl, S., "Playing the accent – comparing striking velocity and timing in an ostinato rhythm performed by four drummers", accepted for publ. in Acta acustica united with Acustica, 2002.

[11] Bernstein, N.A. & Popova T., "Untersuchung über die Biodynamik des Klavieranschlags", Arbeitsphysiologie, 1, pp. 396-432, 1929.

[12] Keil, C., "Motion and Feeling through Music", in Keil & Feld, Music Grooves, Univ. of Chicago Press, pp.53-76, 1994.

[13] Waadeland, C.H., "It Don't Mean a Thing If It Ain't Got That Swing – Simulating Expressive Timing by Modulated Movements", Journal of New Music Research, 30(1), pp. 23-37, 2001.

A SINGING TRANSCRIPTION SYSTEM USING MELODY TRACKING ALGORITHM BASED ON ADAPTIVE ROUND SEMITONE (ARS) PLUS MUSIC GRAMMAR CONSTRAINTS

[1]Chong-kai Wang, [2]Ren-yuan Lyu, [3]Yuang-chin Chiang

[1,2]Department of Electrical Engineering, Chang Gung University, Taoyuan 333, Taiwan

[3]Institute of Statistics, National Tsing Hua University, Hsinchu, 300, Taiwan

[1]ckwang.ee86@nctu.edu.tw [2]rylyu@mail.cgu.edu.tw
[3]chiang@stat.nthu.edu.tw

ABSTRACT

In this paper, an approach for melody tracking is proposed and applied to applications of automatic singing transcription. The melody tracker is based on Adaptive Round Semitones (ARS) algorithm, which converts a pitch contour of singing voice to a sequence of music notes. The pitch of singing voice is usually much more unstable than that of musical instruments. A poor-skilled singer may generate voice with even worse pitch correctness. ARS deals with these issues by using a statistic model, which predicts singers' tune scale of the current note dynamically. Compared with the other approaches, ARS achieves the lowest error rate for poor singers and seems much more insensitive to the diversity of singers' singing skills. Furthermore, by adding on the transcription process a heuristic music grammar constraints based on music theory, the error rate can be reduced 20.5%, which beats all the other approaches mentioned in the other literatures.

1. INTRODUCTION

Automatic musical transcription has been an interesting issue in multimedia applications. A system that is capable of transcribing orchestra input to musical notes is called a polyphonic transcription system. In contrast, a monophonic system takes as input single instrument sound or human singing. Generally speaking, a transcription system for musical instrument performs quite well satisfactorily because the input instrument sounds have contained accurate pitch information and thus are easily decoded. In the case of human singing, the task becomes more difficult due to the fact that a human singer can sing inconsistently and easily drifted away from standard musical scale. Unlike a musical professional, a common person might find it difficult to write down the melody in his mind or in his ears. He can choose to sing to a musical transcription system, but quite rarely he can sing as accurate as an instrument. This will deteriorate a transcription system expecting standard pitch scale input.

A singing transcription system accepts human singing voice, and transcribes it to musical notes. The fact that human singing rarely follows standard musical scale makes the implementation a challenge. There are two basic problems in a human singing transcription system: note tracking and melody tracking. Note tracking is to segment singing waveform into voiced/unvoiced region, for voiced region we mean a waveform segment where pitch period can be estimated. Each pitched region is often regarded as a musical note, and in music score, a syllable of a lyric word often mates one or many music notes. Melody tracking is to map singing pitches into musical notation correctly, even if a poor singer generates a bad melody.

Authors have explored a method for Note tracking using spectral standard deviation (see [1]). This paper reports our singing transcription system, in addition to detail techniques in Melody tracking task. In Melody Tracking module, a statistic model called Adaptive Round Semitones (ARS) and a music grammar constraint called Tune Map is applied to improve the robustness of the system.

This paper is organized as follows. Section 2 is a brief description for old Melody tracking algorithms. Section 3 is for the newly proposed ARS algorithm. Section 4 explores the mathematical calculations of Constraint table and Tune Map model. Section 5 is conclusion.

2. MELODY TRACKING

In this paper, we describe a singing transcription system, which could be divided into two modules. One is for the front-end voicing processing, including voice acquisition, end-point detection, and pitch tracking. The other is for the melody tracking, which maps the relatively variation pitch level of human singing into accurate music notes, represented as MIDI note number. The overall system block diagram can be shown as Figure 1.

2.1. Front-end voice processing

Singing voice utterance is first divided into voiced and unvoiced section. After segmenting the voiced region of waveform, extracting the pitch or fundamental frequency (F_0) of human singing signal follows. Each segmented section will correspond to a musical note. Here, a temporal pitch detection algorithm plus rule-based filtering is used to get the pitch contour.

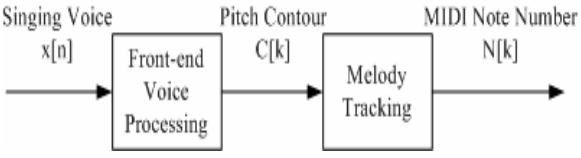


Figure 1: *System block diagram*

2.2. Converting pitch to MIDI number

The Melody Tracking module of the system deals with the average pitch value *f* (frequency in Hertz) for each note and transcribes them into music note representation, which have been defined as MIDI numbers for all possible semitones in the computer music processing domain. The pitch value and the corresponding MIDI number *N* can be related as the following formula:

$$N = 69 + 12 \times \log_2 \frac{f}{440} \tag{1}$$

For example, the note middle A with pitch being 440 *Hz* will be assigned a MIDI number 69, and the note A#, which is one semitone higher than middle A, is then assigned 70 as its MIDI number. On the equal tempered scale, the pitch interval between two consecutive MIDI numbers could be divided into 100 *cents*, e.g., the note middle A is 6900 cents. Thus, 100*N* is called pitch-in-cents.

2.3. Round MIDI and modification

A straightforward approach is to convert pitch-in-cents to MIDI number by rounding the real value of pitch-in-cents/100 to the nearest integer. This is termed as *Round MIDI*.

However, human's pitch generation process is much more unstable than that of musical instruments. An untrained singer rarely generates voice with correct pitch value. It was said that only about 1 in 10000 people claim to have tone-absolute pitch perception/generation [2]. Furthermore, the pitch level can drift upward or downward due to the singer's mood, and thus the pitch interval may not keep constant during singing. The experiments done in the other literatures show that the straightforward Round MIDI algorithm did not work well for melody tracking, thus several improvements have been proposed as follows.

- McNab's *Moving Tuning* [3] and *"Round Intervals"* rounds the difference of pitch in cents between adjacent pitched regions and treats them as the difference of MIDI number between adjacent notes.
- *Haus'* Modified *Round MIDI* [4] estimates an offset from all notes, adjusts all notes with this offset, and then does the same rounding as Round MIDI.

Here, we design a new method called Adaptive Round Semitones (ARS) which was shown to be the most effective among those in the literatures. [5]

3. ADAPTIVE ROUND SEMITONES (ARS)

The ARS was designed on the following three assumptions:

1. Most singers have only tone-relative pitch.
2. The tune scale of human singing voice is not necessarily 100 cents per semitone.
3. The tune scale will probably change with time while singing and be dependent on previous notes.

For example, while a singer singing a rising melody sequence, he will inadvertently sing higher and higher. For a poor singer, the melody sung outward does not always match what he wishes to sing, although the listeners usually can understand what he sang. This is because he has a changeable and unusual tune scale. One important motive of designing ARS is to deal with these issues.

3.1. Statistic model

The block diagram of ARS is shown in Figure 2, where the input is the sequence of the pitch in *cents***,** C[k], and the output is the sequence of the MIDI numbers, or equivalently the music notes, N[k].

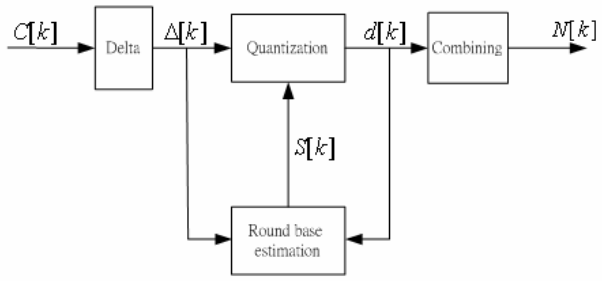


Figure 2: *Adaptive Round Semitone (ARS)*

All the state variables shown in Figure 2 will be described as follows:

$$\Delta[k] = C[k] - C[k-1] \tag{2}$$

$$d[k] = \begin{cases} R\left(\dfrac{\Delta[k]}{100}\right) & 1 \le k \le T \\ R\left(\dfrac{\Delta[k]}{S[k]}\right) & T+1 \le k \le K \end{cases} \tag{3}$$

$$S[k] = \begin{cases} \dfrac{\sum_{m=1}^{T} a[m] \cdot \Delta[k-m]}{\sum_{n=1}^{T} b[n] \cdot d[k-n]} & T+1 \le k \le K \\ 100 & 1 \le k \le T \end{cases} \tag{4}$$

$$N[k] = N[k-1] + d[k] \tag{5}$$

where Δ[k] is the input pitch difference (in cents), d[k] is the difference of the output MIDI numbers, N[k] is the final output MIDI number sequences, and $R(x)$ is a rounding function which rounds off the real number x to the nearest integer, and S[k] is the estimated tuning scale, which is initially set to 100 according to the assumption of equally tempered scale and is dynamically adjusted according to the information provided by the previous music notes.

Here we assume an autoregressive moving average (ARMA) model with parameter set {T; a[1], a[2],…a[T]; b[1], b[2],…b[T]} to estimate S[k]. The parameter set was determined empirically in this paper.

3.2. Rule-based post-processing

Furthermore, a lower and upper bounds for the tune scale S[k] is set to avoid over-tuning. A center-clipping and side-clipping process will be done to compensate for the overly tuning scale. We clip the value if tune scale which is below ±5% and above ±15%.

Besides, some local rules are added to deal with some exceptional cases:

- *Initial effect of tone*: if there are continuous notes identified to the same MIDI number, the tune scale S[k] will be set back to 100 as the initial value.
- *Memory effect of tone*: if pitch of the present note is near the second previous note, they will be identified to the same MIDI number.

4. TUNE MAP MODEL

4.1. Musical grammar constraint

According to well tempered tuning system, an octave is equally divided into twelve notes; a scale is a series of notes selected among these twelve notes. Each of these notes is called a degree, and each degree has its own name designated by a Roman numeral. Most songs are composed conforming to an underline music grammar. Tone distribution of a song melody is not fully randomized. A well-composed song could be clustered in music theory by its tunes, such as C-Major, E-minor…etc. A Major scale is a structure of tones, namely "Do", "Re", "Mi", "Fa", "So", "La", "Si", "Do" (Degree name: Ⅰ, Ⅱ, Ⅲ, Ⅳ, Ⅴ, Ⅵ, Ⅶ, Ⅰ), where the differences in semitones of each continuing tones are 2,2,1,2,2,2,1. Such a tone structure could be looked upon as a grammar to put a constraint on the possibility of notes, which can be followed by another note.

For example, if the pitch differences in semitones of three notes N_1, N_2, N_3, were identified as 1 and 2, then these three notes will be one of the sequences "Mi-Fa-So" and "Si-Do-Re". If there comes a confusing note N_4 which is 1 or 2 semitones higher than N_3. We can easily determine that N_4 is 2 semitones higher according to the Major scale structure distribution.

Form music theory of Bach's well tempered scale, a semitone-level shifting of a sequence melody will not change its inner relation. The singer's singing tune is not unique results from his/her singing key. Equal temperament made possible modern keyboard music with full modulation between all keys. The music grammar constraint is a method that adjusts singer's tune and song's tune in semitone-level.

The idea that adds constraint of music grammar to help track the melody comes from the language modeling technique widely used in speech recognition research, which improve the overall recognition performance significantly.

4.2. The Constraint Table

The Constraint Table lists the rules of melodies based on musical theory. We desired to know the degree name of the first note of a singing utterance. And, according to that, all candidate tones will be constrained. As shown in Table 1, if the degree name of the first note is " Ⅰ " in Major scale. All possible tones (between 2 octaves range) are {0, +2, +4, +5, +7, +9, +11} and {-1, -3, -5, -7, -8, -10} semitones relating to this tone. The tone offset is cyclic repeat every 12 semitones, so the Constraint Table lists 13 Tone offset simply.

Because that Major and Minor scale are cyclic shift in a similar manner, we can handle both Major and Minor scale songs using the same constraint table. Note that the probably degree name is not referring to the tune name, and it is only a symbol presenting the position of the tone in a tune structure. It is like a moving ruler with 7 candidates parting from 2, 2, 1, 2, 2, 2, 1 semitones, moving around the melody line. Aligning the ruler and the singing tones, seeking for fully match.

Tone offset	Probably Degree Name						
0	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ
+1			Ⅲ				Ⅶ
+2	Ⅰ	Ⅱ		Ⅳ	Ⅴ	Ⅵ	
+3		Ⅱ	Ⅲ			Ⅵ	Ⅶ
+4	Ⅰ			Ⅳ	Ⅴ		
+5	Ⅰ	Ⅱ	Ⅲ		Ⅴ	Ⅵ	Ⅶ
+6				Ⅳ			Ⅶ
+7	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	
+8			Ⅲ			Ⅵ	Ⅶ
+9	Ⅰ	Ⅱ		Ⅳ	Ⅴ		
+10		Ⅱ	Ⅲ		Ⅴ	Ⅵ	Ⅶ
+11	Ⅰ			Ⅳ			
+12	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ

Table 1: *The Constraint Table*

4.3. Tune Map

Music grammar constraints described above can be collected to a constraint table (Table 1). It shows the rules of music grammar according to music theory. In tone-relative pitch notion, we are interested in pitch difference of neighboring two notes. We construct a binary tree-like path map called Tune Map (Figure 4) which shows the moving path of notes. There are two kinds of nodes in Tune Map, one is "one-branch-node", and the other is "two-branches-node". Each node refers to a tone; the value on each note refers to the semitone offset between this tone and the first tone of the whole utterance, and the set besides the note refers to the probably degree name of this note.

We make a one-branch-node if the pitch interval rounding error of two neighbor notes is not exceeding ±20%. In the other case, we make it as a two-branches-node, and two of the branches will be connected to the node having tone upper and lower boundary of the rounding scale. The root node refers to the first note of the utterance, and then we add new branches and nodes for next notes. While adding a new node, we should look up the Constraint Table to certificate the node. Certification process doing logic AND for current probably degree name and the

corresponding probably degree name of current tone offset. In the last stage, we can connect a tone path form the survival node to the root node. Finally, the melody of the singing utterance can be tracked.

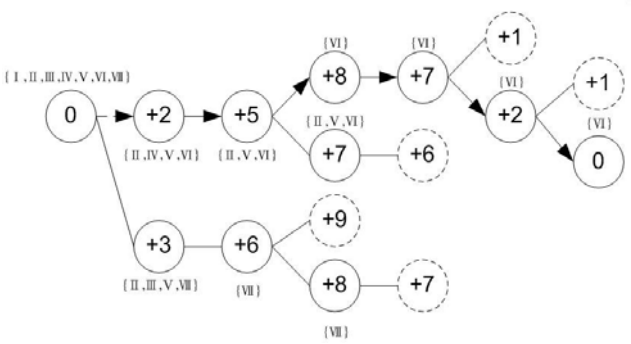


Figure 4: *Tune Map*

4.4. Multi-path of Tune Map

The result of the Tune Map may not be single. That is say, there are more than one path exist in Tune Map. This is because the number of notes of a singing utterance is not many enough to make a fully constraint. Because that, while establishing a two-branches-node in the Tune Map, we choose the nearer one from upper and lower bound of the rounding scale, and put it in the upper node. In the last step of the Tune Map, the uppermost and alive node is the starting node of trace back. Note that if the uppermost node is always alive in each node step, the result is the same as pure Adaptive Round Semitones.

5. PERFORMANCE

Experiments have been carried out with different approaches mentioned in the previous sections for comparison with the proposed one. Testing samples were collected from 13 people. Each person sings 9 pieces of melodies. While singing, the singing key and tempo is free. These testing pieces include melodies from several regions and multiple variations. Total number of testing notes is 2885. We consider them as poor singers if the error rate of Round MIDI algorithm exceeds 30%. There are 4 people (869 notes) in the set of poor singers, and 9 people (2016 notes) in the set of normal singers.

The error rates for each implemented algorithms were listed as in Table 2. , where error rates for all singers, normal singers, and poor singers are listed for all 6 melody tracking algorithms implemented or newly proposed in this paper. One can see that ARS plus music grammar not only achieves the lowest error rates 20.2% for all singers but also achieves the lowest performance degradation from the set of normal singers to the set of poor singers, i.e., least insensitive to the variety of multiple singers. With exception to the ARS plus Tune Map, even though prime ARS does not beat McNab's Moving Tuning algorithm in error rate for all singers, it indeed achieves less performance degradation (5.7%) than the others.

6. CONCLUSION

Music transcription is an interesting and useful technique. In early systems, they are short of any information help from music theory. Authors not only simulating human's performs on singing by an ARMA model but also add constraints from music theory to improve the totally performance. ARS plus Tune Map is a robust analysis algorithm in human singing signal transcription, and it makes "adding musical knowledge" no longer a future work in this kind of system.

7. REFERENCES

[1] Chong-kai Wang, Ren-yuan Lyu, and Yuang-chin Chiang. "An Automatic Singing Transcription System with Multilingual Singing Lyric Recognizer and Robust Melody Tracker," (Submitted to EuroSpeech 2003.)

[2] Profita, J. and Bidder, T.G.,"Perfect Pitch," *American Journal of Medical Genetics*, 29, 763-771, 1988

[3] R.J. McNab, L.A. Smith, and I. Witten, "Signal Processing for Melody Transcription," *Working Paper 95/22, Dept. of Computer Science*, University of Waikato, New Zealand, August 1995.

[4] G. Haus, and E. Pollastri, "An audio front end for query-by-humming systems," *In Proc. of International Symposium on Music Information Retrieval (ISMIR)* ,2001.

[5] Chong-kai Wang, Ren-yuan Lyu, and Yuang-chin Chiang, "A Robust Singing Tracker Using Adaptive Round Semitones (ARS)" *In Proc. of 3rd International Symposium on Image and Signal Processing and Analysis (ISPA03)*, Italy, 2003.

	Round MIDI	McNab's Moving Tuning	Haus' Modified Round MIDI	Round Intervals	**(Prime) Adaptive Round Semitone**	**Adaptive Round Semitone plus Tune Map**
Error rate for normal singers	28.1%	20.7%	29.6%	20.7%	**22.8%**	**19.8%**
Error rate for poor singers	33.6%	25.3%	33.1%	25.3%	**24.1%**	**20.5%**
Error rate for all singers	29.8%	22.0%	31.1%	22.0%	**23.1%**	**20.2%**
Relative Performance degradation from normal to poor singers	19.6%	22.2%	11.8%	22.2%	**5.7%**	**3.5%**

Table 2: *The experimental result for melody tracking algorithms implemented and newly proposed in this paper.*

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

LEARNING TO RECOGNIZE FAMOUS PIANISTS WITH MACHINE LEARNING TECHNIQUES

Patrick Zanon

CSC – DEI, University of Padua
Via Gradenigo 6/a, 35131 Padova, Italy
patrick@dei.unipd.it

Gerhard Widmer

University of Vienna and ÖFAI
Freyung 6/6, A-1010 Vienna, Austria
gerhard@ai.univie.ac.at

ABSTRACT

In this paper we address the question whether it is possible for a machine to learn to identify famous performers (pianists), based on their style of playing. To that end, we extracted a number of features related to performance style from the original audio CD recordings of different Mozart piano sonatas performed by several pianists. Systematic experiments were carried out with various machine learning algorithms. The results show that the computer can learn to identify the performer in a recording with a probability significantly higher than chance. Some pianists exhibited a particularly high degree of recognizability. These findings coincide with another empirical performance style study that we are currently conducting based on novel visualisation techniques.

1. INTRODUCTION

The work presented here is part of a large investigation into the use of novel computational methods for studying basic principles of expressive music performance [7]. One of the questions we study is whether and to what extent aspects of *individual artistic style* can be quantified. And one of the possible approaches to this question is to investigate whether machines can learn to distinguish and recognize different performers based on their style of playing.

Previous research has shown that this seems indeed possible, to a certain extent [5, 6]. However, those studies were limited in many respects, particularly with regard to the data (only 2 pieces) and the musicians involved (piano teachers and students of the University of Music in Vienna). Thus we tried to generalize this research towards the analysis of famous world-class pianists, and work with larger collections of recordings. Early results [10] showed that learning algorithms managed to identify performers with probability above chance, in an n-class discrimination setting. The present paper describes our latest results along these lines. Seven different learning algorithms are applied to the task of identifying the performer in a set audio recordings, by famous pianists, of several Mozart piano sonatas. It will be shown that in a *pair-wise discrimination* setting, quite good results can be obtained.

The paper is organized as follows: Section 2 describes in detail the experimental methodology followed, including a description of the data, a characterization of the performance features extracted from the recordings, and a list of machine learning algorithms tested. Section 3 shows that the learning algorithms can identify performers at a level significantly better than chance, sometimes even with surprising accuracy. The next steps to be performed in this project are then mentioned in Section 4.

ID	Sonata	Movement	Key	Time sig.
kv279_1	K.279	1st mvt.	C major	4/4
kv279_2	K.279	2nd mvt.	C major	3/4
kv279_3	K.279	3rd mvt.	C major	2/4
kv280_1	K.280	1st mvt.	F major	3/4
kv280_2	K.280	2nd mvt.	F major	6/8
kv280_3	K.280	3rd mvt.	F major	3/8
kv281_1	K.281	1st mvt.	Bb major	2/4
kv282_1	K.282	1st mvt.	Eb major	4/4
kv282_2	K.282	2nd mvt.	Eb major	3/4
kv282_3	K.282	3rd mvt.	Eb major	2/4
kv330_3	K.330	3rd mvt.	C major	2/4
kv332_2	K.332	2nd mvt.	F major	4/4

Table 1: Movements of Mozart piano sonatas selected for analysis.

ID	Name	Recording
DB	Daniel Barenboim	EMI Classics CDZ 7 67295 2, 1984
RB	Roland Batik	Gramola 98701-705, 1990
GG	Glenn Gould	Sony Classical SM4K 52627, 1967
MP	Maria João Pires	DGG 431 761-2, 1991
AS	András Schiff	ADD (Decca) 443 720-2, 1980
MU	Mitsuko Uchida	Philips Classics 464 856-2, 1987

Table 2: Pianists and recordings.

2. METHODOLOGY

2.1. The performance data

For the experiments, commercial recordings by six different concert pianists of piano sonatas by W.A. Mozart were collected, and a sizeable number of pieces were selected for performance measuring and analysis. The selected pieces, pianists, and recordings are listed in Tables 1 and 2.

From the audio recordings, rough measurements characterizing the performances were obtained. More precisely, tempo and general loudness were measured at the level of the beats, by using an interactive beat tracking program [2] to find the beat in the audio signal and computing beat-level tempo changes from the varying inter-beat intervals. Overall loudness of the signal at these time points was extracted from the audio signal and is taken as a very crude representation of the dynamics applied by the pianists. No more detailed information (e.g., about articulation, individual voices, or timing details below the level of the beat) is available.

These sequences of measurements can be represented as two sets of performance curves — one representing beat-level tempo,

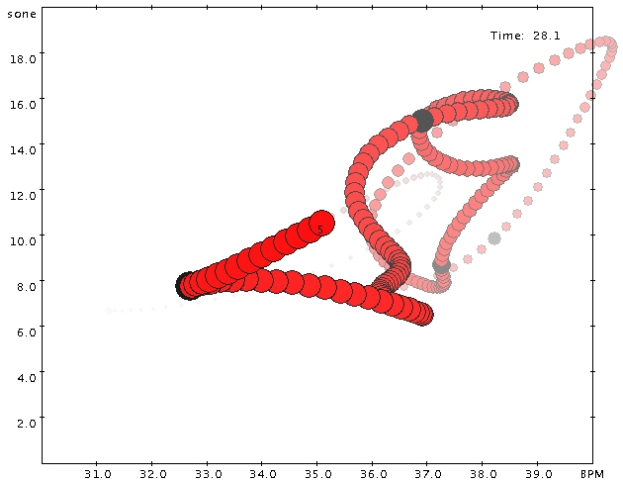


Figure 1: Snapshot of the *Performance Worm* at work: First four bars of Daniel Barenboim's performance of Mozart's F major sonata K.332, 2nd movement. Horizontal axis: tempo in beats per minute (bpm); vertical axis: loudness in *sone*. Movement to the upper right indicates a speeding up (*accelerando*) and loudness increase (*crescendo*) etc. etc. The darkest point represents the current instant, while instants further in the past appear fainter.

the other beat-level loudness changes — or in an integrated two-dimensional way, as trajectories over time in a 2D tempo-loudness space [4]. We have developed a graphical animation tool called the *Performance Worm* [3] that displays such performance trajectories in synchrony with the music. A part of a performance as visualized by the Worm is shown in Figure 1. Note that the display is interpolated and smoothed. For the machine learning experiments reported below, only the actually measured points were used; no interpolation or smoothing was performed.

Thus, the raw data for our experiments is tempo and overall loudness values measured at specific time points in a performance (either every beat according to the time signature or, where beat tracking was performed at lower levels, at subdivisions of the beat). For each measured time point, the following is stored: t_i (absolute time in seconds), B_i (calculated local tempo in bpm), and L_i (loudness level measured in sone).

The raw data so obtained had to be refined in order to be homogeneous and usable in the learning process. In particular, most of the pianists repeated some sections, while others didn't (i.e., Gould). We decided to discard all the non-common repeats, so that the learners would work with comparable data for all the performers, and not with over-represented sections that could affect their output.

2.2. Instances and features

Each measured time point, along with its context, is used as a *training example* for the learning algorithms. In other words, an example or *instance* for the learners is a subsegment of a tempo-loudness trajectory (see Figure 1), centered around a specific time point. Altogether, this procedure results in some 23,000 instances for all the six pianists.

The instances are represented by a set of *features* that are extracted from the raw trajectories. The features were calculated over a sliding window w_i of two bars (the context size), centered on the time point of the instance. Of course, at the beginning and at the end, some of the features were calculated over a window narrower than two bars. The sliding window method produces some redundancy in the data, since the windows were overlapping; this should be an advantage in learning.

For each instance of the original raw data, the following features were computed both for tempo and loudness: the average value within the window $\mu(w_i)$, the standard deviation $\sigma(w_i)$, and the range $R(w_i) = max(w_i) - min(w_i)$. For each of these features, the corresponding normalized ones were also calculated by division by the mean. The normalized features are indicated with lower-case letters of the tempo/loudness subscripts. For example, if $\sigma_B(w_i)$ is the tempo standard deviation, then $\sigma_b(w_i) = \sigma_B(w_i)/\mu_B(w_i)$ is the corresponding normalized version. Additional features were added that represent correlations: $\Sigma_{tB}(w_i)$ is the correlation between the time and the tempo, $\Sigma_{tL}(w_i)$ is the correlation between the time and the loudness, and $\Sigma_{BL}(w_i)$ is the correlation between the tempo and the loudness. *Directness of movement* is a feature that captures aspects of the curvature of a trajectory segment: it measures the ratio between the length of a direct movement from A to B (α) and the length of the actual trajectory between the same points (β) in a two-dimensional space. These features were calculated in different spaces: in the time-tempo space ($\Delta_{tB}(w_i)$), in the time-loudness space ($\Delta_{tL}(w_i)$) and in the tempo-loudness space ($\Delta_{BL}(w_i)$). Finally, some derivatives were calculated for tempo and loudness: the maximum of the absolute value of the derivative ($M\delta(w_i)$), the average of the absolute derivative ($\mu\delta(w_i)$), and the normalized versions too.

Caution had to be taken with some of the extracted performance information. In particular, the features derived from loudness had to be filtered in some way, because they can trivially reveal some of the performers. For example, Gould's recordings are older (1967) than the others (1980-1991), resulting in a significantly lower recording level. That would permit the learners to detect this famous performer simply by absolute loudness difference or by differences in the dynamics range. That is why normalization, especially of the loudness features, is important, and why not all of the above features were then actually used in the learning experiments. Table 3 summarizes all the features, and indicates which ones cannot be used in the learning process.

2.3. The learning algorithms

For our experiments, we selected a representative set of standard machine learning algorithms with different model classes and *biases*. All of these are available in the Waikato Environment for Knowledge Analysis (WEKA) [9].

The following learners were selected: **J48** is a state-of-the art decision tree learner; **NaiveBayes** is a probability-based algorithm that applies Bayes' rule for prediction; **KStar** is a Nearest-Neighbor classification algorithm; **CVR I** (Classification Via Regression) is a 'meta-learner' that induces *linear discriminant functions* for the individual classes and combines these into an n-class classifier by voting; **CVR II** is a 'meta-learner' that induces *M5 model trees* for the individual classes; **JRIP** is a state-of-the-art rule learning algorithm with sophisticated pruning strategies; and **Ridor** is a algorithm for learning so-called 'Ripple-down rules' (basically, rules with exceptions).

All of these algorithms read the same data format and pro-

Operation	Tempo	Loudness	Others
None	B_i	L_i^*	–
Average, St. Dev., Range	$\mu_B(w_i)$, $\sigma_B(w_i)$, $R_B(w_i)$	$\mu_L(w_i)^*$, $\sigma_L(w_i)^*$, $R_L(w_i)^*$	–
Normalization	$b_i(w_i)$, $\sigma_b(w_i)$, $R_b(w_i)$	$l_i(w_i)$, $\sigma_l(w_i)$, $R_l(w_i)$	–
Correlation, Directness	$\Sigma_{tB}(w_i)$, $\Delta_{tB}(w_i)$	$\Sigma_{tL}(w_i)$, $\Delta_{tL}(w_i)$	$\Sigma_{BL}(w_i)$, $\Delta_{BL}(w_i)$
Derivative	$M\delta_B$, $\mu\delta_B$, $M\delta_b$, $\mu\delta_b$	$M\delta_L^*$, $\mu\delta_L^*$, $M\delta_l$, $\mu\delta_l$	–

Table 3: Complete set of attributes and derived features extracted from the data for each instance. *) feature not used by the learner in order not to trivially reveal some of the performers on the basis of the CD recording level.

ID	WEKA name
01	bayes.NaiveBayes -K
02	lazy.kstar.KStar
03	meta.ClassificationViaRegression, -W LinearRegression
04	meta.ClassificationViaRegression, -W trees.m5.M5P
05	trees.j48.J48 -M 50 -C 0.1
06	rules.JRip
07	rules.Ridor

Table 4: Learning algorithms used for pair-wise discrimination.

duce predictive models for discrete classification problems. Table 4 summarizes the algorithms in terms of the precise call (including parameters) by which they are called in WEKA.

2.4. Training and Testing Methodology

For the experiments to be reported here, the original n-class problem was converted to $n(n-1)/2$ two-class discrimination problems, one for each possible pair of pianists. That both gives more insight into the discriminability of various pianists, and is somewhat easier for a classifier than n-class identification. For each pianist pair A-B, the performances by the two respective pianists of the selected training pieces were used for learning, and the task was then to identify the correct pianist in a new test piece, where only recordings by A and B were used for testing.

Recognition accuracy was tested via *cross-validation*. Each of the algorithms was trained on all of the sonata movements except one; the learned classifiers were then tested on recordings of the remaining movement. This process was repeated in a circular fashion, so that each piece served as test piece exactly once for each classifier.

Note that, as explained above, the actual training examples are not entire pieces, but individual time points in pieces, characterized by a set of features. Likewise, the learned classifiers apply to such individual instances, not to entire pieces. To make a classifier predict the pianist for a complete piece, it was applied to all instances making up the piece, and the class (pianist) predicted most often was then chosen as the final prediction for the piece.

3. RESULTS

The results of the cross-validation experiments are summarized in Table 5, which lists the numbers and percentages of correct predictions achieved by the individual classifiers on each pair of performers. For each pianist pair, the classifiers were tested on 24 recordings (12 pieces, played by each of the two pianists). Thus, the maximum possible number of correct predictions is 24, and the *baseline* accuracy — the success rate corresponding to pure guessing — is 12, or 50%. A reasonable learner should at least obtain a recognition accuracy above the baseline.

Looking at the table in terms of learning algorithms first, we note that while all learners obtain recognition rates significantly higher than the baseline, some learners may be better suited to this task than others; the overall prediction accuracies vary between 60 and 70% (Table 5, last row). A closer look shows that not all classifiers perform well or poorly on the same learning problem (pianist pair). That indicates that it may be fruitful to join classifiers into so-called *ensembles* [1] which combine the expertise of a diverse set of classifiers by voting on the class of a new test case.

More interesting is a look at the results in terms of individual pianist pairs. The next-to-rightmost column in Table 5 lists the average recognition accuracy, over all classifiers, for each pair. The figures range from a rather low 54.8% for the pair Barenboim-Uchida (DB-MU) to quite good 72% for the pair Schiff-Batik (AS-RB). These results may not look very exciting — they are far from a perfect recognition rate of 100% — but they are significantly better than the baseline in every case; that indicates that our performance features do contain information that is relevant to the identification of artists. Also, these figures are averages over all classifiers, not all of which are very effective. If we focus only on the best of the 7 learning algorithms — classifier 03 (last column in Table 5) —, the results are much better, with recognition rates ranging from 62.5% (AS-DB, DB-GG, MP-MU) to 87.5% (MP-RB), which we consider quite impressive.

The average recognition rate over all experiments where a given pianist was involved gives a rough measure of the 'distinguishability' of that pianist. Computing this over all classifiers we get the following ranking: Batik (67.8%), Pires (64.8%), Schiff (64.5%), Barenboim (61.9%), Gould (61.8%), and Uchida (58.9%). Our best classifier, 03, yields similar results, but with higher recognition rates: here, the ordering is Pires (75.0%), Batik (74.2%), Schiff (70.8%), Barenboim (69.2%), Gould (68.3%), and Uchida (62.5%).

Mitsuko Uchida seems to be most difficult to recognize (at least in terms of our given features). On the other end of the scale, we find Roland Batik and Maria João Pires as the most easily distinguishable ones. Listening to the performances by Batik (a local Viennese pianist), we find that he indeed sounds quite different compared to the more famous pianists (though he is a highly skilled pianist). Among the more famous pianists, Pires and Schiff are the two that seem most easily identifiable. In fact, classifier 03 achieves its second-best result in the direct discrimination between Pires and Schiff (83.3%).

Interestingly, this last finding coincides with results of another empirical performance style study we are currently conducting [8]. Using novel visualization techniques, we managed to show clear stylistic differences between Pires and Schiff. Our next step will be to identify those features that contribute most to correct prediction, in order to get closer to a formal characterization of the stylistic differences.

	Classifiers									
Pair	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**Ave.**	**Ave.** [%]	**Classifier 03** [%]
AS - DB	16	15	15	14	15	15	14	14.9	61.9	62.5
AS - GG	14	15	17	13	14	14	13	14.3	59.5	70.8
AS - MP	15	16	20	19	18	13	14	16.4	68.5	83.3
AS - MU	13	14	17	17	13	14	14	14.6	60.7	70.8
AS - RB	16	17	16	20	15	19	18	17.3	72.0	66.7
DB - GG	16	16	15	13	13	13	13	14.1	58.9	62.5
DB - MP	17	15	17	14	14	15	13	15.0	62.5	70.8
DB - MU	13	13	17	13	11	13	12	13.1	54.8	70.8
DB - RB	19	20	19	15	14	16	17	17.1	71.4	79.2
GG - MP	15	16	17	16	15	15	16	15.7	65.5	70.8
GG - MU	15	15	16	14	15	14	14	14.7	61.3	66.7
GG - RB	13	17	17	17	13	14	16	15.3	63.7	70.8
MP - MU	13	16	15	14	13	11	13	13.6	56.5	62.5
MP - RB	15	18	21	18	16	15	16	17.0	70.8	87.5
MU - RB	16	17	16	12	14	15	13	14.7	61.3	66.7
Average	15.1	16.0	17.0	15.3	14.2	14.4	14.4	–	–	–
Average [%]	62.9	66.7	70.8	63.8	59.2	60.0	60.0	–	–	–

Table 5: Pair-wise recognition results: classification accuracy in terms of correctly classified pieces. The maximum possible number of correct predictions in each pianist pair is 24 (12 pieces $\times$ 2 pianists); the *baseline accuracy* (the accuracy one would achieve by random guessing) is 50%. Rightmost column: accuracy (percentage) achieved by classifier **03**.

4. CONCLUSION

This paper has presented experimental results that show that machines may be capable of recognizing famous artists on the basis of their style, at least to some extent. The machine learning algorithms achieved recognition rates significantly above the baseline, which indicates that they can pick up some relevant information from the recordings. At first sight, some of the absolute accuracy figures may look rather poor. However, our current training data (performance measurements) are extremely crude and incomplete, comprising only beat-level tempo and beat-level overall loudness. Information about the loudness of individual voices, about articulation, about the timing of individual voices, etc. would be highly relevant, but is hard to obtain from audio recordings.

The next steps on our research agenda are the following. First, we want to identify those features that contribute to the discernibility of individual pianists. That will lead us towards an explicit characterization of aspects of the individual style of pianists. Also, by reducing the set of features to those that are truly relevant, additional improvements in classification accuracy may be achievable. The second important direction of work concerns the refinement of the performance measurements. We are currently working on methods to extract more detailed information regarding timing and dynamics from audio recordings.

We are confident that the current results can still be considerably improved, and that, by continuing this line of research, we may make progress towards better understanding stylistic differenceds between great artists.

5. ACKNOWLEDGMENTS

This research was supported by an ERASMUS scholarship to the first author, the EC project HPRN-CT-2000-00115 MOSART, and the project Y99-INF (START Prize by the Austrian Federal Government). The Austrian Research Institute for Artificial Intelligence (ÖFAI) acknowledges basic financial support from the Austrian Federal Ministry for Education, Science, and Culture.

6. REFERENCES

[1] Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *First International Workshop on Multiple Classifier Systems*. New York: Springer Verlag.

[2] Dixon, S. (2001). Automatic Extraction of Tempo and Beat from Expressive Performances. *Journal of the New Music Research* 30(1), 39–58.

[3] Dixon, S., Goebl, W., & Widmer, G. (2002). The Performance Worm: Real Time Visualization of Expression based on Langner's Tempo-Loudness Animation. *Proc. International Computer Music Conf. (ICMC 2002)*, Göteborg.

[4] Langner, J., and Goebl, W. (2003). Visualizing Expressive Performance in Tempo-Loudness Space. *Computer Music Journal* (in press).

[5] Stamatatos, E. (2002). Quantifying the Differences between Music Performers: Score vs. Norm. *Proceedings of the International Computer Music Conference (ICMC'2002)*, Göteborg, Sweden.

[6] Stamatatos, E. and Widmer, G. (2002). Music Performer Recognition Using an Ensemble of Simple Classifiers. *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI'2002)*, Lyon, France.

[7] Widmer, G. (2001). Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report. *AI Communications* 14(3), 149–162.

[8] Widmer, G., Dixon, S., Goebl, W., Pampalk, E. & Tobudic, A. (2003). In Search of the Horowitz Factor. *AI Magazine* (in press).

[9] Witten, I.H. & Frank, E. (1999). *Data Mining*. San Francisco, CA: Morgan Kaufmann.

[10] Zanon, P. and Widmer, G. (2002). Recognition of Famous Pianists Using Machine Learning Algorithms: First Experimental Results. *Proc. of the XIV Colloquium on Musical Informatics (CIM'2002)*, Florence, Italy, pp. 84-89.

MUSIC PERCEPTION

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

AUTOMATIC TRANSCRIPTION OF MUSIC

Anssi P. Klapuri

Institute of Signal Processing,
Tampere University of Technology, P.O.Box 553, FIN-33101 Tampere, Finland
klap@cs.tut.fi

ABSTRACT

The aim of this tutorial paper is to introduce and discuss different approaches to the automatic music transcription problem. The task is here understood as a transformation from an acoustic signal into a MIDI-like symbolic representation. Algorithms are discussed that concern three subproblems. (i) Estimation of the temporal structure of acoustic musical signals, the musical meter. (ii) Estimation of the fundamental frequencies of concurrent musical sounds. (iii) Higher-level musicological modeling to resolve otherwise ambiguous situations. The emphasis is laid on multiple-F0 estimation. Validation experiments are performed using both synthesized and real-world music signals. Demonstration signals are available at http://www.cs.tut.fi/~klap/iiro/smac/.

1. INTRODUCTION

Written music is traditionally presented as a *musical notation* (score) which comprises the times, durations, and pitches of the sounds that constitute a piece. The aim of music transcription is to discover such a musical "recipe" in an acoustic signal, so that a musician or a synthesizer program can reproduce and modify the original performance. Sometimes a more coarse symbolic representation suffices, for example writing down the chords only.

Automatic transcription of music (AToM) is difficult and includes several subproblems to be discussed in the coming sections. Figure 1 gives an idea of the parts an automatic transcriber, although a fixed structure for such systems does not exist. *Multiple-F0 estimation* refers to a core part which estimates the fundamental frequencies (F0s) of several concurrent sounds. Usually, internal models are needed along with the acoustic signal to perform the analysis. Comparing with speech recognition, *musicological models* play the role of a "language model" in music.

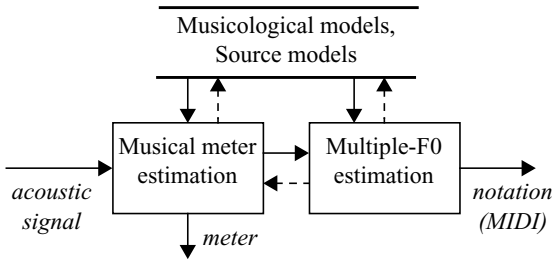


Figure 1. *Relationships of the parts of a transcription system.*

Musical meter characterizes the temporal regularity of a music signal. Musical meter is a hierarchical structure which consists of pulse sensations (periodicities) at different levels [1]. Figure 2 shows the meter at three relatively well-defined levels: *beat* (foot tapping rate), *tatum* (time quantum), and musical *measure* pulse levels. Meter estimation alone has several applications in the synchronization,editing,and analysis of music signals.

In principle, it would be advantageous to perform meter estimation and multiple-F0 estimation in parallel. However, meter estimation can be done relatively robustly and allows the positioning and sizing of the analysis frames in further analysis. For these reasons, it is usually computationally more efficient to perform the two stages in a cascade. Feedback from multiple-F0 stage is needed especially to analyze musical measure pulse which is related to harmonic change rate.

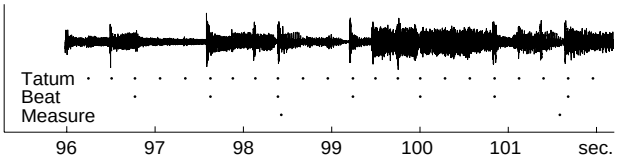


Figure 2. *A musical signal with three metrical levels illustrated.*

2. MUSICAL METER ESTIMATION

Table 1 summarizes some state-of-the-art systems for musical meter estimation. Earlier work has concentrated almost solely on beat tracking, with the few exceptions in [2,3,7]. Most earlier systems have attempted to discover periodicities in symbolic input (MIDI) [2,3,4, see 4 for a review]. Only a few systems have been proposed that process acoustic input signals [4,5,6,7].

Our meter estimation system has been described in [8]. It analyzes meter at the beat, tatum, and measure pulse levels and is able to process acoustic musical signals from all main musical genres. The method consists of three parts. (i) Time-frequency analysis part computes *registral accent signals* which measure the degree of accentuation (emphasis) at four frequency ranges of the input signal. This part is a generalization based on [5] and [6] and has a great impact on performance. (ii) A bank of comb filter resonators is used at each frequency channel to analyze periodicity. The resonators are quite similar to those used by Scheirer in [5]. However, the exact periodicity analysis method is not critical. Among four different periodicity analysis techniques, three performed equally well, and the simplest (comb filters) was chosen. (iii) A musically informed probabilistic model was used to represent dependencies between the different levels of meter and between temporally successive meter estimates. This stage improved the robustness and temporal stability of the system.

3. APPROACHES TO MULTIPLE-F0 ESTIMATION

In this section, we introduce a number of different approaches that have been taken to solve the multiple-F0 estimation problem.

Table 1: *Characteristics of meter estimation systems*

System	Input data	Aim	Target material
Large, Kolen	MIDI	meter	not specified
Goto, Muraoka	audio	meter	pop music, 4/4 time
Scheirer*	audio	beat	mainly 'strong beat'
Dixon*	MIDI & audio	beat	MIDI: all music types
Temperley*	MIDI	meter	all music types
Klapuri	audio	meter	all music types

* Source codes available for download.

Subheadings are provided to improve readability, but it should be noted that any of the cited papers really cannot be put under a single label. Attempts toward multiple-F0 estimation date back to the 1970s. For a brief historical overview, see [21]. In following, an attempt has been made to list the definitive attributes of a representative set of more recent papers.

3.1 Perceptual grouping of frequency partials

Any algorithm that finds the F0s of multiple concurrent sounds is, in effect, also organizing spectral componets to sound sources [9,p.240]. This organization task is called *auditory scene analysis* (ASA) and human auditory system is very good at it. An important step forward in ASA was taken when Bregman pointed out distinct perceptual cues for grouping time-frequency components into sources. The cues (component features) were briefly: proximity in time-frequency, harmonic frequency relationships, synchronous changes, and spatial proximity [9].

Kashino *et al*. brought Bregman's ideas to music scene analysis and also proposed several other new ideas for AToM [10]. The front-end of their system used a "pinching plane method" to extract continuous frequency components from the input data. These were clustered into note hypotheses by applying the above mentioned perceptual rules. *Timbre models* were used to identify the source of each note and pre-stored *tone memories* were used to resolve overlapping frequency components. Chordal analysis was performed based on the probabilities of notes to occur under a given chord. Chord transitions probabilities were encoded into trigram models (Markov chains). For computations, a Bayesian probability network was used to integrate the knowledge and to do simultaneously bottom-up analysis, temporal tying, and top-down processing (chords predict notes and notes predict components). Evaluation material comprised five different instruments and polyphonies of about three simultaneous sounds. The work still stands among the most elegant and complete AToM systems.

A more recent example of the perceptual grouping approach is the PhD work of Sterian [11]. He used Kalman filtering to extract continuous sinusoidal partials and represented the grouping rules as a set of likelihood functions which evaluated the likelihood of observed partials given a hypothesized grouping.

3.2 Auditory-model based approach

The "unitary pitch model" of Meddis and Hewitt has had a strong influence on F0 estimation research [12]. While Bregman's theory focused on what happens in the brain, Meddis and Hewitt modeled the more peripheral (largely physiological) parts of hearing. Although multipitch estimation in sound mixtures was not addressed, research to this direction was inspired, too.

Cheveigne and Kawahara extended the unitary pitch model to the multiple-F0 case. They proposed a system where pitch estimation was followed by the cancellation of the detected sound, and the estimation was repeated for the residual signal [13]. Also a computationally exhaustive joint estimator was proposed. Although evaluation results were shown for a rather artificial data, the proposed iterative scheme was really a successful one. Temporal continuity between frames was not considered.

Tolonen and Karjalainen developed a computationally efficient version of the unitary pitch model and applied it to the multiple-F0 estimation of musical sounds [14]. The spectrum of an incoming sound was flattened using inverse warped-linear-prediction filtering. In pitch computations, only two frequency channels were used instead of the 40–120 channels in the original model, yet the characteristics of the model were mainly preserved. Practical robustness of the system was improved by introducing a FFT-based "generalized autocorrelation" method. Extension to multipitch estimation was achieved by cancelling subharmonics in the output of the model. From the resulting *enhanced summary autocorrelation function*, all F0s can be picked without iterative estimation and cancellation. The method is relatively accurate and can be implemented based on [14].

3.3 Blackboard architecture with auditory front-end

Blackboard architectures were developed to facilitate the integration of different types of knowledge for signal analysis. The *blackboard* is hierarchy of data representations at different analysis (abstraction) levels. The data is common to a set of autonomous *knowledge sources* which operate when requested.

Martin proposed a system for transcribing piano performances of four-voice Bach chorales [15]. In his system, an auditory model was used as a front-end to a blackboard, which employed knowledge about physical sound production, rules governing tonal music, and "garbage collection" heuristics. Support for F0s was raised on a frame-by-frame basis and then combined with the longer-term power envelope information to create note hypotheses. Musical rules favoured F0s in certain intervallic relations.

A more recent model of Godsmark and Brown was particularly designed to facilitate the integration of different auditory organization principles and competition between them [16]. The applied auditory front-end produced "synchrony strands" each of which represented a dominant time-frequency component. These were fused to sound events by extracting features from each strand and by applying Bregman's primitive organization principles. Sound events were further grouped to their respective sources (event "streams") by computing the pitch and timbre proximity between sounds. Musical meter information was used to predict when events will occur and melodic pattern induction to predict recurrent patterns. The model was evaluated by showing that it could segregate melodic lines from polyphonic music and to resolve interleaved melodies. Transcription accuracy as such was not the main goal.

3.4 Signal-model based probabilistic inference

It is possible to state the whole multiple-F0 estimation problem in terms of a signal model, the parameters of which should be estimated. Consider e.g. the model [17]:

$$y_t = \left\{ \sum_{k=1}^{K} \sum_{m=1}^{M_k} a_{k,m}\cos[m\omega_k t] + b_{k,m}\sin[m\omega_k t] \right\} + v_t$$

where K is the number of simultaneous notes, M_k is the number of partials in note k, ω_k is the fundamental frequency of note k, and $a_{k,m}$, $b_{k,m}$ together encode the amplitude and phase of individual partials. The term v_t is a residual noise component.

In principle, *all* the parameters on the right-hand side of the above equation should be estimated, based on the observation y_t and possible prior knowledge about the parameter distributions. As pointed out by Davy *et al.* in [17], the problem is Bayesian since there is a lot of prior knowledge concerning music signals.

Davy and Godsill elaborated the above signal model to accommodate time-varying amplitudes, non-ideal harmonicity, and non-white residual noise [17]. Prior distributions for the parameters were carefully selected. An input signal was first segmented into excerpts where no note transitions occur. Then the parameters of the signal model were estimated in *time domain*, separately for each segment. The main problem of this approach is in the actual computations. For any sufficiently realistic signal model, the parameter space is huge and the posterior distribution is highly multimodal and strongly peaked. Davy and Godsill used Markov chain Monte Carlo sampling of the posterior, reporting that much of the innovative work was spent on finding heuristics for the fast exploration of the parameter space [17]. Although computatinally inefficient, the system was reported to work robustly for polyphonies up to three simultaneous sounds.

Goto has proposed a method which models the *short-time spectrum* of a music signal as a weighted mixture of tone models [18]. Each tone model consists of a fixed number of harmonic components which are modeled as a Gaussian distributions centered at integer multiples of the F0 in the spectrum. Goto derived an expectation-maximization (EM) algorithm which iteratively updates the weights of each tone model and their relative harmonic amplitudes, leading to a maximum *a posteriori* estimate. Temporal continuity was considered by tracking framewise F0 weights in a multiple-agent architecture. Goto used the algorithm successfully to detect melody and bass lines in CD recordings.

3.5 Data-adaptive techniques

In data-adaptive systems, there is no parametric model or other knowledge of the sources. Instead, the source signals are estimated from the data. Typically, it is not even assumed that the sources (notes) have harmonic spectra! For real-world signals, the performance of e.g. independent component analysis alone is poor. However, by placing restrictions for the sources, the data-adaptive techniques become applicable in realistic cases. Such restrictions are e.g. independence of sources and *sparseness* which means that the sources are inactive most of the time.

Virtanen added *temporal continuity* constraint to the sparse coding paradigm [20]. He used the signal model

$$X_t(f) = \sum_{n=1}^{N} a_{t,n} S_n(f) + E_t(f), \qquad (1)$$

which represents the zero-phase power spectrogram $X_t(f)$ of the input as a linear sum of N static source spectra $S_n(f)$ with time-varying gains $a_{t,n}$. The term $E_t(f)$ is error spectrum. An iterative optimization algorithm is proposed which estimates non-negative $a_{t,n}$ and $S_n(f)$ based on the minimization of a cost function which takes into account reconstruction error, sparseness, and temporal continuity. The algorithm was used to separate pitched and drum instruments from real-world music signals.

Another recent example of applying sparse coding to music is that of Abdallah and Plumbley [19].

3.6 Our system

The multiple-F0 estimator proposed by us in [21] is closest to the auditory-model based approach (Sec. 3.2), since psychoacoustics was used as a base of the analysis principles. However, the method is better characterized as problem-solving oriented, focusing on the AToM application, not on auditory modeling.

There is a number of identifiable problems which a multiple-F0 estimator must address in order to resolve real-world music signals. (i) The likelihoods of different F0s must be calculated robustly in the presence of other, co-occurring sounds. (ii) The effect of a true F0 must be cancelled from its harmonics and subharmonics which usually appear as the next-most-likely F0s. (iii) Number of concurrent sounds has to be estimated. (iv) Real-world sounds are often not perfectly harmonic. (v) Generality in regard to different instruments. (vi) Robustness in noise.

Our multiple-F0 method is based on the iterative estimation and cancellation approach (see Sec. 3.2 and [21]). At the estimation stage, problem (i) is addressed by estimating weights (likelihoods) of F0s independently at separate frequency bands, and by letting only a set of selected frequency samples to contribute to the weight, not the overall spectrum. Problem (iv) is addressed by letting the series of partials at subbands be shifted in frequency. Problem (ii) is addressed by the iterative scheme where the spectrum of a detected sound is estimated and subtracted from the mixture, after which estimation is repeated for the residual. Statistical distribution of the features of detected sounds is used to control the stopping of the iteration and thus to estimate polyphony. Points (v) and (vi) are addressed in a preprocessing step which flattens sound spectra and suppresses noise. All the computations are done for a single frame in the frequency domain.

4. MUSICOLOGICAL MODELS

Linguistic information allows speech recognition systems to interpret obscured and ambiguous signals. Musical information is equally important for music transcription. Some systems use such internal models [10,15,16,18], as already described in Sec. 4.

Temperley has presented a very comprehensive rule-based system which models the cognition of basic musical structures [3]. Such a system could be readily used as a post-processor for a musically agnostic transcription system. However, to really benefit of the musical knowledge, it should be utilized already *during* the analysis. Also, while a rule-based model reveals a lot about the human cognition, *probabilistic* models are advantageous in that they evaluate the likelihoods of several candidate analyses.

Consider the following experiment. We represent c*hord unigrams* as 12-bit numbers where each bit signifies the presence/absence of one of the 12 pitch classes (octave equivalence). There are 4096 such unigrams. 359 MIDI pieces were collected and cut into segments where note onsets or offsets do not occur. The harmonic content of each segment was then represented with the appropriate unigram. The probability of occurrence for each uni-

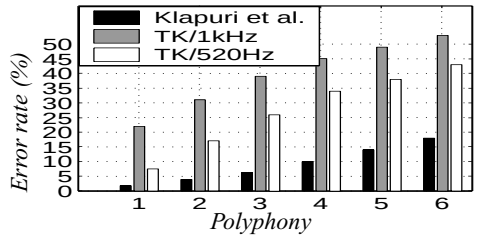


Figure 3. *Error rates of the system of Klapuri and the reference system of Tolonen et al. (with tighter upper limits for F0s).*

gram was computed within the pieces and averaged over all pieces. The results were interesting: among the 30 most probable unigrams were the 12 single notes, seven different major triad chords, five minor triads, and three minor-seventh chords.

The described kind of "brute force" statistical approach has several advantages. (i) The estimated probabilities can be used to rate the likelihoods of several competing F0 mixtures in an AToM system. (ii) The above procedure has *no* heuristic parameters. (iii) No musical expertise was employed, yet the system knows about major and minor triads, the building blocks of Western harmony. (iv) New kind of musical material can be focused simply by retraining the system using the target material.

5. SIMULATION EXPERIMENTS

For meter estimation, we have earlier presented a statistical evaluation of different beat tracking systems using a database of 478 musical pieces and implementations of the original authors. This will not be repeated here. An interested reader is referred to [8].

Reliable comparison of multiple-F0 estimation algorithms is difficult because the systems are typically very complex and source codes are not available. However, a few methods can be replicated. The core EM-algorithm of Goto is implementable based on [18] (but not front-end and post-processing). The EM-algorithm estimates the weights of all F0s, but usually only the predominant F0 was found in our simulations, exactly as claimed by Goto. Also, Tolonen and Karjalainen describe their algorithm and its parametes to sufficient detail to be exactly implementable [14]. Error rates for this reference system (*TK*) and our system are given in Figure 3. As reported by the authors, the method cannot handle "spectral pitch", i.e., F0s above 1 kHz. It was further found out here that the method is best at detecting F0s in the three-octave range between 65 Hz and 520 Hz. Thus, in the simulations the mixtures given to the *TK* method were restricted to contain F0s only below either 520 Hz or 1 kHz, as specified in the results. Given the compactness of the *TK* algorithm, the results are very good.The number of sounds was known to both methods.

Transcription demonstrations for real-world music and synthesized MIDI are at http://www.cs.tut.fi/~klap/iiro/smac/.

6. REFERENCES

[1] Lerdahl, F., Jackendoff, R., *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA, 1983.

[2] Large, E. W., Kolen, J. F., "Resonance and the perception of musical meter". Connection science, 6(1), 1994,pp.177-208.

[3] Temperley, D. *Cognition of Basic Musical Structures*. MIT Press, Cambridge, MA, 2001.

[4] Dixon, S., "Automatic Extraction of Tempo and Beat from Expressive Performances," *J. New Music Research* 30 (1), 2001, pp. 39-58.

[5] Scheirer, E. D., "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.* 103 (1), 1998, pp. 588-601.

[6] Goto, M., Muraoka, Y., "Beat Tracking based on Multiple-agent Architecture — A Real-time Beat Tracking System for Audio Signals," *In Proc. Second International Conference on Multiagent Systems*, pp.103–110, 1996.

[7] Goto, M.,Muraoka,Y., "Issues in Evaluating Beat Tracking Systems," *IJCAI-1997 Workshop on Issues in AI and Music*.

[8] Klapuri, A. P., "Musical meter estimation and music transcription," In Proc. Cambridge Music Processing Colloquium, Cambridge University, UK, March 2003.

[9] Bregman, A.S.,"Auditory Scene Analysis," MIT Press,1990.

[10] Kashino, K., Nakadai, K., Kinoshita, T., and Tanaka, H., "Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitative Information Integration Mechanism," Proc. International Joint Conf. on Artificial Intelligence, 1995.

[11] Sterian, A. D., "Model-based segmentation of time-frequency images for musical transcription," Ph.D.thesis, University of Michigan, 1999.

[12] Meddis, R. and Hewitt, M. J., "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J.Acoust. Soc. Am. 89(6),p.2866–2882, 1991.

[13] de Cheveigné, A. and Kawahara, H., "Multiple period estimation and pitch perception model," Speech Communication 27, pp. 175–185, 1999.

[14] Tolonen, T. and Karjalainen, M., "A computationally efficient multipitch analysis model," IEEE Trans. Speech Audio Processing, Vol. 8, No. 6, pp. 708-716, Nov. 2000.

[15] Martin, K. D., "Automatic transcription of simple polyphonic music: robust front end processing," Massachusetts Institute of Technology Media Laboratory Perceptual Computing Section Technical Report No. 399, 1996.

[16] Godsmark, D. and Brown, G. J., "A blackboard architecture for computational auditory scene analysis," Speech Communication 27, pp. 351–366, 1999.

[17] Davy, M. and Godsill, S. J., "Bayesian harmonic models for musical signal analysis, " In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics VII*, Oxford University Press, 2003.

[18] Goto, M., "A predominant-F0 estimation method for real-world musical audio signals: MAP estimation for incorporating prior knowledge about F0s and tone models," in Proc. Workshop on Consistent and reliable acoustic cues for sound analysis, Aalborg, Denmark, Sep. 2001.

[19] Abdallah, S. A. and Plumbley, M. D., "Sparse coding of music signals," Submitted for Publication.

[20] Virtanen, T., "Sound source separation using sparse coding with temporal continuity objective," International Computer Music Conference, Singapore, 2003.

[21] Klapuri, A., Virtanen, T., Holm, J.–M., "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals". In Proc. COST-G6 Conference on Digital Audio Effects, DAFx-00, Verona, Italy, 2000.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

BRAIN ACTIVITY IN PERCEPTION AND RETENTION IN MEMORY OF THE PITCH OF MUSIC AND SPEECH.

Castro-Sierra, E.[1] & Poblano, H. A.[2]

[1]National School of Music, UNAM, and Hospital Infantil de México Federico Gómez; [2]MS Program in Neurological Rehabilitation, UAM. México, D. F., Mexico

eduardcs@yahoo.com.mx

ABSTRACT

Brain activity stimulated by music and speech has been studied using fMRI and PET. These techniques commonly employ single stimuli which contrast with others in a sequence. Neural activity is elicited when one of the sounds differs from the remaining ones. Such results have led to a distinction between activities of right-hemisphere areas stimulated by music sounds and activities of left-hemisphere areas stimulated by speech sounds. EEG was used to study activities of different frequencies in superior and inferior frontal, temporal and parieto-occipital brain areas of 10 subjects of either sex (age range: 11:11-16:8), native speakers of Zapotec (a Mexican tone language), while responding to a tonal memory test contrasting pairs of musical sequences, a tonal memory test contrasting pairs of speech sequences and a tonal perception test analyzing single speech sequences. α- and β-wave activities in right and left inferior frontal areas correlated with responses to either the music or the speech memory tests. Right parieto-occipital θ-wave activity also correlated with responses to the music test. α- and β-wave activities in left superior frontal, temporal and parieto-occipital areas correlated with responses to the speech perception test. These data point to differential activity of both hemispheres of the brain stimulated by the pitch of tone language samples and to a more localized right hemisphere activity stimulated by the pitch of music samples.

1. INTRODUCTION

Recently, Zatorre et al. [1] have examined the ways in which speech and musical sounds differ from each other in regard to their use of contrasting acoustic cues. Since speech depends on rapidly changing broadband sounds, whereas tonal patterns are slower and employ narrowband sounds, these authors maintain that the auditory cortices in the left hemisphere of the brain are specialized to achieve better temporal resolution (language), whereas the auditory cortices in the right hemisphere of the brain are specialized for better spectral resolution (music). In support of these contentions, they bring forward the results of a number of studies addressing either anatomical or processing asymmetries in the auditory cortex of humans or of other species [2,3], the acoustic uncertainty principle [4] (based on von Heisenberg's 'Uncertainty Principle'), in reference to the impossibility of measuring time and frequency variations simultaneously, and patterns of activation of either left or right auditory cortical areas that vary depending on either the rate of temporal change or the number of spectral elements, respectively [5].

In the first [2] and fourth [5] studies, Zatorre and his associates utilized either anatomical non-functional magnetic resonance (MRI) or functional positron emission tomography (PET) techniques. In the latter, the stimulus parameters that varied were rate of frequency modulation and spectral distribution of elements within a single pattern. Here, the results were that, indeed, corresponding areas of the auditory cortex in either hemisphere showed contrasting responses to one or the other type of acoustic stimulation. From these data, the authors concluded that subtle quantitative differences in the responses at an early step in the processing stream could lead to qualitatively distinct functional roles for higher-order processes. In support of this idea, they argued that neuroimaging studies of linguistic pitch processing in tone languages, such as Thai or Chinese [6,7,8], have demonstrated that left hemisphere structures are well-suited for processing pitch contours in speakers of such languages, whereas non-speakers process the same stimuli using right-hemisphere mechanisms, i.e., as if they were musical.

1.1 Contour tone-language studies

In one of the above mentioned language studies, Klein and her associates [8] used PET to analyze the patterns of changes in cerebral blood flow (CBF) between native Mandarin-speaking and native English-speaking adults when performing same-different judgments of tone in Mandarin word pairs. Mandarin is a contour tone language in which changes in the direction of the fundamental frequency within the vowel of a monosyllabic word provide the latter with a lexical meaning.

When tone discrimination was analyzed relative to a silent baseline, both groups in the study showed common regions of CBF increase. These were observed bilaterally along the superior temporal gyrus, in the lateral cerebellum and the thalamus. Furthermore, unilateral increases were also observed in the right inferior parietal lobule and the left medial frontal and postcentral gyri in both groups. On the other hand, some increases, mainly evident in the left hemisphere, were observed in the superior temporal gyrus, the superior parietal lobule and the cerebellar midline of native Mandarin speakers. Finally, increases in the right hemisphere of native English speakers were observed in the superior temporal gyrus and different regions of the frontal cortex.

On performing a subtraction of tone discrimination minus silent baseline for the Mandarin-speaking group from tone discrimination minus silent baseline for the English-speaking group, however, the following data were found: CBF changes for the Mandarin-speaking group were all in the left hemisphere; in the ventromedial orbital frontal cortex, frontopolar cortex,

pre- and post-central gyri in inferior and superior parietal cortex, and lateral occipitotemporal and middle occipital gyri. In opposition to this, CBF changes for the English-speaking group were observed only in the right hemisphere, to wit, in the right ventrolateral frontal cortex, anterior orbitofrontal gyrus, lateral orbital gyrus, the cingulated region and the superior temporal gyrus.

Regarding these differential activations of the left and right hemisphere of the brain of Mandarin- and English-speaking subjects, respectively, Klein et al. [8] considered that they could reflect differences in perception of natural speech stimuli by two groups of subjects that were either accustomed to phonological, lexical and semantic processing of such stimuli or not accustomed to them, and therefore perceived them as non-words. However, the results could also be due to differences in strategies for processing pitch changes.

2. PRESENT STUDY

An explanation of the data obtained in the former study [8] of differential pitch perception of speech sounds by persons speaking a contour tone language and those not speaking that language, for example, could also be that the stimuli were presented in a sequence made up of two words which could be either the same or different. Thus, the subjects were in a position to make a choice only as to whether the stimuli contrasted acoustically, but not necessarily along a phonological-lexical-semantic continuum. It could be that the same words presented in isolation would have allowed either Mandarin-speaking or English-speaking subjects a more direct possibility for making such linguistic analyses. Such approaches are not easily amenable to fMRI or PET imaging techniques, since the latter usually require comparisons between stimuli. However, they can be implemented with electroencephalographic (EEG) techniques, which allow measurements of brain activity while the subject is carrying out a test of single stimulus perception and analysis.

2.1. Register tone-language studies

There are several tone languages in Central and Southeastern Mexico, which are spoken by Native Mexican communities. Besides being polysyllabic, these languages differ from the contour tone languages of Asia in presenting changes in tone between syllables, which provide lexical meanings, rather than within the vowel of a single syllable. These languages are referred to as register (or level) tone languages.

In 1999, the Laboratory of Psychoacoustics and Auditory Physiology at Hospital Infantil de México Federico Gómez of the Mexican Ministry of Health in Mexico City started a research program to study the development of pitch perception in music and speech among young (6 to 14 years-of-age) monolingual speakers of either Otomi or Zapotec, which are register tone languages spoken in Central and Southeastern Mexico, respectively. The results obtained in a series of studies [9,10] within this program have stressed the importance of an exposure from a very early age to the lexical variations in tone of the languages of these communities that will allow for accurate discrimination, perception and linguistic analysis of the acoustic and phonetic-lexical values of those variations.

2.1.1. Subjects and stimuli

In one of these studies [10], a group of 10 subjects -4 males (age range: 12:2 to 16:2) and 6 females (age range: 11:11 to 16:8)- who were native speakers of Zapotec, right-handed, and without specialized training in music, were subjected to three different psychoacoustic tests: Bentley's tonal memory test contrasting pairs of musical sequences (pairs of 5-note tunes) [11], a tonal memory test contrasting pairs of Zapotec speech sequences (pairs of 2-word phrases) [12] and a tonal perception test analyzing single Zapotec speech sequences (single 2-word phrases) [12]. While each subject was doing the tests, an EEG analysis of θ- (4<8 Hz), α- (8<13 Hz) and β- (13<21 Hz) wave activities in the inferior and superior frontal, temporal and parieto-occipital areas of both hemispheres of his/her brain was carried out with a 16-channel, digital EEG equipment (Vector Bioscience). The speech samples were recorded by an adult male speaker of Zapotec, and the music samples were synthesized with a clarinet (quasi-square wave) spectrum.

2.1.2. Procedures

For the tonal memory test in music, the subjects had to determine which of the 5 notes had changed in pitch after the second presentation of the first tune of each pair. This was supposed to define their ability to retain in auditory short-term memory (A-STM) the pitches in a sequence of 5 notes and to compare them with the pitches in a second sequence of 5 notes.

For the tonal memory test in speech, the subjects had to determine if the two phrases presented in each pair were the same or different depending on the differences or similarities in pitch appearing between them. This was supposed to define their ability to retain in A-STM the pitches in a sequence of 2 words and to compare them with the pitches in a second sequence of 2 words.

For the tonal perception test in speech, the subjects had to determine whether the fundamental frequency movement of each single phrase was upward or downward. This was supposed to define their ability to perform an acoustic, as well as phonological-lexical analysis of the pitches of an isolated sequence of 2 words.

In the course of the EEG analysis, the activities of θ, α and β waves were measured without the subjects' doing any task. Next, while the subjects were carrying out each of the above tests, the same measurements were undertaken again. Finally, the values obtained while the subjects were not doing any task were subtracted from the values obtained while the subjects were carrying out each test. The resulting values served as a measure of the corresponding brain activities.

2.1.3. Results and discussion

The results obtained in the two language tests were uncorrelated with the results obtained in the music test. The results of the two language tests were also uncorrelated with each other.

The univariate correlation coefficient, *r*, comparing the answers to each of the three psychoacoustic tests with θ-, α- and β-wave activities in different areas of the brain, are provided in Table 1.

Certain aspects of the cognition of music and speech in young Zapotec-speaking subjects, such as the discrimination of

fundamental frequency contrasts between either two music samples or two speech samples presented in sequence, would appear to be carried out in the same area of the brain, namely, the right inferior frontal region, where auditory short-term (i.e., working) memory allows the hearer to retain in memory for a short time the difference in fundamental frequency perceived between the second sample and the first sample.

On the other hand, as is generally the case with right-handed people, in these same subjects, different areas of the left hemisphere, to wit, the parieto-occipital region, the temporal region and the superior frontal region, would seem to be active in detecting and localizing an isolated speech sample, and in perceiving and analyzing according to the phonological-lexical rules of their language the fundamental frequency of that same sample. Furthermore, activities directed towards detecting and localizing music samples are apparently present in the right parieto-occipital region of the brain of these subjects.

Test	Brain area/frequency	r
Test of tonal memory in music	Right inferior frontal/α	0.69 (p = 0.042)
"	Left inferior frontal/α	0.81 (p = 0.008)
"	Right parieto-occipital/θ	0.78 (p = 0.01)
Test of tonal memory in speech	Right inferior frontal/β	0.70 (p = 0.036)
Test of tonal perception in speech	Left superior frontal/α	0.79 (p = 0.012)
"	Left temporal/α	0.69 (p = 0.038)
"	Left parieto-occipital/β	0.70 (p = 0.038)

Table 1: *Correlations of answers to the Test of Tonal memory in Music, Test of Tonal Memory in Speech and Test of Tonal Perception in Speech with θ-, α- and ß-wave activities in different areas of the brain of the subjects of the Zapotec-language study.*

These data appear to run somewhat counter to the considerations provided by Zatorre and his associates [1] about the left hemisphere of the human brain being better adapted to handle fast, broadband stimuli, such as are to be found in speech, and the right hemisphere being more specialized to perceive slow, narrowband stimuli, such as are to be found in music.

It could be that, insofar as tone languages are concerned, a brief speech sample may be perceived as a tonal stimulus, not really different from a musical one, when comparing it to another speech sample with the same phoneme sequence but with a different tonal register or contour, and both be discriminated and held in auditory working memory in the right inferior frontal area of the brain.

In opposition to this, and in agreement with the results of the present study, this same short sound stimulus is likely to be perceived as a speech sample when appearing in isolation, and be analyzed in different areas of the left hemisphere of the brain.

Moreover, a relatively longer musical sample would appear to be perceived as a tonal stimulus by these same subjects, and be analyzed, mainly, in different areas of the right hemisphere. It is also possible that with greater musical experience, subjects with similar tone language backgrounds might analyze musical samples of the type that were employed in this study in the left hemisphere of the brain, in the same manner as speech sound stimuli, since they could be using intervals instead of contours to recognize melodies [13]. Indeed, it is imaginable that this tendency might become more evident among speakers of register-tone languages than among speakers of contour-tone languages.

Lastly, it is also feasible to picture an auditory development of children speaking tone languages with variations in the localization of either speech or music discrimination and analysis in either the left or the right hemispheres of the brain, and a tendency for global (bi-hemispheric) processing among younger children later leading to a tendency for more local (left- or right-hemisphere) processing among older children speaking these languages [14].

To be sure, EEG techniques do not allow study of activation of anything but the outer portions of the frontal, temporal, parietal and occipital lobes of the brain in response to acoustic stimulation of different sorts; activities of, for example, the mesial temporal lobe, with structures such as the hippocampus, the amygdala and the perirhinal and entorhinal cortices, or of the thalamus, the hypothalamus and the basal nuclei, which are so important for auditory discrimination, perception, memory and emotions, are completely left out of EEG studies. Nevertheless, the temporal precision with which such electrophysiological analyses of activity of the external parts of the brain may be carried out seem to make them an indispensable adjunct to morphologically detailed, but temporally imprecise functional imaging techniques, such as fMRI or PET.

3. REFERENCES

[1] Zatorre, R.J., Belin, P. and Penhune V.B., "Structure and function of auditory cortex: music and speech", TICS, 6:37-46, 2002.

[2] Penhune, V.B. et al., "Interhemispheric anatomical differences in human primary auditory cortex: probabilistic mapping and volume measurement from magnetic resonance scans", Cereb. Cortex, 6:661-672, 1996.

[3] Hauser, M., "The sound and the fury: primate vocalizations as reflections of emotion and thought". The Origins of Music, Wallin, N. et al., eds., MIT Press, Cambridge, pp. 77-102, 2000.

[4] Joos, M., "Acoustic phonetics", Lang. Monogr., 23:1-37, 1948.

[5] Zatorre, R.J. and Belin, P., "Spectral and temporal processing in human auditory cortex", Cereb. Cortex, 11:946-953, 2001.

[6] Gandour, J. et al., "Pitch processing in the human brain is influenced by language experience", NeuroReport, 9:2115-2119, 1998.

[7] Gandour, J. et al., "A cross-linguistic PET study of tone perception", J. Cogn. Neurosci., 12:207-222, 2000.

[8] Klein, D. et al., "A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers", NeuroImage, 13:646-653, 2001.

[9] Castro-Sierra, E., Romero-Garibay, D. and Gómez-Gama, M., "Development of pitch perception and memory, and of general psychological abilities in children speaking tonal languages", J. Acoust. Soc. Amer., 107(5, Pt. 2): 2801, 2000.

[10] Castro-Sierra, E., Poblano, H.A. and Gómez-Gama, M., "Development of pitch perception in children", ISMA Mexico City Proc., 2001.

[11] Bentley, A., Musical Ability in Children and its Measurement, Harrap: London, 1966.

[12] Castro-Sierra, E., Development of Pitch Perception in Children, Unpublished doctoral dissertation. Stanford University, 1989.

[13] Peretz I., "Auditory agnosia: a functional analysis". Thinking in Sound. The Cognitive Psychology of Human Audition. McAdams, S. & Bigand, E., eds., Clarendon Press, Oxford, pp. 198-230, 1993.

[14] Moses, P., Roe, K., Buxton, R.B., Wong, E.C. Frank, L.R. and Stiles, J., "Functional MRI of global and local processing in children", NeuroImage, 16:415-424, 2002.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

LOOKING AT PERCEPTION OF CONTINUOUS TEMPO DRIFT - A NEW METHOD FOR ESTIMATING INTERNAL DRIFT AND JUST NOTICEABLE DIFFERENCE.

Sofia Dahl and Svante Granqvist

Department of Speech Music and Hearing
Kungl Tekniska Högskolan
sofia@speech.kth.se, svante@speech.kth.se

ABSTRACT

The method proposed here investigates if there is such a thing as an internal representation of a "steady tempo" - and whether this representation itself is free from tempo drift. The method uses a modification of the method for Parameter Estimation by Sequential Testing (PEST). Several click sequences are presented to the listener in each test and depending on the listener's response (correct or incorrect) the magnitude of the tempo drift is modified for the next presentation. The method does not rule out the possibility that the internal "clock" can have an inherent tempo drift. Such a drift would mean that some listeners could perceive an increasing tempo when, in fact, it is decreasing, and vice versa. Preliminary results indicate that some listeners tend to place their answers "biased" towards either increasing or decreasing tempo. The results also indicate that these listeners appear to be consistent in doing this. Thus we would like to propose a model of detectability of continuous (linear) tempo drift based on a person's *Internal Drift* (ID), which can be zero or biased in either direction. Surrounding this ID is an interval corresponding to twice the *Just Noticeable Difference* (JND).

1. INTRODUCTION

In many musical genres tempo changes are used intentionally to mark the structure of the piece. For instance, tempo tends to be increased in the beginning of a phrase but decreased towards phrase endings. However, even when the intention is to keep a steady tempo, measured timing data usually display both short-term and long-term variations. These unintended fluctuations and tempo changes include gradual changes of tempo over time, *tempo drifts*. In studies of finger tapping Madison reported linear drifts in both directions (increasing and decreasing) ranging between 0.05 and 0.3 % per interval [1]. Measured timing data from drummers also display drifts in tempo, when estimated as linear the drift was typically about 0.2 % per interval [2].

The measured tempo drifts in tapping and drumming data bring up interesting questions about how small gradually introduced tempo changes that can be perceived. It seems reasonably to assume that even very small continuous changes ought to be audible if enough time passes. However, this would require some kind of representation of the original tempo to compare the changing tempo with.

If there is such a thing as an internal representation of a "steady tempo", can we assume this representation to be isochronous? Madison has pointed out that isochrony is simply the special case where no monotonous drift is present and it does not need to have any special status [3]. This considered, there is a possibility that the internal reference itself may not be free from drift, and some earlier investigations have pointed in this direction. Vos et. al investigated perception of tempo drifts and found that subjects responses were biased by the the base tempo. For a base tempo of 240 *beats per minute* (BPM) subjects tended to respond that tempo was accelerating, while a base tempo of 60 BPM generated more responses that tempo was decelerating. For both these extreme tempi subjects' responses were biased *even with no tempo drift present*. For the intermediate tempo 120 BPM (nominal beat separation 500 ms), subjects perceived as many decelerations as accelerations and no bias was found [4].

As illustrated by the example above, the *physical tempo drift*, i.e. the physical lengthenings or shortenings of subsequent intervals in a stimulus, does not necessarily have the same direction as the perceived tempo drift. Such bias in the direction of tempo drift has a consequence on how the smallest perceivable drifts is measured.

In this paper we depart from the assumption that there is an internal reference that we use for comparison when judging tempo, without ruling out the possibility that this reference may have an inherent tempo drift. While we assume that there is a range of continuous tempo drifts that are too small to be perceived, we take into consideration that range may not be centered around isochronous physical tempo. We propose that this span of non-discriminable drift can be interpreted as twice the *Just Noticeable Difference* (JND), centered around an internal reference; the *Internal Drift* (ID).

2. METHOD

2.1. Model and definitions

If the perception of tempo drift is modelled as an internal drift of the subject and a band around this internal drift in which no tempo drift can be detected, the tempo drifts generated by our proposed method will converge to the positive or negative limits (just noticeable drifts) of this band. In the method the physical tempo drifts are generated by constant and continuous change of each *inter-onset interval* (IOI), that is, the interval between the onset of each click is consistently either increased or decreased by adding or removing a fixed number of samples.

The method uses a modification of Parameter Estimation by Sequential Testing (PEST), implemented in the custom made computer program "Beat". During a test session the program presents stimuli consisting of a number of clicks with either increasing tempo (successive shortening of IOIs) or decreasing tempo (successive lengthening of IOIs). After each completed playback of a stimulus the listener is asked to indicate the direction of the tempo

drift by clicking one of the buttons “increasing” or “decreasing”. Each clicking of a button is equivalent to a *response*. This procedure is repeated a number of times and the program adjusts the physical drift before each stimulus playback in order to arrive at the just noticeable drift. Two correct responses in a row are required to assume that the response was correct, while one incorrect response is considered an incorrect answer (see Figure 1). This causes the tempo drift to converge towards the 75 % correct responses level.

Each test session consists of several series running in parallel, each series consisting of several played back stimuli (normally between 20 and 60). Each test series is assigned a nominal tempo (in BPM), a stimulus duration (playback time in seconds), an initial drift (% of nominal initial IOI), an initial drift step, and a designation as either ‘increasing’ or ‘decreasing’. Because of the designation, the correct response to a series designated ‘decreasing’, is always “decreasing” - also if the physical drift is increasing. In this way, the physical drift of a series can shift sign and cross the zero-line as the response curves converge around a perceived isochronous tempo.

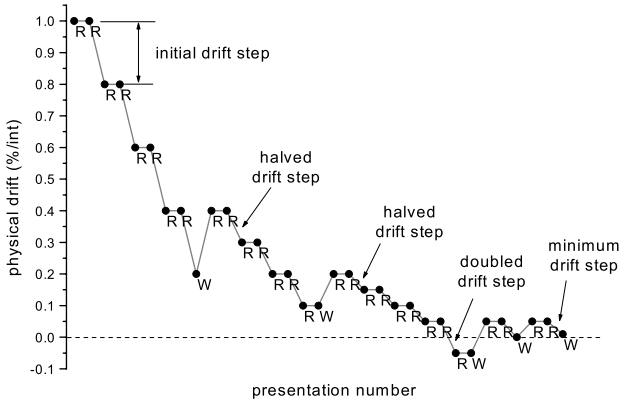


Figure 1: *Schematic illustration of a series and the associated responses from the subject. The first stimulus has a physical drift of 1 %/interval. After two correct responses (R) the physical drift is decreased by the initial drift step (0.2 %/int.). Later an incorrect response (W) is given and the drift is increased by the same drift step. After two correct responses following the incorrect response, the drift step is halved. The drift step never becomes smaller than the minimum drift step. Because of the designation of the series as ‘decreasing’ the response “decreasing” will always be considered correct, also if the physical drift should cross the zero-line.*

The Beat program runs on PCs with Windows and normal soundcards. The tempo (IOI) drift is achieved by inserting or removing a whole number of zero samples at the end of each click sound. The number of samples inserted/removed is kept exactly the same throughout each stimulus in order to achieve a smooth tempo drift.

3. LISTENING TEST

To test the perception of small tempo drifts and investigate whether listeners show ID $\neq 0$ a listening test was performed. Seven subjects, aged 25-57 years (mean 35.6 years) participated in the test. All subjects had received training in playing musical instruments or singing, their experiences ranging between 2 and 40 years.

The subjects were tested individually in three listening sessions of 40 minutes each. In order to expose the subjects to a mix of stimuli with both increasing and decreasing tempo drift, each session consisted of six parallel series; three designated ‘increasing’ and three designated ‘decreasing’. The order in which series were to be played back was decided semi-randomly, so that the number of turning points in the six series were synchronized. In this way the amount of transitions from correct-incorrect or incorrect-correct answers for were approximately equal for each series throughout the test session. To reduce the risk of subjects being more inclined to use the right or the left button for their responses, the placement of the response buttons ‘increasing’ and ‘decreasing’ was randomized before each test session.

The subjects were instructed that they would hear a number of click sequences, 5 s long, with either increasing or decreasing tempo drift. For each presented stimulus the subject was to carefully pay attention to the direction of the drift and indicate this by clicking the corresponding button on the screen. All ten intervals in a stimulus had to be played before it was possible to respond. The listeners were not aware of the test design or that the method was adaptive. Each test session was interrupted after 40 minutes.

The nominal start tempo was 120 BPM, (nominal beat separation 500 ms), but the actual initial tempo that was used for each stimulus was randomized by the same amount as the total tempo change, from beginning to end, for the stimulus. Each stimulus consisted of ten intervals ($\sim$ 5 s) which were played after an initial silence with a randomized duration of 0.5 to 1 s. The maximum allowed drift step was 0.2 % of initial IOI and the minimum allowed drift step was set to 0.04 % of initial IOI. The sampling frequency used was 44100 Hz, which gives a resolution in the IOI drift at 120 BPM of approximately 0.0045 %.

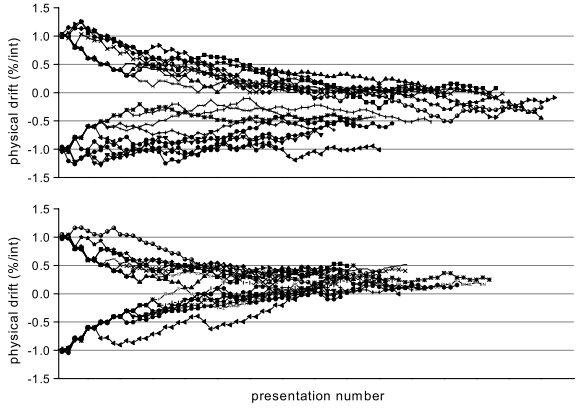


Figure 2: *Examples of series produced for two different subjects with different internal drift and different JND. Each panel presents 18 series originating from three test sessions. Each of these sessions consisted of three series with an initial physical drift of -1 %/interval (designated ‘increasing’), and three series with an initial drift of 1 %/interval (‘decreasing’). Both subjects display non-zero internal drifts. Subject GS (top) has an internal drift corresponding to a increasing tempo. By contrast, subject AA (bottom) has an internal drift corresponding to a decreasing tempo with a smaller JND than subject GS.*

4. RESULTS

Figure 2 shows all of the produced series for two of the subjects participating in the three test sessions. In each panel the nine series starting with a drift of 0.1 %/interval (designated 'decreasing'), and the nine series starting with a drift of -0.1 %/interval (designated 'increasing') are seen. Although the series originate from three different sessions the subjects are remarkable consistent in their judgements.

Many of the subjects showed bias in their judgements of the drifts. For instance, the 'decreasing' series for subject GS (top panel, Figure 2) all approach, or even cross, the zero-line for the physical drift, indicating that this subject has an internal drift corresponding to an increasing tempo. The span between 'decreasing' and 'increasing' series is rather large, indicating large JND. By contrast the produced series for subject AA (bottom panel) cluster closely together on the positive side of the physical drift, indicating an internal drift corresponding to a decreasing tempo and a smaller JND.

In the analysis of data the average endpoints across the 'increasing' and 'decreasing' series respectively, were taken as giving the span of non-discriminable drift centered around the internal drift (ID). The span was interpreted as twice the Just Noticeable Difference (JND).

4.1. Internal drifts

The results from the analysis of the Internal Drifts can be seen in Figure 3. The grand average ID, across subjects and sessions, was not significantly different from 0. Between subjects the IDs were significantly different ($F[6, 2.341] = 40.45, p < .001$, Greenhouse-Geisser compensated), and five of the subjects had an ID significantly different from 0. There were no significant differences between test sessions.

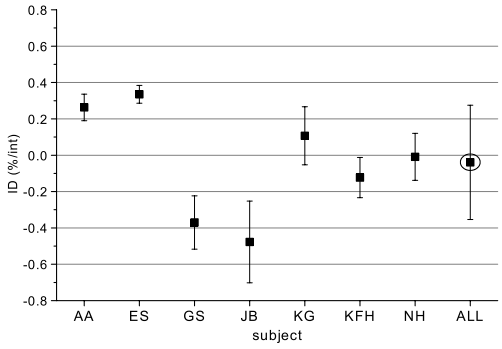


Figure 3: *The resulting internal drifts (ID) for the seven subjects. The vertical error bars show the standard deviation. The grand average across all subjects (encircled data point to the right) was -0.04 %/interval, which was found not to differ significantly from 0.*

4.2. Just Noticeable Differences

The JNDs relative to the found ID for the seven subjects can be seen in Figure 4. The mean JND, across subjects and sessions, was 0.27 %/interval, corresponding to 1.35 ms/interval. Like for the IDs the JNDs were significantly different between persons ($F[6, 1.962] = 11.79, p < .001$, Greenhouse-Geisser compensated).

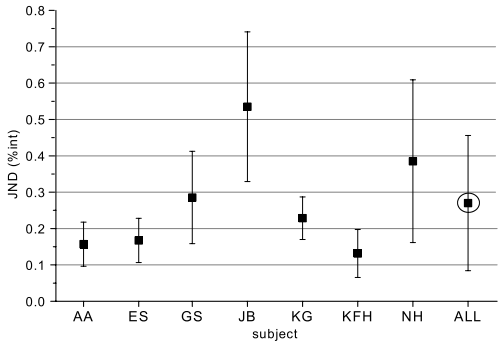


Figure 4: *The resulting just noticeable differences as related to the internal drift (see Figure 3) for the seven subjects. The standard deviation is shown by the vertical error bars. The grand average across all subjects (encircled) was 0.27 % per interval, equal to 1.35 ms/interval.*

5. DISCUSSION AND CONCLUSIONS

Although the subjects in our study clearly differ in their opinions of a "steady beat", with IDs covering a range of drift from about -0.5%/interval to +0.4%/interval, the grand average ID still ends up almost at zero (see Figure 3). We have, however, so far only investigated the tempo 120 BPM, the same tempo which Vos et. al found no bias for in their study [4]. Even so, the method appears to be successful in estimating the internal drift. Subjects proved to be very consistent in their responses, there were no significant differences in ID between series or test sessions. More tempi and more subjects are likely to help in shedding some light on whether the ID may be dependent on tempo.

The found values for the JNDs correspond to measured production data from drumming [2] and tapping [1]. However, when comparing to previously reported estimates of JND for drift and tempo detection (see surveys in [3] and [5]) they are surprisingly small. For a tempo with a nominal beat separation of 500 ms (120 BPM) these JNDs have been estimated to about 2 %. There may be several reasons for these discrepancies:

First, our model places the JND in relation to the ID of each subject. Considering that the found IDs proved to be non-isochronous for five of the seven subjects this makes comparisons to other studies somewhat hazardous. However, if our grand average JND is added to the largest ID for a subject, it still does not amount to more than 4 ms/interval (corresponding to 0.8 %/interval).

Second, many studies have concentrated on the detection threshold for stepwise tempo changes and perhaps not as often has continuous tempo drift been studied. However, a continuous tempo change of 0.2 %/int. over 10 intervals results in a total tempo change of 2 % over the whole sequence. It is possible that the accumulated tempo change over several intervals causes a detection, even when the change for each interval does not.

Third, many other studies have used pairwise comparisons or

adaptive methods asking whether a change was heard or not. We chose the Forced Directional Response paradigm used by Vos et. al [4], asking the subjects to indicate the direction of drift. The fact that the subjects were informed about the presence of tempo drift, and that they were forced to make their best guess regarding the direction may have contributed to the lower JNDs.

Some experiments by Madison exemplify differences between methods. Using an adaptation of PEST, and asking whether drift had been detected (yes or no), several experiments showed that for 9 intervals at a tempo of 120 BPM, the JND was about 2 %/interval (or 10 ms/int.). In a following experiment the method was changed and subjects were instead asked to modify the tempo by pressing keys ('faster', 'slower') until they perceived it to be isochronous. The subjects could use up to 30 intervals to make their judgement before adjusting or confirming. The method also allowed the subjects modifications to cross the zero level, although this was rare. The results showed that the subjects' Point of Subjective Equality was about 0.2 % at a tempo of 120 BPM. The mean number of intervals used by the subjects across all trials varied from 8.1 to 16.8.

The number of intervals that subjects are allowed to hear also affects the JND. Drake and Botte [6], and Madison [3] have reported that detection thresholds decrease with increasing number of intervals presented. However, in many investigations of threshold for tempo detection, rather short stimuli durations have been used, usually between 4 and 6 intervals. In a musical context listener are also used to gradual changes occurring over more than a measure, and can be assumed to use more than 6 intervals to detect changes in tempo.

To conclude, the method seems to be successful in estimating the internal drift. There were differences between subjects, of which five displayed internal drifts significantly different from zero, either in the positive and negative direction. The found JNDs were small when compared to earlier studies and will need to be investigated further through more listening experiments.

6. ACKNOWLEDGEMENTS

The authors are indebted to all the subjects participating in the test, and to Joakim Westerlund for the statistic analysis.

7. REFERENCES

[1] Madison, G., "Variability in Isochronous Tapping: Higher Order Dependencies as a Function of Intertap Interval", Journal of Experimental Psychology: Human Perception and Performance, 27(2), 2001, 411-422.

[2] Dahl, S., "The Playing of an Accent - Preliminary Observations from Temporal and Kinematic Analysis of percussionists", Journal of New Music Research, 29(3), 2000, 225-234.

[3] Madison, G., "Functional Modelling of the Human Timing Mechanism", Published doctoral dissertation, Uppsala University, Sweden, 2001.

[4] Vos, P. G., van Assen, M., and Franek, M., "Perceived tempo change is dependent on base tempoand direction of change: Evidence for a generalized version of Schulze's (1978) internal beat model", Psychological Research Vol. 59, 1997, 240-247.

[5] Friberg, A. and Sundberg, J. "Time discrimination in a monotonic, isochronous sequence", J. Acoust. Soc. Am., 98(5), 1995, 2524-2531.

[6] Drake, C. and Botte, M-C., "Tempo sensitivity in auditory sequences: Evidence for a multiple-look model", Perception & Psychophysics, 54(3), 1993, 277-283.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

WHAT CAN THE BODY MOVEMENTS REVEAL ABOUT A MUSICIAN'S EMOTIONAL INTENTION?

Sofia Dahl and Anders Friberg

Department of Speech Music and Hearing
Kungl Tekniska Högskolan
sofia@speech.kth.se, andersf@speech.kth.se

ABSTRACT

Music has an intimate relationship with motion in several aspects. Obviously, movements are required to play an instrument but musicians move also their bodies in a way not directly related to note production. In order to explore to what extent emotional intentions can be conveyed through musicians' movements only, video recordings of a marimba player performing the same piece with the intentions Happy, Sad, Angry and Fearful, were recorded. 20 observers watched the video clips, without sound, and rated both the perceived emotional content as well as movement cues. The videos were presented in four viewing conditions, showing different parts of the player. The observers' ratings for the intended emotions showed that the intentions Happiness, Sadness and Anger were well communicated, while Fear was not. The identification of the intended emotion was only slightly influenced by the viewing condition, although in some cases the head was important. The movement ratings indicate that there are cues that the observer use to distinguish between intentions, similar to the cues found for audio signals in music performance. Anger was characterized by large, fast, uneven, and jerky movements; Happy by large and somewhat fast movements, Sadness by small, slow, even and smooth movements.

1. INTRODUCTION

Musical performances are often enjoyed visually as well as aurally. It is not unusual to see the audience' necks stretched in attempt to follow the musicians' movements. That it would only be the actual sound producing movements that interest us seems unlikely, for these movements are often too small or too fast to be seen properly. However, musicians move also in ways that are not directly related to the production of notes. These movements have been shown to be able to convey information about the expressive intent of performances. For instance, in studies by Davidson [1][3], subjects were about equally successful in rating music performances according to their expressive intent (deadpan, projected or exaggerated) regardless if they were allowed to only listen, only watch, or both watch and listen. The musically naive were even better in recognizing the intent in the watch-only mode, compared to the other modes [3].

The ability of observers to obtain information regarding emotional intent (affect) from movements only has been well documented, not only for music performances but also for other settings, such as dancing [8], drinking, or knocking [4]. Work has also been dedicated to what kinds of movement characteristics that provide the pieces of information that observers use in order to distinguish between performances. Some suggestions of such *movement cues* have been made by DeMeijer and Boone and Cunningham [5] [6] [7]. For instance, actors' movements were associated with Joy when their movements were fast, upward directed, with arms raised, whereas the optimal movements for Sadness were slow, light downward directed, with arms closed around the body [5] [6].

That the direction of movement and the position of the arms seem to be of such importance is interesting in perspective of Davidson's work. Musicians' arm and hand movements are primarily involved in the sound production, and expressive movements used by observers to discriminate between performances must therefore either appear in other parts of the body, or *coincide* with the actual playing movements. Davidson [2] found that observers were not able to identify the expressive intention from the hand movements only, while the head movements seemed to be of greater importance.

In analysis of music performances audio cues, such as tempo, sound level etc., have been found to characterize emotional coloring [9][10]). For example, a Happy performance is characterized by a fast mean tempo, high sound level, staccato articulation, and fast tone attacks, while a Sad performance is characterized by a slow tempo, low sound level, legato articulation and slow tone attacks. It seems reasonable, then, to assume that the body movements in the performances contain cues corresponding to those appearing in the audio signal.

The questions for this study were the following: (1) How successful is the overall communication of each intended emotion? (2) Are there any differences in the communication, depending on intended emotion, or what part of the player the observers see? and (3) How can perceived emotions be classified in terms of movement cues?

2. EXPERIMENT

A professional percussionist was video recorded when performing a short piece of music with the intentions Sadness, Anger, Happiness and Fear, on the marimba. The piece chosen was a practice piece from a study book by Morris Goldenberg: "Melodic study in sixteens". This piece was found to be of a suitable duration and of rather "neutral" emotional character, allowing for the different interpretations.

From the video recordings, stimuli clips were generated showing different parts of the player in four *viewing conditions*: *full* (showing the full image), *nohands* (the player's hands not visible), *torso* (the player's hands and head not visible), and *head* (only the player's head visible). A video editing software was used to cut out the stimuli clips for the four viewing conditions using a cropping

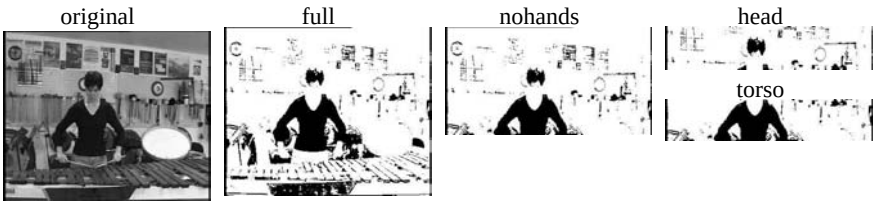


Figure 1: *Original (far left) and filtered video images exemplifying the four viewing conditions used in the test: full, nohands, head, and torso.*

filter. A threshold filter was also used so that facial expressions would not be visible (see Figure 1). Based on the original eight video recordings a total of 32 (4 emotions x 2 performances x 4 conditions) video clips were generated. The duration of the video clips varied between 30 and 50 s.

Twenty subjects watched the video clips individually and rated the emotional content on a scale from 0 (nothing) to 6 (very much) for the emotions Fear, Anger, Happiness, and Sadness. The subjects were also asked to mark how they perceived the movements. The ratings were done on bipolar scales (from 0 to 6) for the cues:

Amount:	none	-	large
Speed:	fast	-	slow
Fluency:	jerky	-	smooth
Distribution:	uneven	-	even

3. RESULTS

3.1. Measure of achievement

From the emotion ratings a measure of how well the intended emotion was communicated to the listener was computed. The *achievement* was defined as the similarity between the *intended* ($\mathbf{x}$) and the *rated* ($\mathbf{y}$) emotion, for each video presentation. Both $\mathbf{x}$ and $\mathbf{y}$ are vectors that consist of four numbers representing Fear (F), Anger (A), Happiness (H), and Sadness (S). For the intended emotion Happy $\mathbf{x} = [F\ A\ H\ S] = [0\ 0\ 1\ 0]$ and the maximum achievement would be for a rating of $\mathbf{y} = [F\ A\ H\ S] = [0\ 0\ 6\ 0]$. The achievement $A(\mathbf{x}, \mathbf{y})$ for a specific presentation is defined as

$$A(\mathbf{x},\mathbf{y}) = \frac{1}{C}\frac{1}{n}\sum_{i=1}^{n} \overbrace{(x_i - \overline{x})}^{\text{intention}}\ \overbrace{(y_i - \overline{y})}^{\text{rating}}$$

where $\mathbf{x}$ and $\mathbf{y}$ are arrays of size n (in our case $n = 4$), and $\overline{x}$ and $\overline{y}$ are the mean values across each array. C is a normalization factor to make the "ideal" achievement equal to 1. In this case, given that $\mathbf{x}$ can only take the values 0 and 1, and $\mathbf{y}$ can be integer values between 0 and 6, $C = 1.125$. A negative achievement value would mean that the intended emotion is confused with other emotions, and zero is obtained when all possible emotions are ranked equal. We assume that an achievement significantly larger than zero implies that the communication of emotional intent was successful. In practice, the achievement measure is the same as the average of the covariance between the intended and rated emotion for each presented video clip, with a normalization factor included.

Figure 2 shows the mean achievement for all eight performances presented according to intended emotion, viewing condition and performance. The 95 % confidence intervals are indicated by the vertical error bars. The figure illustrates that the player was able to convey most of the intended emotions to the observers in all viewing conditions. Sadness, Happiness and Anger were all well communicated, while Fear received low, sometimes negative, achievement.

In order to facilitate comparisons with other results the proportion of correct identifications were calculated by converting the ratings to "forced choice" answers. The conversion was made strictly, meaning that only the answers were the intended emotion received the highest ratings were considered as "correctly" identified. The proportion of correct identifications for each intended emotion are indicated by the small black squares above each bar in Figure 2. The proportion of correct responses follow the same pattern as achievement, with the highest values for the intention Sadness (95 % correct), followed by Anger, Happiness, and Fear. Despite the fact that the responses where the intended emotion was rated equal to another emotion were treated as "incorrect", the correct identifications are well above chance level (25 %) in most cases.

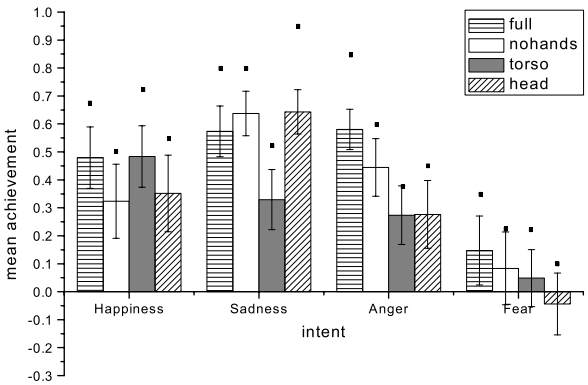


Figure 2: *Mean achievement for the four intended emotions and viewing conditions averaged across the first and second performance of each intended emotion. Each bar shows the mean achievement for one emotion and viewing condition, full (horizontally striped), nohands (white), torso (grey), and head (diagonally striped), averaged across 20 subjects and two performances. The error bars indicate 95 % confidence interval. Performances with the intentions Happiness, Sadness, and Anger received ratings in correspondence with the intention, while the Fearful performances were hardly recognized at all. Above each bar a small black square indicate the relative proportion of correctly identifications, as calculated from the highest rated emotion for each stimulus response.*

3.2. Influence of viewing conditions and emotions

To reveal the importance of the differences between the intended emotions and viewing conditions, the achievement measures were subjected to a 4 conditions x 4 emotions x 2 performances repeated measures ANOVA. The analysis showed main effects for intended emotion $[F(3, 57) = 33.65, p < 0.0001]$, and viewing conditions $[F(3, 57) = 9.54, p < 0.0001]$, and significant results for the two-way interactions: viewing condition x emotion $[F(9, 171) = 4.46, p < 0.0001]$, and emotion x performance $[F(3, 57) = 2.86, p < 0.05]$.

Although the main effect of viewing condition was significant, the effect was surprisingly small, see Figure 2. Initially one would hypothesize that seeing more of the player would provide the observer with more detailed information about the intention. The achievement values would then be ordered from high to low for the three conditions full, nohands and head, and similarly for full, nohands and torso. Such a "staircase" relation between the viewing conditions was only observed for the intention Anger.

The significant interaction between emotion and viewing condition seems to be due to differences in the Sad and Angry intention. For the Sad intention the head seems to be of highest importance in perceiving the intended expression. All the conditions where the head is visible (full, nohands, and head) received high ratings for Sadness with mean achievements from 0.57 to 0.64, while torso rated a much lower mean achievement of 0.32. For Anger, the full condition received the highest Anger ratings, while the conditions torso and head seem less successful in conveying the intention, particularly in the first performance.

Overall, the mean achievements proved to be very similar for the player's two performances of each intention. The intention Fear was rated as other emotions to a higher extent in the second performance, resulting in negative achievement.

3.3. Movement cues

Figure 3 shows the mean ratings of the movement cues for each intended emotion. The different movement cues; Amount (none - large), Speed (fast - slow), Fluency (jerky - smooth) and Distribution (uneven - even), received different ratings depending on whether the intended expression was Happy, Sad, Angry, or Fearful. Note that high ratings correspond to large amounts of movement, slow speed, smooth fluency, and even distribution, while low ratings correspond to small amounts of movement, fast speed, jerky fluency, and uneven distribution.

The intentions Happiness and Anger obtained similar rating patterns. Both Anger and Happiness seem to display large movements, but the Angry performances are somewhat faster and jerkier compared to the Happy performances. In contrast the ratings for the Sad performances display small, slow, smooth and even movements. The ratings for Fear are less clear-cut, but tend to be somewhat small, fast, and jerky. A similar pattern was found when investigating how the subjects related the emotions to the movement cues. The correlation between the rated emotions and the ratings of movement cues is shown in Table 1. According to the table, Anger is associated with large, fast, uneven, and jerky movements; Happy with large and somewhat fast movements, Sadness with small, slow, even and smooth movements, and Fear with somewhat small, jerky and uneven movements. However, since the communication of Fear failed, its characterization is questionable.

Differences in cue ratings for different viewing conditions were, in general, small. For the intentions Happy and Sad and partly for Anger, the cue ratings are closely clustered. Again, the head seems to play a special role. When a rating stands out from the other viewing conditions it is either for the head or for the torso. Since the latter is the only condition where the head is not visible, it can in fact also be related to the head's movements.

	amount	speed	fluency	distrib.
Happiness	0.40	-0.27	-0.15	-0.12
Sadness	0.32	0.60	0.50	0.38
Anger	0.31	-0.48	-0.54	-0.44
Fear	-0.24	-0.01	-0.13	-0.11

Table 1: *Correlations between rated emotions and rated movement cues. All correlations, except between Fear and speed, were statistically significant ($p < 0.01, N = 603$).*

4. DISCUSSION

The results show that the four intended emotions were communicated successfully, with the exception of Fear. The most successfully conveyed emotion seems to be Sadness.

While there generally were surprisingly small differences between viewing conditions, the head seemed to be very important for correctly identifying the Sad intention. The only viewing condition where the head was not visible, torso, received much lower Sadness ratings than the other conditions for both performances with the Sad intention. Our visual inspections of the stimuli clips revealed no extraordinary features in the movement of the head, but for the Sad performances there do seem to be less and slower movements in the vertical direction compared to the other intentions.

Our results for the ratings of movement cues resemble the cues used by young children in the study by Boone and Cunningham [8]. They reported that the children used more force and rotation and a higher tempo when portraying Happiness and Anger than they did for Sadness and Fear. Their cues force and rotation correspond well to our cues for amount of movement and speed. The children also used fewer shifts in movement patterns for Sadness than for the other emotions, something that bears similarities to our cues for fluency and distribution.

There is also a strong resemblance between these movement cues and the audio cues used in expressive music performances. The most evident connection seem to be between movement speed and musical tempo, but also the similarities between amount of movement and sound level, or fluency and articulation, seem clear.

5. CONCLUSIONS

Our results show that the intentions Sadness, Happiness, and Anger were conveyed through musician's movements only, while Fear was not. The identification of the intended emotion was only slightly influenced by the viewing condition, although in some cases the head was important.

The movement cues used in the communication have similarities to the cues found for audio signals in music performance. Anger was characterized by large, fast, uneven, and jerky movements; Happy by large and somewhat fast movements, Sadness by small, slow, even and smooth movements. Further research could reveal whether the movement cues reported here would apply also for other performers and instruments.

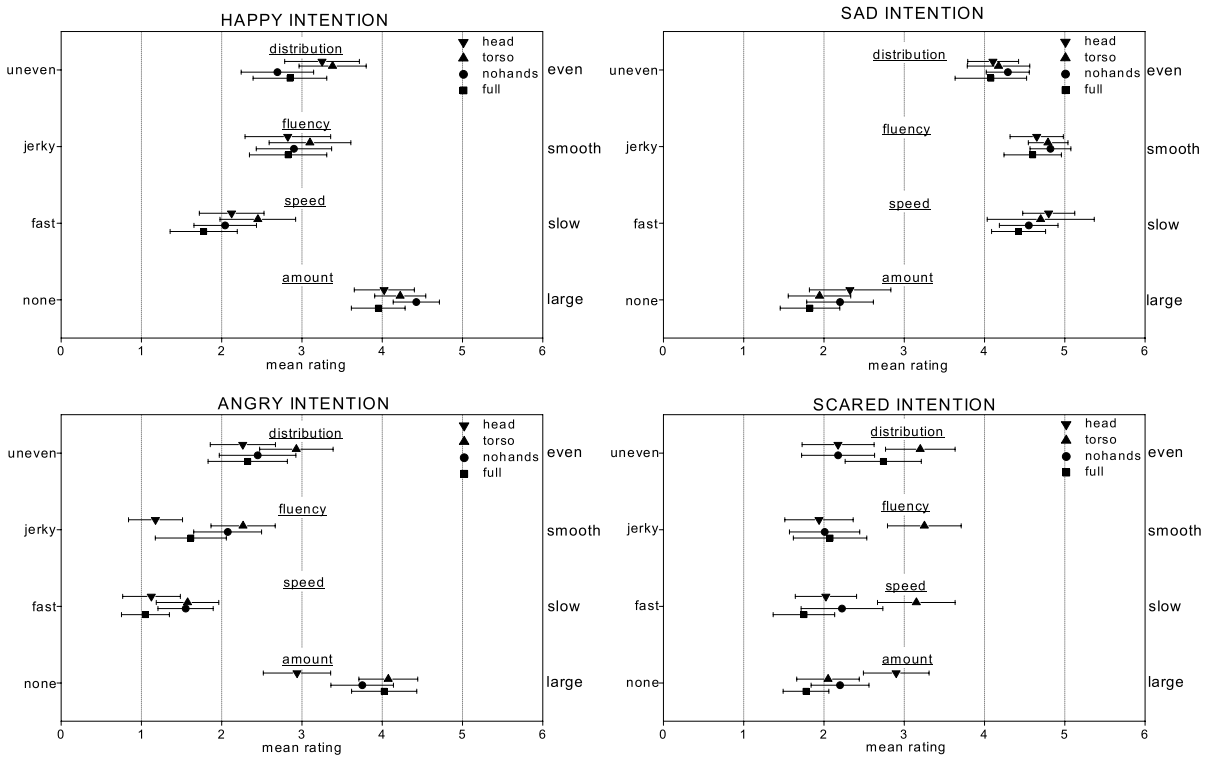


Figure 3: *Ratings of movement cues for each intended emotion and viewing condition. Each panel shows the mean markings for the four emotions averaged across 20 subjects and the two performances of each intended emotion. The four viewing conditions are indicated by the symbols: full (square), nohands (circle), torso (pyramid), and head (top-down triangle). The error bars indicate 95 % confidence interval.*

6. ACKNOWLEDGEMENTS

The authors would like to thank Alison Eddington for the marimba performances and all persons participating as subjects in the viewing test. This work was supported by the European Union (MEGA - Multisensory Expressive Gesture Applications, IST-1999-20410; http://www.megaproject.org/)

7. REFERENCES

[1] Davidson, J. W.,"Visual perception and performance manner in the movements of solo musicians", Psychology of Music, Vol. 21, 1993, 103–113.

[2] Davidson, J. W.,"What type of information is conveyed in the body movements of solo musician performers?", Journal of Human Movement Studies, Vol. 6, 1994, 279–301.

[3] Davidson, J. W., "What does the visual information contained in music performances offer the observer? Some preliminary thoughts", In Steinberg, R. (Ed.) Music and the mind machine: Psychophysiology and psychopathology of the sense of music, Heidelberg: Springer, pp. 105–114, 1995.

[4] Pollick, F. E., Paterson, H. M., Bruderlin, A., and Sanford, A. J., "Perceiving affect from arm movement" Cognition, 82(2), 2001, pp. B51-B61.

[5] De Meijer, M., "The contribution of general features of body movement to the attribution of emotions" Journal of Nonverbal Behavior, Vol. 13, 1989, 247–268.

[6] De Meijer, M.,"The attritution of aggression and grief to body movements: The effects of sex-stereotypes" European Journal of Social Psychology, Vol. 21, 1991, 249–259.

[7] Boone, R. T., and Cunningham, J. G., "Children's decoding of emotion in expressive body movement: The development of cue attunement", Developmental Psychology, Vol. 34, 1998, 1007–1016.

[8] Boone, R. T., and Cunningham, J. G., "Children's expression of emotional meaning in music through expressive body movement", Journal of Nonverbal Behavior, 25(1), 2001,21–42.

[9] Gabrielsson, A., and Juslin, P. N., "Emotional expression in music performance: Between the performer's intention and the listener's experience", Psychology of Music, Vol. 24, 1996, 68–91.

[10] Juslin, P. N., "Cue Utilization in Communication of Emotion in Music Performance: Relating Performance to Perception", Journal of Experimental Psychology: Human Perception and Performance, 26(6), 2000, 1797-1813.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

TIME ESTIMATION: ISOCHRONOUS VERSUS ACCELERATED SEQUENCES

Diego Dall'Osto, Santiago Estaún Ferrer

Department of Psychology of Education[1]
Universitat Autònoma de Barcelona

diego@dallosto.com Santiago.Estaun@uab.es

ABSTRACT

The duality *regularity-irregularity* describes the framework of the present initial study, as a general relationship regarding time and rhythm in music.

The present work focuses on the influence of the temporal content of musical stimuli on the perception of duration.

Starting to study the effect of gradual processes of temporal deviation on musical time perception, only acceleration will be investigated in a pilot experiment.

1. INTRODUCTION

Rhythm, as a major characteristic of musical time, can make perceivable some basic and regular pulsation levels, or, differently, it can slide, by retarding or anticipating listeners' expectations.

There are several manners to realize *regularity* and *irregularity* in music. Performers, who modify the musical pace with expressive intentions, underlining some precise moments of musical pieces, introduce different amount of *irregularity* within the metronomic *regularity*. Composers can vary durations and proportions by introducing gradual temporal changes, like accelerations or decelerations, so to create sort of goal-directed processes.

But, how listeners to music do perceive those temporal deviations? A starting point could be to know whether the perception of duration is influenced by the content of musical stimuli.

2. REGULARITY AND IRREGULARITY

The use of the *regularity-irregularity* concept here is uniquely meant to indicate a general temporal relationship: thinking of that duality more as a continuum between two extreme points rather than a rigid opposition.

A basic distinction is needed between isochronous sequences of time intervals, resulting in periodic or regular rhythms, and gradual processes of time deviations, like acceleration and deceleration.

Regularity is supposed to be what a listener attempt to infer at the beginning of a musical piece, when musical events start to occur in time (Drake 1998). *Regularity* extraction seems to be used by listeners to create *expectancies* about next events and to predict when future events in the piece are likely to occur (Jones & Boltz 1989). *Regularity* creates *expectancy* and involves *predictability*, confirming *expectancies*.

Irregularity, as perceived deviation in time, does not allow listeners to detect equally separated frames of events, because of the lack of periodicity. Disruption of the *regularity* of events, as one expects they should occur in time, by gradually decreasing or increasing time intervals, might influence the perception of time and duration. *Irregularity* implies uncertainty about next events, because of lower degrees of *predictability* (Zakay, 1992).

We consider *acceleration* a special case within the dual category *regular - irregular*. It is related to gradual processes and perceptual continuity that seems to be implied in a series of progressively changing durations.

Gradual changes in time, as acceleration and deceleration, break down the periodicity, so that the regularity of rhythm, strongly related to the creation of expectancies in time, no longer can be perceived.

Players can influence the musical time, performing 'expressive timing', with deviations from the mechanical clock, depending on structural and emotional demands.

We can also find types of composed gradual temporal changes, defined and fixed by musical notation.

The idea of an implicit 'elastic time' in music writing, including properly composed accelerations and decelerations, has been more and more developed in several contemporary musical styles.

The expressiveness of time variations regards the feeling of tension or relaxation, hesitation, expectation, surprise and so forth.

It has been suggested that confirmation and violation of expectancies is critical to aesthetic experience and the expression of emotional meaning in music, and that artful deviations of expectancies are just as important for one's overall interest level and musical enjoyment (Meyer, 1992).

[1] The experiment was realized as part of the Ph.D. "Perception, Communication and Time" (Dept. of Psychology of Education, Universitat Autònoma de Barcelona, Spain), directed by Dr. Santiago Estaún Ferrer, with the contribution of MC and T reference BSO2001-2008.

Starting to study the effect of gradual processes of temporal deviation on musical time perception is the goal of the presented experiment.

3. PERCEPTUAL PRESENT AND SHORT TIME MEMORY

'Perceptual present', or 'subjective present' (also 'specious present'), is a concept meaning that the psychological present time is not a dimensionless point. It is the horizon of finite duration within which we are able to perceive groups of sensorial events.

There is no general agreement about the real duration of the 'subjective present' within the short time memory, being reported to be in a range from few seconds up to 180 sec.!

Fraisse (1956) found a normal range of 3-5 sec. and there are recent indications of a window of temporal integration up to 3 sec. (van Noorden & Moelants, 1999). Other indications regard an upper limit within a range of about 8 to 10 sec. with a maximum limit of 20 sec. 'under extreme conditions' (Kramer, 1988).

A crucial aspect of the experiment is related to the duration of each single stimulus and the overall interval of time to be used for pairs of stimuli. Consequently, single and overall durations had to take into account the possible problems of memory overload, in case of too large sequences.

Considered time intervals were supposed to lie within ranges that are congruent with the process of perceptual grouping in music, or, supposedly to be related with some plausible extents of 'perceptual present'.

Exploration of the 'perceptual present' regards the subjective ability for grouping together events within time intervals: grouping and chunking is thought to influence the estimation of time.

4. EXPERIMENT

The main question of the present pilot experiment regards whether regular rhythms, produced by isochronous intervals, and accelerating sequences are perceived as equally long.

Generally it is well established that non-temporal tasks interfere with time judgments (Brown et al., 1992), with the possible exception of highly coherent musical structures (Boltz, 1998).

Consequently, starting to investigate the *regular-irregular* time duality requires focusing on temporal information, in order to possibly restrict the effects of 'non-temporal' information (e.g. melody and timbre).

In a comparison task experiment, with simple musical stimuli, the attention focused on the acceleration effect on the perception of time intervals.

4.1. Subjects

Subjects were previously told that judgments on overall time intervals would be asked. Precisely they had to estimate the second sequence of each presented pair of stimuli, choosing between longer, shorter or equal than the first sequence.

Subjects were 20 musicians: 15 Conservatory of music's students (thanks to the Conservatory of Music of Mantua, Italy) and 5 professional musicians, with age between 18 and 50 years.

4.2. Stimuli: sound and sequences

Two short percussion sounds have been used, with low and high pitch, with same intensity and very closed durations.

Two low-pitched cues mark the beginning and the end of a stimulus (two low-pitches every sequence), while the high-pitches form the remaining internal pulses (see Figure 1).

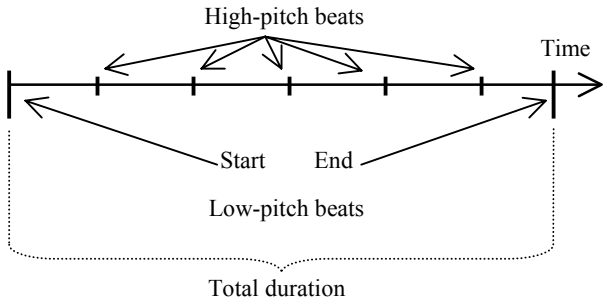


Figure 1: *Stimuli structure*

The structural boundaries of the stimuli (start/end low-pitch) should have facilitated the perception of grouping, in absence of other musical references, as melody, harmony, etc.

4.3. Duration of stimuli

Each trial used pairs of isochronous and accelerated stimuli, equal in duration, with short and long durations (closed to 1:2 ratio).

Pairs of stimuli have of two kinds of overall durations:
9.8 sec. (4.4 sec. +1sec. silence +4.4 sec.)
17.84 sec (8.42sec. +1sec. silence +8.42sec.)

Durations of 4.4 and 8.42 seconds would be the windows of the actual 'perceptual present', while longer durations of 9.8 and 17.84 sec. would be the actual short-term memory buffers, within which a present stimulus has to be compared with a stored and recalled version of it.

4.4. Isochronous sequences.

Two periods have been chosen of 400 msec and of 441 msec as the reference beat rates for the isochronous intervals sequences.

Values around 400-600 msec are recognized as characterizing the threshold between Long and Short duration categories in perceptual judgments (Fraisse, 1956).

Time periods of 400 and 441 msec are also within the 'spontaneous tapping' range, below a threshold of 1000 msec.

Also it is within the observed 'optimal tempo sensitivity' range, between 400 and 800 msec, for which people better react to small tempo changes.

With an upper limit of 1000 msec interval they are also associated with the 'indifference interval', for which people's duration estimations are most veridical; and for which perceptual and performance timing tasks are sufficiently correct (Marek et al., 2000). Shorter durations show systematic overestimation while longer durations would be underestimated (Clarke, 1999).

The difference of 41 msec between the two period intervals (441 and 400 msec) represents a small perceivable difference in the pace rate: corresponding to 150 BPM (440 msec) and 136 (441 msec) respectively.

The difference of 14 BPM corresponds to a deviation of about 9-10% within the range of 4-11% reported as the Just Noticeable Difference range for musical *tempi* (Levitin and Cook, 1996).

Possibly it allows the renewing of the subjects' attention between different stimuli pairs, with a minimum effect of novelty and without suddenly surprising them with greater differences in beat rates.

4.5. Accelerating sequences

A general choice of having same overall durations and same number of pulses was made for accelerations with respect to isochronous sequence. An equal overall duration is obviously needed for testing the influence of the acceleration on the perception of time intervals.

The numbers of pulses for each sequence were 12, for the 4.4 sec sequence and 21, for the 8.82 sec sequence.

In order to realize a more gradual acceleration, with better-distributed intervals and a perceptually consistent process, the final choice was to realize an acceleration similar to which it could be realized by means of the traditional music notation.

A simple way to realize accelerations using that proportional system is by progressively subtracting subdivision of unit values.

For example, if we take a musical unit, for instance a semiquaver (𝅘𝅥𝅯) we can produce a simple sequence like the following:

11𝅘𝅥𝅯, 10𝅘𝅥𝅯, 9𝅘𝅥𝅯, 8𝅘𝅥𝅯, 7𝅘𝅥𝅯, 6𝅘𝅥𝅯, 5𝅘𝅥𝅯, 4𝅘𝅥𝅯, 3𝅘𝅥𝅯, 2𝅘𝅥𝅯, 1𝅘𝅥𝅯.

If we translate it in terms of intervals of time, taking for instance 40 msec as semiquaver, a plausible musical acceleration is just obtained.

Taking into account the constraints of the perceptual discrimination and grouping, the general range of inter onset intervals (IOs) between successive pulses is the following: minimum interval of 40 msec, maximum interval of 1500 sec.

One has to take into consideration that with values below a threshold of 40-50 msec our hearing starts to enter a range where audio events tend to fuse in unique audio streams, producing sustained instead of separated tones (Warren, 1999).

Also we know that durations longer than 1500-2000 msec between events start to be perceived as separated and unrelated; only below that threshold intervals can be processed as grouped together.

The general formula for calculating decreasing durations is the following:

$$D = (n-X)\ U \qquad (1)$$

n = total number of time intervals
X = interval index, from 0 to n –1
U = time unit (minimum time interval)

For the duration of 4402 msec we have n=11 and U=66.7 msec., while for the duration of 8820 msec, n=20 and U=42 msec.

The following 8 pairs were presented to the 20 subjects:
4.4 sec. Isochronous and 4.4 sec. Isochronous,
4.4 sec. Acceleration and 4.4 sec. Isochronous,
4.4 sec. Isochronous and 4.4 sec. Acceleration,
4.4 sec. Acceleration and 4.4 sec. Acceleration,
8.42 sec. Isochronous and 8.42 sec. Isochronous,
8.42 sec. Acceleration and 8.42 sec. Isochronous,
8.42 sec. Isochronous and 8.42 sec. Acceleration,
8.42 sec. Acceleration and 8.42 sec. Acceleration.

Each pair was presented 5 times.

5. RESULTS

The table below shows the results (Table 1).

Pairs	Response			%		
	Longer	Shorter	Equal	Longer	Shorter	Equal
	Incorrect		Correct	Incorrect		Correct
4-4	11	14	70	12%	14%	**74%**
4-4ac	17	48	30	18%	**49%**	33%
4ac-4	55	12	28	**57%**	13%	30%
4ac-4ac	16	11	68	17%	12%	**71%**
8-8	45	10	40	**49%**	10%	41%
8-8ac	23	41	31	25%	**42%**	33%
8ac-8	61	11	23	**63%**	12%	25%
8ac-8ac	37	6	52	39%	6%	**55%**

Table 1: *Responses and Percentages*

Subjects generally tend to perceive the accelerated sequences as shorter than the correspondent periodic sequences (underestimation), when matching isochronous-accelerating or accelerating-isochronous pairs, for both durations of 4 and 8 sec.

Complementarily the isochronous sequences are generally perceived as longer than the accelerated ones (overestimation). Enlargements are slightly more evident when matching accelerating-isochronous pairs with duration of 8 sec. ('8ac-8', Table 1).

Longer durations judgments tend to be overestimated with 49 % of incorrect 'longer' judgments with respect to 41% of 'equal' correct judgments ('8-8', Table 1). So it seems that the duration identity is better perceived for shorter than for longer durations ('4-4' and '8-8' pairs, Table 1).

Two identical accelerating sequences are instead perceived as equal in duration in the majority of responses. But also in this case a tendency to overestimate the 8 sec. accelerating stimuli can be observed, with incorrect responses of 17% ('4ac-4ac') and 39% of subjects ('8ac-8ac' pair, Table 1).

With the 4 sec accelerating stimuli pair ('4ac-4ac' pair, Table 1) incorrect responses are distributed between 'longer' (17%) and 'shorter' (12%) while correct 'equal' responses clearly prevail (71%).

Subjects have been generally able to recognize the equal duration of stimuli pairs, both accelerated and isochronous, with the exception of the isochronous longer pair ('8-8' pair).

If we compare the results for larger durations ('8-8' and '8ac-8ac') we see that while the majority of responses is correct in case of identical accelerating sequences, incorrect responses unexpectedly prevail in case of identical isochronous sequences.

Being pairs of the same type made of copies of the same stimulus (i.e. the stimulus was exactly repeated), one could suppose that their identical structure was recognized instead of their overall duration.

Such a supposition is suggested by comparing the results of larger durations ('8-8' and '8ac-8ac', Table 1): while the majority of responses is correct in case of identical accelerating sequences, incorrect responses unexpectedly prevail in case of identical isochronous sequences. In fact stimuli of 8 sec. duration seem to be better recognized when accelerated than when isochronous (55% and 41% respectively).

Instead of a more difficult process of perceiving the distance in time between the first and the last pulses, subjects could perceive the identical relationship between successive inter onsets. So the hypothesis is that the whole duration of accelerating stimuli could have been inferred from the unfolding sequence of intervals.

In general, data show a minor proportion of correct answers for 8 sec. than for 4 sec. pairs. This could be due to problems of memory overload, suggesting that the overall interval of 8 sec, at least with this kind of stimuli, may not be perceived as psychologically 'present', because of a major effort within the short-time memory. The majority of subjects reported more difficulties while judging the longer audio sequences.

6. CONCLUSIONS

Acceleration seems to influence the estimation of given durations, inducing the sensation that accelerated sequences are shorter than those containing isochronous intervals of time.

In our case the type of stimulus content mostly was temporal, so the internal distribution of time intervals should have had a major influence on the perception, being the main source of information.

Researchers have suggested that time perception could be related to the amount of processing activity: events that are high in complexity require more storage space in memory, lengthening the perceived duration. Considering our pairs of stimuli, the isochronous sequences seem to require less storage space in memory, because of their periodicity (less complex).

Nevertheless the shortening effect, induced by the acceleration, seems generally stronger than the supposed lengthening, possibly caused by the higher amount of storage space.

The condition of judging time intervals does not seem a normal one, unlikely listeners to music attentively think of it, like in this experiment. Subjects have generally found difficult to estimate and compare intervals between the first and the last beat of each sequence of pulses, suggesting that the total durations were in general conditioned by limitations of memory.

The feeling of uncomfortableness was stronger when subjects had to judge two long durations, as confirmed with the 8 seconds stimuli pairs, for which results show the lower correctness in responses.

On the other hand some subjects had a sort of feeling about the internal structure of accelerations, when compared to the regular sequences: they perceived a sort of 'compensation' principle, independently from the judgments they finally gave.

It is worth to recall that time intervals of the accelerated sequences were actually 'compensated'. In order to have the same number of pulses within the same interval of time, the initial IO intervals were longer than the beat period of the isochronous sequences, successively compensated by the gradual decreasing. Consequently, our accelerated sequences always started slower than the isochronous ones.

Further experiments should involve non-musicians, but if we consider that possible significant relationships for musicians have been found, a consequent supposition (to be verified in further experiments) could arise that musical training might not significantly improve the correctness of the requested task.

7. REFERENCES

[1] Boltz, M.G., "The processing of Temporal and Non-temporal information in the remembering of event durations and musical structure", J. Exp. Psychology, vol.24, No.4, 1998.

[2] Brown, S.W. et al., "Attention, multiple timing and psychophysical scaling of temporal judgments", in F. Macar et al. (Eds.), Time, Action and Cognition bridging the gap, pp. 129-140, Kluver Academic Publisher, Dordrecht, 1992.

[3] Clark, E.F., "Rhythm and Timing in Music", in Deutsch, D. (Ed.), The Psychology of music, pp. 473-500, Academic Press, New York, 1999.

[4] Drake, C., "Psychological processes involved in the temporal organization of complex auditory sequences: universal and acquired processes" Music perception, Vol. 16, No. 1, pp. 11-26, Fall 1998.

[5] Fraisse, P., Les structures rythmiques Etude psychologique, Érasme, Paris, 1956.

[6] Jones, M.R. and Boltz, M., "Dynamic attending and response to time", Psychological Review, 96, 459-491, 1989.

[7] Kramer, J., The time of music, Schirmer books, New York, 1988.

[8] Levitin, D.J., and Cook, P.R., "Memory for musical tempo: Additional evidence that memory is absolute", Perception and Psychophysics, 58(6) 927-935, 1996.

[9] Marek et al., "Tempo change: Timing of simple temporal ratio", in Desain, P., Windsor, L. (Eds.), Rhythm perception and production, 143-156, Lisse, Sweets and Zeitlinger, 2000.

[10] Meyer, L.B., Emozione e Significato in Musica (C. Morelli, trad.), Il Mulino, Bologna, 1992.

[11] Van Noorden L., Moelants D., "Resonance in the perception of musical pulse", Journal of new music research, Vol.28, No. 1., Sweets and Zeitlinger, Lisse, 1999.

[12] Warren, R.M., Auditory Perception A New Analysis and Synthesis, University Press, Cambridge, 1999.

[13] Zakay, D., "On prospective time estimation, temporal relevance and temporal uncertainty", in F. Macar et al. (Eds.), Time, Action and Cognition Toward bridging the gap pp. 109-118, Kluver Academic Publisher, Dordrecht, 1992.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

TIMBRAL SEMANTICS AND THE PIPE ORGAN

Alastair C. Disley and David M. Howard

Department of Electronics
The University of York

acd114@ohm.york.ac.uk

ABSTRACT

Words used to describe timbre are usually difficult to define or relate to a measurable phenomenon. This paper attempts to establish if any such words are objective with a common understanding, or if they are all subjective. The pipe organ is used because it is both a complex timbral synthesizer and a non-electronic instrument from which repeatable samples can be taken.

A number of adjectives were gathered from English speaking subjects. The most common words were selected and used without specific definition as rating scales in comparative listening tests. Comparison with spectral analyses revealed possible cues for some words, and audio examples were synthesized to test these theories. Some adjectives have emerged as having degrees of common understanding across a majority of subjects.

1. INTRODUCTION

Adjectives are frequently used to describe sound, but little research has been done into how the words people use relate to measurable acoustic phenomena. It is not clear whether any such words are idiosyncratically subjective, i.e. with meaning only to a given person, or objective across a number of people with common understanding. Previous research on adjectives describing aural and other sensations is summarized in Rioux [1]. Rioux also gathered descriptive words from an organ-builder [2], categorized them and conducted listener tests [3]. That research involved single pipes, and the words gathered mostly referred to the speech of a pipe.

The use of pipe organs in both this and Rioux's experiments is because it is one of the few acoustic instruments to be an additive timbral synthesizer, where the player controls the timbre by combining pipes of different pitches and individual timbres into one chorus. Thus with its variety of sounds, different timbres can be synthesized without going outside the realms of a single acoustic musical instrument. It is also possible to take repeatable samples from a pipe organ because of the nature of its sound generation and control, something that is impossible on most other acoustic instruments.

Listeners do not usually hear pipes in isolation, so in order to gain meaningful results applicable to common situations, all audio samples of pipe organs used in this experiment were of complete choruses of pipes, played in chords as might be heard in a hymn. The words to be used in this test were chosen by frequency of occurrence in a wide scale listening experiment, to ensure that they were words that would be relevant to the listeners.

2. EXPERIMENTAL PROCEDURE

2.1 Adjective gathering

Forty-five listeners were played recordings from four different organs and asked to describe what they were hearing with no other prompting. The adjectives they used were then categorized and counted, including all derivatives, and the most frequently used were selected for further study. Listeners were players of or keen listeners to the pipe organ, gathered from a request on the internet mailing list piporg-l.

The most common words used were: "bright", "clear", "flutey", "reedy", "thin", "full", "balanced" and "warm". "Reedy" was not studied further as organ pipes are also divided into two types, flue and reed, which could cause confusion. In all experiments described in this paper, listeners were not given definitions of any terms to avoid the disturbance of any presence or lack of prior common understanding.

2.2 Initial experiments

As part of a seminar on this research, 75 people took part in a large scale listening test using the terms "bright" and "clear", which were the most common. All listeners were English speakers, most as their first language, and self-divided into three groups according to their musicality (none, some or lots). They were also asked if they were regular listeners to or players of the pipe organ. In this and all other experiments in this paper, testing took the form of comparison tests with listeners marking a scale from 1 to 11, such that an answer of 6 would indicate either "neither" or "both equal".

In the first case, the principal chorus of an organ in a reverberant acoustic was played with and without its mixture stop (a mixture is a stop that adds several higher harmonics to the ensemble).

Listeners on average thought that the mixture added some slight brightness, but made no difference to clarity.

In the second example, the 8' principal choruses of two different organs were played with their mixtures, one a 19th century instrument (the Schulze in Doncaster Parish Church) in a moderately reverberant environment, and the other a 20th century instrument (the Grant, Degens and Bradbeer at York University) in a drier acoustic. The latter instrument was thought to be both brighter and clearer, and Fourier analysis revealed that its higher harmonics were of greater amplitude, suggesting that as might be expected, the perception of brightness is directly related to the strength of higher harmonics. Clarity might have an inverse relationship with the amount of reverberation.

Results were also analysed against amount of musical training. Although there was little difference in average results between the groups, as musical training decreased, standard deviations increased, indicating a decrease in common understanding of words. Interestingly, the subset of listeners who indicated familiarity with the pipe organ had the largest standard deviation of all.

The next five most common adjectives, "flutey", "thin", "full", "balanced" and "warm", were used in a listener test conducted over the internet with 23 subjects, all of whom were familiar with the pipe organ as regular performers or enthusiastic listeners. All were native English speakers: thirteen were American, eight were English and two were Canadian. All but one were male, and age and geographical data were collected, but these appeared to have no significant bearing on results. Samples were presented in uncompressed 16-bit stereo WAV format, sampled at 22kHz, as a compromise between download time and maximum resolution. Following comments from previous such tests, reverberation was not cut off early.

Four recordings of different organs were used, with in each case the same hymn-type extract played on the 8' principal chorus. Acoustic analyses of the four organs are given below, with steady-state Fourier analyses (4096 point Hamming window) simplified to harmonics only, and shown at ten points across the keyboard range. The principal chorus of all four organs includes 8', 4' and 2' stops. Heslington (figure 1) includes a Twelfth, all other organs include a Mixture. St. Chad's (figure 3) is in a very reverberant acoustic, and like Heslington is a 19th century instrument. Port Sunlight (figure 2) is an early 20th century instrument and has a more muted chorus typical of its time. The University of York organ (figure 4) is also typical of its later 20th century date, by which time the fashion was for stronger mixtures and weaker 8' stops.

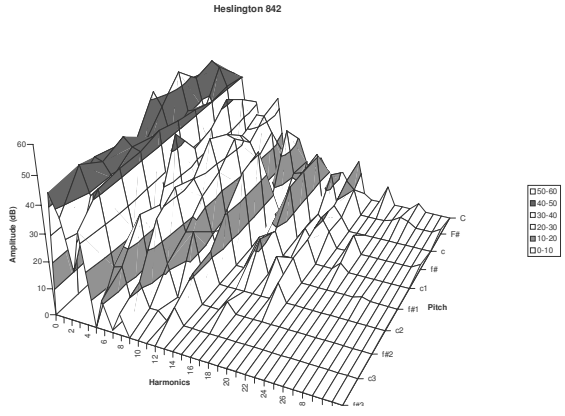


Figure 1: *Principal chorus of Heslington*

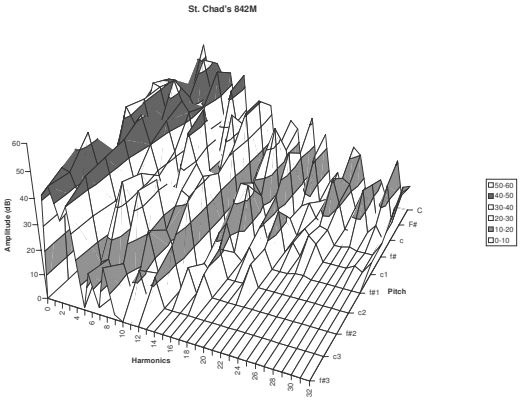


Figure 3: *Principal chorus of St. Chad's*

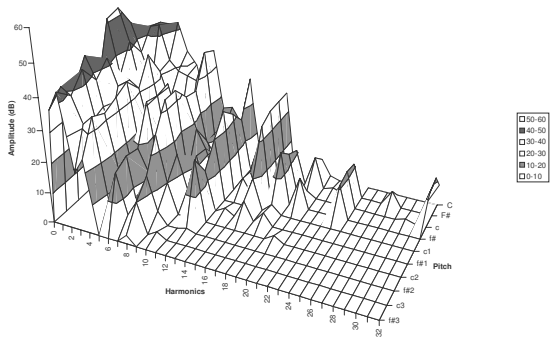


Figure 2*: Principal chorus of Port Sunlight*

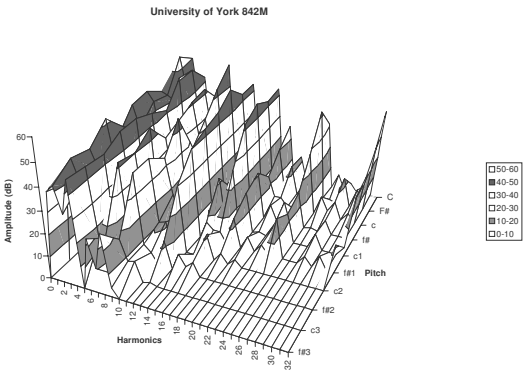


Figure 4: *Principal chorus of University of York*

2.3 Results and analysis

For some of the adjectives tested, the mean in many comparisons was near the central value of 6. For the term "balanced", the mean was always within 0.91 of that central value, and standard deviations varied from 2.14 to 2.61, indicating some difference of opinion on what "balanced" meant, but also that no organs in the test were perceived to be unbalanced. Full again had a mean always within 0.96 of 6, and standard deviations varied from 2.23 to 3.01. Full as an adjective seems to have less common understanding than any other word used.

The word "thin" produced its strongest results (means biased 1.83 and 1.52 from the centre, standard deviations 2.1 and 2.04 respectively) when it was applied to the University of York instrument compared with Port Sunlight and St. Chad's. Possible explanations of this from the acoustic analyses are the stronger high and weaker low harmonic components of the University of York's ensemble, which each require separate testing.

The word "flutey" appeared to follow the instrument with weaker higher harmonics. The strongest bias from the centre was 2.09, with a standard deviation of 2, when the Jack Lyons organ was compared with the Port Sunlight instrument. Standard deviations varied from 1.88 to 2.15, suggesting a degree of uncertainty among participants.

The word "warm" produced the greatest agreement among subjects, with standard deviations varying from 1.34 to 1.5 in tests involving the Jack Lyons instrument, which was clearly perceived to be not warm in comparison with other instruments. The other three instruments were perceived to be warmer by between 2.26 and 2.65 from the central value. However it appears that warmth is not solely a result of less strong higher harmonics, as St. Chad's was perceived to be slightly warmer than Heslington (by 1.52, with a standard deviation of 1.65). This may be because of St. Chad's greater reverberation.

These results suggest that an organ with stronger higher harmonics will be perceived as less flutey and less warm. Perception of a sound as "thin" may be related to weaker lower harmonics, and/or stronger higher harmonics. Warmth also seems to be affected by the strength of higher harmonics, but may also be affected by the amount of reverberation present. In cases where the amount of reverberation is similar, the results for warmth and flutey are also similar.

2.4 Verification of theories

To develop these theories, a further set of comparative listener tests was developed for the same subject group, of whom 14 took part. The samples were developed to establish whether adding stronger higher harmonics made people describe the sound as less "warm", less " flutey", and more "thin". The samples also needed to test whether reducing the strength of lower harmonics produced a description of more "thin". Thirdly they needed to test whether adding reverberation produced an increased perception of "warmth". The words "balanced" and "full" were rejected at this stage, and replaced by the terms "bright" and "clear" that had been used in the first listening test.

In order to place an identical sound in two acoustic environments, the four samples for this test were produced on an electronic organ, a Wyvern B235 using Bradford synthesis technology [4]. The two reverberant environments were simulated on a Yamaha S-REV 1 convolution-based device. The more reverberant acoustic used its preset program "St. John the Divine, no. 3", with the other three samples using preset "Warm Wooden Church no. 1". Timbral manipulation of the samples needed to be within the range of sounds that a pipe organ might produce, so synthesis of timbres was achieved by addition and removal of stops. Stonger higher harmonics were produced by adding a mixture to a basic chorus of 8', 4' and 2' principals. Replacing the 8' Principal with an 8' Stopped Flute for the fourth sample produced a reduction in lower harmonics.

Results from this final test were more consistent, with an average standard deviation of 1.74, compared with 2.18 for the previous test. When the sample with the 8' stopped flute was compared with that with the 8' Principal, it was described as more "thin" (by 2.08 from the mean in both more and less reverberant examples). "Flutey" gave significant results when a pair of samples included that with the mixture, and means varied between 1.46 and 2.38 towards the samples without the mixture. When the mixture was added, perception of brightness clearly increased, by 3.54 and 3.62 from the mean both with and without reverb (standard deviations 1.13 and 0.96 respectively).

Subjects gave more varied answers when the sample including the mixture was compared with the sample containing the stopped flute – although the average perception of brightness was 2.00 towards the mixture sample, the standard deviation was 2.68. The perception of "warm", like " flutey", was significant in those samples compared with the mixture. The stopped flute sample was perceived to be slightly less warm than those other samples that lacked a mixture. In all cases clarity appeared to follow brightness, although to a lesser extent, and with a slight bias away from the more reverberant samples and towards both the mixture and stopped flute samples.

3. CONCLUSIONS

It was suggested in section 2.2 that the addition of a mixture would increase the perception of brightness, and that clarity was related to less reverberation. The tests in section 2.4 appear to support these theories. Listeners appear to have a common understanding of brightness as being an increase in strength and/or number of higher harmonics. The perception of the sample with reduced lower harmonic strength suggests that this increase is measured relative to the strength of lower harmonics. It is more difficult to make such a declaration for clarity, as reverberation is not a simple linear phenomenon, but there is at least a distinct degree of common understanding as to its meaning.

"Full" appears to have no common understanding among subjects, and "balanced" was also less well understood, although

both may have greater understanding outside the limitations of this experiment. "Flutey" was used to describe instruments with weaker higher harmonics when they were compared with a mixture, but was less consistently used when neither sample had a mixture.

In the final test, "Thin" did not appear to be related to an increase in strength of higher harmonics, but to weaker lower harmonics in the sample including the Stopped Flute. "Warm" was also related slightly to greater strength of lower harmonics, but was most strongly associated with the sample in a pair exhibiting weaker high harmonics. Perception of a sound as "warm" also increased with the amount of reverberation.

In summary, most of the common words provided by the initial gathering of adjectives have proved to have degrees of common understanding for certain circumstances. Some words, particularly "bright" and "warm", have good common understanding across all circumstances used in this report, within a set of listeners for whom English is a first language and who are familiar with the pipe organ.

Although reverberation has been mentioned as having an effect for several words studied, the complex nature of reverberation means that its precise effect in any circumstance is hard to quantify without further study. The first large-scale experiment indicated that as musical training decreased common understanding of terms also decreased, which suggests that these terms might be introduced subconsciously as part of the growth of a person's musical understanding. Future research could also investigate how without specific formal training, the common understanding identified in this paper been gained.

4. ACKNOWLEDGEMENTS

The experiments described in this paper could not have been completed without the willing assistance of listeners from piporg-l and the University of York. Thanks are also due to the churches, organizations and their staff who permitted recordings of their pipe organs to be made.

5. REFERENCES

[1] Rioux, V., Sound Quality of Flue Organ Pipes, PhD thesis, Chalmers University of Technology, Gothenburg, 2001, Chapter II.

[2] Rioux, V. and Västfäll, D., "Verbal descriptions of flue organ pipe sounds (I): a qualitative analysis", in [1], pp. 99-115.

[3] Rioux, V. "Methods for an objective and subjective description of starting transients of some flue organ pipes – integrating the view of an organ-builder", Acustica, Vol. 86, No. 4, pp. 634-641, 2000.

[4] Comerford, P. J., "Simulating and organ with additive synthesis", Computer Music Journal, Vol. 17, No. 2, pp. 55-65, 1993.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

HARDNESS RECOGNITION IN SYNTHETIC SOUNDS

Bruno L. Giordano, Karin Petrini

Department of General Psychology, University of Padova, Italy
bruno.giordano@unipd.it

ABSTRACT

Sound source recognition investigates recovery of different features of the objects, whose interaction lead to the generation of the acoustical signal. Among them material type have received particular attention, while recovering of material properties, such as hardness, have been scarcely considered. Hardness plays a significant role in the musical field too, especially for percussion instruments, where resonating objects of variable hardness are struck with mallets of variable hardness. Comparison of previous results on hardness recognition point toward the perceptual independence of the resonator and exciter properties. This issue was addressed in four experiments conducted on stimuli synthesized with a physical model, which allowed independent manipulation of the exciter and resonator properties. Free identification and forced choice tasks have been used to investigate the ability of listeners to discriminate variations in the exciter from variations in the resonator. Scaling tasks have been used to investigate the relationship between the synthesis parameters and the hardness estimates of the exciter and of the resonator. Free identification and forced choice data reveal a bias toward the interpretation of the acoustical signals in terms of features of the resonating object. Hardness scaling results reveal the perceptual dependence of exciter and resonator properties, although strong individual differences are found.

1. INTRODUCTION

A relatively recent field in auditory perception investigates recognition of the features of the sounds source [1]. The most studied class of signals is that of impact sounds [2], [3], generated by the interaction between a highly damped object, the exciter or hammer, and a vibrating object, called resonator or sounding object. Particular interest have received the study of categorization or discrimination of the material type of the resonator [4], [5], [6], [7]. Wildes and Richards [8] hypothesized that material could be uniquely specified, at the acoustical level, by means of the $tan\phi$ coefficient, which measures the degree of anelasticity of the material of the resonator, defined as

$$tan\phi = 1/\pi f t_e \quad (1)$$

where f is the frequency of vibration and t_e is the time it takes for the amplitude to decay to $1/e$ of its starting value. Different perceptual studies found material categorization to be strongly influenced by this acoustical parameter (rubber and wood chosen for higher $tan\phi$ values than glass and steel)[5], [4]. Other acoustical parameters, modulated by physical features extraneous to the material type of the resonator, were found to influence recognition of this sound source feature: the decay time of signal amplitude, as modulated by external damping of the resonator, and signal frequency, as modulated by the size of the resonator [6]. In a further experiment Giordano [6] demonstrated recognition of material type of the resonator to be independent of the material of the exciter, although the range of variation in the elastic properties of the used exciters was too low to induce a relevant timbral variation in the generated signals. Scaling of a mechanical material property, namely mallet hardness, was investigated by Freed [9]. Different cooking pans of variables size were struck with mallets of variable hardness. Participants were found to properly scale the hardness of the mallets, independently of the size of the resonator. Hardness estimates were found to be influenced by signal amplitude, spectral centroid, and amplitude decay velocity. On the basis of these results, recognition of the material properties of the resonator can be hypothesized to be independent of the features of the exciter, and viceversa. This was first assessed by investigating the ability of listeners to distinguish variations in the resonator from variations in the exciter features. Both free identification and forced choice tasks were used. Second, scaling of the hardness of both the exciter and the resonator was conducted, in synthetic stimuli generated by manipulating the properties of the resonator as well as those of the exciter. These topics are of interest to the musical field too, particularly for the field of percussion instruments, where resonators made of a large variety of materials (e.g., metallic and wooden instruments, membranophones, lithophones, glass and ceramic instruments) are struck with objects of variable hardness (e.g., felt and wooden mallets, metallic sticks, hands) [10]. Then research on hardness scaling could partially highlight the problem of the perceptual quality of percussion instruments, scarcely considered in timbre perception research.

2. SYNTHESIS PARAMETERS

Stimuli were generated with the impact model by Avanzini et al. [11]. Three parameters of the model were investigated. The first two were related to the resonator: the frequency of the lowest mode f, and $\tau = 1/\pi tan\phi$. The first parameter can be thought as modelling the geometrical properties of the resonator (e.g., length). The τ parameter models the material of the resonator. High τ values correspond to hard or stiff materials (steel, glass). Low τ values correspond to soft or elastic materials, such as rubber. The third synthesis parameter was related to the interaction between the exciter and the resonator: the force stiffness coefficient k, directly related to the elastic properties of the exciter. Low k values may represent rubber or felt mallets. High k values may represent stiff mallets (metal, hard woods).

3. FREE IDENTIFICATION

A free identification experiment was conducted in order to assess how variations in the resonator or in the exciter parameters were described or interpreted by listeners. A set of 16 stimuli was syn-

thesized combining two values for the f parameter (50 and 800 Hz), two values for the τ parameter (30 and 200), and four values for the k parameter (10^2, 10^4, 10^6, and 10^8). Stimuli were presented in eight series of four sounds. In half of them only the resonator parameters were varied, while the k parameter remained constant, in the other half the k parameter varied and the resonator parameters were kept constant. Stimuli were presented via headphones in a silent room. The order of the series within the experiment, as well as the order of the stimuli within each series, was randomized. Nine participants were presented each series once and were asked to describe what was varying within each series. No information concerning the nature of the stimuli was given.

3.1. Results

Descriptions were categorized in three groups: generic descriptions (e.g.,"four sounds together"), descriptions based on the perceptual features of the sounds (e.g.,"high pitch", "bright"), descriptions which focused mainly on sound source features (e.g., "finger tapping on wood"). This latter was divided in two subgroups: those which described a constant resonator (e.g., "a xylophone sound with different decay velocities"), and those which described different resonators (e.g., "a drum, a bell..."). Table 1 shows the proportion of occurrence of each category.

Varied	**Category**			
parameter	G	P	Cr	Vr
Resonator	0.166	0.027	0.111	0.694
Exciter	0.027	0.111	0.583	0.278

Table 1: *Proportion of choosing each description category, for the variable resonator and variable exciter series. G = generic, P = perceptual, CR = constant resonator, VR = variable resonator.*

Overall listeners tended to describe sounds mainly in terms of sound source features (83% of the descriptions), rather than in terms of perceptual features alone. Notably the constant object descriptions where more frequent when the k parameter was varied, while the variable object descriptions were more frequent when the resonator parameters where varied. This indicated that a variation in the exciter features did not generate a variation in the recognized resonator features, and viceversa. This conclusion can be biased by the fact that, differently from what observed for the resonator, the exciter was rarely mentioned in the descriptions. When a constant object description was given, listeners tended to describe the variation within the series either in terms of perceptual qualities (e.g., differences in the attack quality) or in terms of a change in the action exerted over the resonator (e.g., a variation in the external damping).

4. FORCED CHOICE CATEGORIZATION

A forced choice task was used to asses wether listeners were able to distinguish variations in the resonator from variations in the exciter. Stimuli were selected from a synthesis space defined by the combination of five equally log-spaced levels for f, τ, and k. Table 2 shows the levels used for each parameter.

Sixteen series of stimuli were built by varying only one of the three investigated parameters, either in ascending or descending order. In half of them $[(f_a, \tau_2, k_2)$, (f_a, τ_4, k_4), (f_2, τ_a, k_2), $(f_4, \tau_a, k_4)]$, a resonator parameter was varied, in the other half $[(f_2, \tau_2, k_a)$, (f_2, τ_4, k_a), (f_4, τ_2, k_a), $(f_4, \tau_4, k_a)]$ the k coefficient was varied (the subscript a denotes the use of all the available parameters within the series.). All series were judged once by each participant, and were presented in randomized order. Two types of instructions were given. In both cases participants were given a definition of resonator and of exciter, and an example of exciter/resonator interaction was described verbally (a baked clay dish struck with a wooden ball). Additionally a sub-group of participants was presented auditory examples of variations in the resonator and in the exciter properties, using real rather than synthetic sounds. They were asked to tell wether the presented series was generated by a variation in the resonator or in the exciter. Twelve participants were given verbal-only instructions; nine participants were presented the verbal+auditory instructions.

	Levels				
Parameter	1	2	3	4	5
f (Hz)	50	100	200	400	800
τ	10	20	40	80	160
k	$5e^6$	$3.3e^7$	$2.24e^8$	$1.495e^9$	$1e^{10}$

Table 2: *Synthesis parameters values used for the investigated stimuli.*

4.1. Results

Table 3 shows the proportion of choosing the response "variation of the resonator" for the variable exciter and variable resonator series, and for the two types of instructions.

Varied	**Instructions**	
parameter	Verbal	Verbal+Auditory
Resonator	0.604 (0.673–0.535)	0.694 (0.770–0.619)
Exciter	0.573 (0.643–0.503)	0.403 (0.483–0.323)

Table 3: *Proportion of choosing the "variable resonator" response category for the variable resonator and exciter series, and for the verbal and verbal+auditory instructions.* 95% *confidence intervals are given in parentheses.*

An overall tendency of participants to attribute the change in the auditory stimuli to a variation in the resonator was observed. This tendency was particularly strong when only verbal instructions were given, where the proportion of choosing the response "variable resonator" in the variable exciter series was higher than chance level and not significantly different from the same proportion for the variable resonator series. This tendency was weakened by the use of auditory examples. These results are consistent with those collected with the free identification procedure. Listeners revealed a bias toward the interpretation of the perceptual variations in terms of features of the resonator, rather than in terms of features of the exciter. The effect of the auditory examples supports this conclusion, as it shows that listeners do not pay attention to the perceptual variations associated to a change in the features of the exciter until are instructed in this sense.

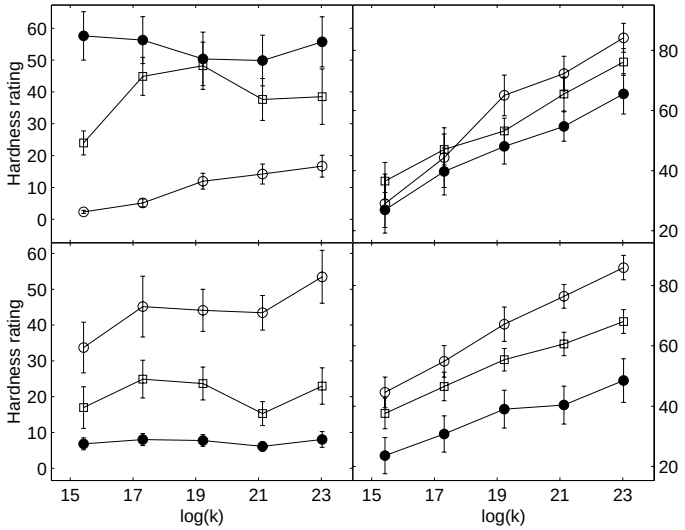


Figure 1: *Average exciter hardness ratings as a function of the k coefficient. Variable f set: top row, variable τ set: bottom row. Filled circles: f_1 (top) τ_1 (bottom), empty squares: f_3 (top) τ_3 (bottom), empty circles: f_5 (top) τ_5 (bottom). Left column: cluster 1, right column: cluster 2. Error bar = ± 1 SE*

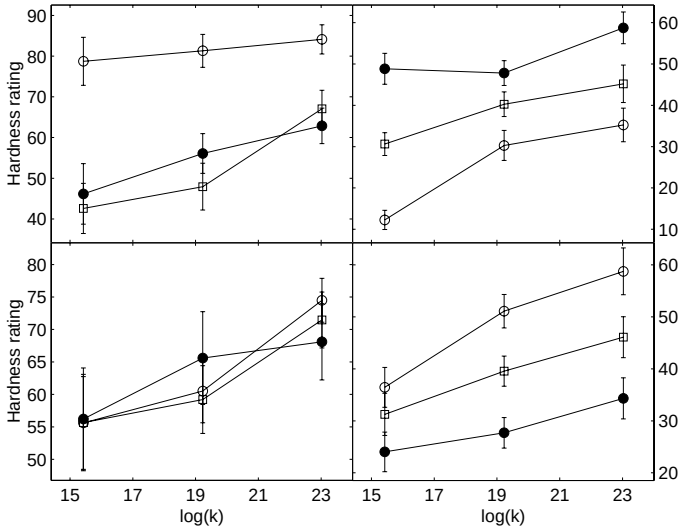


Figure 2: *Average resonator hardness ratings as a function of k. Filled circles: f_1 (top row) τ_1 (bottom row), empty squares: f_3 (top row) τ_3 (bottom row), empty circles: f_5 (top row) τ_5 (bottom row). Left column: cluster 1, right column: cluster 2. Error bar = ± 1 SE*

5. HARDNESS SCALING

Hardness scaling experiments were conducted in order to test the perceptual independence of the resonator and of the exciter properties. Two experiments were conducted. Stimuli were selected from the synthesis space investigated in the previous experiment. Scaling of the hardness of the resonator was investigated with a set of 27 stimuli, given by the combination of the first, third and fifth level of each parameter. Scaling of the hardness of the exciter was investigated with two sets of 15 stimuli. In both cases all the five levels of k were used. In the variable f set τ was fixed at the third level, and three f values were used (f_1, f_3, f_5). In the variable τ set f was fixed at f_3 and τ assumed the following values: τ_1, τ_3, τ_5. Scaling of the hardness of the resonator and of the exciter was performed by two different groups of 18 participants. The variable f and variable τ sets were presented in counterbalanced order across participants. Stimuli were presented in randomized order within each session. Every stimulus was judged three times by each participant.The verbal+auditory instructions of the previous experiment were given. Participants were asked to rate the hardness of the exciter/resonator on a 1 (very soft)–100 (very hard) scale.

5.1. Results

In order to evaluate the presence of different response profiles, average hardness ratings were subjected to cluster analysis. An agglomerative hierarchical clustering procedure, with average linkage between clusters as grouping criterion and a quadratic Euclidean distance measure was used. This analysis was performed separately on the variable τ, on the variable f, and on the resonator hardness data sets. Two clusters of participants were extracted from each data set, by considering only the two furthest ones, joined in the last stage of the agglomerative procedure. This analysis individuated two clusters of nine participants in the variable f data set, one cluster of seven participants and one of eleven in both the variable τ and in the resonator hardness data sets. Despite the cluster analysis was performed separately for the variable τ and f data sets, the division of participants among clusters was almost the same for both these sets. In particular all the seven participants of the first cluster in the variable τ data set were included in the first cluster of the variable f set. Not surprisingly their response profiles for both the sessions had strong similarities, as shown below.

Figures 1, and 2 show average hardness estimates as a function of the k coefficient, for the different data sets, and for all the extracted clusters. Average ratings were analyzed with a repeated measures ANOVA. A separate analysis was performed for each cluster of participants.

Hardness estimates for the first cluster in the variable f set were significantly influenced by f ($F_{2,16} = 38.67$, $p < 0.001$), by its interaction with k ($F_{8,64} = 5.70$, $p < 0.001$), but not by the one-way effect of k ($F_{4,32} = 2.28$, $p = 0.083$). Analysis of the simple effects of the k variable revealed that its effect on the hardness estimates was significant only for the intermediate f level ($F_{4,32} = 5.80$, $p = 0.001$). For this reason it can be concluded that hardness estimates were determined almost only by signal frequency, higher frequency leading to higher hardness estimates. Responses, for the second cluster of the same data set, were influenced only by k ($F_{4,32} = 55.88$, $p < 0.001$), higher k values being associated to higher hardness estimates. Signal frequency, in other words, did not influence exciter hardness estimates. In the variable τ set, hardness estimates for the first cluster were influenced by τ $F_{2,12} = 34.94$, $p < 0.001$) and by its interaction with k ($F_{8,48} = 2.19$, $p = 0.045$). As with the variable f set, the one-way effect of the k coefficient was not significant ($F_{4,24} = 2.24$, $p = 0.095$). Analysis of the simple effects of the k parameter revealed that it slightly influenced hardness estimates only for the

highest τ level ($F_{4,24} = 2.91$, $p = 0.043$). For this reason it can be concluded that participants belonging to the first cluster focused mainly on the τ parameter, discarding almost completely the k parameter as with the variable f set. Within this cluster hardness estimates increased as a function of the τ coefficient. On the contrary responses in the second cluster were influenced by the one-way effect of both τ and k variables ($p < 0.001$ in both cases), and by their interaction ($F_{8,80} = 2.17$, $p = 0.038$). As in the first cluster an increase in the τ parameter was associated to an increase in the exciter hardness estimates. The same effect was found for the k parameter. Overall participants belonging to the first cluster of both data sets ignored the k coefficient when estimating the hardness of the exciter, focusing on the parameters related to the resonator, although instructed otherwise. The second cluster of participants, instead, gave results highly consistent with those collected by Freed [9] on real sounds, where hardness estimates were found independent on signal frequency. Hardness estimates of the exciter in both clusters were found influenced by the material of the resonator, as modelled by the τ coefficient.

Resonator hardness estimates for the first cluster (seven participants) were slightly influenced by f ($F_{2,12} = 4.01$, $p = 0.046$), and by k ($F_{2,12} = 4.18$, $p = 0.042$). The other effects were non-significant ($p > 0.08$ in all cases). Higher frequencies were given higher resonator hardness estimates. As found with the previous experiment, an increase in k was associated to an increase of the hardness estimates. The absence of a significant effect of the τ parameter for this cluster is counterintuitive, and contradicts the already known effects that this parameter has on material type categorization. The low significance of the effects f and k variables, makes more probable the hypothesis that this group of participants gave unreliable estimates. In the second cluster (eleven participants) responses were significantly influenced by the one-way effect of all the synthesis parameters ($p < 0.002$ in all cases), as well as by the two-way interactions between k and f, and between k and τ ($p < 0.02$ in both cases). An increase in both k and τ was associated to an increase in the resonator hardness estimates. The same response pattern was found in the second cluster for the variable τ set. The effects of f and τ on resonator hardness scaling, are consistent with results gathered on categorization of material type [4], [5], [6], and establish a link between results collected with largely differing experimental techniques (scaling and categorization).

6. CONCLUSIONS

Three different experimental techniques were used to address the perceptual independence of the resonator and exciter features. Free categorization and forced choice tasks revealed a bias toward the interpretation of the variation among stimuli in terms of a variation in the resonator. Confusion among these two objects was investigated by means of hardness scaling of both the exciter and the resonator. Across almost all clusters, independent of wether the exciter or the resonator hardness was scaled, the k and τ parameter were found to have similar effects. Both these synthesis parameters model material properties, so that it can be concluded that material properties of the exciter and of the resonator are confused each other. A large variability was associated to the effects of the f parameter, which modelled the geometrical properties of the resonator. This was associated either to an increase or decrease of the resonator hardness, and to an increase or, consistently with results by Freed [9], to no effects on the exciter hardness. Acoustical analyses will have to be performed to highlight the dependence of hardness estimates of the exciter and of the resonator on the material properties of both. Plausibly the equivalence of the effects of these sound source features on the recognized hardness is due to similar effects on relevant acoustical signal properties.

7. ACKNOWLEDGMENTS

This work was supported by the European Commission, under contract IST–2000–25287.

8. REFERENCES

[1] McAdams, S., Recognition of sound sources and events, in McAdams, S. and Bigand, E., eds., Thinking in sound: the cognitive psychology of human audition, pages 146-198, Oxford University Press, 1993.

[2] Kunkler-Peck,A.J.,Turvey, M.T., “Hearing shape”, Journal of Experimental Psychology: Human Perception and Performance, 26(1):279-294, 2000.

[3] Lakatos, S., McAdams, S., Causse, R., “The representation of auditory source characteristics: simple geometric form”, Perception & Psychophysics, 59(8):1180-1190, 1997.

[4] Klatzky, R.L., Pai, D.K., Krotkov E. P., “Perception of material from contact sounds”, Presence: Teleoperators and Virtual Environment, 9(4):399-410, August 2000.

[5] Rocchesso, D.,Ottaviani, L., Fontana, F., Avanzini, F., Size, shape, and material properties of sound models, in Rocchesso, D., and Fontana, F., eds., The sounding object, Mondo Estremo, Firenze, 2003.

[6] Giordano, B.L., Material categorization and hardness scaling in real and synthetic impact sounds, in Rocchesso, D., and Fontana, F., eds., The sounding object, Mondo Estremo, Firenze, 2003.

[7] Lutfi, R. A., Oh, E. L., “Auditory discrimination of material changes in a struck clamped bar”, J. Acoust. Soc. Am., 102(6):3647-3656, 1997.

[8] Wildes, R., Richards W., Recovering material properties from sound, In W. Richards, editor, Natural Computation, pages 356-363. MIT Press, Cambridge, MA, 1988.

[9] Freed, D.J., “Auditory correlates of perceived mallet hardness for a set of recorded percussive events”, J. Acoust. Soc. Am., 87(1):311-322, 1990.

[10] Rossing, T.D., Science of Percussion Instruments, World Scientific, New Jersey, 2000.

[11] Avanzini, F., Rath, M., Rocchesso, D., Ottaviani, L., Low level models: resonators, interactions, surface textures, in Rocchesso, D. and Fontana, F., eds., The sounding object, Mondo Estremo, Firenze, 2003, available at `http:\\www.soundobject.org`

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

CATEGORICAL PERCEPTION OF MICROTONAL INTERVALS IN SIMPLE TONAL CONTEXTS

Erkki Huovinen

Department of Musicology
University of Turku, Finland
erkhuo@utu.fi

ABSTRACT

The categorical perception of melodic intervals was studied in a discrimination experiment that used pure tone intervals between 252.6 and 505.3 cents, following complex context tones of lower fundamental frequency after a 3-s pause. The pitch relationship between the context tones and the stimulus intervals was altered in 63.2-cent increments. Despite instructions to ignore the context tone, even such simple "tonal contexts" had some significant effects on subjects' capability of detecting pitch shifts of 63.2 cents in the lower tone of the stimulus interval. Detection performance showed asymmetries that could be explained with reference to categorical perception of the intervals between the context tone and the tones of the stimulus interval. Such effects were especially marked within pitch ranges that corresponded to the pitch-class of the context tone and the fifth above it. Moreover, performance was improved when the pitch-class relationship between the context tone and the tone to be detected roughly corresponded to an interval of (minor or major) third. Thus, even a very simple musical context may affect the categorical perception of intervals in ways that can only be explained by taking the properties of the context into account.

1. INTRODUCTION

"Categorical perception" referes to the fact that along some physical continua, discrimination of equally spaced stimuli is better at category boundaries than within categories. At present, the categorical perception of musical intervals is a well documented phenomenon (for a review, see Burns, 1999). It is especially important for melodic intervals for which there are no psychoacoustic cues of purity such as beats and roughness. Besides more traditional laboratory experiments, there have been studies in which the intervals to be identified or discriminated have appeared within a tonal context, being presented after a melody (Wapnick, Bourassa, & Sampson, 1982) or a scale (Tsuzaki, 1991). The main interest in these studies has been whether the subjects' performance is improved by embedding the stimuli in larger contexts. Burns (1999, p. 235) sums up the present state of knowledge by stating that "intervals apparently still are perceived in a categorical manner in context, and the accuracy of identification and discrimination does not appear to be markedly different in context or in isolation."

This may be true, but the problem is that the categorical perception of intervals might still reveal differences depending on the way that the embedding has occurred. Especially, what we do not know is whether and how the effects of an established tonal context vary according to the pitch relationship of the stimulus interval in relation to a perceptual tonal center, or, for instance, a perceptually salient bass tone. Given the basic idea of tonal hierarchies (cf. Krumhansl, 1990), it seems plausible to hypothesize that identification and discrimination would differ according to the "tonal stability" of the pitch range within which the stimulus occurs in relation to an established tonal center. Experiments have suggested that Western listeners are generally adept in hearing even highly unconventional melodic stimuli in terms of tonal centers, and that they often use subjectively consistent criteria for finding a tonal center, which is subsequently used to organize the auditive input (Huovinen, 2002). In the study at hand, the idea was to present a series of interval discrimination tasks in a systematically conceived set of intervallic relationships with respect to a tonally salient bass tone, which could be used to emulate a local tonal center. It was hypothesized that even such elementary "tonal contexts" would elicit significant differences in the categorical perception of the actual stimulus intervals.

The pitch materials for the experiment were derived from the 19-tone equal temperament (19-ET; see, *e.g.*, Mandelbaum, 1961). Perhaps the most convenient way to intoduce oneself to this system is by imagining a diatonic scale in which each whole-tone is divided into three intervallic units of 63.2 cents and each semitone into two such units. If we assign pitch-class numbers 0 through 18 to the 19 pitch-classes of the system, we get a diatonic scale that comprises pitch-classes 0, 3, 6, 8, 11, 14, and 17. Another lucky property of 19-ET is that its thirds (minor third 315.8 cents, major third 379.0 cents) are better approximations to their pure counterparts (315.6 and 386.3 cents) than what we have in 12-ET. The present study can also be read as a cognitive study of the possibilities within 19-ET, but the main reason for using the system was that it provided a non-arbitrary tonal reservoir with a manageable number of pitch-class relationships and a good mixture of familiar diatonic intervals and what might easily be judged their out-of-tune variants. In the following, interval categories inherent in 19-ET are referred to by the number of intervallic units. For example, int5 and int6 are the minor and major thirds of this system, respectively, and int8 and int11 are the fourth and the fifth.

2. METHOD

2.1. Subjects

43 subjects aged between 19 and 54 (M = 26.6, SD = 7.0) participated in the experiment. 32 of them were students of musicology at the University of Turku, Finland. Others were professional musicians or people having music as their hobby. On average, they had been actively engaged in making music for

16.4 years (SD = 8.2), and they had received instruction in music theory or ear training for an average of 8.7 years (SD = 8.9). Five of the subjects reported having absolute pitch.

2.2. Apparatus

Both the generation of the stimuli and the experiment itself were carried out using a Power Macintosh G4 computer. The stimuli were generated by the sound synthesis software program csound, and the experiment was conducted using the experimetal laboratory software SuperLab. The signals were played through Sony MDR-P180 headphones.

2.3. Stimulus materials

The experiment used a two-alternative forced choice discrimination paradigm. Each trial consisted of a pair of melodic intervals (X–Y) in the upward registral direction, both preceded by a lower context tone (L). The context tones were composed of the first 16 partials of a sawtooth wave, and they had a duration of 1 s, including a (linear) growth of 100 ms and a similar decay. The upper tones in the melodic intervals were pure tones of a 200 ms duration, including growth (10 ms) and decay (10 ms). Each trial had a total duration of 9.8 s, consisting of the following temporal sequence: L, pause (3 s), X1–Y, pause (1 s), L, pause (3 s), X2–Y. The only difference in the two repeats of the sequence (L, pause, X–Y) was that the lower tone of the second melodic interval, X2, could be shifted up or down with respect to X1, or it could remain at the same pitch.

In the construction of the trials, the tone L was always taken to represent pitch-class 0. The size of the melodic interval X1–Y1 varied from 4 to 8 units of 19-ET, giving five *basic intervals* of 252.6, 315.8, 379.0, 442.1, and 505.3 cents. The pitch shifts for the lower tone of the melodic interval always corresponded to the smallest intervallic unit of 19-ET, which is 63.2 cents. For a basic interval of 5 scale steps (315.8 cents), X2–Y might thus be changed to a 4-step interval (252.6 cents) or a 6-step interval (379.0 cents). The transposition level of the upper tone Y, or *Y-level*, was systematically varied between pitch-classes 0–18, so that the pitch distance between L and Y was three octaves plus 0–18 units of 19-ET. This arrangement gave a 5 (basic intervals) x 3 (pitch shift conditions) x 19 (Y-levels) design, resulting in 285 trials for each subject.

For presentation in the experiment, pitch-class 0 was fixed to C of standard tuning (C4 being 261.6 Hz), and the pitch of tone L was variously transposed for different trials. One set of transpositions was created by associating with each Y-level one of the 19 possible transpositions within a one octave range (Y-levels 0–18 corresponding to transpositions of 0, 2, 4, 6, 8, 10, –7, –5, –3, –1, 1, 3, 5, 7, 9, 11, –6, –4, and –2 units of 19–ET, respectively). Taking as a starting point for transpositions of tone L the pitch C2 (so that the fundamental frequency of the complex tone was 65.4 Hz), L thus varied between what would be a slightly sharp G1 and a G2. In order to balance out possible effects of transposition, a second set of transpositions was created, in which each trial was in a tritone relationship to the corresponding trial in the first set. Here, L was kept within the registral range of G1–G2 by transposing it either 9 units up or 10 units down. 23 of the subjects heard all stimuli in the first set of transpositions, and 20 heard them in the second set of transpositions.

2.4. Procedure

The 285 trials were divided in a quasi-random manner to 8 stimulus blocks. Within each block, the order of trials was randomized. All subjects were tested individually in two sessions, A and B, using blocks 1–4 in session A and blocks 5–8 in session B. In order to balance out any effects of novelty and fatigue, the presentation order of the randomized blocks was shuffled by allowing the subjects to choose their starting block from a list. The other three blocks were then run in numerical order, jumping from block 4 back to block 1 (*e.g.*, starting from A3, and then A4, A1, A2), and the second session used a corresponding order. The number of subjects starting from each block was thus 6, 6, 7, 8, 3, 3, 5, and 5 for blocks 1–8. Both of the two sessions included three short breaks during which the subjects saved their performances on the computer. The duration of each session was approximately 45 minutes.

Subjects received written instructions so that they were able to carry out the experiment alone with the computer whenever they wanted, with the restriction that the two sessions, A and B, would have to be scheduled for separate days. The task of the subjects was to listen to each pair of melodic intervals and to press the "1" key if they thought that the two intervals sounded exactly alike, and the "0" key if they thought that there was any difference whatsoever between them. They were told that the "bass tone" should be ignored. There were four practice trials, during which the subjects could and to adjust the volume to a comfortable listening level.

3. RESULTS

The analysis of the results was begun by counting the proportion of hits (correctly detected changes in the melodic interval) separately for upward and downward changes, for each combination of basic interval and Y-level. The proportion of false alarms (incorrectly reported changes) was similarly determined using only the trials with no change. For each combination of basic interval and Y-level, the two hit rates were used separately, together with the false alarm rate, to determine d' scores according to signal detection theory (see Wickens, 2002). This transformation has the effect of eliminating response bias. In the present context, only some of the more general results will be reported. I will concentrate on differences in performance due to the basic intervals, and on the effects related to Y-levels, which was the main focus of the study.

3.1. Detection of the five basic intervals

Disregarding the effect of the context tones for a moment, the results concerning the detectability of the five basic intervals were not surprising. Table 1 shows the mean d' scores for each basic interval, averaged over the separate d' scores for 38 different combinations of Y-levels and registral directions. As one would expect, the d' scores for the non-diatonic int4 and int7 are lower than those for the diatonic intervals. Interestingly, the minor third (int5) has a larger score than the major third (int6), which could, perhaps, be understood by taking into account the relative closeness of int5 to its pure equivalent.

int	mean d' (all)	sd (all)	mean d' (up)	mean d' (down)	t-test
4	0.87	0.49	1.24	0.50	t = 7.36 (p < 0.001)
5	1.40	0.42	1.53	1.26	t = 2.06 (p = 0.046)
6	1.12	0.34	1.14	1.10	t = 0.38 (p = 0.706)
7	0.89	0.39	0.76	1.02	t = –2.12 (p = 0.040)
8	1.36	0.47	1.09	1.64	t = –4.40 (p < 0.001)

Table 1: d' scores for the basic intervals int4–int8.

The table also shows d' scores that were averaged over the 19 Y-levels separately for the downward and upward changes. The differences in d' between upward and downward conditions were shown by two sample t-tests to become increasingly significant towards the ends of the range of interval sizes (see the last column). However, whereas the upward shifts of the lower tone X were more readily detected in the case of the smaller basic intervals, for larger basic intervals it was the detection of downward shifts that was facilitated. In other words, making the small intervals smaller seems to trigger category change more easily than making them larger, and the converse is true of the larger basic intervals. Such effects cannot be explained with reference to consonance, because neither type of change was more significantly detected for int6, even though its larger neighbor, int7, does not, in theory, correspond to any familiar consonant interval. Thus it seems that for the five basic intervals used in the experiment, relative detectability of upward and downward changes varies systematically with size. Alternatively, one might hypothesize that thirds in general and major thirds in particular represent characteristic sounds that are especially well recognized by Western listeners. Given that int7 may have been accepted as a type of third (which is not implausible for melodic intervals composed of pure tones), int6 becomes the category center for the larger category of thirds. If this is correct, we can say that, in general, changes away from the paradigmatic major third were detected more easily than changes towards it.

3.2. The effect of the context tone

Differences in the d' scores due to various Y-levels (pitch-class relationships between L and Y) are shown in figure 1, in which the context tone L has been taken as pitch-class 0 (ignoring the transpositions in the experiment). The figure plots the mean d' scores for each of the 19 pitch-class levels of X1, X2, and Y, averaged over the five basic intervals and (in the case of Y) over the upward and downward changes. It seems that the context tone did have some systematic influence on the ease of detecting pitch shifts. Looking at the curves for both of the pitches X1 and X2, we may firstly see that performance improves from the area around pitch-class 0 to the area around pitch-class 5. It then gradually falls off again towards the upper part of the octave. This seems to indicate that on the side "above the tonal center" there is a marked "third area" in relation to the context tone: an area where tonally meaningful differentiations are easy to establish. Secondly, if we assume that the subjects have indeed tended to interpret the context tone as a tonal center, it is remarkable that they have been rather flexible in making categorical distictions that have involved the pitch-class equivalents of this context tone (see the low d' score for X2 at pitch-class 0). The general picture that we get is that categorical distinctions between closely similar intervals are facilitated when the moving pitches are located within a suitable distance above the contextually emphasized tone (on the pitch-class circle).

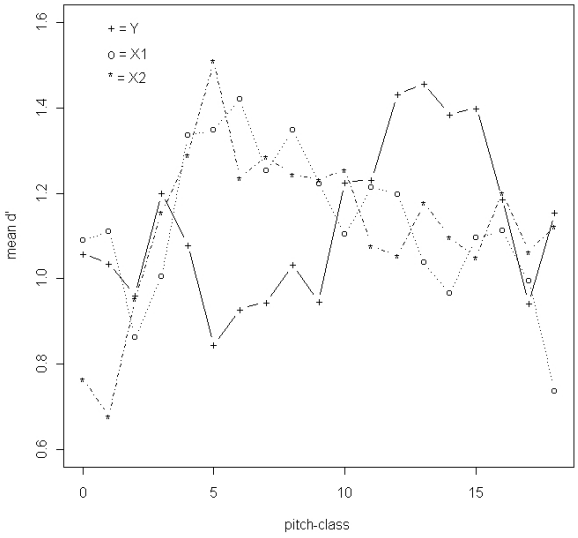


Figure 1: Mean d' as a function of the pitch-class of the upper tone Y and the pitch-class of the lower tones X1 and X2, when the context tone is taken to represent pitch-class 0.

The curves for X1 and X2 are reflected in the somewhat opposite tendencies observed for the upper pitch Y. The large values of d' at pitch-classes 12–15 are an effect of the interval sizes 4 through 8 that were used for the experiment: if performance was good when X1 was located in the third area, it was also good when Y was in the "sixth area". The converse dependence is of course logically possible, but it seems wiser to construct explanations which favor that range on the pitch-class circle in which the actual pitch shifts were located.

Despite the promising outlook of figure 1, an ANOVA indicated that the overall influence of the pitch-class relationships with respect to the context tone failed to be significant. This was true for the pitches X1 [$F(18,171) = 1.395$, $p = 0.139$] and X2 [$F(18,171) = 1.647$, $p = 0.054$], as well as for Y [$F(18,171)$, $p = 1.649$]. For some conditions, theoretical significance is reached, however, when we look at the upward and downward changes separately. This seems to be because the different context tone effects for the two registral conditions have cancelled each other out for some particular pitch-class relationships. Thus, the influence of the pitch-class relationship between L and X1 (and between L and X2) was, indeed, more significant for upward shifts [$F(18,76) = 1.752$, $p = 0.048$], although not for downward shifts [$F(18,76] = 1.051$, $p = 0.416$]. Similarly, the relationship between L and Y had a significant

influence only for the upward shifts [$F(18,76) = 2.009$, $p = 0.019$].

As an example, figure 2 shows the case of X1. The white circles connected by the solid line represent the detectability of downward shifts from each pitch-class, when pitch-class 0 once again is taken to correspond to the context tone. For this line, large values of d' thus indicate a relatively sharp category boundary on the lower side of the corresponding pitch-class. Conversely, the "+" signs stand for the detectability of upward movements originating from a given pitch-class, and large d' values now indicate a category boundary closely above the pitch-class. The figure appears to reveal areas of heightened perceptual acuity that correspond to important tonal relationships. The relatively high scores between pitch-classes 4 and 8 conforms well to ideas concerning the difference between the two scale-steps for minor and major thirds (the difference between pitch-classes 5 and 6), their separation from their immediate surroundings (4–5 and 6–7), and the differentiation of this "third area" from the "fourth area" at pitch-class 8.

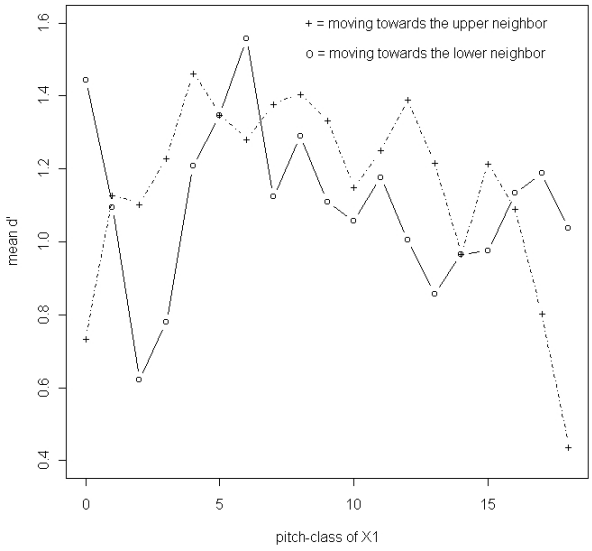


Figure 2: Mean d' as a function of the pitch-class of X1, shown separately for upward and downward shifts between X1 and X2 (pitch-class 0 corresponds to the context tone).

Although the overall difference between the above curves is not significant (t-test, $p > 0.5$), the d' scores for some particular pitch-class relationships appear to have been strongly dependent on the registral direction of the pitch shifts. Quite surprisingly, performance has been worst for pitch-class 18 moving upwards to 0, which can only be understood by saying that tones corresponding to pitch-class 18 have been interpreted as pitch-class equivalents of the context tone L. The most interesting phenomenon that becomes evident in figure 3 is, indeed, the flexibility of the areas around pitch-classes 0 and 11 (the context tone and its upper fifth). In both cases, the d' scores corresponding to upward shifts show a rising tendency from the lower neighbor of pitch-class 0 or 11 through this pitch-class itself to its upper neighbor, and in the downward shifts there is local peak of d' scores at 0 and 11, the two upper neighbors having increasingly lower scores. These observations can be interpreted by introducing the common-sense idea that various scale-steps may have more or less central representatives. On hearing X1, the listeners may often have interpreted even less representative members in terms of the category, and in such circumstances, subsequent "corrections" at X2, while theoretically "better in tune", have been difficult to detect. By contrast, when a more central member of the category has been hit first, deviations from it have been relatively easier to detect. What is especially interesting is that in downward intervallic motion, the locally highest d' scores for both the "tonal center" area and the "fifth" area do appear at pitch-classes 0 and 11, respectively, but that in upward motion the highest local peaks are situated at the nearest upper neighbor. The lower category boundaries thus seem to be closer than the higher ones to the theoretically best representatives of both of these categories.

4. CONCLUSIONS

A musical context may affect the categorical perception of melodic intervals in different ways according to the tonal relationships between the actual interval to be judged and the context tones. In the present experiment, the context tones were separated from the the stimulus interval by pauses of 3 s (and 1 s between the two presentations), but even then, the categorization of the stimulus interval (X–Y) was affected by the categorization of other intervals between the context tones and the tones of the stimulus intervals (L–X, L–Y). The results for both types of categorizations were partly asymmetrical: changes between a central and a less central category member were easier to detect when the central member was presented first. These findings are in line with other related results concerning the asymmetry of similarity ratings in tonal contexts (*e.g.*, Krumhansl, 1979).

5. REFERENCES

[1] Burns, E., "Intervals, Scales, and Tuning", In D. Deutsch (Ed.), The Psychology of Music, 2nd ed., Academic Press, San Diego, pp. 215-264, 1999.

[2] Huovinen, E., Pitch-Class Constellations. Studies in the Perception of Tonal Centricity (Acta Musicologica Fennica 23), Finnish Musicological Society, Turku, 2002.

[3] Krumhansl, C. L., "The Psychological Representation of Musical Pitch in a Tonal Context", Cognitive Psychology, 11:346-374.

[4] Krumhansl, C. L., Cognitive Foundations of Musical Pitch. Oxford University Press, New York, 1990.

[5] Mandelbaum, J., Multiple Division of the Octave and the Tonal Resources of 19-Tone Temperament, Unpublished doctoral dissertation, Indiana University, 1961.

[6] Tsuzaki, M., "Effects of the Preceding Scale on Melodic Interval Judgment in Terms of Equality and Size, Music Perception, 9(1): 47-70, 1991.

[7] Wapnick, J., Bourassa, G., and Sampson, J., The Perception of Tonal Intervals in Isolation and in Melodic Context, Psychomusicology, 2(1): 21-37, 1982.

[8] Wickens, T. D., Elementary Signal Detection Theory, Oxford University Press, New York, 2002.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

TOWARDS A VIRTUAL SINGER : CONTRIBUTION TO THE DEVELOPMENT OF THE TEACHING OF SUNG DICTION THROUGH THE INTERMEDIARY OF A FUNCTIONALISED SYSTEM

Monique Nahas, Hervé Huitric, Jocelyne Kiss

Department of Art and Technologies of Image, INREV
University Paris 8, France..
mnahas@worldonline.fr,KissJocelyne@aol.com

ABSTRACT

Learning to sing, whether within the construction of the physical vocal pattern, or the interpretive preparation of a piece of music, proceeds initially through observation and then imitation of a given physical model. One of the principal difficulties in this mimetic process frequently lies in the different perceptions [1] of the singer and the listener. Without attempting to compensate directly for this problem, we wish to propose an interactive system that reacts to exterior data, thus facilitating the creation of a model of the singer and at the same time real time simulation of his acts.

We are offering, within the framework of this undertaking, to create a basic virtual animation. It is important at this stage to define the nature of the work envisaged, which will not consist of an exhaustive project. We intend more specifically to describe the means for encoding the vocal expression and its relation to the facial expression in order to translate a feeling.

We will then proceed to the development of a computer generated synthetic singer, adopting as our method the use of a process of interactive phonation represented by a synthetic actor with simplified expressions. Initially this phonation will be limited to the expression of a small number of phonemes and that this part of the process could become the basis of further concrete developments and constitute the foundation for various potential extensions.

1. INTRODUCTION

The act of singing has been described by many performers [2] as a juxtaposition of "internal physical sensations". According to R. Miller [3] these enable the performer, to a certain extent, to escape from the problem of listening to one's self during the phonation. The majority of beginners are frequently doubtful concerning the idea that they "don't hear themselves". The teaching of singing is therefore based on the teaching of these sensations that allow accurate production of sound.

The formalisation of these impressions implies recourse to a metaphorical and visual representation in a model of the mistakes made by the student and also of the ideal example. The principal difficulty in this mimetic exercise often resides in morphological differences between model and student. The project of using an interactive synthetic computer-generated singer does not represent an alternative for the singing coach but does provide a systematic teaching aid that acts directly with the image and the voice of the student. The creation of a synthetic computer-generated singer requires a restrictive definition of the action of the singer, since this is strictly influenced by aesthetic criteria and differs according to cultural practice.

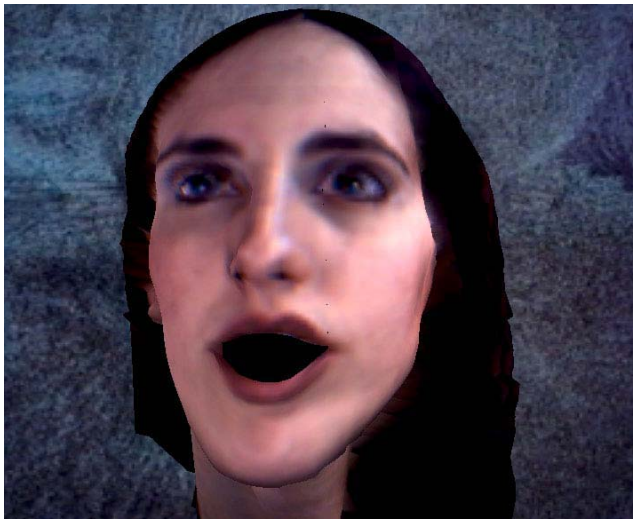


Figure 1: *Example of virtual singer's animation create by Monique Nahas and Hervé Huitric.*

We will therefore limit ourselves solely to the Western practice of Italian opera that is also referred to as *Bel Canto*. Of course, we could equally attempt to define the act of singing by adopting the viewpoint of the physiologist and taking.as our point of departure the observations of the phoniatrician [4] on this subject. However, it seems to us that this would serve to obscure the extremely expressive character of the singing act, and we refer therefore to the famous definition of the singer given by Charles Panzera: "*The singer presents the particularity that his voice testifies to the reciprocal action of the body and the mind.…The singer is his own instrument, vocal technique is his base."* [5].

This project also necessitates a certain degree of caution concerning the use of the term "synthetic". Clearly, by synthetic image we are referring to an image generated by a computer, either directly, or through the manipulation of a previously produced image using some kind of peripheral. In using the term interactivity what we envisage is *"a type of relation between two systems which results in the behaviour of one system modifying the behaviour of the other"* [6].

The creation of a synthetic computer-generated singer equally necessitates a choice concerning the methodological direction to adopt [7], which in itself leads to an initial problem of representation. Firstly, we intend to produce a comparative study intended to establish a taxonomic classification of the relationships between facial expressions and the production of sounds.

Next, we will encode each of these sound-image associations according to aesthetic criteria linked to the teaching of singing. Then we will attempt to put into perspective the various possibilities created by this project.

2. ENCODING OF THE VOCAL EXPRESSION

2.1. Creation of the automaton

If we consider the computer modelling of a synthetic automaton reproducing the act of singing as a mediatory function between the concrete act of vocalisation and another, more abstract act, forming itself simultaneously from a symbolic vocabulary and from ideal objects, it would then appear that this constitutes an effective means of arriving at an understanding of certain processes pertaining to the structure of the singing act and to artistic expression[8].

The principal difficulty in such a system resides in the balancing of the shares that must be accorded to the specific inherent characteristics of this act/gesture and the necessary generalisation. Thus we can identify an initial problem concerning the orientation of the computer-generated singer we wish to create, since it would be, in effect, utopian, to envisage at the present stage of research a complete system capable of reproducing total interactivity capable of being used in the perfect staging of a particular context, the interpretation of an Aria, for example. The initial presentation of the synthetic automaton resides therefore in a series of behaviour activated by the actions of the user.

2.2. The different conceptual orientations

If, in an ideal conception, we find ourselves in the presence of a source of sound and visual material producing events as varied as possible – this may be concretely envisaged as a library [9] of sung phonemes linked to appropriate facial expressions – and we opt for automation of the choice of vocal expressions, we can then, in this situation, have recourse to three types of "filter", which are complementary in nature.

The first "filter" could be similar to a collection of symbolic units [10] a corpus of rules resulting from a thorough analysis of the sound-image-facial expression dependency, may lead to a systematic adaptation of an aesthetic context in terms of stage or theatre interpretation. An "interesting" automaton will naturally possess an extensible palette of options, of possibilities for decisions, of prepositional parameters, allowing wide and flexible possibilities for the creation of syntactic articulations.

It must be noted that even an exhaustive set of rules cannot reasonably be expected to make explicit the richness of a style, and that their exclusive use presents the danger of monotony, and even of predictability [11].

Consequently, the second will essentially be stochastic in nature, that is to say that it will contain decisional "filters" based on the laws of probability and the frequency of appearance. Thus, certain stochastic laws may require that the automaton makes infinitesimal variations around a particular expression, in the canonical pronunciation of the sounds, in the frequency and velocity of the timbres produced, etc and we can imagine that, in a general manner, these random constraints could intervene at different levels of theatrical organisation, for example in the choice of rules of coordination or of their parameters.

This complexity is due to the very nature of the recursive character of a system of decisions operating on another. Nevertheless, at this stage, even if the random modules result in unexpected changes in the sequence of facial expressions, sounds and events, the production thus automated, even in relation to certain symbolic units, will not correspond, except extremely fleetingly, to the particular hopes and expectations within the framework of an artistic creation [12].

With this in mind, in order to improve our automaton we will therefore have recourse to a third type of "filter", capable of integrating the stylistic characteristics inherent in a work, an author, a period or a genre, in order to enable it to accept or reject the sound sequences presented, outside of any clearly formulated analytic framework, according to an act not based on simple contingency. We can see then, that this decisional level [13] may be introduced with the help of the notion of interactivity [14], which can be constructed from the starting point of deterministic controls, or bringing random corrections.

It is equally important to note that this pragmatic conception suggests a modular construction, which presents a number of advantages. The most important of these is undoubtedly linked to current advances in robotic technology, which favour the resolution of a complex task through numerous independent units rather than through one overall calculation centre controlling all the necessary functions.

We will not envisage the production of an optimal synthetic computer-generated singer whose creation is beyond the limits of this project. What we are proposing here is the creation of a basic example. We will leave to one side random and connectionist interactivity, and concentrate on the creation of the basis, which, as we have explained, will be capable of extension and generalisation.

3. THE VOCALITY BETWEEN IMAGES AND SOUNDS

The creation of a computer-generated synthetic singer proceeds from a process of mimicry [15] to a greater or lesser degree stimulated by the real world. It would appear that there is a second fundamental problem to be considered, this one relating to the pertinence of the elements or parameters to be taken into account in the modelling, parameters that relate not only to the descriptive field, but also to those that are the concern of expressiveness, or the aesthetic intention. The putting in place of a simplified elementary typology stemming from the structure of song with a view to constituting a collection of initial considerations appropriate for a taxonomy of the modelling of a virtual singer[16].

The aim of this section is not so much to exhaustively study the multiple modalities of the phenomenon of the emission of sound; rather it is to develop an initial methodology capable of discovering useful extension within the virtual conception in the domain of lyrical song. We will now describe the study of three types of sound attack, as an illustrative example of the correspondence of sound and facial expression.

There are three types of attack, or commencement of vocal emission, corresponding to specific positions of the vocal chords. We normally distinguish: the soft attack, the hard attack, and the balanced attack. Each corresponds to a specific modelling [17].

3.1 The hard attack

In two varieties, glottal attack and glottal occlusion, it is a term characterising a vocal sound defined by the adduction or the tightening of the vocal chords prior to phonation. In this case the activity of the vocal muscle begins in force before the sound emission and is distinguished by its power compared to the other two fundamental types of attack.

This is because the glottis closes completely and then opens suddenly as the sound is produced. This effort is accompanied by specific facial expressions although control on the part of the part of the singer can reduce these effects. The "hard" attack provokes a kind of brief congestion or even strangulation in the neck (often more visible in a male subject because of the presence of the Adam's apple) resulting in the opening of the mouth (also occasionally the forward projection of the chin) allowing the briefly stifled sound to take its place in the oral cavity and emerge. This action also produces a slight backwards movement of the head.

3.2 The soft attack

Conversely, the soft attack [18] is accompanied by a respiratory flow before the emission of the sound as if the singer were preceding it with an aspirated introduction. It corresponds, especially when it is temporally drawn out, to a coming together of the vocal chords along the length of their paramedian line without obturation of the glottis.

One sometimes observes a low level of sound emission accompanied by a powerful flow of breath. As in the hard attack, the sound emission is accompanied by characteristic hypertonic movement of certain facial muscles that the singer can attempt to control in order7 to reduce the visible signs. One generally notices a tendency to exaggerate the respiratory action, forcing the singer to draw in the chest and tighten the muscles of the neck. The balanced attack corresponds to an intermediate activity between the extreme positions of the "whisper" and the "grunt". It consists of the balanced working of the larynx controlling the sound accumulations during the course of a pronounced phrase. It requires phonatory "calibration" to act rapidly in the course of a song or a discourse and this aspect concerns not only the commencement of the vocal emission but also continues throughout the declamation.

This phonatory control, which is similar to a "pre-tuning" of the activity of the emission, depends essentially on the correct management of air before the sound attack and to a lesser degree on the musculature of the larynx, which constitutes the principal voluntary contribution to its action during a sung or spoken phonation [19].

2.3 The balanced attack

The balanced attack corresponds to an ideal commencement of sound emission, in which each muscle of the neck relax as if the subject were yawning. As the speech therapist B.D Wyke describes: "This process of pre-phonatory tuning involves not only the participation of the intrinsic muscles of the larynx, but also the participation of the intercostals and abdominal muscles, as well as the external laryngeal muscles. It goes into action immediately after every voluntary inspiration taken between the phrases". [20] Schematically speaking, the balanced attack is situated at an intermediate stage between the soft and hard attack.

3. CONCLUSIONS AND PERSPECTIVES

The fulfilment of this elementary project can be viewed as the starting point of a system to aid voice coaching and the teaching of singing, but it could equally be directed towards the production of interactive scenes involving virtual singers.

The analysis of facial expressions and the vocalic formants could lead to extremely realistic expressiveness or to sophisticated artistic effects. A far-reaching project could be based on the fundamental principles of the structure of song in order to develop complex scenography aimed towards interactive operatic creation [21].

Another interesting direction flowing directly from this project concerns the domain of interactivity. We have seen just how tedious it can be using a peripheral such as a keyboard to animate ambitious or complex interactions. It is also interesting to consider the possibility of extended interaction between the user and one or more virtual singers: this situation would undoubtedly find an application in video games [22], whether for educational or purely entertainment purposes, but this construction can also be seen from an instrumental point of view. Through the intermediary of sophisticated software the

user would be enabled to sing 'virtually' in the manner of a genuine instrumentalist.

Lastly, as we explained earlier, just as this type of virtual performance has helped to experiment with certain artistic ideas in choreography, this interactivity could also provide useful tools in the sphere of operatic creation. The necessary condition for development in this direction remains the development or practicable, easy to use interactive software. It appears that the connectionist orientation may in the medium term provide a seductive solution to this technical difficulty: the user would directly integrate with the performance, either thought the use of his own voice, or through his facial expressions. One could even envisage a mixed system correlating the two types of approach.

4. REFERENCES

[1] De Bonis M., Nahas M., Chimeric emotional faces : a tool for investigating facial perception of emotions. Smiling mouth and fearful eyes. 8th European Conference, Facial expression of emotion and meaning, Sarrebruck, Germany. 1999.

[2] Reid C. Bel Canto : Principles and Pratices. NewYork. Coleman-Ross. 1950 ; and Christy (van) A. Expressive Singing. 3 éd. Dubuque IA. wm. C. Brown. 2 Vols. 1974.

[3] Miller R. et Shutte H. The effets of Tongue position on spectra in singing. The NATS Bulletin 37, 3 : 26-27.34.1981

[4] Ladefoged P. Elements of acoustics Phonetics. Chicago : The University Press. 1962.

[5] Panzera C. L'Art de Chanter. Editions Costallat. Paris 1957. p.7.

[6] Barda J., Dusantes O., Notaise J., in, Dictionnaire du multimédia, AFNOR. 1995

[7] Huitric, H., Nahas, M., Rioux M., Domey J., 1990 "Facial Image synthesis using skin texture recordings", The Visual Computer,6, 613-626.

[8] Scotto Di Carlo, N., et Guaïtella, I., Facial Expressions in Singing : A pilot study, Communication au XIII° Congrès International des Sciences Phonétiques. Stockholm, 1995, I.

[9] Peterfalvi J-M., Recherches expérimentales sur le symbolisme phonétique. Paris. Centre national de recherche scientifique. 1970.

[10] Haugland, J., Mind Design. MIT Press. 1981. and alsoVarela F.J., Connaître les sciences cognitives. Trad. P. Lavoie. Paris. Seuil. 1989. p.35-42

[11] Kohonen T., Laine P., Tiits, Torkkola K., A Nonheuristic Automatic Composing method, in, Music and Connectionnism. Massachusetts Institute of Technology. 1992.

[12] Scherer, K.R., & Kappas, A. Primate vocal expression of affective state. Dans D.Todt, P.Goedeking, & D. Symmes (Eds.), Primate vocal communication (pp. 171 - 194). Berlin: Springer-Verlag.1988.

[13] Simon, H.A, Sciences de l'artificiel. 2ème édition 1981. Cambridge Mass. the MIT Press. French traduction Sciences des systèmes, sciences de l'artificiel. Paris. Dunod. 1991.

[14] Angel, E. Interactive Computer Graphics. Addison-Wesley. 2000.

[15] Scotto Di Carlo, N., Analyse sémiologique des gestes et mimiques des chanteurs d'opéra, Semiotica,1973, IX,

[16] Ekman, P., And Friesen, W.V., Manual for the Facial Action Coding System, Palo Alto: Consulting Psychologists Press.1977

[17] Kappas, A., Hess, & Scherer, K.R. Voice and Emotion. in, .R. Feldman & B. Rimé(Eds.), Fundamentals of nonverbal behavior. p. 200-238. New York : Cambridge University Press. 1991

[18] Weede, R. Intelligent care of the singing voice. The Etude December. 65 : 679.

[19] Cohn, M.A., Wertz, J.M, Perrott, M. A., And Parrott, D.J., A Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression. Journal of Nonverbal Behavior, 25, p.167-186. University of Pittsburgh

[20] Wyke, B.D. Laryngeal neuromuscular control systems in singing : a review of current concepts. Folia Phoniatrica 26,1.1974

[21] Paquelin, D., Conception d'un environnement d'apprentissage interactif en fonction des attentes des usages, Université d'Avignon et des pays de Vaucluse, Avignon, 1999.

[22] Pecchinenda, A., Kappas, A., & Smith, C.A., Effects of difficulty and ability in a dual-task video game paradigm on attention, physiological responses, performance, and emotion-related appraisal. Présenté au Thirty-Seventh Annual Meeting of the Society for Psychophysiological Research, Cape Cod, Massachusetts. et culturelle. L. 1997

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

MUSICAL SOUND PARAMETERS REVISITED

B. Kostek, P. Zwan, M. Dziubinski

Department of Sound & Vision Engineering
Gdansk University of Technology

kido@sound.eti.pg.gda.pl

ABSTRACT

Recently, a new standard MPEG7 was established. A set of parameters was defined in order to represent musical sound as a multimedia object. These so-called low level features are related to time and frequency domain of musical sounds, as well as to audio waveform parameters. Among others, the following parameters were specified within the standard: log attack time, temporal centroid, spectral flatness and spectrum spread, spectral centroid, harmonic variation, etc. Some of these parameters are related to human cognition of musical sounds whereas others not or the relationship between those parameters and their perceptual meaning is not yet defined. Therefore the principal aim of this paper is to carry on a discussion as to which of the parameters defined within MPEG7 standard could be related to the musical timbre. This is done by means of listening tests. Additionally, the set of parameters is used in experiments consisting in automatic recognition of musical instruments sound and separation of musical duets. Results of experiments are described and conclusions drawn and included.

1. INTRODUCTION

In the recently standardized MPEG7 it is defined that audio signals can be described at different abstraction levels: from the lowest level (primitives - such as for example: temporal or audio spectrum centroids, spectrum flatness, spectrum spread, inharmonicity, etc.) to the highest level (related to semantic information) [1][2]. In this context semantic information is related to textual information on audio such as titles of songs, composers, etc. On the other hand, there is a way to change primitives to higher abstraction level, namely semantics. This means that, for example, *Brightness* can be expressed either in numbers (spectral centroid) or in words (bright, dark, etc.). Therefore there is a need to study a relationship between objectively extracted information from sound and subjective notions of timbre.

The notion of musical sound timbre refers to work done by Grey [3] and later by McAdams and Winsberg [4], Wessel [5], and many others [6][7][8][9][10][11], only a few listed here. In most papers dealing with perceptual space it may be seen that researchers tried to limit this space to three or two parameters. It is obvious that in such a case dimensions can be easily interpreted and presented as a two- or three-dimensional projection, however, in authors' opinion derived from the results of processing multidimensional feature vectors describing musical sound characteristics, there is no need to limit the perceptual space to such a small number [12]. Computer processing can easily deal with multidimensionality of feature vectors [8][12][13][14]. Moreover, such processing is now highly needed for automatic queries within sound digital archives. More dimensions can help to distinguish between particular instruments or musical instrument groups.

The paper includes the results of experiments aiming at testing the influence of signal representation description on the effectiveness of music identification. The parameterization process used features of musical sounds contained in MPEG-7 [1], as well as features based on wavelet analysis of musical sound transients [13]. It should be remembered that parameters contained in the MPEG-7 standard were selected arbitrarily, hence the need to test the quality of the parameters in a process of automatic identification of musical instrument classes. The experiments will be based on a decision-making system employing learning algorithms.

2. FEATURE VECTORS

2.1. Parametrization

A multimedia database created at the Department of Sound and Vision at the Gdansk University of Technology [12] was used in the feature vector effectiveness analysis. Musical instruments were selected randomly, they were such as: clarinet (*staccato*), trumpet (*staccato*), violin (*forte nlegato*), sax (alto), (*staccato*), trombone (tenor), (*staccato*), cello (*forte nlegato*). Whole register of each instrument was considered in the analysis.

A so-called Spectrum Peak Analysis (SPA) was applied for the pitch detection purpose [15]. Estimating pitch contour was performed in block processing, i.e., signal was divided into blocks with equal widths, whereas overlap could be time varying. Each block was weighted by the Hann window. The algorithm was supported by a Neural Network, and in the post-processing phase a time domain pitch contour correction was applied. The effectiveness of the implemented pitch detection algorithm was very high for all instruments (regardless of articulation and dynamics) and no octave errors occurred in the process of fundamental frequency estimation (see for example Fig. 1).

The parametrization process started from the calculation of the MPEG-7-based descriptors. They were such as follows:

- *Brightness*;
- *Contents of odd harmonics*;
- *Harmonic Spectral Spread*;
- *Harmonic Spectral Deviation Descriptor*;
- *Harmonic Spectral Centroid Descriptor*;
- *Audio Spectral Flatness* - spectral Power Density Function calculated on the basis of the *N*-point FFT algorithm;

- *Attack Time*, defined as the time during which the envelope changes its value from 0.1 to 0.9 of the maximum value;
- *Temporal Centroid Descriptor.*

To the listed above parameters a fundamental frequency was added, thus the format of the MPEG-7-based feature vector consisted of 9 parameters.

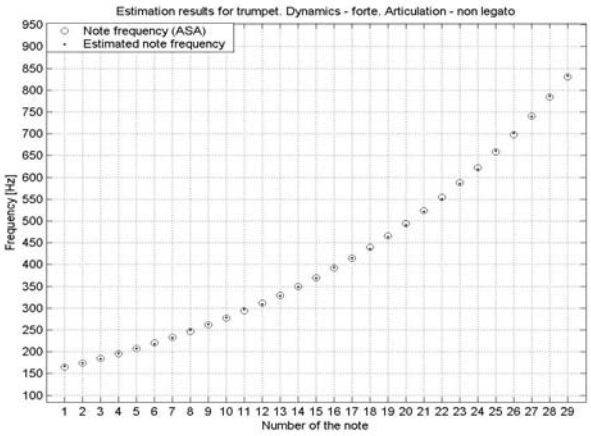


Figure 1*: Pitch estimation results for trumpet (articulation: non legato, dynamics: forte, range: E3 - G5#).*

Because most MPEG-7-based parameters were calculated for the steady state of sound, the decision was made to enhance the musical sound description by adding descriptors based on the transient phase of sound. Earlier research on wavelet parameters for the transient state of musical sound Department of Sound and Vision allowed for easy calculation of such parameters [13]. In the second case, feature vectors (FVs) consisted of 8, both energy- and time-based, wavelet parameters, and fundamental frequency. The third type of feature vector encompasses both MPEG-7 and wavelet-based parameters.

2.1. Testing Feature Vector Effectiveness

To test feature vectors for their usefulness for musical instruments classification process, an artificial neural network algorithm was used (ANN). The neural network had three layers with adequate: 9, 9 or 17 neurons in the input layer, 14 neurons in the hidden layer and 6 neurons in the output layer. Feature vectors were fed to the network input. Sounds were divided into two groups of the same size. The first one was a training set for the neural network, while the remaining feature vectors were used in testing. Activation functions were unipolar, and the neurons at their outputs had values from 0 to 1. In the classification process, the winning neuron had to be determined, i.e. one whose output value was the highest. The mean square error was calculated during the training process and this process was stopped when the error would continue to increase in 5 consecutive training cycles. It should be remembered that such a procedure is heuristic and algorithm settings were established basing on carried out experiments.

Tab. 1 shows recognition effectiveness of the MPEG-7-based descriptors in [%]. Each line in the table stands for the recognized input object, and the columns show the ANN recognition effectiveness. The highest score obtained for a particular instrument are given by the table diagonal (correct recognition of the class of the instrument). As seen from Table 1, apart from cello sounds, there were some problems with proper recognition of instrument class. Especially, it happened for trumpet and saxophone sounds. These instrument sounds were confused with each other.

Table 1: Recognition effectiveness of the MPEG-7-based FVs

[%]	clarinet	trumpet	violin	sax	trombone	cello
Clarinet	**84.2**	5.3	5.3	0.0	5.3	0.0
Trumpet	0.0	**53.8**	0.0	46.2	0.0	0.0
Violin	6.3	0.0	**93.8**	0.0	0.0	0.0
Sax	0.0	23.1	0.0	**46.2**	30.8	0.0
trombone	0.0	17.6	0.0	0.0	**82.4**	0.0
cello	0.0	0.0	0.0	0.0	0.0	**100.0**

Table 2 presents results of testing the effectiveness of the wavelet-based feature vectors. It can be observed that musical instrument sounds were properly classified for most instrument classes. There were some errors in recognition of trombone, violin and sax sounds, but in overall better results were obtained than in the case of the MPEG-7-based feature vectors.

The combined FFT and wavelet representation ensures very good recognition results for each group of instruments (see Table 3). This is because the parameterization process covers both the transient and steady state of sound, and the resulting features of the feature vector complement each other.

Table 2: Recognition effectiveness of the wavelet-based FVs

[%]	clarinet	trumpet	violin	sax	trombone	cello
clarinet	**94.7**	0.0	0.0	0.0	5.3	0.0
trumpet	0.0	**100.0**	0.0	0.0	0.0	0.0
violin	6.3	0.0	**87.5**	0.0	0.0	12.5
sax	7.7	15.4	0.0	**76.9**	0.0	0.0
trombone	0.0	0.0	0.0	5.9	**88.2**	5.9
cello	0.0	0.0	10.0	0.0	0.0	**90.0**

Table 3: Recognition effectiveness of the MPEG-7 and wavelet-based FVs

[%]	clarinet	trumpet	violin	sax	trombone	cello
clarinet	**94.7**	0.0	0.0	5.3	0.0	0.0
trumpet	0.0	**92.3**	0.0	7.7	0.0	0.0
violin	0.0	0.0	**93.8**	6.3	0.0	0.0
sax	0.0	0.0	0.0	**100.0**	0.0	0.0
trombone	5.9	5.9	0.0	0.0	**88.2**	0.0
cello	0.0	0.0	0.0	0.0	0.0	**100.0**

3. INDIVIDUAL PARAMETER CHARACTERISTIC

To assess sound timbre or sound quality listening test sessions are often organized. Two important assumptions were made in tests carried out in the multidimensional perceptual scaling of musical timbre. Sound stimuli were synthesized, and they were equalized for perceived pitch, loudness and duration. This was done in order to get reliable results during timbre scaling tests. However, in the presented study it is assumed that relationship between subjective descriptors and objectively derived parameters will serve for better quantization of numerical

values. In such a case testing should be done using natural sound stimuli.

The feature extraction process was performed for the six instruments mentioned before and in addition, six other instruments (bassoon, French horn, English horn, piano, oboe, viola) were parametrized. All instrument sounds were presented with three differentiated articulations. This was done in order to know the range of values possible for these classes of instruments. In Fig. 2 histogram of *Spectrum Flatness* parameter values is shown.

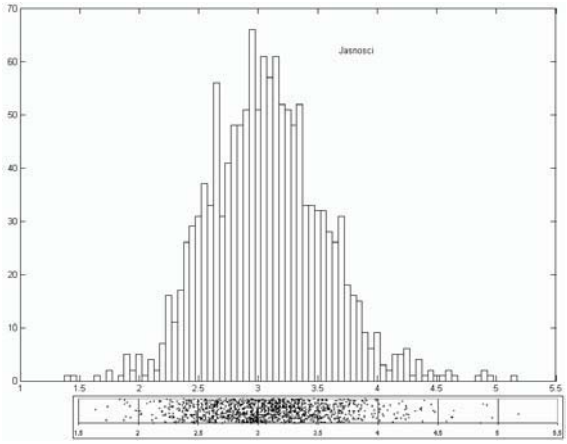


Figure 2*: Histogram of the Brightness parameter values.*

Subjective evaluation aiming at assigning semantic expressions to numerical ones can be performed employing paired comparison test. In such a case a listener is asked, which of the two presented sounds is, for example, brighter or more noise-like, etc. Assuming that the average number of sounds being parametrized for a particular instrument is 30, this means that in the worst case there would be 30x12x3=1080 sounds to be tested within particular parameter. Such a task seems impossible to perform. Therefore, another way to test the relationship between numerical values and semantic description was thought. First, discretization of the each parameter value domain should be done based on fuzzy logic principles [16]. This means that the whole domain of numerical values can be divided into a few intervals. This can be done using *Fuzzy Quantization Concept*, recently introduced by one of the authors [12]. Such intervals would create so-called membership functions, for example, low, medium and high. The listening tests should answer the question whether the cut-off points were properly defined. The problem is not only to assign ranges to parameter domains using word descriptors. One can easily imagine that experts would unanimously decide what does it mean high or low pitch of a certain musical instrument. The subject's task is to indicate for each sound a three- or five-point scale, which of the given terms applies to the stimulus. Such tests are now in progress at the Sound & Vision Eng. Department of the Gdansk University of Technology.

4. SEPARATION OF MUSICAL DUETS

Another domain, which is also related to multidimensional representation of musical sounds, is separation of musical instruments. There is a need to evaluate the quality of the separation process. This can be done either by listening to individual sounds after separation or introducing objective measures of the process. However, when sound patterns (sounds before mixing) are not available, objective evaluation of the quality of sounds can be done by calculating sound parameters and compare them with parameter values that are derived from musical instrument sound patterns. This should be done using multidimensional parameter vectors, because there is still no clear answer, which of these parameters are most important in the timbral context. In addition, such evaluation should be performed automatically. Therefore, the evaluation of musical instrument sound quality after the separation is done according to Fig. 3. The left side of Fig. 3 corresponds to the description contained in Section 2.1. The ANN is trained on feature vectors extracted on musical instrument sound patterns. The right side of the figure is related to the separation process, which is shown in detail in Fig. 4 [17].

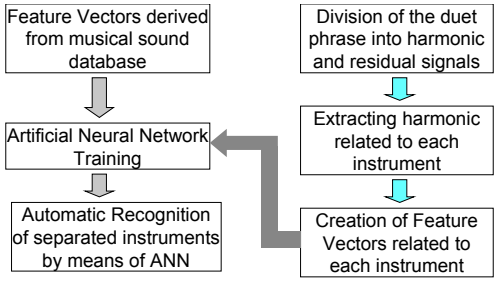


Figure 3: *System analysis.*

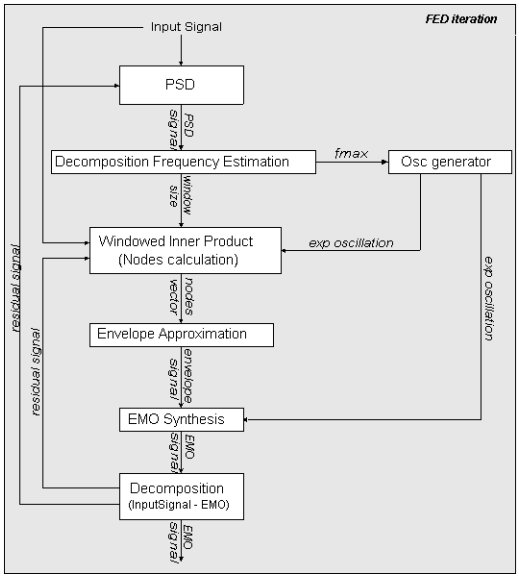


Figure 4: *Block diagram of the separation process.*

After the decomposition process is finished, feature vectors are extracted and fed to the ANN input. The task of the ANN is to answer to which of the instrument group these sounds belong [17]. In addition, some metrics can be calculated for parameters contained in the feature vectors. Figure 5 presents 3-D projection of parameters (*Brightness*, *Content of even harmonics – Hev*, and *Time*) of two sounds (piano and saxophone) before and after the separation process. Such projection serves only for visualization purpose. A comparison of multidimensional feature vectors is done employing the Euclidean metric. Calculation of

the distance between feature vectors before and after separation process gives the insight into the musical duet decomposition quality.

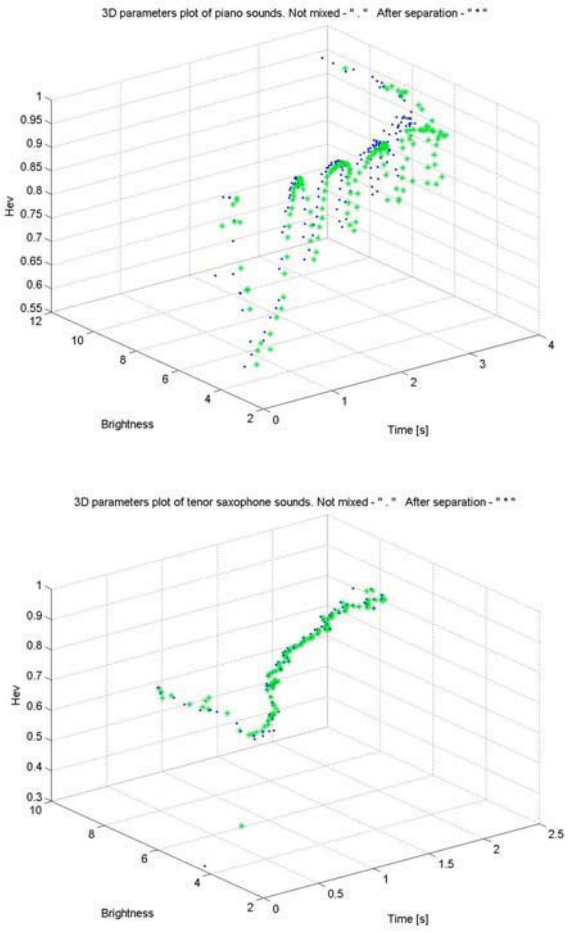


Figure 5: 3-*D projection of Parameter Space for saxophone and piano sounds.*

4. CONCLUSIONS

In this paper it was shown that analysis based on multidimensional feature vectors of musical instrument sounds could be applicable not only to the automatic recognition of musical instruments, but FVs use was extended to evaluation of sound quality of sounds after the separation process.

Acknowledgement

The research is sponsored by the Committee for Scientific Research, Warsaw, Grant No. 4T11D 014 22, and the Foundation for Polish Science, Poland.

5. REFERENCES

[1] http://www.meta-labs.com/mpeg-7-aud

[2] Lindsay A.T., Herre J., MPEG-7 and MPEG-7 Audio – An Overview, J. Audio Eng. Soc., 49, 7/8, 589-594, 2001.

[3] Grey J.M., Multidimensional perceptual scaling of musical timbres, J. Acoust. Soc. Am., 61, 5 (1977) 1270-1277.

[4] McAdams S., Winsberg S., Multidimensional scaling of musical timbre constrained by physical parameters, J. Acoust. Soc. Am., 105 (1999) 1273 (A).

[5] Wessel, D.L., Timbre Space as a Musical Control Structure, Computer Music J., 3, 2 (1979) 45-52.

[6] De Bruijn A., Timbre-Classification of Complex Tones, Acustica 40 (1978), 108-114.

[7] Iverson P., Krumhansl C.L., Isolating the dynamic attributes of musical timbre, J. Acoust. Soc. Am., 94, 5 (1993).

[8] Jensen K., The timbre model, Workshop on current research directions in computer music, Barcelona, Spain, (2001).

[9] Pollard H.F., Jansson E.V., A Tristimulus Method for the Specification of Musical Timbre, Acustica 51 (1982), 162-171.

[10] De Poli G., Piccialli A., and Roads C., Eds., "Representations of Musical Signals", MIT Press, Cambridge, MA, 1991.

[11] Pratt R.L., and P.E. Doak, A subjective rating scale for timbre, J. Sound and Vibration, 45, 3 (1976) 317-328.

[12] Kostek B., Soft Computing in Acoustics, Applications of Neural Networks, Fuzzy Logic and Rough Sets to Musical Acoustics, Studies in Fuzziness and Soft Computing, Physica Verlag, Heidelberg, New York 1999.

[13] Kostek B., Czyżewski A., Representing Musical Instrument Sounds for their Automatic Classification, J. Audio Eng. Soc., 49, 9, 768 – 785, 2001.

[14] Herrera P., Amatriain X., Battle E., Serra X., Towards Instrument Segmentation for Music Content Description: A Critical Review of instrument classification techniques, Proceedings Int. Symp. on Music Information Retrieval, ISMIR'2000, Plymouth, MA, (2000).

[15] Dziubiński M., Kostek B., High Accuracy and Octave Error Immune Pitch Detection Algorithm, Archives of Acoustics (*in print*), 2003.

[16] Kostek B., "Computing with words" Concept Applied to Musical Information Retrieval; Int. Workshop on RSKD, vol. 82, 4 (2003) Warsaw, Poland.
http://www.elsevier.nl/locate/entcs/volume82.html

[17] Kostek B., Czyżewski A., Dziubiński M., Decomposition of Duet Instrument Sounds; ISMA'2002, pp. 292 - 301, Meksyk, Meksyk, 9.12.2002 - 13.12.2002.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

PITCH MEASUREMENTS VERSUS PERCEPTION OF SOUTH INDIAN CLASSICAL MUSIC

Arvindh Krishnaswamy

Center for Computer Research in Music and Acoustics
Dept of Electrical Engineering, Stanford University
arvindh@ccrma.stanford.edu

ABSTRACT

Certain pitch inflexions used in South Indian Classical (Carnatic) music are discussed. It is proposed that these inflexions should be considered as time-varying, inflected intervals, equal in importance to constant-pitch intervals. Transient notes and inflexions are also introduced and examined. Possible reasons for some misconceptions relating to the number and the nature of the musical intervals used in Carnatic music today are also suggested. Sample pitch tracks are presented to illustrate our arguments.

1. INTRODUCTION

It should hardly be surprising if theories not based on experimental evidence are eventually discovered to have serious problems. It should also not be alarming if inadequate or incorrect system models yield confusing or contradictory results. We believe that many historic and contemporary theories of Indian music that propose divisions of the octave into more than 12 constant-pitch notes suffer from both of these pitfalls *when applied to present-day Indian music*.

There is a prevalent notion among practitioners and fans of Indian music that more than 12 musical intervals are used, where some of these intervals are "microtonal." While we agree with this belief "in spirit," we argue that the very definition of the term "interval" needs to be expanded before application to Indian music today.

While we would strongly object to the idea that Carnatic music uses more than 12 *distinct* constant-pitch notes, we would still claim that it would be inaccurate to say that Carnatic music uses 12 notes only: how would one then account for the Ga_3 in Aarabhi?

We believe that certain pitch inflexions (gamakams) used in Carnatic music are "equal in stature" to the 12 basic constant-pitch notes. Our assertion with respect to some of these gamakams is that they are *not* mere ornamentations: is there any possibility at all that the Ga_3 in Aarabhi can be played without an inflexion from Ma_1? Would it still be Aarabhi ragam if this Ga_3 was played similar to the constant-pitch Ga_3 in Hamsadwani? Can this same Ga_3 in Aarabhi be held (sounded) indefinitely? The answer to all three questions is negative.

Many people seem to readily associate the term "gamakam" with embellishments or ornamentations. However, this characterization implies an optional nature to these inflexions. We stress that playing the Ri_1 in Gaulai in the phrase "$pmgmR$;" without an oscillating, inflected pitch from Sa is neither optional nor tolerable. While in certain phrases a particular type of gamakam *has* to be used when rendering a particular note, in some other situations the musician may have many possible gamakams to choose from or may even decide not to use any gamakam at all. In these cases, the gamakams involved may be regarded as ornamentations. However, the distinction between what is optional and what is required may not always be clear since gamakams have become an integral part of Carnatic music today.

Intonation was found to be "variable" even among top musicians in Carnatic music [1]. Given the caliber of some of these musicians, it should be understood that there is indeed *some* flexibility in intonation that is musically insignificant in Carnatic music. But poor intonation (apaswaram or pisiru) can indeed be glaring and annoying at times, and so there are obviously some implicit limits in the range of acceptable intonation. These limits may depend on a lot of factors such as the note being played, its duration, the melodic context, the ragam, the underlying tempo and the instrument being used. Sometimes, a note that is vocalized or notated on paper may not directly correspond to the pitch that is actually sounded to make a phrase sound correct and acceptable [1].

Despite these intonational variances, an important point to note is that the pitch inflexions frequently used in Carnatic music are by no means "drowned" or obscured. Rather, these inflexions are distinctly audible (or visible on graphs of pitch) and can be easily recognized and reproduced by a sufficiently trained Carnatic musician.

In this paper, continuing the work described in [1], we present additional pitch tracks and examine a few pitch inflexions in more detail. We also introduce the ideas of "inflected intervals," "transient notes" and "transient inflexions" to more appropriately categorize the different "musical atoms" used in Carnatic music. Possible reasons as to why certain people might still believe that more than 12 distinct constant-pitch notes are used in Carnatic music are also suggested.

In the graphs of pitch tracks presented in this paper, the solid horizontal lines with circles at both ends on each graph correspond to an interval value from [2]. The equal temperament intervals are also depicted with solid lines but without the terminating circles. Relative pitch with respect to Sa is plotted in cents versus time. The time-gaps which occur in some plots are due to instances in the associated sound where pitch is not well-defined. These instances include periods of silence or too low a signal amplitude and "unvoiced" sounds (like when the violin bow changes direction). The STFT-based pitch tracking method described in [1] was used to produce all the graphs in this paper.

2. PITCH INFLEXIONS

Figure 1 shows two examples of Ri_1 in the ragam Saaveri. In the plot on the left, we have a time-varying pitch whose "base" is Sa

(0 cents). The distinct feature which differentiates this pitch track from any constant-pitch note is the presence of the inflexions going up from Sa. And it is not a vibrato-like pitch variation centered around Sa or a constant-pitch Ri_1 either. Roughly speaking, this pitch track can be segmented into "flat portions" and "spikes."

The variability in the shapes and peak heights of the three spikes shown should be noted. If we calculated the average value (in cents) of these spikes, we would end up roughly in the quartertonal range or less. It is clear, even by simple visual inspection, that neither the peak heights nor the average values of these spikes are in any sense consistent with the interval values stated in [2] (which are also indicated in the graph). Quartertones (constant-pitch intervals bisecting the semitones) are not used in Carnatic music, and it is possible that they may even sound strange or unmusical to many Carnatic musicians. On the whole, we argue that this (roughly) one second pitch track can't be substituted with any constant-pitch interval, with the same time duration, to achieve a similar and *acceptable* perceptual effect.

The differences in the peak heights of the spikes are not really important in the overall perception of this sound: the pitch is slewing rapidly within a spike, and lingers close to its peak value only for a very small time duration, which is really too short to be perceptually significant. The pitch sensation felt when listening to this sound is probably more influenced by the overall effect of each spike.

The variances in the spike shapes or heights can be noticed, of course, with careful and repeated listening, but as long as the *perceived* smoothness and consistency of the inflexion is within musically acceptable limits, no audience would complain. Determining exactly what these limits are may be a very difficult task, but with enough empirical data, it may be possible to come up with a range that would work well enough for applications like computer sound analysis and synthesis.

On the right in Figure 1 another time-varying Ri_1 is shown, obtained from the same recording as before. The spikes look a little higher and wider. (Or they may even be interpreted as being inverted or "downward spikes" from around 100 cents or as symmetric oscillations between two constant-pitch note levels). Indeed, the corresponding audio also sounds "higher in pitch."

3. TIME-VARYING, INFLECTED INTERVALS

The Western concept of an interval may not be sufficient or appropriate to explain some of the "musical atoms" used in Carnatic music like those in Figure 1. The idea of a time-varying, inflected interval may be required to accommodate them, rather than simply dismissing them as ornamentations.

Though the time-varying nature of such inflexions can be easily detected by the human ear, they may also have an interesting perceptual effect if a listener chooses to place such sounds into categories alongside the constant-pitch notes. Obviously the inflexions in Figure 1 will end up, if classified based on mean pitch value for example, in-between Sa and the constant-pitch Ri_1.

However, we stress that while a classification based on mean value may indeed be very convenient, and informative to some extent, a single numerical value alone will fail to capture the true nature and the perceptual effect of these inflexions. Rather, we feel that the following dimensions need to be considered, at a minimum, when analyzing or characterizing such inflexions: pitch and time. The amplitude (loudness) modulation of a phrase or inflexion may also be musically important. Different parts of an inflexion (like the spikes and flat regions for example) may also scale in time non-uniformly and independently, if the underlying tempo of the phrase in consideration is varied.

It should be possible to come up with a set of parametric models for these inflexions. Perhaps the two inflexions in Figure 1 can be represented as variations of the same underlying pitch inflexion model. Or it may be more convenient, perhaps even necessary, to introduce more than one model for different inflexions between Sa and the constant-pitch Ri_1 depending on the ragam, melodic context or intended effect of the inflexion.

Of course this idea of inflected intervals might make it very inconvenient to compare Carnatic music to other musical systems in the world, like Western music. However, this inconvenience should be the least of one's concern when trying to model Carnatic music accurately.

4. TRANSIENTS

In fast passages, or even in phrases with a slow tempo but where a note appears only briefly, the ideas of a transient note and a transient inflexion are useful. We distinguish between these two categories, and we also classify them separately from "prolonged intervals" like constant-pitch intervals or inflexions like those in Figure 1.

The distinction made between a transient note and a transient inflexion is as follows: while a transient note was *meant* to be played by the musician without any gamakams, a transient inflexion can be described as one cycle of an inflexion or one spike. The first and third occurrences of Ni_2 shown in Figure 8 are examples of transient inflexions. The perceived pitch of Ni_2 in these cases is higher than Da_2 but lower than a constant-pitch Ni_2. However, it is possible that some people may fail to realize the time-varying nature of this inflexion and hear it as a lower-pitched version of a constant-pitch Ni_2. The situation sometimes could become even more confusing when such inflexions are halted near their peak or decreased in loudness immediately after reaching the peak. Doing so makes an inflected note appear just like a transient note to a casual listener. Another cause of confusion may be fingering techniques that some instrumentalists may use to produce such inflexions. For example, in the phrase "$rgrs$" in Sri ragam, a violinist may use two different fingers to play the Ri_2 and the Ga_2. This may lead to the impression that a lower-pitched version of a constant-pitch Ga_2 is being played. However, this Ga_2 is more accurately classified as another example of a transient inflexion. Nevertheless, we certainly do not disagree that the pitch sensation felt when listening to this Ga_2 is lower than that of a prolonged constant-pitch Ga_2 appearing in some other ragams like Kapi.

The two most important features of a transient inflexion are that (1) its pitch or intonation is very much tied to the base or starting note (which could be above or below the inflexion) and (2) the pitch sensation perceived will be in-between two of the 12 basic constant-pitch notes. Gamakams are used so often that transient notes may be particularly rare when a note is followed by another one just one semitone higher or lower. In these cases, the second note may frequently be played with an inflexion from the first. The boundary between a transient inflection and a transient note may not *always* be distinct. To distinguish between the two types of transients without some musical insight or knowledge of what the musician intended to do may be difficult. It is also possible that certain musicians are not even aware that they are using an inflexion sometimes. But in some cases, it is very clear what is played:

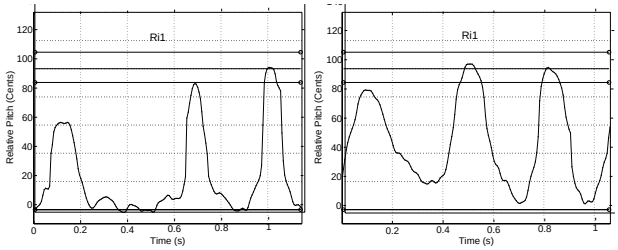


Figure 1: *Lalgudi Jayaraman - Violin - Saaveri - Inflected* Ri_1

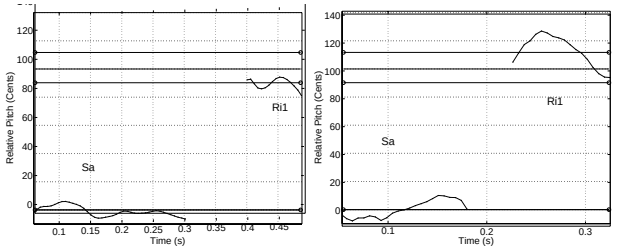


Figure 2: *Lalgudi Jayaraman - Violin - Saaveri - Transient* Ri_1

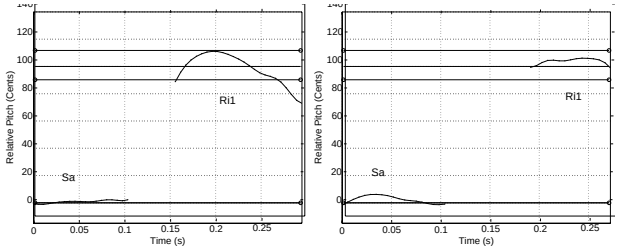


Figure 3: *Lalgudi Jayaraman - Violin - Saaveri - Transient* Ri_1

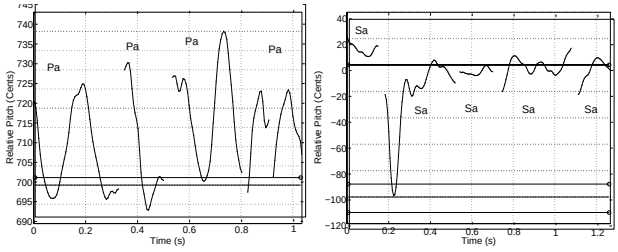


Figure 4: *Lalgudi Jayaraman - Violin - Saaveri*

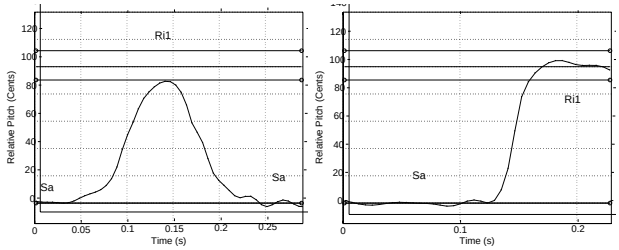


Figure 5: *U. Srinivas - Mandolin - Thodi*

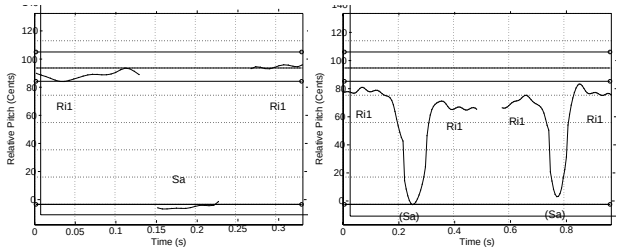


Figure 6: *T.N.Krishnan - Violin - Thodi*

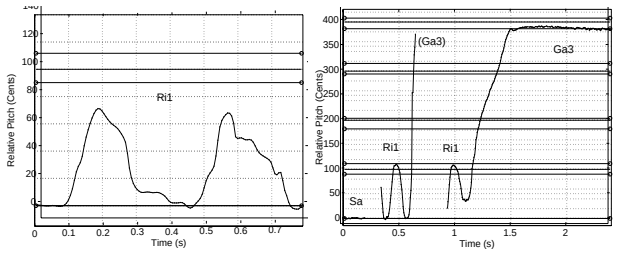


Figure 7: *M.S.Gopalakrishnan - Violin - Malayamaarutham*

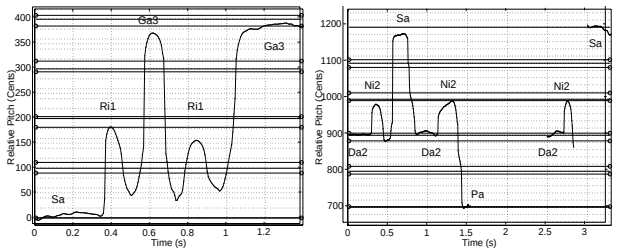


Figure 8: *N.Ramani - Flute - Poorvikalyani (L) & Kamas (R)*

the Ga_3 in Aarabhi certainly always belongs to the transient inflexion category.

Figures 2 and 3 show four transient Ri_1 examples, drawn from a 20 second audio clip of Saaveri. As can be seen, intonation is variable, but precise intonation may be unnecessary especially in such fast passages.

5. THE SAME NOTE IN DIFFERENT RAGAMS

Different versions of the same note may appear in one ragam. For example, in Saaveri, there are various types of an upward-inflected Ri_1 from Sa, there is a constant-pitch Ri_1 played in certain other phrases, and then there is yet another version of Ri_1 which involves an inflexion upward from the constant-pitch version of Ri_1, similar to what is shown in Figure 8. Incidentally, though the base of the Ri_1 inflexions shown in Figure 8 is much lower than 100 cents, this observation simply illustrates the flexibility in intonation discussed before. Based on our musical knowledge and experience, we claim that this is not intended to be a centered vibrato around a constant-pitch Ri_1, as it may appear at first.

Different ragams using Ri_1, like Thodi or Malayamaarutham, also may use many different versions of Ri_1. For example, Figure 5 shows a Sa-based inflected Ri_1 appearing in Thodi and Figure 7 shows the same type of inflexion appearing in Malaya-

maarutham. Figures 5 and 6 contain more examples of Ri_1 appearing in Thodi.

The constant-pitch version of Ri_1 is similar in all ragams using Ri_1, allowing for the ever-present variability in intonation [1]. Similarly, there are common inflected intervals among different ragams as well. Though the tempo or the shape of inflexions may be adjusted by artists to suit different phrases, it is not correct in general to say that a particular inflexion is unique to one ragam alone.

Why do some people feel that the inflected Ri_1 appearing in Saaveri or Gaulai is unique in some sense? The explanation might be that the frequency of occurrence of or the prominence given to such inflexions may be different in various ragams. Furthermore, the overall tempo and phrasing of certain inflexions may also be affected by the ragam. Certain ragams may also require a particular kind of inflexion in a phrase while other ragams may require an alternate inflexion in a similar phrase. There are also ragams which may use only a limited subset of the possible note variations, like Aarabhi employing only the inflected Ga_3.

6. OTHER SOURCES OF CONFUSION

In addition to previously discussed reasons, there may be additional sources of confusion that could lead people to a misunderstanding of the nature of the notes used in Carnatic music. We now discuss some possibilities.

6.1. Current popularity of the 22-sruthi theory

There have been many authors on Indian music theory who have put forth differing ideas [3]. But the most popular theory today seems to be the idea of "22 sruthis" [2]. This theory seems to have "mass appeal" for many reasons. It purports a connection to ancient Indian texts on music, which gives it a certain mystic charm. It also seems to accommodate the different pitch sensations felt when listening to the same note in different phrases and ragams by offering, for 10 of the 12 basic notes, two variations each. Also, a list of 22 rational numbers is both simple to grasp and appears mathematically rigorous at the same time. It would therefore be very difficult to even imagine that such an attractive theory could be invalid, when many people have promoted it with conviction.

6.2. Vocalization of inflexions using one syllable only

When a sound similar to those shown in Figure 1 is sung, only one syllable, "Ri" is used to vocalize it, instead of "Sa - Ri - Sa - Ri - Sa - Ri - Sa." This may perhaps lead an unwary person to believe that a constant-pitch note is being sung which has a lower pitch than the "usual" Ri_1. This phenomenon is probably more pronounced in transient inflexions like the Ga_3 in Aarabhi, where the time-varying pitch inflexion is sung using only the syllable "Ga."

6.3. Unfamiliarity with "small" intervals

Many people may not be aware of how small some interval differences in the "22-sruthi" theory are, compared to the variances in intonation of even top musicians. They may not even know what a 20-cent interval difference sounds like. Even the intonation of Sa and Pa can vary considerably as shown in Figure 4 without causing any distress to listeners. It is clear from the audio that at least the Pa in the left plot is modified intentionally by a vibrato-like ornament. The downward spike in the plot of Sa is due to a "janta" (double note) phrasing and is also intentional.

6.4. Dearth of empirical pitch data and plots

The unfortunate effects of the dearth of empirical data on Indian music can't be overstated. Most people simply haven't been confronted with compelling evidence, that is easily understood or digested, to dispute various incorrect beliefs.

7. PAST, PRESENT AND FUTURE WORK

Jairazbhoy and Stone [4] have done some remarkable, painstaking work, given the nature of pitch tracking equipment then available to them. Though they analyzed North Indian music only, their results and conclusions are similar to what we have observed with respect to Carnatic music. Near the end of his article, Jairazbhoy makes an important point which agrees with our own observations and conclusions: current-day Indian musicians seem to use the term "sruthi" to indicate variations of a note due to pitch inflexions. Levy's work [3] is also noteworthy: the comprehensive review of historical theories is very impressive and points made in various discussions are very insightful. While we believe that these two publications are very significant pieces of work on Indian music thus far, the notable missing elements are: (1) high-resolution, detailed pitch tracks and (2) the explicit categorization and treatment of inflexions separately from constant-pitch notes. Though these authors were obviously aware of the pitch inflexions appearing in Indian music, and even commented on them, their main goal seems to have been to disprove the validity of certain theories. We believe they accomplished this goal successfully. Since none of the theories they were examining dealt directly with pitch inflexions, they certainly weren't obliged to consider them separately either. It was also clear that they didn't intend to provide an alternate, detailed and comprehensive theory of North Indian music.

While Levy suggested that the North Indian music recordings he had analyzed seemed to be based on a flexible system of 12 semitones, we have explicitly stated that Carnatic music uses more than 12 "musical atoms" by elevating certain pitch inflexions to the status of musical intervals. We believe that doing so would eventually lead to a more accurate model of Carnatic music. Finally, though our findings are incompatible with lists of 22 or more numerical interval values like those given in [2], we would point out that such lists do indeed contain some very reasonable values like $3/2$, $4/3$ and $5/4$, but we defer the discussion of such theoretical intervals to a future report.

8. REFERENCES

[1] Krishnaswamy, "Application of pitch tracking to south indian classical music," in *Proc IEEE ICASSP*, 2003.

[2] Sambamurthy, *South Indian Music*, vol. 4, The Indian Music Publishing House, Madras, India, 1982.

[3] Levy, *Intonation in North Indian Music: A Select Comparison of Theories with Contemporary Practice*, Biblia Impex Private Limited, New Delhi, India, 1982.

[4] Jairazbhoy and Stone, "Intonation in present-day north indian classical music," *Bulletin of the School of Oriental and African Studies*, vol. 26, pp. 119–132, 1963.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

RELATIONSHIPS BETWEEN MUSICAL AUDIO, PERCEIVED QUALITIES AND MOTORIC RESPONSES – A PILOT STUDY

M. Leman[1], V. Vermeulen[1], L. De Voogdt[1], A. Camurri[2], B. Mazzarino[2], G. Volpe [2]

1 IPEM - Dept. of Musicology, Ghent University, Blandijnberg 2, 9000, Ghent, Belgium
2 InfoMus - DIST, University of Genova, Viale Causa 13, 16145, Genova, Italy
Marc.Leman@rug.ac.be

ABSTRACT

The paper describes a pilot experiment in which quantitative and structural relationships between auditory stimuli, perceived qualities and motoric responses to music are investigated. In a previous study [16] a 3-dimensional semantic space of perceived qualities was extracted from the data of a group of hundred students that evaluated sixty musical excerpts (each thirty seconds long). Stepwise multiple regression then revealed significant relationships between these three dimensions and audio cues. Valence adjectives correlate with number of onsets, tempo and harmonicity. Dominance was related to roughness and loudness. Less good results have been obtained with Arousal but this is explained by the fact that the method is calling for cognitive evaluation whereas arousal adjectives may call upon personal emotional engagements. The present pilot study aimed at extending this methodology to motoric behavioral responses. For this aim, cognitive and motoric responses to all sixty musical excerpts were recorded with a computer mouse from seven subjects. The study aimed at finding correlations between motoric behavior and cognitive evaluation, as well as between motoric behavior and audio features. Although significant correlations were found between perceived qualities and audio cues, no significant correlations have been found between perceived qualities and motoric cues, nor among motoric and audio cues. This may be due to the kind of feature extraction used for the motoric behavioral data: although they proved to be useful in other contexts, they were inappropriate for summarization over the entire musical excerpt. Despite the negative outcome at this stage of the research, the described experiment revealed new methodological and technical issues which can probably be solved. On a broader scale we performed this study as an initial step in a research program on the motoric annotation of perceived musical qualities. In the near future similar related experiments will be carried out using 2-D an 3-D haptic and digital interfaces.

1. INTRODUCTION

Experimental work on music and its perceived qualities aims at clarifying the underlying semantics in terms of maps or qualitative spaces induced from listener responses. The assumption of the present study is that semantic spaces of perceived qualities in music can be correlated with (i) features relating to physical characteristics of the musical signal, and with (ii) features related to human motoric behavior in response to the musical signal. While evaluations pertain to cognitive assertions about the music, motoric behavior is very much a sub-cognitive behavioral response to music, often related to gait patterns (Cfr. [8]).

Up to now, the main focus of the research has been on the relationship between audio cues and perceived musical qualities. In the categorical approach, where perception is cast using a set of basic emotions such as fear, sadness, joy, anger, relationships have been found between the intended emotions or moods and local acoustical features of melodies ([6],[7],[12]). In the dimensional approach, researchers tend to deal with natural music rather than melodies. The perceived qualities of music are studied using a differentiated set of adjectives from which factors of the underlying emotive/affective space are deduced. Research in this area has been conducted by many authors since the 1930ies ([9],[10],[11],[13],[17],[18],[20]) but the price of working with natural (polyphonic) music is often a less good understanding of the role of audio structural features.
The paper describes a study in which quantitative and structural relationships between natural auditory stimuli, perceived qualities, and motoric responses to music are investigated. The first part of this paper describes how the semantic space was obtained with a large group of listeners, and how individual recordings of a smaller group of listeners have been set up. The second part describes what kind of features, acoustic and motoric, were extracted. The third part discusses relationships between the three datasets.

2. SPACE OF PERCEIVED MUSICAL QUALITIES

2.1. Group Experiment

The structure of perceived emotive/affective qualities in a large set of musical stimuli was addressed using the method of semantic differentials. Hundred students from the Faculty of Letters and Arts (section Art Sciences, second year) from Ghent University participated in the experiment (see [16] for more details). The mean age was twenty-one years, 73% were females and about 60% of the subjects did not have any active musical experience. Subjects were asked to fill in a form on which to rate the musical excerpts using a set of semantic differentials. The test used adjectives in Dutch language, (also translated into English and Italian). Out of 60 different musical excerpts, chosen from many musical genres (pop, rock, ethnic, jazz, classical, ...) and on the basis of an assumed wide variety of emotional expressiveness, each student evaluated twenty-four

randomly selected pieces (24*100 = 2400 responses). In total, each excerpt was evaluated forty times. The evaluation was done with fifteen bipolar adjectives using a 7-point scale. Factor analysis, using maximum likelihood estimation and varimax rotation, was performed to reveal underlying structure in listener responses. The inter-correlations between the fifteen semantic differentials revealed three dimensions. The first dimension refers to the contrasting poles of adjectives such as Carefree and Anxious, Gay and Sad, Desperate and Hopeful, and Positive and Negative. Given the bipolar nature of the adjectives, it means that the adjectives Carefree, Gay, Hopeful, Positive are on one side of the axis, and Anxious, Sad, Desperate, Negative on the other side. The adjectives come down to the distinction between favorable and unfavorable qualifications. The dimension is therefore related to the valence of perceived qualifications. This dimension explains 19.5% of the total variance in the data. The second dimension refers to the contrasting poles of Tender, Calm and Fragile versus Bold, Restless, Powerful. The adjectives come down to the distinctions based on power, control and aggression. The dimension is called dominance and it explains 19% of the total variance. The third dimension refers to the contrasting poles of Moving, Exciting, Pleasing, Passionate, versus Indifferent, Boring, Annoying, and Restrained. The adjectives relate to arousal and this dimension explains 18% of the total variance.

2.2. Individual Experiments

In order to carry out the correlation analysis on an individual basis, a second experiment was set up with individual subjects which rated all sixty musical excerpts using the above fifteen semantic differentials. The conditions were similar as in the first experiment. The responses of each individual were then mapped onto the 3-dimensional space (Valence, Dominance, Arousal). Each factor was then taken as the to-be-explained variable in the analysis, first, of the relationships with the automatically extracted auditory-based features, secondly, of the relationships with the features extracted from motoric behavior. Two experiments then aimed at studying correlations between space of perceived qualities and audio cues, and between space of perceived qualities and motoric cues.
In the first experiment eight subjects from Ghent University evaluated all sixty audio fragments and their cognitive profile was projected onto the 3-dimensional space. In the second experiment seven subjects from Genova University evaluated the sixty audio excerpts using the same set of adjectives. They were then also asked to move a computer mouse while listening to the excerpts. No particular indications have been given to the subjects, i.e., they were free to move the mouse as they preferred. The only available visual information was the pointer of the mouse in an empty screen (i.e., no lines nor icons were drawn on the screen, just the blue background). The audio excerpts have been presented to each subject in a random order with 0.5 s of silence in between two subsequent excerpts.

3. FEATURE EXTRACTION

Low-level features have been obtained from the musical audio using the real-time auditory model of MEGASE (www.megaproject.org). In the present context, only loudness extraction, roughness extraction, virtual pitch extraction, and centroid extraction have been considered. The loudness extractor is based on a low-pass filter on the amplitude in each auditory filter band, and then summed over all bands. The roughness is the amplitude after a high-pass filter on the amplitude. Pitch and pitch prominence are computed from an FFT in each band. Over a window of frames, onsets are detected by comparing the average loudness with the peak loudness. The centroid is the center-of gravity of the amplitudes in the bands. Though each feature is extracted in a continuous way over the total duration of an excerpt, only basic statistics has been taken into account. As such, each audio cue is just one single numeric value for the entire musical excerpt. Single value descriptors are here justified on the basis of homogeneous expressiveness in each of the 60 excerpts.

Algorithms for analysis of motoric trajectories included in the EyesWeb Expressive Gesture Processing Library (Cfr. [4]) (see also www.megaproject.org) have been employed to extract motion descriptors from the gathered trajectories. In particular, the following motion cues have been computed: Length in pixels of the mouse trajectory during each excerpt. Average, standard deviation, and peak value of speed during each excerpt, measured in [pixels/s]. The asymmetric backward numeric derivative has been used to calculate velocity from the trajectory data. Average, standard deviation, and peak value of the module of acceleration during each excerpt, measured in [pixels/s2]. Directness index, calculated as the ratio between the length of the straight trajectory connecting the first and the last point of the mouse trajectory and the sum of the lengths of each segment constituting the trajectory (i.e., the length of the mouse trajectory as described above). Angle (in polar coordinates) of the average displacement during each excerpt, giving indication about the overall direction along which the mouse moved. These features have been chosen since they were grounded on theories of expressive non-verbal communication through human movement (e.g.,[1], [14], [15], [19]). The feature have been successfully employed in human movement analysis in other contexts, e.g., [21] used some of them for classifying arm movements along Laban's Effort dimensions (Cfr.[14], [15]), and in artistic scenarios, e.g., in [4] an experiment is described in which some of these features are employed for classifying dance fragments in terms of basic emotions (anger, fear, grief, joy).

4. DATA ANALYSIS AND RESULTS

4.1. Correlation of perceived qualities with audio cues

A stepwise (both directions) multiple regression analysis reveals that Valence has a significant correlation with the number of onsets and the inter onset interval. Dominance shows a very significant correlation with roughness in seven out of the eight cases. Arousal, however shows no significant correlations. See [6] for more detailed analysis and further results with manually annotated data.

4.2. Correlation of perceived qualities with motoric cues

A first analysis based on bidirectional multilinear regression does not help in in finding relations between the cues computed from the mouse trajectories and the space of perceived qualities

(Valence, Dominance and Arousal). Similarly, correlations with any of the audio cues do not emerge in this preliminary study. Explanations of such results suggest that the choice of motion cues (mostly related to aspects related to the shape and kinematics of the curves) should be reconsidered. Although these cues function well in other applications (Cfr. [3]), these are clearly unable to summarize the possible dynamic features of the movement in this experiment. Higher-level motion descriptors more closely related to the real movement perfomed by listeners in this experiment will be taken into account in a second phase of this project. they include cues describing aspects on fluency (i.e., how much the movement is fluent, hesitating, bound/unbound according Laban's Effort terminology, the degree of segmentation in movement), impulsiveness.

5. CONCLUSIONS

The study aimed at finding correlations between perceived affective qualities in music, audio cues and motoric behavioral responses. Although perceived qualities and audio cues show clear systematic relationships, the comparison with motoric responses has not led to any results thus far. Nevertheless, the authors have gained insight into the precise nature of the motoric features needed. The current experiment shows that simple statistical operations on the lowest-level motion descriptors (e.g., average of speed along each excerpt) are not enough meaningful for describing the behavioural movement response to music. Instead higher-level motion descriptors should be taken into account. Our ongoing research program will moreover embody continuous statistical data analysis tools and optimalization techniques enabling a back-and-forth approach between extraction of relevant motoric cues and cognitive respons data or structural audio features.

Acknowledgements

We thank Matteo Ricchetti, Olivier Villon, and the students Michele Bruzzone and Marco Scarfò at the DIST-InfoMus Lab for their help in this work.

6. REFERENCES

[1] Boone R. T., Cunningham J. G., "Children's decoding of emotion in expressive body movement: The development of cue attunement",Developmental Psychology, 34, 1007-1016, 1998.

[2] Kleinen, G. (1968), *Experimentelle Studien zum musikalischen Ausdruck.* Hamburg: Universität Hamburg

[3] Camurri A., De Poli, G. & Leman, M. (2001), Megase: a multisensory expressive gesture applications system environment for artistic performances. In Proceedings of CAST01 // *living in mixed reality – conference on communication of art, science and technology, september 21-22*/ Schloss Birlinghoven,Sankt Augustin, Bonn: Fraunhofer-Institute for Media Communication IMK.

[4] Camurri A., Mazzarino B., Volpe G., "Analysis of Expressive Gestures in Human Movement: The EyesWeb Expressive Gesture Processing Library", in Proc. XIV Colloquium on Musical Informatics (CIM), Florence, Italy, May 2003.

[5] Camurri A., Lagerlof I., and Volpe G., "Emotions and cue extraction from dance movements", International Journal of Human Computer Studies, in press.

[6] Canazza, S., De Poli, G., Rinaldin, S. & Vidolin, A. (1997). Sonological analysis of clarinet expressivity. In M. Leman (Ed.), *Music, Gestalt and computing : Studies in cognitive and systematic musicology.* Berlin, Heidelberg,: Springer-Verlag.

[7] Dannenberg, R. & De Poli, G. (1998). *Synthesis of performance nuance.* Special issue of Journal of New Music Research.

[8] Friberg, A., Sundberg, J., and Frydén, L. (2000) Music from motion: Sound level envelopes of tones expressing human locomotion. *Journal of New Music Research* , 29 (2000), No 3, 199-210

[9] Gabrielsson, A. (1973), Adjective ratings and dimension analysis of auditory rythm patterns. *Scandinavian Journal of Psychology*, 14, 244-260.

[10] Hevner, K. (1936), Experimental studies of the elements of expression in music. *American Journal of Psychology, 48*, 246-248.

[11] Imberty, M. (1976), *Signification and meaning in music.* Groupe de Rechereches en Sémiologie Musicale, Faculté de Musique, Université de Montréal, Canada.

[12] Juslin, P. (2000), Cue utilization in communication of emotion in music performance: Relating performance to perception. Journal of Experimental Psychology – Human Perception and Performance, 26(6), 1797-1892

[13] Kleinen, G. (1968), *Experimentelle Studien zum musikalischen Ausdruck.* Hamburg: Universität Hamburg.

[14] Laban R., Lawrence F.C., "Effort", Macdonald&Evans Ltd. London, 1947.

[15] Laban R., "Modern Educational Dance", Macdonald & Evans Ltd. London, 1963.

[16] Leman, M., Vermeulen, V., De Voogdt L., Taelman, J. & Moelants, D.(2003), ("in preparation"), Correlation of perceived qualities and auditory features of music.

[17] Lindström, E., (1997). Impact of melodic structure on emotional expression. In A. Gabrielsson (Ed.), *Proceedings of the third triennial escom conference* (p. 292-7). Upsalla University.

[18] Nielzén, S., & Cesarec, Z. (1981), On the perception of emotional meaning in music. *Psychology of music, 9*, 17-31.

[19] Wallbott H.G., "The measurement of Human Expressions", in Walbunga von Rallfer-Engel (Ed.), Aspects of communications, pp. 203-228, 1980.

[20] Wedin, L. (1972), Multidimensional scaling of emotional expression in music. *Swedish Journal of Musicology, 54*, 1-17.

[21] Zhao L., "Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures", Ph.D Dissertation, University of Pennsylvania, 2001.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

USER-DEPENDENT TAXONOMY OF MUSICAL FEATURES AS A CONCEPTUAL FRAMEWORK FOR MUSICAL AUDIO-MINING TECHNOLOGY

Micheline Lesaffre[1], Marc Leman[1], Koen Tanghe[1], Bernard De Baets[3], Hans De Meyer[4] and Jean-Pierre Martens[2]

1 Department of Musicology (IPEM), Ghent University, Blandijnberg 2, 9000-GHENT, Belgium,
2 Department of Electronics and Information Systems (ELIS), Ghent University
3 Department of Applied mathematics, Biometrics and Process Control, Ghent University
4 Department of Applied Mathematics and Computer Science, Ghent University
Micheline.Lesaffre@rug.ac.be

ABSTRACT

Musical audio-mining technology allows users to search and retrieve music by means of content-based text and audio queries. Though these systems are promising, there is a need for a better understanding of the role of user preferences and user profiles. The development of taxonomies for different aspects of a musical audio-mining system aims at bridging the gap between system development and user interactive interfacing. In the first part of this paper, the need for user-dependent taxonomy development is addressed and an experiment in spontaneous user behavior is described, based on 72 users and 1148 vocal queries. Statistical analysis provides insight into the characteristics of vocal querying and possible useful concepts. In the second part of the paper, it is described how categories and concepts from the statistical analysis have been used for the refinement of taxonomies that address user interactive interfacing and feature extraction.

1. INTRODUCTION

Search and retrieval of information is a core activity of the information society and music is a product of interest to many members of this society. Hence the need for advanced music information retrieval (MIR) systems, which allow users to specify musical content interactively [1]. For this aim, reliable automatic annotation and processing of musical content is needed (see the proceedings of ISMIR 2000-2).

The extraction and processing of musical content from musical audio is called musical audio mining. Figure 1 shows the general architecture of an audio-based music retrieval system. It consists of a target audio-database (left), a query interface (right), and a similarity-matching engine with optional parts that account for users profiling. The task is to retrieve the wanted musical audio files using information provided by the query.

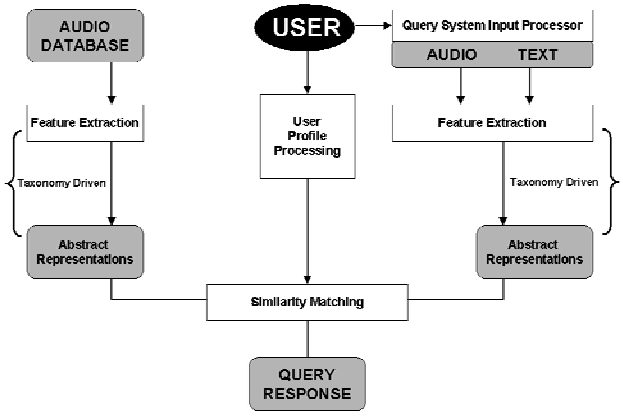


Figure 1: *General architecture of an audio-based MIR system. Left: the target structures with audio database and abstract representation. Right: the query structures, with query input and abstract representations. In the middle, user with particular preferences and profile. Bottom: query response (e.g. in the form of a list of titles)*

A main problem, however, is how users want to interact with those systems. Although there is general agreement that the user-centered approach is a key area of music information retrieval, few studies have been undertaken that concentrate on user behavior [2]. Thus far, a variety of systems have been developed that can automatically extract information from an audio signal [3, 4] but less attention has been paid to the user-friendly characteristics of this interaction.

Progress in developing user-friendly musical audio-mining depends on a number of factors, among which: (i) careful analysis of spontaneous user query behavior, (ii) the development of taxonomies, that is, sets of meaningful concepts and relationships between concepts dealing with musical content, (iii) the implementation of a system that uses this

network of concepts and interrelations for automatic annotation and processing of musical content, (iv) the design of a user-friendly query interface that deals with the spontaneous ways in which users tend to interact with music.

This paper highlights the approach of the MAMI-project in several of these domains. In the first part of this paper, the need for user-dependent taxonomy development is addressed and an experiment in spontaneous user behavior is described, based on a large-scale experiment with 72 users and 1148 vocal queries. Statistical analysis provides insight into the characteristics of vocal querying and possible useful concepts. In the second part of the paper, it is described how categories and concepts from the statistical analysis have been used for the refinement of taxonomies that address user interactive interfacing and feature extraction.

2. A TAXONOMY OF MUSIC DESCRIPTORS

2.1. Why a Taxonomy?

Music is characterized by multiple levels of content descriptors, from low-level acoustical features up to high-level structural descriptions and qualitative specifications. Taxonomies aim at providing descriptor labels for these many levels, which is useful for two reasons: (i) provide a common language of descriptors to developers of audio-based MIR systems (they tend to come from very different disciplines), (ii) provide a common language to users that interact with these systems (also users tend to have different backgrounds). The development of a taxonomy is challenging and dynamic: refinement is needed in view of application results. The role of taxonomies is to make bridges between system development and user interactive interfaces.

Until now, taxonomies for music description have not been conceived of as operational instruments related to audio mining. The music industry, Internet music retailers, copyright companies and many others have designed music taxonomies for genre classification [5], but most of these taxonomies are created to meet the particular aims of the industries rather than the users, or guide their consumers, without having any foundation within a network of musical relevant concepts. As a result of this particularity, these taxonomies show many constraints and inconsistencies. They are not very useful in audio mining where instrumental and user-friendly taxonomies are needed.

Instrumental user-friendly taxonomies are meant to define content, structure and boundaries as a framework for the systematic development of musical audio mining. They deal with the interdisciplinary background of researchers working in audio-mining as well as with user behavior, and they have to cope with the multi-facetted ways in which music can be described.

2.2. What kind of Taxonomy?

The taxonomies worked out within the context of the MAMI project are multi-leveled, and multi-dimensional. Distinctions have been made, for example, between description levels, concept categories, features and methods [6]. Low (physical and sensorial), mid (perceptual) and high description levels (formal and expressive) point to the degree of abstraction in which content is related to the audio signal. Concept categories are classes of items with common attributes (e.g. genres, styles, sound sources, and structural entities). Features are properties extracted from the musical signal (e.g. note duration, pitch). Methods are related to how users are dealing with content specification (e.g. vocal: singing, humming, textual).

It is not our intention to work on a single taxonomy for all problems, rather to develop different taxonomies dedicated to particular aspects of the MIR system.

As mentioned before, a taxonomy serves both the developers and the users, although users may be the driving factor in taxonomy design. A user-oriented taxonomy aims at specifying concept categories that can deal with the broad diversity of how users deal with music. In view of Figure 1, a user-friendly taxonomy is related to the query-interface part. Such a taxonomy should draw on the different query statements which users tend to formulate spontaneously.

In what follows, a coarse taxonomy, defined in a previous phase of the MAMI-project [7], has been refined on the basis of a study of how users tend to deal with audio-based MIR systems. We first describe the results of that study, and then describe its effect on the development of our taxonomy.

3. SPONTANEOUS USER BEHAVIOR

A query-by-voice (QBV) experiment for studying spontaneous user behavior has been set up [8] with 72 subjects. Their task was to reproduce a tune from memory. The experiment resulted in a large database of 1148 vocal queries that was annotated and statistically analyzed [2]. The main results are briefly reviewed.

3.1. Characteristics of vocal querying

Vocal querying has defined characteristics. Results of the statistical analysis indicate that a user-friendly MIR system should be able to deal with people starting up to 2 seconds after the start of the recording and reproducing a few seconds of a tune up to 30 seconds or more. The mean query length however is around 14 seconds.

The ideal system should also cope with changes in query method such as singing syllables, singing text, humming and whistling. Sometimes subjects prefer to tap along with the drum or make spoken comments while performing a query. More than half of the queries is performed using only one method, the rest varies between 2 and 6 methods. Most common query methods are singing lyrics and singing syllables.

The results suggest that any system that concentrates on one single query method will confine the spontaneous way in which the user tends to describe the musical content. Our experiment showed that less than half of the participants produce whistled queries and a particular group even tends to provide long whistled queries of high quality. In contrast to this, most existing systems require the use of specific syllables for the ease of query processing [9]. Yet the kind of syllables requested from

users are often the wrong ones. Syllables like [ba] or [fa], for example, hardly occur in spontaneous vocal querying.

Three quarter of the query collection is performed in a melodic way, which supports the known wisdom that melodic content is to be regarded as a major salient feature of vocal queries [10].

3.2. User groups

The analysis of query methods suggests the categorization of users into different groups. This categorization depends on age, gender, musical background and memory type. Five user profiles were found. The first relies on singing text, the second prefers singing syllables, the third divides its query time between two methods, the fourth mixes several methods and the fifth (small group) prefers whistling.

Significant effects on the methods used were found between musicians and non-musicians. For musicians, text occupies a significantly smaller position than for non-musicians. As to gender, men tend to use a larger variety of syllables than women do. There is also clear influence of the memory type on the query method that changes from textual dominance to syllabic dominance with a growing importance of short-term memory. Remarkable is the sudden increase in percussive queries when long-term memory of the song is no longer present.

Conclusion: In this study of spontaneous querying, a large variety of strategies was observed but the experiment generated guidelines that provided information for the refinement of user-oriented taxonomies.

4. TAXONOMY REFINEMENT

4.1. User-Oriented Taxonomy

The experiment on spontaneous user behavior provides guidelines for the development of user-oriented taxonomies, which serve as bridges between user preferences and needs, and system development.

The taxonomy depicted in Figure 2 has been used for developing an interactive user interface for querying. Query methods are categorized into four approaches called standard info, melody and harmony, timing and rhythm, loudness, timbre and subjective qualities. Standard info requires only textual input about standard meta-data such as the title of the piece, name of the composer, orchestra director, or record company. Apart from that, content specific information can be given about melody, harmony, rhythm, loudness and timbre. These categories allow the specification of structural descriptions. Inputs may be based on specifications of structural descriptors (e.g. selecting pre-defined verbal descriptors related to pitch and rhythm), audio queries and audio examples (e.g. recording of a voice or an instrument, example), graphisms (e.g. make a drawing of tempo evolution) or motorical actions (e.g. tapping the beat, handle a slider).

Subjective qualities, related to perceived affective qualities of the music, up to now, are described by a limited set of bipolar adjectives (e.g. sad-gay, tender-brutal, boring-exiting). Input can be given by moving a slider, for example.

CATEGORY	SUBCATEGORY	QUERY METHOD
STANDARD INFO		textual list selection
MELODY & HARMONY	MELODY	notated melody recording (audio, MIDI) musical excerpt
	CHORD PROGRESSION	notated chord progression recording (audio, MIDI) musical excerpt
	TONALITY	notated tonality recording (audio, MIDI) musical excerpt
TIMING & RHYTHM	TEMPO	BPM, tempo tapping recording (audio, MIDI) musical excerpt, drawing
	TIMING / STRUCTURE	textual list selection
	DRUM PATTERNS	notated drum pattern recording (audio, MIDI) musical excerpt
LOUDNESS		list selection drawing
TIMBRE		list selection
SUBJECTIVE QUALITIES		selection of qualities moving a slider

Figure 2: *Categories and subcategories of a user interface, with related query methods that are facilities for vocal input, verbal specifications in terms of low to high level structural descriptions, and qualitative descriptions.*

4.2. Constituent music-related concepts

A major part of taxonomy development is devoted to musical feature extraction and analysis. This taxonomy (see Figure 3) includes six basic categories: melody, harmony, rhythm and timbre, dynamics and expression. Pitch, as a unit in time and space, is constituent for melody and harmony. Melody, harmony and rhythm attach to spatio-temporal structural properties. Timbre is part of the concept class that describes sound sources or sonic material. The Dynamics category relates to aspects of loudness and the many aspects of movement and musical gesture. Dynamics levels and accents are natural indicators for emotional mood as well. As shown in the figure, the categories relating to audio-structural properties are also distinguished according to physical, sensorial, perceptual and formal descriptors. Many researchers are currently studying the automated extraction of these audio-structural features.

Expression, on the other hand, is a high level inter-subjective category which cannot be directly extracted from the audio, but which can be modeled on the basis relationships between extracted structural (cognitive) features and perceived qualities (affects, emotions). Investigation on expressive qualities reveals interesting relationships between structural and semantic descriptions [6]. It has been shown that the listener's perception of qualitative musical content tends to be inter-subjectively consistent and that semantics can connect with automatically extracted low and mid level features.

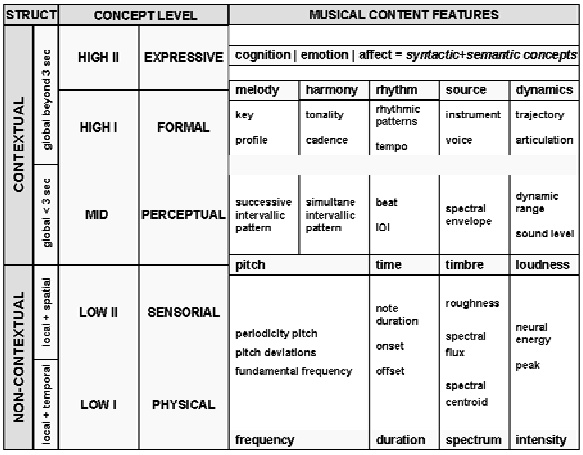


Figure 3: schematic overview of a taxonomy for feature extraction

5. TAXONOMY AND SYSTEM IMPLEMETATION

The development of the MAMI audio-based MIR system draws on the development of taxonomies such as the ones described above. As to system implementation, two distinctions can be made. (i) A query interface is currently under construction which allows users to specify musical content in terms of vocal melody inputs, musical excerpts and verbal descriptors, based on the taxonomy of Figure 2. The main task is to transform the query input in a representation that is appropriate for search and similarity matching. Search results are reported as a hierarchical list ranked in order of decreasing relevance. Audio feedback allows the user to check whether the engine returned the required music. (ii) Underlying this interface is a library of modules (C++ classes) designed in a clear relationship to the basic concepts of the taxonomy.

5. CONCLUSIONS

Structures found in spontaneous vocal queries were helpful for the refinement of a taxonomy that addresses the design of MIR systems. Taxonomies offer a list of well-defined concepts and relationships between concepts. They provide structure to developers and to users whose backgrounds may be very different. The main requirement for a MIR system is that it should offer the user as much flexibility as possible in performing a search on a music database. Concepts in the taxonomy should allow designers and users to deal with the broad diversity of preferences, needs, and expectations.

Although the work thus far has been focusing on the query part of the MIR system, definitions of the triangular dependencies between query, target and low-level features derived from the signal are urgently needed. The MAMI taxonomies provide steps towards MIR system development that focus on the user, making audio mining more efficient and attractive.

ACKNOWLEDGEMENTS

This research and the QBV experiment were conducted in the framework of the MAMI project, which is funded by the Flemish Institute for the Promotion of Scientific and Technical Research in Industry. The authors gratefully acknowledge Dirk Moelants and Gaëtan Martens for their assistance.

6. REFERENCES

[1] Downie, S. J., "Music information retrieval" in Annual Review of Information Science and Technology 37, ed. Cronin Blaise (Medford, NJ: Information Today, 2003), pp. 295-340

[2] Lesaffre, M., Moelants, D. and Leman, M., "Spontaneous user behavior in vocal queries for audio-mining." submitted, 2003.

[3] McNab, R. J., Smith L. A., Bainbridge D. and Witten I. H., "The New Zealand Digital Library MELody inDEX", D-Lib Magazine (1997), pp. 11-18

[4] Prechelt, L. and Typke R. "An interface for melody input" in ACM Transactions on Computer-Human Interaction (2001), pp. 133-149

[5] Pachet, F. and Cazaly, D. "A Classification of Musical Genre." Proceedings of Content-Based Multimedia Information Access (RIAO) Conference, Paris, France, 2000

[6] Leman, M., Vermeulen, V., De voogdt, L., Taelman J., Moelants D., and Lesaffre M. "Correlation of gestural musical audio cues and perceived expressive qualities" submitted, 2003

[7] Leman, M., Clarisse, L., De Baets, B., De Meyer, H., Lesaffre, M., Martens, G., Martens, J., and Van Steelant, D. "Tendencies, perspectives, and opportunities of musical audio-mining." In A. Calvo-Manzano, A. Pérez-López, & J. S. Santiago (Eds.), Forum Acusticum Sevilla 2002, 16-20 september, 2002. Madrid: Sociedad Española de Acustica - SEA. (Special issue of Journal Revista de Acustica Vol XXXIII, no. 3-4), 2002.

[8] Lesaffre, M., Tanghe, K., Martens, G., Moelants, D., Leman, M., De Baets B., De Meyer H. and Martens J.P. " The MAMI Query-By-Voice Experiment: Collecting and annotating vocal queries for music information retrieval" proc. ISMIR 2003

[9] Pauws, S. "CubyHum: A fully operational Query by Humming System" in Proceedings of the Third International Conference on Music Information Retrieval, ed. Michael Fingerhut (Paris: IRCAM – Centre Pompidou, 2002), pp.187-196

[10] Chai, Wei. "Melody Retrieval on the Web" M.S. Thesis, MIT Media Lab, 2001.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

CONSISTENCY IN LISTENER'S RATINGS AS A FUNCTION OF LISTENING TIME

Guy Madison and *Björn Merker*

Department of Psychology
Uppsala University
guy.madison@psyk.uu.se

ABSTRACT

We examined the consistency and similarity of adjective ratings of 10 music examples (ME) across 7 different example durations. Although the effect of ME was generally statistically significant for all 14 adjectives and all durations from 0.5 to 16 s, the agreement between the relative ratings of MEs was poor for some adjectives. Consistency, in terms of *F* ratios, did not increase with duration, which seems to contradict the hypothesis that more information entails better judgements. We discuss possible explanations for these results, and conclude that further examination of the validity of adjective ratings is warranted.

1. INTRODUCTION

Ratings of adjectives describing possible experiential properties is one of the most common dependent variables in music perception research. The consistency of such ratings is typically high, allowing statistically significant results to be based upon relatively small numbers of listeners (e.g., 10-30). For reviews see [1, 2]. The wide range of sample durations employed in the literature (from a few seconds to several minutes) raises the question of how sample length might relate to the property one is attempting to measure. At least in theory, sample length might exert a significant influence on listener judgements via a number of conceivable dimensions of musical structure. These include, but are not limited to, potential effects of redundancy on listener confidence, sample length effects on listener access to cognitively significant musical content, and the dynamics of making judgements under uncertainty.

These issues have so far not received systematic study, and even the existence of a sample length effect as such has not been empirically established. As a first step in filling this gap we have asked how the validity and reliability of adjective ratings might be affected by sample durations spanning five octaves in length, from 0.5 to 16 s. We used *F*-ratios to determine consistency among listeners as a function of music sample durations, and changes in mean ratings as indicators of sample duration effects on listener judgements. We obtained significant effects for some durations and adjectives, and discuss possible implications of these results for the use of adjective rating scales in music research.

2. METHOD

2.1. Participants and task

The 105 listeners were between 18 and 55 years of age, approximately evenly divided between men and women, and had a wide range of musical experience. Each person sat alone in a room and rated the appropriateness of 14 adjectives on a scale from 0 to 10 on a computer, which recorded the ratings and played the music examples through headphones. The adjectives are listed in Table 1, and at least 10 of them are common in previous music perception research.

A critical part of the instruction was to rate every adjective carefully even if uncertain, that 0 corresponds to "not at all appropriate", 10 to "very appropriate", and that 5 therefore corresponds to "moderately appropriate", and that no value may be thought of as "don't know".

2.2. Music examples (ME)

All 38 examples consisted of instrumental music selected from commercially available CDs recorded between 1990 and 1999. They comprise a wide range of different groove-based musics which can be described as jazz, African, African-American, and South-European. We deliberately selected unfamiliar music so as to minimise possible associations to a certain time, context, or value that previous exposure to the music might represent.

2.3. Design

The music examples on which the present study is based were either 0.5, 1, 2, 4, 6, 8, or 16 s in duration, and each level of duration was given to 15 out of the 105 listeners. In order to avoid that listeners adopt different general strategies as a trivial adaptation to hearing only examples of a certain duration, these examples were embedded among 28 other items with a duration between 8 and 19 s. They thus make up slightly more than 1/4 of the total number of items to which participants were exposed, and were randomly distributed among this total of 38 items.

3. RESULTS AND DISCUSSION

3.1. Mean ratings

Main effects of duration are not necessarily related to the issue of listeners' consistency. Given that there is a difference among the music examples in the properties that listeners judge, we would expect listeners to distribute the ratings across the range 0-10. If we further assume that listeners are more uncertain for very short durations, and hence give ratings close to 5, then no difference in mean rating should be expected as a function of duration.

Nevertheless, duration produced a significant effect on mean ratings for four rating scales, as detailed in Table 1, although this effect was much smaller than that of music example (ME).

Rating scale	**Duration** *df* = 6, 98	**ME** *df* = 9, 882	**Dur × ME** *df* = 54, 882
Driving		12.02	
Simple		11.10	1.95
Floating		12.25	
Happy	2.82	26.23	1.70
Rocking		15.25	
Walking		8.42	
Solemn		12.16	
Intensive		17.12	
Calm		32.11	
Movement inducing		10.28	
Good	2.34	20.82	1.45
Rapid	4.71	44.00	1.57
Bouncing		10.67	
Swingy	3.55	24.14	
Number of effects	4	14	4

Table 1: *F-ratios for statistically significant effects of duration, music example (ME), and their interaction, according to 14 mixed (7 duration – between × 10 ME - within) ANOVAs.*

Since a given effect is more likely to reach significance for a within-subjects variable than for a between-subjects variable, this may account for much of the difference in effect size between duration and ME. The difference can be gauged by the *F*-ratios corresponding to a 0.05 probability for rejecting the null hypothesis, which are approximately 2.2 for duration, 1.88 for ME, and 1.35 for their interaction.

What might account for these main effects? Perhaps there was an overall difference in, say, perceived happiness between these 10 ME and the other 28, and more so as duration increased. However, visual inspection of all rating means revealed that all but three increased slightly (typically 0.5-1 units) with duration, and it is unlikely that such a set of disparate properties should be greater for this group of 10 ME. Figure 1 demonstrates this tendency for all the included ratings except Calm. Therefore, it seems more plausible to interpret this result as a tendency to make lower ratings under uncertainty, which is corroborated by the observation that mean rating is positively correlated with the *F*-ratio for ME across all 14 scales and 7 durations ($r = 0.30$, $N = 98$, $p < .05$).

The interaction between Duration and ME is more pertinent for the issue of consistency, since it indicates that the relative ratings among the ME were affected by duration.

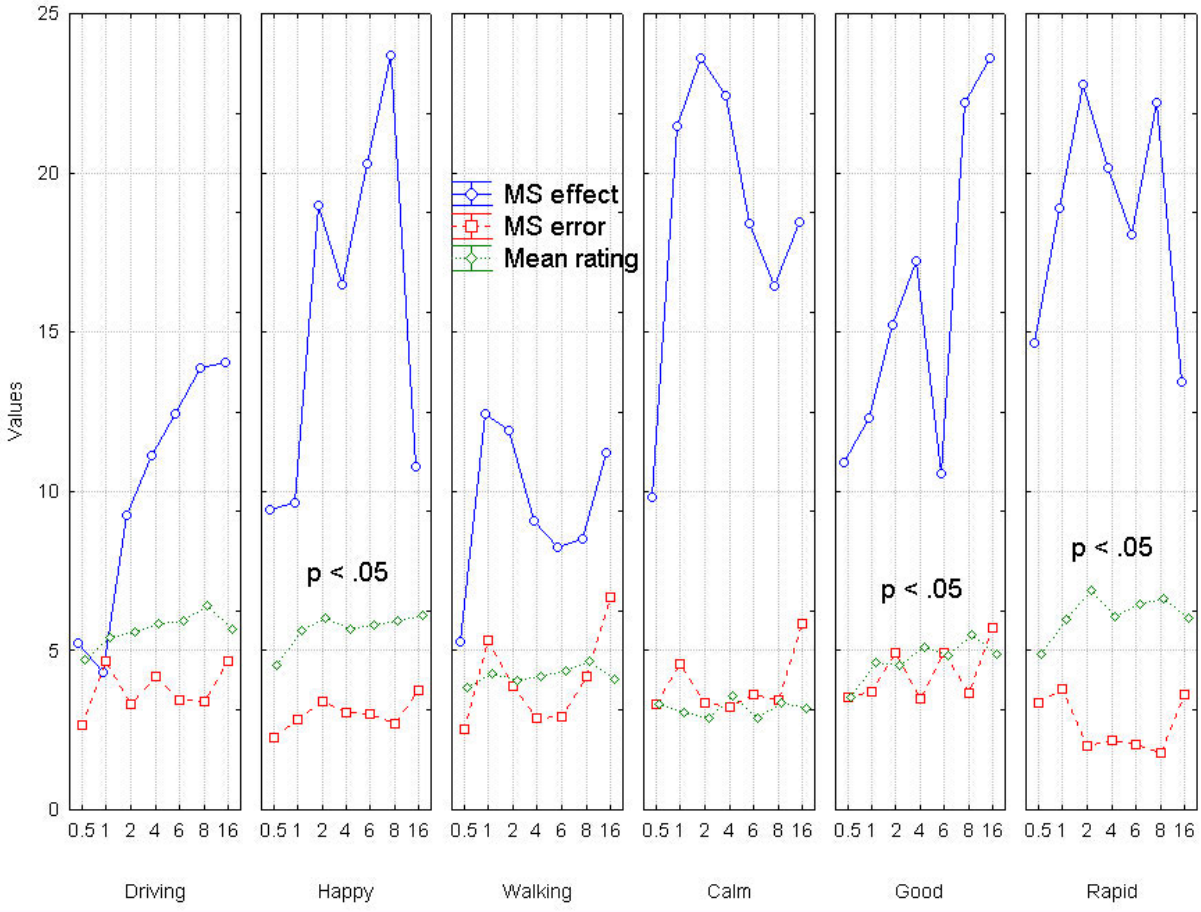


Figure 1: *Mean ratings and mean squares for both effect and error for six selected rating scales, according to within-subjects ANOVAs for the effect of music examples.*

This only emerged for four scales, three of which also yielded main effects. We therefore selected these three for a more detailed analysis, joined for comparison by the three rating scales with the lowest *F*-ratio for the effect of duration, namely Calm, Driving, and Walking.

3.2. Consistency among listeners: Comparison of mean squares

Our approach to estimate listeners' consistency is based on the partitioning of variance provided by within-subjects one-way (10 ME) ANOVAs. The *F* estimate of statistical significance is the ratio between two variances: the systematic effect mean squares ($MS_{eff.}$) due to variation in a dependent variable, and the residual error ($MS_{err.}$) due to variation among experimental cases. In this experiment $MS_{eff.}$ thus reflects how differently each group of 15 listeners rated the music examples, while $MS_{err.}$ expresses the difference among the listeners.

Considering $MS_{eff.}$ and $MS_{err.}$ separately allows us to distinguish among different interpretations of the listeners' behaviour. The null hypothesis concerning ME – corresponding with $MS_{eff.}$ being smaller than $MS_{err.}$ – is already rejected for most combinations of duration/group and rating scale. Specifically, Calm and 4, 8, and 16 s, and Driving and 1 s subtract 4 non-significant effects from 42 (7 durations × 6 rating scales), with 38 ANOVAs remaining significant.

We might predict, for example, that the smaller amount of information provided by shorter examples makes listeners reluctant to distinguish between them, leading to both small $MS_{eff.}$ and $MS_{err.}$, with $MS_{eff.}$ increasing as a function of duration. On the other hand, listeners might feel inclined to guess in this situation, presumably about the continuation of the interrupted music. Since individual experiences in this case is likely to play a larger role, this would predict small $MS_{eff.}$ and large $MS_{err.}$, with the latter decreasing as a function of duration.

Figure 1 plots mean rating, $MS_{eff.}$, and $MS_{err.}$ for each of the six selected rating scales. Significant effects of duration on mean rating is indicated with $p < 0.05$ (according to Table 1) for Good, Happy, and Rapid. First, the functions of duration are generally not monotonic, which may to some extent be due to differences between groups. However, a potential trend should be evident across groups in spite of infelicitous sampling. One consistent tendency across all scales is that $MS_{eff.}$ is lower for 0.5 s, and this can also be said about 1 s for Happy, Walking, and Good. Second, there is a tendency for $MS_{err.}$ to increase with duration, as suggested for all rating scales but Rapid. This stands in stark contradiction to the hypotheses outlined above. Third, $MS_{eff.}$ is highest for intermediate rather than maximum durations for Happy, Calm, and Rapid, also contrary to expectation.

Thus, the only prediction that found support is $MS_{eff.}$ being smallest for short durations. But with the exception of Driving, this was only true for 0.5 and 1 s, with no further consistent increase between 2 and 16 s. Recall that the listeners in almost all conditions made a significant distinction between the MEs. Did they make this distinction on the same basis for 1 s and 16 s? This is the question we address in the next section, in which we also return to some of these results.

3.2. Consistency among music examples: Do people rate the same property across different durations?

$MS_{eff.}$ informs us about the overall variance attributable to the differences in ratings among ME, but it tells us nothing about the content of this variance. For example, if MEs 2, 4, 6, and 8 are rated high relative to MEs 1, 3, 5, and 7 for 2 s duration, but low for 4 s duration, this will make no difference in $MS_{eff.}$ If we think about the 10 ME ratings as a profile, we can gauge the similarity of different profiles by computing the correlation coefficient between them. This estimate will be insensitive to the absolute "height" of two profiles, but will be sensitive to their shape, i.e. the sequence of peaks and throughs. This similarity is even better expressed by the amount of explained variance – the squared correlation coefficient - in other words the variance common to the two sets of ME ratings. Figure 2 depicts this quantity for adjacent levels of duration, i.e. 0.5 and 1, 1 and 2, 2 and 4 etc. Each correlation is based on the mean ratings for each ME and duration across 15 listeners.

It is to be noted that these correlations express a different property than that tapped by previous measures: thus no relationship with the effects of duration or ME is apparent, for example. The plots for Driving and Good indicate that there is no agreement at all between 0.5 and 1 s, which might suggest that people give inconsistent or indistinct ratings for 0.5 s. However, the main effects of ME were significant for both these rating scales, which means rather that the listeners made different, but equally consistent, ratings. Rapid and Calm demonstrate the highest agreement across durations, albeit slightly smaller between 0.5 and 1 s.

Strikingly absent from the plot is the predicted gradual increase in agreement with longer duration, except for Happy. Instead, Driving and Good, and to some extent Calm, Rapid, and Walking suggest that the agreement is highest between sets of relatively short durations, and then decreases with longer durations. This interpretation is supported by the plots of $MS_{eff.}$ in Figure 1, which also reach their highest values for intermediate durations, and by the fact that $MS_{err.}$ does not generally increase for shorter durations. If anything, the tendency for $MS_{err.}$ is to increase for longer durations. In other words, listeners seem to agree more when there is less to take stock of.

Finally, the conspicuous upward leaps between 1 and 2 s, and between 2 and 4 s might indicate that listeners for these durations tap the same properties, but resort to other properties as duration increases.

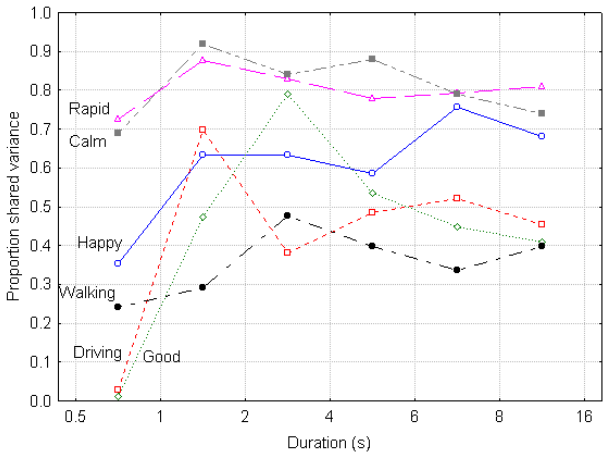


Figure 2: *Proportions of shared variance between adjacent levels of duration for the 10 mean ratings of music examples across 15 listeners.*

4. CONCLUSIONS

Beyond all differences between conditions reported above lingers an as yet uncommented null result. We had expected duration to have a substantial effect upon consistency in the range 2-16 s within which fundamental musical structure develops [3]. Instead, we found that ratings were equally consistent regardless of duration in this range. This consistency was underscored by similar ratings for some adjectives, such as Calm, Rapid, and to some extent Happy. It thus appears possible that listeners tend to rely on the kind of surface information that is equally well extracted from a 1 s sample as from a 16 s sample, even though the information of the latter is available. One may of course argue that these adjectives stand for low-level properties that can correctly be gauged with very short samples. However, that the agreement generally decreases between 8 and 16 s, for example, suggests the influence of some other mechanism.

A different concern is raised by cases of low agreement in ratings across all durations, exemplified by Driving, Good, and Walking. In this case, less than half the variance is explained by the ME "profile", while the differences between the MEs remain well beyond limits for statistical significance. Again, shouldn't access to more information with longer duration lead to greater agreement? At some point adding yet another stanza can not improve our judgement of how driving or good we find the music. Perhaps the durations up to 16 s are so short that vital information for the judgements is added for every level. This may be the case for Good, but seems less likely for Driving and Walking.

Thus, although the present results are not unequivocal, they prompt us to consider carefully the nature of adjective ratings. Watt and Ash [4] let listeners make forced choices between adjective pairs in response to 15-20 s music samples. Whereas some of these adjectives corresponded with spontaneous descriptions of music, such as joyful and sad, many pairs such as good vs. evil and female vs. male had no apparent relation to music. The important finding in light of the present results was that the consistency in the listeners' choice of words was unrelated to their applicability to music, but was substantially higher for words that apply to describing persons. One implication of this result is that the properties that people attribute to words may be more or less arbitrary. In other words, faced with the perhaps artificial task to choose the appropriateness of a word we are opportunists and base our choice on whatever property that seems to serve the demand of the task. This may be the mechanism responsible for increasing $MS_{err.}$ with duration. As more information becomes available judgements are based on more complex associations, and the more complex and remote these associations are, the larger are the individual differences.

In any case, the researcher with access to ratings for only one listening duration – which is the normal case – would feel confident in reporting this as a true effect of the experimental manipulation. Leaving the clearly problematic case in which examples are of different duration aside, we argue that some of the apparent consistency is in many cases spurious. Thus, further examination of the validity of adjective ratings is warranted.

5. REFERENCES

[1] Gabrielsson, A. & Juslin, P. N. (2002). Emotional expression in music. In H.H.Goldsmith, R. J. Davidson, & K. R. Scherer (Eds.), Handbook of affective sciences. New York: Oxford University Press.

[2] Schubert, E. D. (2001). Continuous measurement of self-report emotional response to music. In P.N.Juslin & J. Sloboda (Eds.), Music and emotion (pp. 393-414). New York: Oxford University Press.

[3] Gabrielsson, A. & Lindström, E. (2001). The influence of musical structure on emotional expression. In P.N.Juslin & J. Sloboda (Eds.), Music and emotion: Theory and research (pp. 223-248). New York: Oxford University Press.

[4] Watt, R. J. & Ash, R. L. (1998). A psychological investigation of meaning in music. Musicae scientiae, 2, 33-54.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

VERBAL DESCRIPTION OF MUSICAL SOUND TIMBRE IN CZECH LANGUAGE

Ondřej Moravec, Jan Štěpánek

Department of Musical Acoustics
Sound Studio of the Faculty of Music, AMU, Prague
moravec@hamu.cz, stepanek@hamu.cz

ABSTRACT

The words used for description of musical sound timbre were acquired. The research was carried out among people with active relation to the music (instrument players, conductors, composers, sound engineers etc.). Each of the respondents have filled out the questionary composed of two parts. In the first part personal background of the respondent (respondent profile) was collected, in the second part respondent wrote down particular expressions he uses for timbre description, synonymic and antonymic relations among them. Common frequency vocabulary and frequency vocabularies over the selected respondent classes were created from the data. Similarity of vocabularies and their differences in dependency on the respondent class are studied.

1. INTRODUCTION

The timbre of musical sound represents a subjective phenomenon which reflects the acoustic properties of musical signals in our consciousness in a complex manner, and which together with pitch and loudness represents a fundamental psychoacoustic quantity. The majority of contemporary researches regard timbre as multidimensional subjective quantity. In many timbre studies main features (categories, dimensions, factors) of some specific class of musical sounds are searched from the results of listening tests.

Method used for the interpretation of results is dependent on the method of listening test. Listened sound can be identified as like to a specific musical instrument [1]. When dissimilarities are judged in pairs of sounds, external interpretation of results may use either acoustic characteristics of listened signals [2] or spontaneous verbal descriptions collected in different listening of the same sound context [3]. It is also possible to select verbal attributes in advance [4]. The main problem in this case is the representativeness and method of selection: is it possible to use approved set of attributes or is it necessary to start new selection for every new context of sounds?

Our research group proposed a project leading to the selection of verbal attributes used for the description of timbre in Czech language. The basic aims of this project are:

- search for dimensions of perceptual timbre space and for their verbal descriptions,
- finding when all musicians share common perceptual space or certain groups (professional, age, etc.) use specific dimensions described by specific words.

The methods used in this study are based on sound-context free collection of verbal attributes and their evaluation used in pilot experiment with a group of violin players [5]. The results of the first stage of the project are summarised in this contribution.

2. METHOD

The research was carried out using questionary composed of two parts. In the first part respondents have described their personal profile – their age, years of musical experience, sex, the highest musical education, current and last music-related profession, position, instrument(s) they play etc. The second part of the questionary was divided into three subsections. In the first subsection respondents wrote down in free order the words and expressions which they use for the description of musical sound timbre, in the second one they wrote down groups of synonyms and in the third one they wrote down groups of antonyms.

The first part of the questionary was used for the description of population sample and for the discrimination of respondents into classes according to selected criteria. The second part as a whole was used for building of common vocabulary and partial vocabularies of individual respondent classes. The second and the third subsections of that part will be used for determination of word links in the next stage of the project.

The research was carried out among students and professors of Faculty of Music of Academy of Performing Arts in Prague, members of several symphonic and chambre orchestra from different regions of the Czech Republic, teachers from musical schools and people from recording studios. These groups overlap in some cases, e. g. professors of the musical faculty are often also members of chambre orchestra or solo instrument players. In the course of research several methods of population sample selection were tried [6]. The first one, contacting letter to the orchestra officials with request for cooperation remains with very poor response. Public addressing of musicians during orchestra rehearsal gave better but always unsufficient return of questionaries. The most effective way of collection was snowball technique and personal contact with individual respondents.

3. RESULTS

Finally 120 respondents sent back filled questionaries. Distribution of respondents according to their main profession is displayed in Table 1 (120 respondents = 100%). Some respondents have stated more then one profession. Profession with highest priority (if stated) or profession with lower overall occurence was used in such cases.

Table 2 shows distribution of respondents according to the instruments they play (109 respondents = 100 %). There are also included respondents who did not state playing the instrument as their main profession. Distribution is rather disproportionate but it approximately corresponds with some exceptions to the distribution of instrument players in symphonic orchestra. In the following analysis the division to only main instrument groups (bow, wind, keyboard) is used in order to reach sufficient amount of respondents in individual classes. Respondents wrote

down 1 964 different words and expressions in total. Each respondent wrote at least one its own "new" word, so only 230 words (i. e. 12 %) from common vocabulary have relative frequency greater then 5 % (the word was used at least by 6 respondents).

Main profession	Respondents	%
teacher	41	34.2
composer	6	5.0
conductor	3	2.5
instrument player	53	44.2
other	17	14.2

Table 1: *Respondents according to their main profession.*

Instrument	Respondents	%
violin	30	27.5
viol	9	8.3
violoncello	7	6.4
double bass	2	1.8
flute	6	5.5
oboe	1	0.9
clarinet	6	5.5
bassoon	3	2.8
trumpet	6	5.5
trombone	1	0.9
french horn	1	0.9
organ	4	3.7
piano	30	27.5
other	3	2.8

Table 2: *Respondents - instrument players according to instrument they play.*

Class		Respon-dents	Relative frequency of words			
			>= 5%	>= 25%	>= 30%	>= 50%
All		120	230	30	25	5
Sex	Male	87	214	36	23	6
	Female	33	237	35	25	6
Instrument	Bow	48	206	28	22	5
	Wind	24	173	36	21	10
	Keyboard	34	279	25	19	5
Age	< 30	37	274	27	20	5
	30 - 50	40	369	41	30	8
	> 50	43	186	31	26	8

Table 3: *Number of respondents in selected classes and number of words in partial vocabularies with relative frequency greater than indicated level.*

Expression	All			Rank in class		
	f_{abs}	f_{rel}	Rank	Bow	Wind	Keyb.
ostrý – sharp	94	78.3	1	1	1	1
temný – gloomy	79	65.8	2	5	2	2
měkký – soft	78	65.0	3	2	3	4
jasný – clear	75	62.5	4	3	4.5	4
sametový – velvety	61	50.8	5	7	6	4
kulatý – round	58	48.3	6.5	14.5	4.5	8
jemný – delicate	58	48.3	6.5	9.5	11.5	6.5
tupý – unpointed	55	45.8	8	11	8	10
tvrdý – hard	54	45.0	10	9.5	8	15.5
světlý – bright	54	45.0	10	16.5	11.5	10
drsný – harsh	54	45.0	10	12.5	19.5	6.5
sladký – sweet	53	44.2	12	4	15	22
plný – full	51	42.5	13	12.5	8	15.5
tmavý – dark	46	38.3	14.5	21.5	15	15.5
hrubý – rough	46	38.3	14.5	7	15	37
teplý – warm-hot	43	35.8	16	21.5	23.5	15.5
zářivý – radiant	42	35.0	17	16.5	45	22
vřelý – warm-hearty	40	33.3	18.5	7	31	78
čistý – clear	40	33.3	18.5	26	15	19
barevný – colored	38	31.7	20.5	19	45	12
zvonivý – ringing	38	31.7	20.5	39.5	19.5	10
průzračný – lucid	36	30.0	23.5	24	65.5	15.5
úzký – narrow	36	30.0	23.5	39.5	8	37
široký – wide	36	30.0	23.5	32	15	37
chladný – cool	36	30.0	23.5	14.5	31	61.5
kovový – metallic	34	28.3	26.5	32	31	22
studený – cold	34	28.3	26.5	27.5	31	37
svítivý – shining	32	26.7	28	48	19.5	37
zastřený – blurred	31	25.8	29	19	23.5	28.5
hladký – smooth	30	25.0	30	39.5	45	28.5

Table 4: *Thirty the most frequently used* w*ords from common vocabulary with their relative frequency and rank (including ties) in common and partial vocabularies of selected instrument group classes.*

Table 3 shows numbers of respondents in selected classes built on the following criteria:

- all respondents
- sex (male, female)
- instrument group (bow, wind and keyboard instruments)
- age (under 30, from 30 to 50, over 50 years)

and number of words with relative frequencies f_{rel} >= 5 %, f_{rel} >= 25 %, f_{rel} >= 30 %, f_{rel} >= 50 % for each class. All relative frequencies are computed from word occurence in partial vocabulary of respective class.

In Table 4 thirty the most frequent words from common vocabulary are presented. Each table row contains the Czech word together with its approximate English equivalent, absolute and relative frequency of word in common vocabulary and rank of occurence of word in common and partial vocabularies of classes of bow, wind and keyboard instrument players.

Table 5 gives overwiev of prominent differences in rank of words in partial vocabularies in comparison to common

vocabulary (class All) extracted from Table 4. Rank difference of words is greater then five positions.

Word rank against All	Class		
	Bow	**Wind**	**Keyboard**
more used	sweet rough warm-hearty cool blurred	narrow wide shining	ringing
less used	round bright dark clear ringing narrow wide shining smooth	delicate harsh warm, hot radiant warm-hearty colored lucid cool smooth	hard sweet rough warm-hearty narrow wide cool cold shining

Table 5: *Words from partial vocabularies with prominent difference in rank in comparison to common vocabulary (rank difference greater than five positions).*

Table 6 shows differences between partial vocabularies of selected instrument classes (bow versus wind instruments, bow versus keyboard instruments and wind versus keyboard instruments). All words in table have relative frequency greater then 25 % in class with more frequent use, words with difference in relative frequency in respective partial vocabulary greater then 20 % are listed.

Δf_{rel}	**Bow - Wind**	**Bow – Keyboard**	**Wind - Keyboard**
1. > 2.	warm, hearty (+ 20.8 %) sweet (+ 20.8 %)	warm-hearty (+ 33.0 %) sweet (+30.1 %) rough (+ 25.3 %) cool (+ 22.1 %) honeyed (+ 21.5 %)	narrow (+ 29.5 %) hard (+ 20.8 %) full (+ 20.8 %) warming (+ 20.5 %)
1. < 2.	narrow (- 29.2 %) round (-25.0 %)	dry (- 20.8 %) ringing (- 20.2 %)	- - -

Table 6: *Words from partial vocabularies with differences in relative frequencies greater then 20 %*

4. DISCUSSION

Sound-context free collection of verbal attributes used for the description of musical sound timbre resulted to rich set of words and expressions. Many respondents have stated not only the words describing sound timbre as property of musical instrument (e.g. *sharp*, *bright*, *round*), but also the words encompassing artistic expression into sound timbre (e. g. *sweet*, *warm*, *cool*) which may have influence on the definition of timbre space dimensions. Such words have usually smaller relative frequency then words describing sound timbre as instrument property so their effect is not extremely significant.

As Table 3 shows, the decrease of number of words with increasing frequency is well balanced among all indicated respondent classes. It can be mentioned from Table 4 that the four words most frequently used by all respondents are also most frequently used by respondents in indicated classes. On the other hand, there is no word written by all respondents.

Even if professional musicians have very complex experience with all kinds of musical sounds, results from Tables 5 and 6 suggest that respondents of individual classes prefere some expressions. For example bow instrument players use more frequently the words *sweet* and *warm-hearty* in contrast to both groups of wind and keyboard instrument players; the word *narrow* is preferred by wind instrument players.

5. CONCLUSIONS

The set of words selected in the questionary shows its richness and multiformity and also strong view of musical professionals. The most frequently used words were not a surprise for us; the majority of them is known from spontaneous verbal description of timbre. Even if simple translation into English (or any other language) cannot be precisely correct, their closest English equivalents are known from many published studies of musical sound timbre and thus they confirm internationality of the perception of musical sound and its verbal description.

Further stage of the project will be focused on the reduction of the number of representative expressions, definition of links among them and on more detailed description of differences among classes of music professionals.

6. ACKNOWLEDGEMENTS

The research was supported by Grant Agency of the Czech Republic (Project No. 202/02/1370).

7. REFERENCES

[1] Wedin, L., Goude, G. "Dimension analysis of the perception of instrumental timbre", Scandinavian Journal of Psychology, 13 (3), 228-240, 1972.

[2] Grey, J. M. "Multidimensional perceptual scaling of musical timbres", J. Acoust. Soc. Am., 61 (5), 1270-1277, 1977.

[3] Štěpánek, J., Otčenášek, Z., Melka, A. "Comparison of five perceptual timbre spaces of violin tones of different pitches", CD-ROM of Joint Meeting 137th ASA, 2nd EAA Forum Acusticum, 25th DAGA Berlin, 5aMUb5, 1999.

[4] Bismarck, G. von "Timbre of steady sounds: A factorial investigation of its verbal attributes", Acustica, 30, 146-159, 1974.

[5] Melka, A., Štěpánek, J., Otčenášek, Z. "Czech and German verbal description of violin sound properties: multidimensional analyses of survey data", ACUSTICA – acta acustica, 82, Suppl. 1, Antwerpen, 214, 1996.

[6] Ferjenčík, J. Úvod do metodologie psychologického výzkumu (in Czech language), Praha 2000.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

DEVELOPMENT OF A LANGUAGE FOR SPECIFYING SAXOPHONE TIMBRE

Arne Nykänen and Örjan Johansson

Department of Human Work Sciences
Division of Sound and Vibrations
Luleå University of Technology
971 87 Luleå, Sweden

arne.nykanen@arb.luth.se

ABSTRACT

When writing requirement specifications for musical instruments, or when writing music, specifying the sound of the instrument is of greatest interest. Today's notation of music leaves many aspects of sound open for interpretation. This implies that it is necessary to know a music style and time typical ideal to be able to perform music in the way intended by the composer.

To investigate how musicians use verbal descriptions of musical sounds, interviews have been made with saxophone players. The results have been analysed with respect to how frequent specific words have been used. Sounds described by the most popular words have been analysed to search for physical aspects of the sounds which could be coupled to the description. Since a large number of words have been used to describe the sounds, and only a few of them were used frequently by different musicians, the conclusion is that there is a lack of common language to describe sounds. New ways to describe sounds are desirable, since a larger set of descriptions will increase the possibility to choose a convenient set of descriptions for musical sounds. Therefore, one new way to describe sounds, based on vocal mimicking, is suggested and evaluated.

1. INTRODUCTION

Interviews with saxophone players have shown that they use a large number of words to describe the timbre of the saxophone. Most of the words seem to be used by individuals, but during the interviews, 9 adjectives occurred more frequently.

When studying typical frequency spectra of the saxophone, they show some similarities to the human voice. The spectrum has a strong fundamental, it is falling approximately 10 dB/octave, and has strong formant-like characteristics. The human voice also has a strong fundamental, and the spectrum is falling approximately 6 dB/octave. The vowel sounds of the human voice are to a large extent determined by the formants [1]. Hence, the saxophone sound could be expected to resembles vowel sounds. Therefore, the use of vowel-similes have been tested as descriptors of the saxophone sound.

The aim of the investigation was to find correlations between adjectives frequently used by saxophone players, the use of vowel-similes, and acoustically measurable characteristics of the saxophone tone.

2. METHOD

2.1. Interviews with saxophone players

To find which adjectives that are commonly used by Swedish saxophone players to describe the timbre of the saxophone, ten subjects have been interviewed. The interviews were made in the conditions and the order stated below:

1. The respondent was asked which adjectives he/she uses to describe the timbre of the saxophone, without playing or listening to any examples.

2. The respondent was asked to pick the most convenient adjectives to use to describe the timbre of the saxophone from a list, without playing or listening to any examples. The adjectives in the list had been selected from scientific papers considering timbre of musical instruments [2], [3], [4], [5], [6], [7].

3. Different recordings of saxophone solos were played to the respondent, who was asked to describe the timbre by own words.

4. Different synthetic saxophone sounds were played to the respondent, who was asked to describe the timbre by own words.

5. The respondent was asked to describe the timbre he/she is striving for, by own words.

6. The respondent was asked to play different saxophones and describe the timbre of them.

The interviews were performed in Swedish. The adjectives most commonly used among the subjects can be found in Table 1, together with a translation to English.

Swedish	English translation
Stor	Large
Fyllig	Rounded
Rå	Rough1
Varm	Warm
Mjuk	Soft
Nasal	Nasal
Kärnfull	Coreful
Vass	Sharp1
Botten	Bottom

Table 1: *Most common adjectives in use among the saxophone players participating in the interview.*

2.2. Recoding of typical saxophone sounds

Saxophone sounds have been recorded with an artificial head (Head Acoustics HMS III), and played back to the subjects in the listening tests by using equalised headphones. To get sounds that differ in characteristics, recordings have been made by two different saxophone players, each playing on two different alto saxophones, giving four possible combinations. The saxophone players were asked to play a tune, to achieve playing conditions that are similar to ordinary music playing.

2.3. Listening tests

Two groups of subjects participated in the listening tests. One group consisted of 10 saxophone players, and the other group of 6 subjects who were experienced listeners but did not play the saxophone. The listening test was divided into two parts:

In **listening test 1**, a 7 second long part of a melody was played to the subject. The subjects were asked to estimate the timbre of one specific note (Ab_4, f=415 Hz) in the middle of the melody. 4 different sounds were used, one of each possible combination of saxophone and saxophone player.

In **listening test 2**, a 0.5 second long cut from one single note was played (Ab_4, f=415 Hz). The cut was placed so that both the attack and the decay was removed. Instead, the tone was ramped up and down with a ramp-up and –down time of 0.075 s. 12 different sounds were used, three of each possible combination of saxophone and saxophone player.

In each part, two different questionnaires were used. In questionnaire 1, the subject was asked to estimate how well the timbre quality was described by each of the 9 adjectives in Table 1. In addition to this, the subject was also asked to estimate the total impression of the timbre of the judged note.

In questionnaire 2, the subject was asked to estimate how well each of three different psycho-acoustic quantities and 9 vowel-similes described the sound (Table 2).

Swedish	English translation
Skarp	Sharp2
Rå	Rough2
Tonal	Tonal
likt A, som t.ex. i ”mat”	A-like [ɑ]
likt E, som t.ex. i ”smet”	E-like [e]
likt I, som t.ex. i ”lin”	I-like [i]
likt O, som t.ex. i ”ko”	O-like [u]
likt U, som t.ex. i ”lut”	U-like [ʉ]
likt Y, som t.ex. i ”fyr”	Y-like [y]
likt Å, som t.ex. i “båt”	Å-like [o]
likt Ä, som t.ex. i ”färg”	Ä-like [æ]
likt Ö, som t.ex. i ”kör”	Ö-like [œ]

Table 2: *Psycho-acoustic quantities and vowel-similes used in questionnaire 2. To give a hint of how the vowels should be pronounced, an example of a word was given in the following way: like A, as for example in “car”. In the English translation, the phonetic description of the vowel has been given instead.*

The estimations of the perceptual qualities were made on an 11-grade scale, ranging from e.g. “not at all sharp” (=0) to “extremely sharp” (=10), except for the estimation of the total impression, which was made on an 11-grade bi-polar scale ranging from “extremely bad” (=-5) to “extremely good” (=+5).

2.4. Measurements of physical characteristics

For each sound stimulus, the following acoustical characteristics have been measured: loudness (according to ISO532B), roughness [8], sharpness [9], tonality [8], frequencies for 6 formant-like humps in the frequency spectrum, the spectral width, and the fundamental frequency. In addition to these quantities, the extent of formant-like characteristics has been estimated by visual inspection of smoothed FFT-spectra (see Table 3).

2.5. Multivariate analysis

Principal Component Analysis (PCA) was used to find relationships between acoustical properties, psycho-acoustic indices and the estimated perceptual qualities. The estimated perceptual qualities listed in Table 1 and 2, and the properties in Table 3 was used as variables.

Name	Description
Sound	Type of sound: 0=**listening test 1** 1=**listening test 2**
Fund	Fundamental frequency
F-char	Formant like characteristic: 0=hardly no formant like characteristic at all 1=week formant like characteristic 2=strong formant like characteristics
F1	Frequency of formant 1
F 2	Frequency of formant 2
F3	Frequency of formant 3
F4	Frequency of formant 4
F5	Frequency of formant 5
F6	Frequency of formant 6
Width	Spectral width: The order number of the highest partial with a SPL above 20 dB
Loudness	Loudness (according to ISO532B)
Roughness	Roughness [8]
Sharpness	Sharpness [9]
Tonality	Tonality [8]
Listener	Type of litener: 0=listener is not saxophone player 1=listener is saxophone player
Sax	Type of saxophone: 0=saxophone of make 0 1=saxophone of make 1
Subject	Type of subject 0=musician 0 1=musician 1

Table 3: *Names and descriptions of variables used in the principal component analysis, in addition to the subjective qualities estimated in the listening tests.*

3. RESULTS

The PCA of data from **listening test 2** gave 6 significant dimensions. The significance, together with an interpretation of the four most important dimensions can be found in Table 4.

Dim	r^2X	Interpretation
1	0.16	Best described by the psycho-acoustic index, Sharpness, which have a negative loading. "Sharp", "Rough", "Soft" and "Warm" are the most important estimated qualities. A bi-polar scale from "sharp" to "soft" gives a good description of this dimension.
2	0.14	Best described by the psycho-acoustic indices, Tonality and Roughness. All vowel similes have negative loading in this dimension, indicating that a vowel-like sound is characterised by high Tonality and low Roughness. The adjective "Coreful" (Swe. "Kärnfull") seems to be used to describe Tonality. The differences between the saxophones are best described by this dimension.
4	0.08	The fourth dimension is mainly described by formant frequency 1 and 3. The differences between the musicians is best described by this dimension.
6	0.05	The sixth dimension is mainly described by formant frequencies 1, 2, 3 and 6. The saxophone used is strongly correlated with this dimension.

Table 4: *The significance of the 4 most important dimensions found by PCA of the data from* **listening test 2**, *together with an interpretation of each dimension. The variables listed in Table 1, 2 and 3 have been used.*

A PCA of the combined data from **listening test 1** and **2**, was made to check how the results were influenced by the type of sound. The result showed that the two main dimensions were independent of the type of sound. Only two dimensions differed between the combined case and the case when only **listening test 2** was used. These dimensions were mainly built up by the type of sound and vowel-similes. The vowel-similes are probably influenced by the length of the judged note and the context in which it is presented (e.g. attack, decay, the preceding- and the following note). To simply the interpretation of the results, the remaining discussion will be based on the PCA of the data from **listening test 2**.

In Figure 1, the loadings of the two most significant dimensions are shown. Figure 2 shows the dimensions that best describe differences between the two saxophone players, and Figure 3 shows the dimensions that gives the description of the differences between the two saxophones.

4. DISCUSSION

The two most significant dimensions in the analysed data (Figure 1) are well described by the psycho-acoustic indices Sharpness (dim. 1) and Roughness (dim. 2). The perceptual qualities that best match dimension 1 are the Swedish adjective "vass" (Eng. "sharp"), and its opposite "mjuk" (Eng. "soft"). The Swedish adjective "skarp" (Eng. "sharp"), which is commonly used in psycho-acoustics seem to have the same meaning as "vass". Dimension 2 is best described by the perceptual quality "kärnfull" (Eng. "coreful"). The Swedish adjective "rå" (Eng. "rough"), which is commonly used in psycho-acoustics, seems to have a different meaning for listeners who are not trained in psycho-acoustics. "Rå" has almost no loading on the roughness dimension (dim 2), but a quite strong loading on the sharpness dimension (dim. 1).

The perceptual qualities "fyllig" and "varm" (Eng. "rounded" and "warm"), as well as all vowel-similes and total impression, can be described by a combination of low sharpness and low roughness.

Dimension 2 and 6 (Figure 3) gives the best description of the differences between the two saxophones. Dimension 2 can be described by the bi-polar scale, roughness-tonality, and dimension 6 reflects the differences in the frequencies of formants 1, 2 and 6. Dimension 6 is mainly related to acoustically measured quantities. The perceptual quality that fits best is "Ä-like".

The two saxophone players differs mainly in dimension 2 and 4 (Figure 2). In dimension 4, the influence of the player is most pronounced. The dimension is correlated with formant frequency 1, 3 and 6. There are no perceptual qualities that explains dimension 4 well. The best description is given by the vowel-similes, "O-like", "Ä-like" and "Å-like". Dimension 2 only reflects differences between the musicians slightly.

Roughness seems to represent two dimensions (dim. 2 and 3). An explanation to this could be that the roughness of the samples in this experiment is not normally distributed. There are two groups of sounds, one with lower Roughness and one with higher.

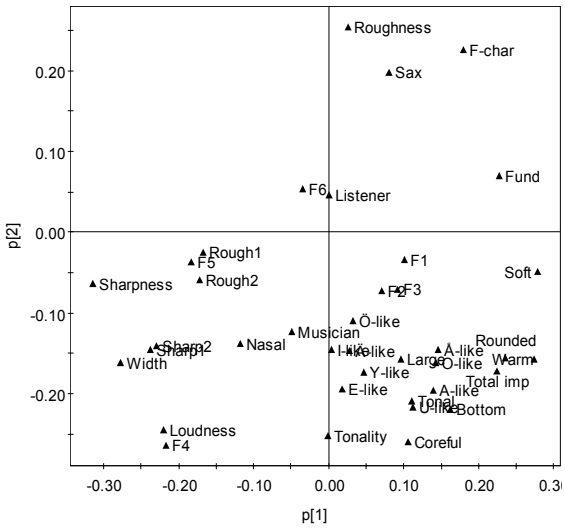


Figure 1: *The two most significant dimensions in the PCA of* **listening test 2**

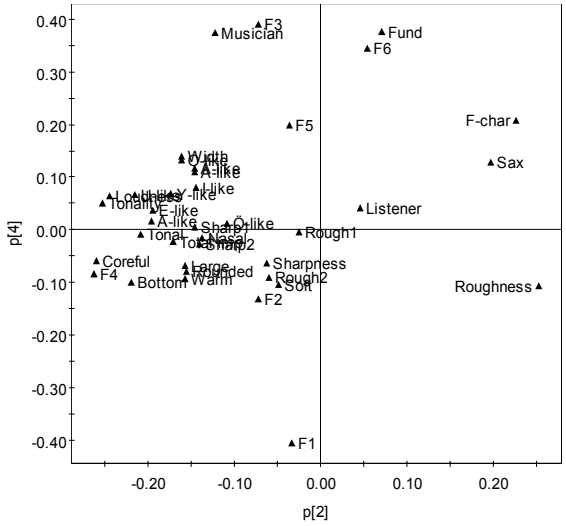


Figure 2: *Dimensions 2 and 4 reflects the main differences between the two saxophone players.*

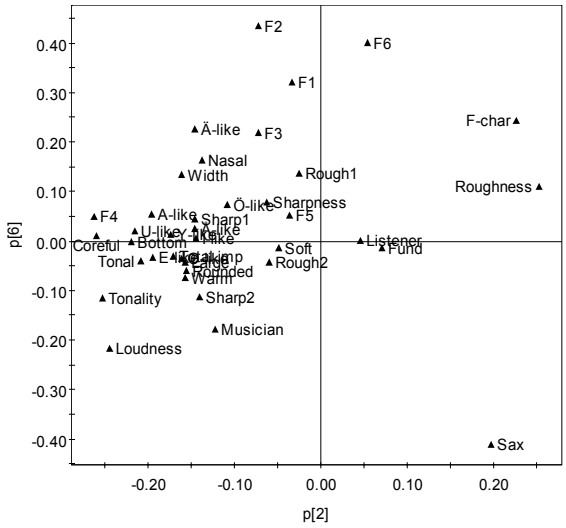


Figure 3: *Dimensions 2 and 6 reflects the main differences between the two saxophones.*

5. CONCLUSIONS

The psycho-acoustic indices, Sharpness, Roughness and Tonality, describes a considerable part of the timbre of the saxophone. The psycho-acoustic index, Sharpness, correlates well with the perceptual qualities "skarp" and "vass" (Eng. "sharp"). The psycho-acoustic indices, Roughness and Tonality, are each others opposites, mainly due to how they are defined. The perceptual quality that has the strongest correlation with this dimension is "kärnfull" (Eng. "coreful"). The adjective "rå" (Eng. "rough") has week correlation with this dimension. This indicates that either training is required to be able to describe sounds by the use of this adjective, or that a more convenient description has to be found.

The saxophone has a rough timbre. Still, a good total impression is correlated with tonality. This indicates that a good total impression requires that the timbre has a good balance between roughness and tonality.

The two saxophone players in this experiment did not differ in the Sharpness or Roughness dimensions. Still there is a clear difference between the timbres of the two players. This difference could be explained by differences in the formant- like characteristics. There seem to be some correlation between vowel-similes and the formant-like characteristics, but it is ambiguous. One reason for this could be that, even though the formant-like characteristics of the saxophone tone reminds about the human voice, it still differs from the vowels. The formant-like characteristics probably builds up patterns that are easy to recognise, even if they are difficult to describe. A suggestion for further investigations, is to study the effect of training on the ability to describe and distinguish between timbres with different formant-like characteristics.

6. REFERENCES

[1] Rosner, B.S. and Pickering, J.B., "Vowel Perception and Production", Oxford University Press, Oxford, 1994

[2] Von Bismarck, G., "Timbre of steady sounds: a factorial investigation of its verbal attributes", Acustica, Vol. 30, 1974, pp. 146-159.

[3] Pratt, R.L., Doak, P.E., "A subjective rating scale for timbre", J. Sound and Vibration, Vol. 45, 1976, pp. 317-328.

[4] Gridley, M.C., "Trends in description of saxophone timbre", Perceptual and Motor Skills, Vol. 65, 1987, pp. 303-311.

[5] Kendall, R.A and Carterette, E.C., "Verbal attributes of simultaneous wind instrument timbres. 1. VonBismarck adjectives", Music Perception, Vol. 10, 1993, pp. 445-468.

[6] Rioux, V., "Methods for an objective and subjective description of starting transients of some flue organ pipes - integrating the view of an organ-builder", Acta Acustica, Vol. 86, 2000, pp. 634-641.

[7] Stephanek, J., Otcenasek, Z., "Psychoacoustic aspects of violin sound quality and its spectral relations", Proc. 17th International Congress on Acoustics, Rome, Italy, 2001

[8] Aures, W., "Ein Berechnungsverfahren der Rauhigkeit", Acustica, Vol. 58, 1985, pp. 268-280.

[9] Von Bismarck, G., "Sharpness as an attribute of timbre of steady sounds", Acustica, Vol. 30, pp. 159-192.

Influence of duration of tone stationary part on perception of starting transient

Zdeněk Otčenášek, Jan Štěpánek, Václav Syrový

Sound Studio of the Faculty of Music, Academy of Performing Arts Prague
otcenasek@hamu.cz, stepanek@hamu.cz, syrovy@hamu.cz

ABSTRACT

Sounds of different types of organ pipes were used for the study of an influence of duration of the tone stationary part on starting transient perception. Long sounding tone recordings were truncated from the end in successive steps from original duration about 2 s to the length of 80 ms. Attack transients of tones remains unchanged, truncated tone ends were equally modified with level decrease to silence (fade out). Sets of tones derived from the same original tone were judged on dissimilarity in pairs, focusing on perception of only initial transient part. Results are discussed for tones with various types of transient and stationary part.

1. INTRODUCTION

When getting the stimuli for psycho-acoustic listening tests ready for musical acoustic purposes, it is often necessary to manipulate the recordings of the real musical instrument sounds. One kind of the manipulation can be the shortening of the time duration of the sounds, the other changing of the initial and final transient parts. Using the manipulation not suitable for the case, the results of those tests might be affected by a systematic error. For example when the sound is too short, then the timbre perception mostly defines the transient part of a sound and can even totally reduce the perception of stationary part.

While creating the project with an aim of organ pipe voicing, it was required to compare tones having various transient parts and also judge the stationary part. The question was how long could be those tones for such listening test.

The aim of this work is the subjective evaluation of the effect of the tone shortening on the perception of the initial transient of the tone.

2. METHOD

The recordings of the concert organs in Martinu hall were used for this experiment. The recordings were made for stops of the 1st manual (Spitzgambe 8´, Trompete 8´, Superoctave 2´) and the selected types of transients (see Figure 1 to 4). The recordings were made using two omnidirectional microphones BK 4005 with a base of 175 mm. The microphone was situated inside the organ in the position where the organmaker stays during the pipes voicing. The sounds were digitalized and saved into a PC in a format of 44.1 kHz / 24 bits.

The duration of the recorded tones was about 4 seconds. For experiment they were shortened on 2.0, 1.6, 1.3, 1.0, 0.7, 0.5, 0.4, 0.3, 0.25, 0.2, 0.18, 0.16, 0.15, 0.14, 0.13, 0.12, 0.11, 0.10, 0.09, 0.08 s (see Figure 5).

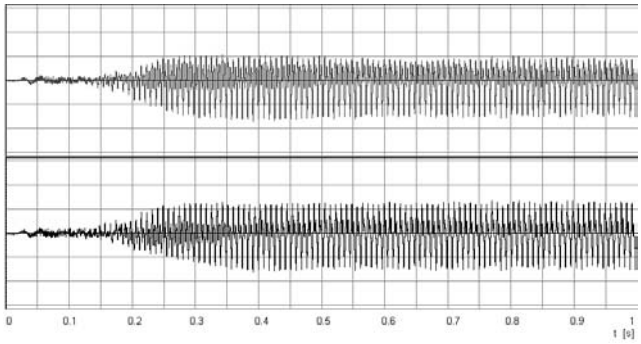


Figure 1: *Spitzgambe 8´, tone B2.*

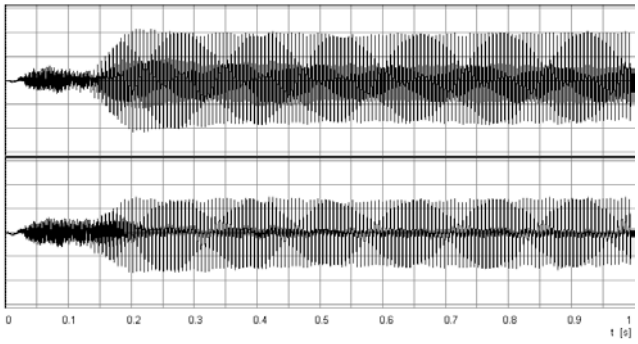


Figure 2: *Spitzgambe 8´, tone E3.*

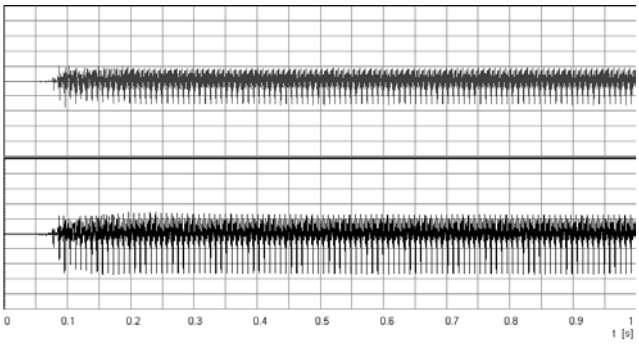


Figure 3: *Trompete 8´, tone B2.*

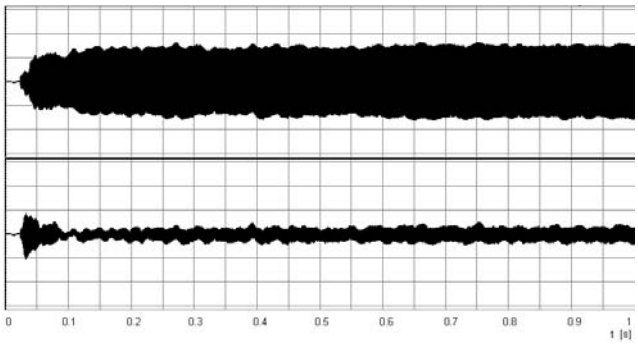


Figure 4: *Superoctave 2′, tone B5.*

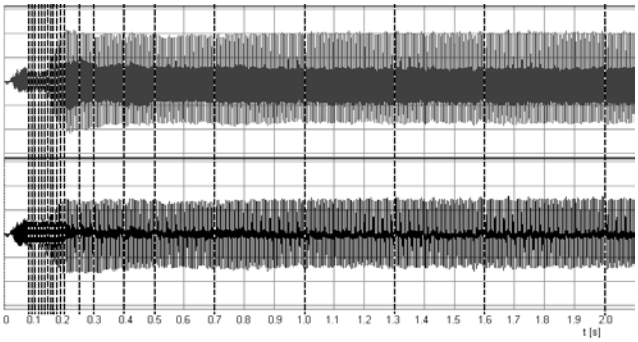


Figure 5: *Positions where the shortening of the tone was made.*

The final part of all tones was treated with a 30 ms fade out (see Figure 6).

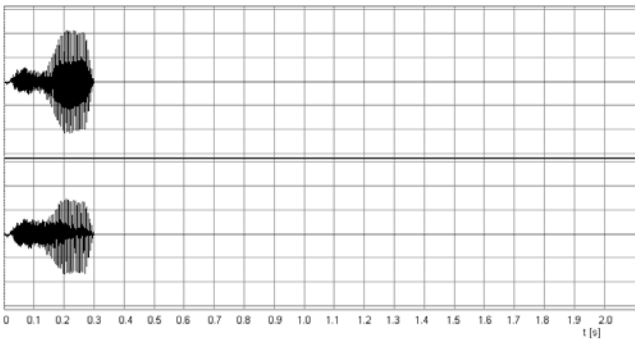


Figure 6: *An example of the 30 ms fade out, Spitzgambe 8′, tone E3, duration 0.3 s.*

The estimation of physical transient duration was made from time course of each tone (see Figure 7).

	physical transient duration
Spitzgambe 8′, tone B2	0.45 s
Spitzgambe 8′, tone E3	0.22 s
Trompete 8′, tone B2	0.19 s
Superoctave 2′, tone B5	0.11 s

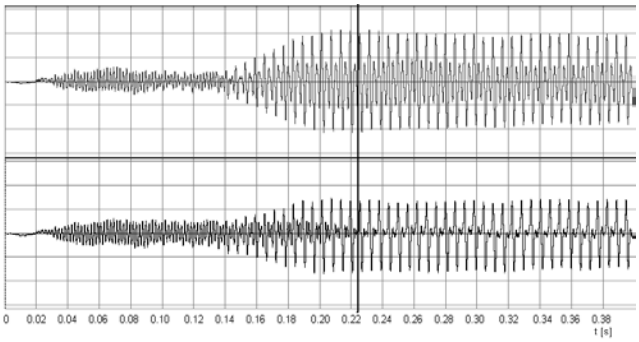


Figure 7: *An example of physical transient duration estimation (220 ms), Spitzgambe 8′, tone E3.*

The listening test was done with Sennheiser headphones HE 60 and headphone amplifier HEV 70. The loudness was the same as during the recording. All pairs of the shortening modifications of each tone were listened. The dissimilarities in the initial transient were judged on scale from 0 (no dissimilarity) to 5 (maximum dissimilarity).

3. RESULTS

For the identification of an appropriate tone length, to be used in future listening tests of the comparison of the initial transient, such criterion was established that the tone elongation above this appropriate length does not influence the value of the judgement. On a graph of results in Figures 8 to 11 this length is a start of an area with the same shade of gray.

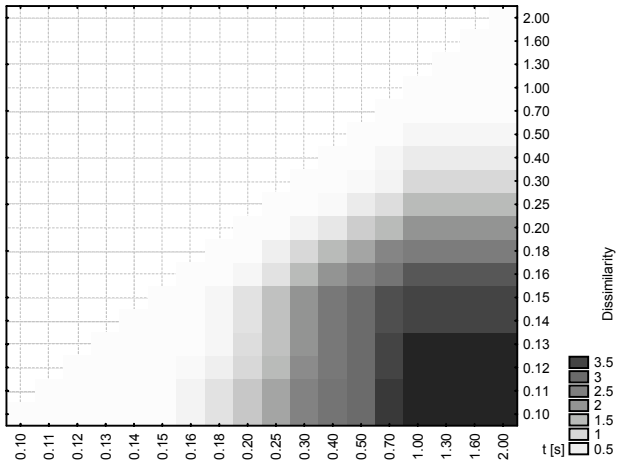


Figure 8: *Resulting dissimilarity pair test judgment for Spitzgambe 8′, tone B2.*

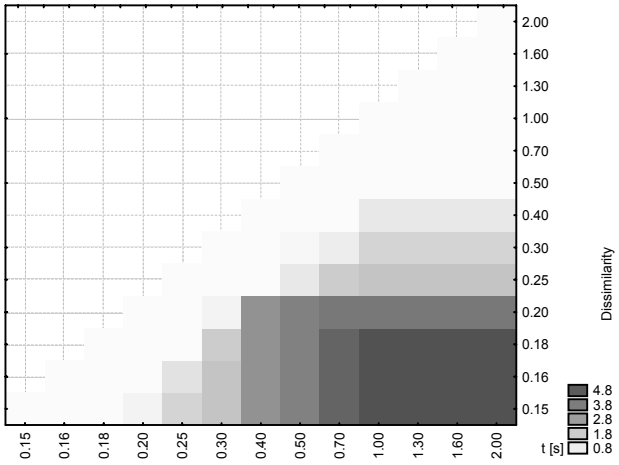


Figure 9: *Resulting dissimilarity pair test judgment for Spitzgambe 8′, tone E3.*

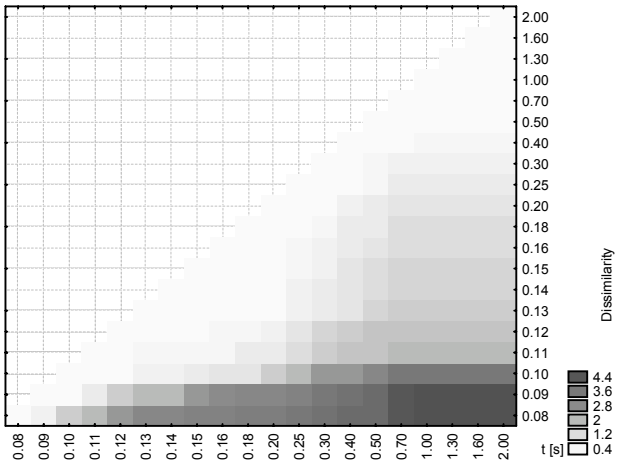


Figure 10: *Resulting dissimilarity pair test judgment for Trompete 8′, tone B2.*

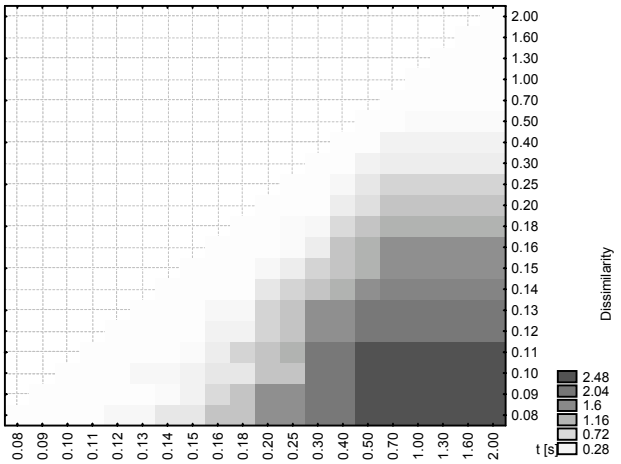


Figure11: *Resulting dissimilarity pair test judgment for Superoctave 2′, tone B5.*

4. DISCUSSION

Even though none of the physical transients exceeded length of 0.45 s (typically 0.2 s), the stabilization of judgement began mostly at a tone with the length of 0.7 s, even for the length of the transient 0.45 s the dissimilarity judgement is stabilized at 1.0 s. The respondent justifies his judgement by explaining that only by a longer tone he was sure with the end of the transient.

At a tone length comparable with the length of the duration of the physical transient, the dissimilarity of the perception is smaller with the shorter tones, because in both compared tones the expectation remains that the tone transient may continue after its end.

Psychologically the perception of the transient is longer about such duration of the stationary part at which the listener finds out that the initial transient was finished. "About how much longer" and "how dependent and on what" would be possible to state on the basis of another, more complex experiment. For the preparation of a tone for listening tests, this has to be taken into account.

5. ACKNOWLEDGEMENTS

The Ministry of Education and Youth of the Czech Republic (Project No. 511100001) supported the research.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

PERCEIVED INFLUENCE OF CHANGES IN MUSICAL INSTRUMENT DIRECTIVITY REPRESENTATION

Felipe Otondo and Brent Kirkwood

Ørsted-DTU, Acoustic Technology, Technical University of Denmark, build. 352, Ørsteds plads, DK-2800, Kgs. Lyngby, Denmark

fo@oersted.dtu.dk and bck@visi.com

ABSTRACT

The directivity representation of musical instruments in room acoustic simulations has shown to be significant in the distribution of the calculated acoustical parameters in a room. Different directivity representations of musical instruments were used to create pairs of room acoustic auralizations in order to test for perceived changes in the sound. Listening tests were designed and conducted with an emphasis on the perception of the acoustical attributes of room simulations. Results show that changes in the directivity representation of the source can influence the perceived sound in auralizations and that these perceived changes are more pronounced in terms of loudness and reverberance

1. INTRODUCTION

The radiation of musical instruments has been studied mostly considering frequency averages of the directivity over the performing pitch range of tones [1]. Recent studies have shown that the directivity of musical instruments change, in each frequency band, for different tones played [2, 3, 4] and investigations using room auralizations have shown that changes in the directivity of sound sources can be audible by a listener in a room [5]. The aim of this study has been to investigate the perceptual importance of the changes in the directivity of musical instruments in a room using room acoustic simulations as the reference.

2. DIRECTIVITY MEASUREMENTS

2.1. Method

The instruments used for the directivity measurements were a trumpet, a Bb clarinet and a French horn. The instruments were measured at a similar musical intensity using a simultaneous 13 channel setup in an anechoic chamber, at steps of 45° in the horizontal and vertical axes, 24-bit quantisation, 44.1 kHz sampling rate, in octave bands from 125 to 8000 Hz, and at a distance of 1.5 meters from the physical centre of the source. Short samples of isolated tones over the whole pitch register were used to build the averaged directivity at different octave bands and also to obtain specific directivities for particular tones in the same octave bands.

2.2 Results

Measurements showed variations between the directivities of individual tones in a given octave band. These variations were greater for the clarinet and the French horn than for the trumpet. Figure 1 and 2 show the measured directivities for three specific tones and the calculated averaged directivity over the whole pitch range at the 1000 Hz octave band. Results are shown for both, the horizontal and vertical axes.

3. DIRECTIVITY IN A ROOM

3.1. Room acoustic simulations

As a way to investigate how the changes of the directivity for the studied musical instruments affect the sound in a room and relate these changes to the perceived changes in auralizations, room simulations were carried out with one specific measured directivity and the calculated averaged for each instrument. The simulations were done using the software ODEON [6], assuming the same sound power for the sources, in a model of the Elmia concert hall in Jönköping, Sweden. The acoustical parameters considered for the simulations were the sound pressure level (SPL), the clarity factor (C80), the lateral energy fraction (LF80), and the early decay time (EDT). Two different directivity representations were used for each instrument as a basis of comparison for the simulations: a particular directivity of the instrument when playing a specific tone and the octave band averaged directivity over the whole pitch performing range. The specific tones selected for each instrument were: C4 for the trumpet, B3 for the French horn and C#4 for the Bb clarinet.

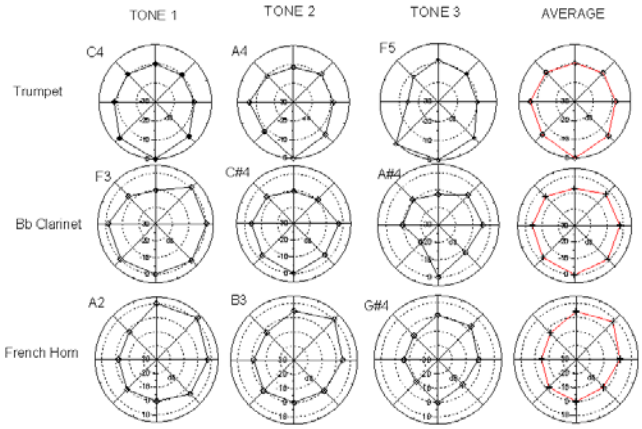


Figure 1: *Specific and averaged directivities for the trumpet, the clarinet and the French horn at the 1 kHz octave band in the horizontal axis. The fronts of the instruments are facing downwards, and in the case of the French horn, the bell is facing right. The averaged directivities correspond to the average over the whole pitch performing range of the instrument.*

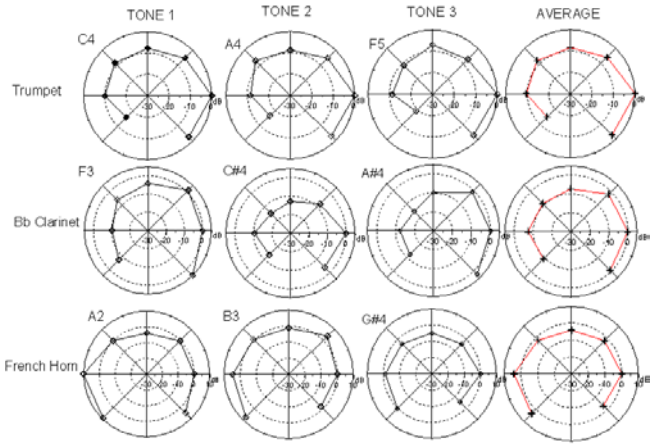


Figure 2: *Specific and averaged directivities of the trumpet, the Bb clarinet, and the French horn at the 1 kHz octave band in the vertical axis. The right sides of the figures correspond to the fronts of the instruments. The averaged directivities correspond to the average over the whole pitch performing range of the instrument.*

3.2. Results

The results of the room simulations indicate distinctive sound distributions in the room that can be linked to the original directivities of the instruments. Figure 3 shows grid-responses of the SPL in the 1000 Hz octave band simulated in the hall for the three instruments using the directivity of a specific tone and the averaged directivity over the whole pitch performing range. The difference in the spatial distribution of the acoustical parameters in the room showed to be greatest for the SPL and C80, less pronounced for LF80, and negligible for the EDT.

In order to compare the magnitude of the acoustical parameters at a certain position, for each instrument, with the two directivity representations, a location in the audience was chosen at 9 m from the source, 3 m to the right of the central axis of the hall. This position was also considered for the listening tests with auralizations in the final stage of the study.

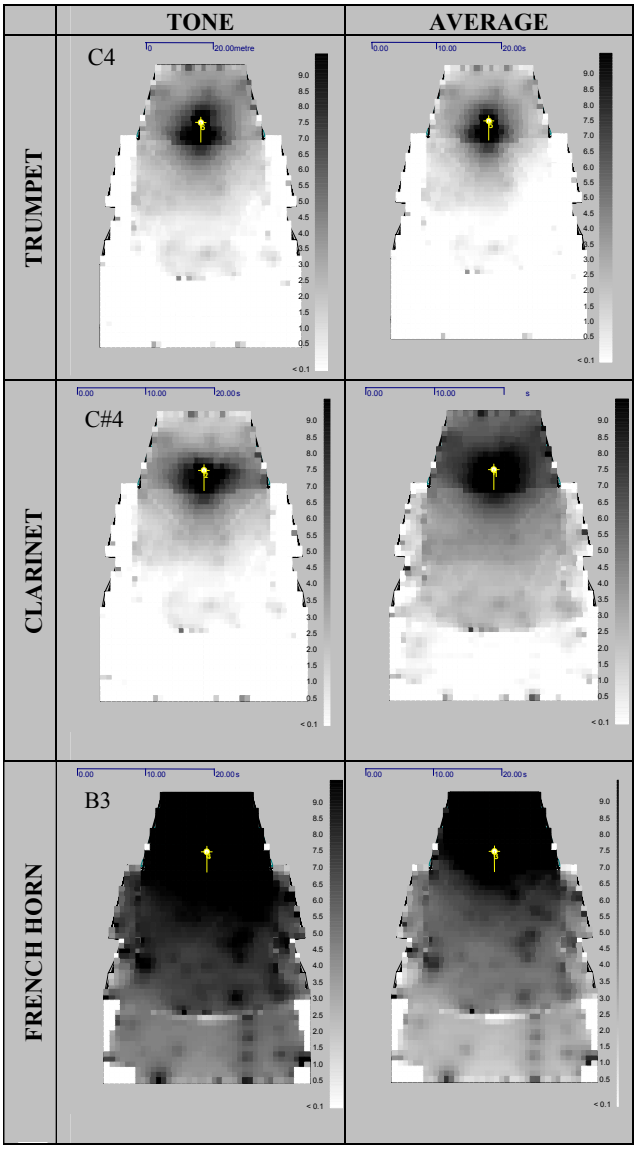


Figure 3*: Grid-responses of the sound pressure level produced by the three musical instruments in the 1 kHz octave band, simulated in the Elmia concert hall with the directivity of a tone and the averaged directivity. The scale is relative and shown from 0 to 10 dB with white and black as the minimum and maximum values, respectively.*

Table 1 shows the results in three octave bands of the simulation of the acoustical parameters at the selected position in the room for each directivity of the three instruments. In most of the cases the results of the comparisons proved that there existed a considerable difference in the acoustical parameters for the different directivity representations. As in the spatial distribution of the acoustical parameters, the difference are greater for the SPL and the C80, less noticeable for the LF80, and minor for the EDT.

Trumpet						
Frequency (Hz)	**500**		**1000**		**2000**	
Directivity Representation	C4	Average	C4	Average	C4	Average
SPL (dB)	13.7	6.7	8.6	6.8	7.2	6.6
C80 (dB)	0.1	0	1	2.1	2.8	2.0
LF80	0.28	0.20	0.14	0.2	0.16	0.18
EDT (s)	1.78	1.79	1.60	1.52	1.36	1.41
Bb Clarinet						
Frequency (Hz)	**500**		**1000**		**2000**	
Directivity Representation	C#4	Average	C#4	Average	C#4	Average
SPL (dB)	1.3	7.1	2.8	5	7.4	7.4
C80 (dB)	3.6	0.3	1.9	1.6	0.2	2.1
LF80	0.10	0.21	0.10	0.19	0.16	0.17
EDT (s)	1.29	1.82	1.49	1.58	1.54	1.4
French Horn						
Frequency (Hz)	**500**		**1000**		**2000**	
Directivity Representation	B3	Average	B3	Average	B3	Average
SPL (dB)	4.2	4.4	9.7	7.3	15	4.3
C80 (dB)	0.8	0.8	-0.3	-0.1	-0.5	1.2
LF80	0.16	0.17	0.22	0.21	0.32	0.17
EDT (s)	1.79	1.79	1.97	1.92	1.77	1.61

Table 1: *Room Acoustic parameters simulated for the three musical instruments in the position of the listener.*

4. LISTENING EXPERIMENTS

4.1. Method

Listening experiments were designed with the goal of testing the audibility of differences between the two directivity representations of each instrument. Using the software ODEON [6] pairs of room acoustic auralizations were created using these directivity representations in the previously described position in the Elmia concert hall. Short melodies of approximately 10 s played on the three musical instruments were recorded anechoically and used for the comparisons in the listening tests.

A forced choice paired comparison method was used for the listening tests, which consisted of eleven test subjects. The subjects were presented with pairs of auralizations, created for each of the instruments, through Sennheiser HD 250 headphones. After a training session the listeners were asked to make a qualitative comparison between the two auralizations and select one according to five perceived acoustical features: loudness, perceived reverberance in the hall, clarity, ease of source localisation and naturalness of instrument timbre. Each of the comparisons was tested twice as a part of a balanced random test sequence.

4.2. Results

The results of the experiments were first analysed using the McNemar test in order to determine the level of randomness of the data [7]. This method was used to separate the data provided by test subjects who were inconsistent from those who were consistent in their answers. A test was also used to determine which results were statistically significant compared to a threshold of 0.95 [8].

The results of the three instruments were analysed separately for each of the five parameters. Table 2 shows the significant results according to the McNemar test.

Instrument Parameter	**Trumpet**	**Bb clarinet**	**French Horn**
Loudness	C4	Average	B3
Reverberance	C4	Average	No preference
Clarity	No preference	No preference	B3
Localization	No preference	No preference	No preference
Timbre	No preference	No preference	No preference

Table 2: *Statistically significant results of the listening tests, according to the McNemar test. The directivity pattern that was found to be "favoured" for each parameter is shown as the average directivity or the directivity of a tone. In the cases where no conclusions could be made from the analysis, "No preference" is stated.*

As shown in Table 2, the results show that all test subjects could hear a difference in loudness between the two compared auralizations. The audibility of the changes of in reverberance proved to be significant for two of the three instruments and the audibility of the clarity was significant only for the French horn. The audibility of differences in timbre and the localisation of the source did not prove to be significant.

4.3. Discussion

The results of the listening tests have shown that there are changes in the directivity representation that were perceived by the listeners. Comparing the results of the simulation of the acoustical parameters of Table 1 with the results of the listening tests of Table 2, one can see that the only parameters that proved to have the same general tendencies in both cases were the sound pressure level and the loudness. The comparison between the results for the perceived reverberance and the simulated EDT showed some correspondence that could also be linked with the SPL. However these comparisons between the subjective reverberance parameter and the calculated EDT parameter are unclear due to the EDT fluctuating across frequency bands. The

level of the sound could have helped the subjects to hear better the sound decays in the room. The discrimination of the clarity of the listening tests for the French horn did not prove to be strongly correlated with the simulations of the C80 in the room. Not much information can be obtained from the other simulated acoustical parameters that could be related to the results of the listening tests.

5. CONCLUSIONS

Room acoustic simulations using averaged and specific directivities of tones of musical instruments have shown that the directivity has a direct influence on the distribution of acoustical parameters in a room. Listening tests with auralizations using different directivity representations showed that the directivity changes were mostly audible in the perceived loudness and that some of these results could be correlated to the simulated acoustical parameters in the room. Further investigations could consider alternative types of directivity representation where the changes of the directivity of the source could be considered as in a performance situation.

6. AKNOWLEDGMENTS

This work has been financed by the DOREMI European project (IST-2001-39158) [9].

7. REFERENCES

[1] Meyer, J., Acoustics and the performance of music, Verlag Das Musikinstrumenter, Frankfurt/Main, 1978.

[2] Otondo, F., Rindel, J.H. and Christensen, C.L., "Directional patterns and recordings of musical instruments in auralizations", Proc. Workshop on Current Research Directions in Computer Music, Barcelona, Spain, pp. 230-232, 2001.

[3] Otondo, F. and Rindel, J.H., "New method for the representation of musical instruments in auralizations Proc. International Computer Music Conference, Göteborg, Sweden, pp. 248-250, 2002.

[4] Otondo, F. and Rindel, J.H., Caussé R., Misdariis O. and De la Cuadra P., "Directivity of musical instruments in a real performance situation", Proc. International Symposium on Musical Acoustics, Mexico city, Mexico, pp. 312-318 (CD-ROM), 2002.

[5] Dalenbäck, B.-I., Kleiner, M. and Svensson, P. "Audibility of Changes in Geometric Shape, Source Directivity, and Absorptive Treatment-Experiments in Auralization", J. Audio Eng. Soc., Vol. 41, 1993, pp. 905-913.

[6] "The Odeon home page." http://www.dat.dtu.dk/~odeon

[7] Milton, J.S., An introduction to probability and statistics: principles and applications for engineering and the computing sciences, McGraw-Hill, New York, USA, 1990.

[8] Kirkwood, B., Audibility of changes in source directivity for room acoustic auralizations, Internal report, Acoustic Technology, Ørsted-DTU, Technical University of Denmark, 2003.

[9] "The DOREMI project web page." http://www.at.oersted.dtu.dk/~doremi

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

EFFECTS OF GROUPING CUE ON THE PITCH SHIFT OF A PURE TONE INDUCED BY OTHER TONES

Tamotsu Shirado* and Masuzo Yanagida**

*Communications Reserach Laboratory **Doshisha University
*shirado@crl.go.jp **myanagid@mail.doshisha.ac.jp

ABSTRACT

It is known that pitch of the pure tones can be shifted when they are partially masked by other tones. This phenomenon has been explained principally by place models, such as models that assume a certain interaction between the excitation patterns of each component in auditory peripheral.

The purpose of our study was to examine whether the pitch shift of a pure tone induced by other tones is affected by some central mechanism such as auditory stream segregation.

In our study, we measured the pitch of a pure tone(800 Hz) presented simultaneously with a complex tone consisting of three components of lower frequencies(600 Hz, 400 Hz, and 200 Hz). The target pure tone with the complex tone was presented under two conditions. In the first condition, there were preceding complex tones, each of them being exactly the same as the complex tone presented with the target tone. These preceding tones promote the target tone to segregate from other tones. In the second condition, there were preceding complex tones, each of them consisting of components of 600 Hz, 590 Hz, and 580 Hz. Under this codition, the target tone was not promoted to encourage segregation. The place model predicted that the amount of pitch shift of the target tone would be same under both conditions, or smaller under the first condition than the second condition. The experimental results showed, however, that the amount of pitch shift of the target tone was larger under the first condition than in the second condition. These results are contradictory to the prediction that the place model provided, and suggest that some central mechanism plays an important role in our experiments.

1. INTRODUCTION

Pitch perception is an important function of human beings in listening to sounds such as music, speech and so on. Since most sounds in nature are complex tones consisting of several pure tones[1], pitch perception models for pure and complex tones must be established to reveal the psychology involved in music.

A number of models have been proposed so far to address this point, and most of them can be roughly divided into place models and temporal models[2, 3]. Place models assume that pitch is determined by the excitation patterns of the basilar membrane, in which tone frequencies are encoded tonotopically. In contrast, temporal models assume that pitch is determined by temporal patterns of neural pulses in which tone frequencies are encoded.

A number of psychological experiments corresponding to these models have been performed. From these experiments, it was found that the pitch of pure tones can be affected by other tones(called "pitch shifts"). Up to now, the mechanism of the pitch shifts of pure tones has been explained principally by place models[4, 5, 6, 7]. However, pitch shifts of pure tones that cannot be explained by the place models have been found recently[8]. As a consequence, a hybrid place and temporal models has been applied to explain them. This hybrid model assumes mechanisms in auditory peripheral system. On the other hand, evidence has been found that some central mechanisms, such as auditory stream segregation, affect the pitch perception.

In our study, experiments were performed to examine whether pitch shift of a pure tone induced by other tones is affected by auditory stream segregation.

2. PITCH SHIFTS OF PURE TONES

2.1. General tendency of pitch shifts of pure tones induced by other tones

Experiments have revealed that pitches of pure tones can be affected by other pure tones[4], i.e., by the narrow bands of noises[5, 6, 9] presented simultaneously, and also by tones presented in advance[10].

The common tendency among them is that the pitches are always repelled from other tones, e.g., the pitch shifts of pure tones induced by a pure tone of lower frequency are always positive, i.e. of a higher frequency.

Place models described above has principally applied to explain such pitch shifts as follows.

2.2. Pitch shifts in pure tones through a place model[5]

Let us assume the the pitch of a pure tone(tone A) of frequency f is shifted by another pure tone(tone B) of lower frequency presented simultaneously.

In this case, the pitch shift of tone A can be explained by masking interaction between tone A and tone B. The right and left patterns in Fig. 1 respectively show tone A and tone B masking patterns. The peak of the right pattern corresponding to tone A is around frequency f. These masking patterns reflect excitation patterns of the basilar membrane by these tones[3]. We can see that the left (and also right) pattern has a characteristic in that the slope of the left portion is steeper than that of the right portion around the peak[2].

Here, the following assumptions are introduced:

- The loudness of tone A is proportional to the area that lies above the overlapping region of the masking patterns, i.e., the residue pattern of the right pattern subtracted from the left pattern in Figure 1;,
- The tone A pitch is principally determined by the peak of the residual pattern.

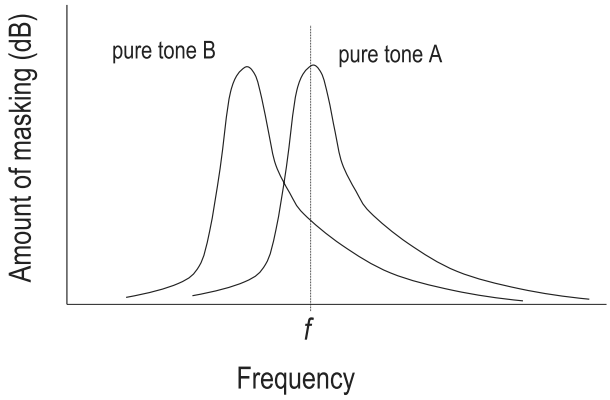


Figure 1: Excitation (masking) patterns.

Figure 2 shows the residual pattern of the right pattern subtracted by the left pattern in Figure 1. We can see that the peak of the residual pattern moves higher than frequency f because of the character of the masking pattern described above.

Therefore, the tone A pitch shifts to a higher frequency.

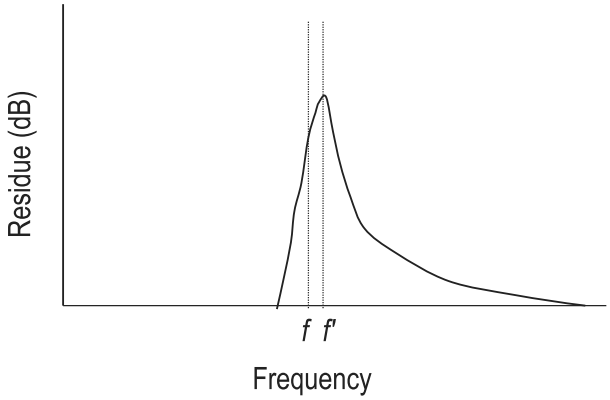


Figure 2: Residue of the right pattern subtracted from the left pattern in Figure 1.

2.3. Pitch shift of a pure tone that cannot be explained by place models

Recently, an experiment revealed a pitch shift of a pure tone that cannot be explained by place models[8]. Frequencies of a pure tone were mistuned either positively, or negatively in a harmonic complex tone whose lower and higher harmonic components also exist. A place model predicted that the pitch of the pure tone would always shift higher whether the pure tone was mistuned higher or lower in the harmonics[7].

The experimental results, however, showed that the pitch of the pure tone shifted lower when it was mistuned lower. These results contradicted the prediction of the place model. A temporal model also failed to give an explanation that was aconsistent with the results. Finally, a hybrid model of a place and temporal mechanism in auditory periphery was introduced.

2.4. Purpose of our experiments

As described above, some mechanisms in auditory periphery have been assumed in most previous models for explaining pitch shifts of pure tones by other tones.

On the other hand, there is considerable evidence that some central mechanism, such as auditory stream segregation [11], also takes part in hearing.

The purpose of our study was to examine whether the pitch shift of a pure tone induced by other tones is affected by auditory stream segregation.

Hereafter, we will refer a pure tone whose pitch shift is of interest to us in our experiments as a “target tone”. The pitch of a target tone presented simultaneously with other tones wws measured under two conditions. Under the first of these(condition 1), the target tone was promoted so as to segregate it from other tones. Under the second condition(condition 2), no such promotion was introduced.

3. EXPERIMENTS

A university student(Y.T.) majoring in piano performance participated in the experiments.

3.1. Stimuli

Corresponding to the two conditions described above, two tone series patterns were prepared as stimuli of the experiments.

Figure 3 shows pattern (pattern 1) corresponding to condition 1.

Each bar represents a 100-ms pure tone including 5-ms rise/fall raised-cosine ramps. All the pure tones had the same amplitude, starting with the sine phase. At the beginning of pattern 1, four harmonic complex tones separated by 100 ms of silence were presented. Each tone consisted of first, second, and third harmonic components of fundamental frequency 200 Hz. Following these tones, a complex tone was presented. The complex tone consisted of the 800 Hz target tone and a harmonic complex tone that was exactly the same as each of the preceding harmonic complex tones. The target tone was identical to a fourth harmonic component of the harmonic complex tone. Finally, a matching tone separated by 500 ms was presented to measure the pitch of the target tone.

It was expected that the target tone would be promoted so as to segregate it from the harmonic complex tone presented simultaneously, because it was expected that the latter would be to be captured into an auditory stream of the preceding tone series.

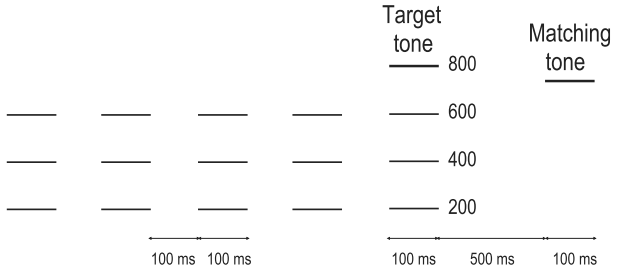


Figure 3: pattern1.

Figure 4 shows the pattern(pattern 2), which corresponds to condition 2.

The preceding complex tones differed between the pattern 2 and the pattern 1. Each preceding complex tone consisted of pure tones of 600 Hz, 590 Hz, and 580 Hz. The harmonic complex tone (including the target tone) that followed was exactly the same as that in pattern 1.

In this case, the target tone was not promoted so as to segregate it from the harmonic complex tone presented simultaneously.

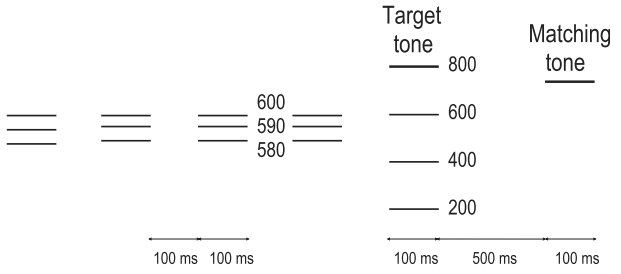


Figure 4: pattern2.

3.2. Pitch shift predicted by place model

Now let us consider experimental results predicted by the place model described above.

First, simultaneous masking to the target tone by other tones is expected to be the same between both patterns because the harmonic complex tones presented simultaneously with the target tone were exactly the same for both patterns.

Second, if some forward masking to the target tone by the preceding tone series is assumed, the masking in pattern 2 is expected to be larger than that in pattern 1 because the components (600 Hz, 590 Hz, and 580 Hz) in each preceding tone of pattern 2 were closer to the target tone than the components (600 Hz, 400 Hz, and 200 Hz) in each preceding tone of pattern 1 on the whole.

Therefore, if the some forward masking to the target tone exists, the amount of pitch shifts is expected to be smaller in pattern 1 than that in pattern 2.

Finally, the total amount of pitch shifts is expected to be the same between pattern 1 and pattern 2, or smaller in pattern 1 than that in pattern 2.

These considerations are summarized in Table 1.

Table 1: Pitch shift predicted by place model.

	Pattern1	Pattern2
Complex tone presented with target tone simultaneously	same	
Amount of simultaneous masking to target tone	same	
Amount of pitch shifts by simultaneous masking	same	
Frequencies of components comprsing preceding tones vis-à-vis target tone component frequencies	farther	closer
Assumed amount of forward masking to target tone	smaller	larger
Assumed amount of pitch shifts by forward masking	smaller	larger
Total amount of pitch shifts	smaller or same	larger

3.3. Procedure

The method of adjustment was applied to measure pitches of the target tones.

Series of tone stimuli was presented to the subject diotically via headphones (STAX Λ NOVA) in a silent room.

The subject was able to adjust the matching tone frequency. She was required to match the pitch of the matching to that of the target tone. She was able to call up the tone stimuli as often she liked until the pitch of the matching tone was satisfied.

Eleven trials were performed for each pattern, with a rest period taken every two or three trials.

4. EXPERIMENTAL RESULTS

Figure 5 shows the mean pitch shifts of subject Y.T. averaged over eleven trials for each pattern. The mean pitch shift is around 0.9% for pattern 1, and around zero% for pattern 2. A statistical test shows the mean pitch shift for pattern 1 is significantly larger than that for pattern 2.

The difference between these values is around 0.9%, which is larger than the frequency difference limen around 800 Hz[3].

5. DISCUSSION

Figure 5 shows that the amount of pitch shift of the target tone is larger in pattern 1 than that in pattern 2.

As shown in Table 1, the place model predicted that the pitch shifts of the target tones would be the same between

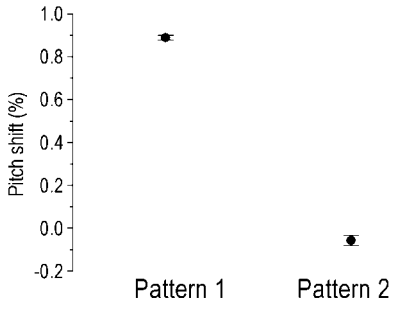


Figure 5: Mean pitch shifts of subject Y.T. in frequency percent

pattern 1 and pattern 2, or smaller in pattern 1 than the pattern 2.

Therefore, the place model cannot explain the experimental results, instead, some central mechanism seems to play an important role in pitch perception of the target tone in our experiments.

Unfortunately, we have not yet found any mechanism that can provide an explanation consistent with our experimental results. At best, we can note several ideas to explain the experimental results as follows.

The pitch shift around zero% for pattern 2 suggests an effect of harmonic binding by the harmonic complex tone presented simultaneously such that a pitch of each harmonic component whose frequency is precisely tuned in a harmonic complex tone is enforced to coincide with the frequency.

Adjacency of the components in pattern 2 that fall into an identical critical band may also have affected the experimental results.

Only one subject participated in the present experiment. It is well known that no small deviation in pitch perception exists among different listeners[3]. Thus, several subjects will need to participate in the experiments to show the universality of the characteristics suggested in our experiments.

Experimental stimuli that do not contain any preceding tones must also be examined to obtain a standard value of the pitch shift of the target tone presented with the harmonic complex tone.

In any event, larger scale and finer experiments must be performed to verify the validity of the present results.

6. CONCLUSIONS

We measured the pitch of a pure tone presented simultaneously with a complex tone under two conditions. Under the first condition, there were preceding tones so as to promote the target tone to segregate it from the complex tone. There were also preceding tones under the second condition, but they did not include any such promotion.

A place model predicted that the amount of pitch shift of the target tone would be the same for both conditions, or smaller for the first condition than for the second condition.

The experimental results, however, showed the amount of pitch shift of the target tone was larger for the first condition than that for the second condition. These results contradict the prediction made by the place model, and suggests that some central mechanism plays an important role in our experiments.

Accordingly, we plan to conduct larger scale and finer experiments in the near future.

ACKNOWLEDGMENT

This work was partially supported by the grant to the research project at Doshisha University entitled "Intelligent Information Science and Its Applications to Problem Solving in Engineering Fields" from the Ministry of Education, Culture, Sports, Science and Technology, Japan. The authors would like to thank Dr. Isahara, group leader of Keihanna Info-communication center, Communications Research Laboratory for his kind support to our work.

7. REFERENCES

[1] Handel, S., Listening - An Introduction to the Perceptual of Auditory Events -, MIT Press, 1989.

[2] Hartmann, W.M., "Pitch, periodicity, and auditory organization", J. Acoust. Soc. Am. Vol. 100(6): 3491–3502, 1996.

[3] Moore, B.C.J., An Introduction of the Psychology of Hearing, Academic Press, 1989.

[4] Zwicker, E. and Fastl, H., Psychoacoustics - Facts and Models, Springer-Verlag, 1990.

[5] Egan, J. P. and Meyer D. R., "Changes in pitch of tones of low frequency as a function of the pattern of excitation produced by a band of noise", J. Acoust. Soc. Am., 22(6): 827–833, 1950.

[6] Houtsma, A.J.M., "Noise-induced shifts in the pitch of pure and complex tones", J. Acoust. Soc. Am. 70(6): 1661–1668, 1981.

[7] Terhart, E., Stoll, G. and Seewann, M., "Algorithm for extraction of pitch and pitch salience from complex tonal signals", J. Acoust. Soc. Am. 71(3): 679–688, 1982.

[8] Hartmann, W.M. and Doty, S. L., "On the pitches of components of a complex tone", J. Acoust. Soc. Am. 99(1): 567–578, 1996.

[9] Stoll, G., "Pitch shift of pure an complex tones induced by masking noise", J. Acoust. Soc. Am. 77(1): 188–192, 1985.

[10] Rakowski, A. and Hirsh, I.J., "Post-stimulatory pitch shifts for pure tones", J. Acoust. Soc. Am, 68(2): 467–474, 1980.

[11] Bregman, A.S., Auditory Scene Analysis, MIT Press, 1990.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

LISTENER COMMON AND GROUP PERCEPTUAL DIMENSIONS IN VIOLIN TIMBRE

Jan Štěpánek, Zdeněk Otčenášek

Sound Studio of the Faculty of Music, Academy of Performing Arts Prague
stepanek@hamu.cz, otcenasek@hamu.cz

ABSTRACT

Results of listening tests of violin tones are studied with respect to listeners and their perceptual models. Five sets of violin tone recordings (pitch B3, F#4, C5, G5, D6) were used. Twenty experienced listeners – violin players – assessed dissimilarities in timbre in pairs of tones.

The results of five listening tests (individual dissimilarity matrixes) were separately processed using latent class approach (CLASCAL). This approach yields to perceptual spaces of common dimensions shared with all listeners and defines listener classes (groups). The perceptual models of CLASCAL groups were also studied. The results revealed that the stability is higher in group perceptual models then in common models of all listeners. Affinity of dimensions and models across groups showed potential existence of additional perceptual dimensions in some class models.

1. INTRODUCTION

1.1 Timbre studies

Usually the aim of study of musical sound timbre is a detection of distinctive features of specific sound context and the search for their acoustic correlates. More technically, a (multidimensional) perceptual space of sound stimuli is constructed and attempts are made to interpret its individual dimensions using adequate spectral characteristics. Subjective data are collected in listening tests using different methods that allow the use of Factor analysis [1] or Multidimensional scaling of different kind [2, 3] for the evaluation of results. Some studies have been focused only on stationary parts of the sounds [1, 2] or on both transient and stationary parts [3]. As for listeners, one group or more groups of judges (e.g. musicians and nonmusicians in [1]) are apriori chosen to participate in listening tests.

1.2. Listener classes

A question arises whether the listeners are in higher or lower concordance in their judgements or even whether there exist groups of listeners with different perceptual models. Latent class approach (CLASCAL MDS) [4, 5] results in a perceptual space of common dimensions shared by all listeners, but it also determines classes of listeners based on the similarity of individual perceptual models (aposteriori defined groups of listeners); the group models differ in weights of common dimensions.

In this contribution the study of violin timbre focused on stationary part of sounds is described. The results of listening tests (individual dissimilarity matrixes) were processed using latent class approach applied on weighted Euclidean model (CLASCAL) and extended CLASCAL model [4] (overview of models see in [6]). The first results of these analyses – common perceptual space dimensionality and main source of the differentiation of listeners into classes – were described in [7]. Further the classical Euclidean model and extended Euclidean model [6] were applied separately on a set of individual dissimilarity matrixes of each class (group) of listeners. Common and group perceptual spaces and dimensions were further compared.

2. METHOD

Five sets of violin tone recordings (pitches B3, F#4, C5, G5, D6) were used in the study [8]. Attack and decay transients were unified to weaken their influence on judgements. Seventeen tones for each pitch were listened in headphones and judged. Twenty experienced listeners – violin players (four Academy professors and sixteen students) assessed dissimilarities in timbre in pairs of violin tones with marks 0, 0.5, 1, ..., 4.5, 5.

2.1. Latent class approach

Five sets of individual dissimilarity matrixes were separately processed using latent class approach applied on weighted Euclidean model (CLASCAL) and extended CLASCAL. Latent class approach (CLASCAL MDS) solves two optimization tasks:

1. To fit stimuli dissimilarities into distances of Euclidean space of appropriate low dimension.
2. To establish the appropriate number of (latent) classes of listeners and to add each individual listener to one of these classes (aposteriori grouping).

Optimal model selection is an iterative process in which the solution of the first and the second optimization task is alternately improved.

The new version of the CLASCAL program with a bootstrap procedure [5] was used. It enables an application of latent class approach with both weighted Euclidean model (each stimulus is described by its coordinates in common dimensions; each class of listeners has its own weight of every dimension) and extended weighted Euclidean model (each stimulus is described moreover by its specificity value, which indicates the existence of a stimulus feature not shared with other stimuli; set of all specificities is weighted separately for each class of listeners).

The CLASCAL MDS procedure generally yields to a model, which best fits dissimilarity data. Every discussed model will be denoted here as $C_iD_jS_k$: *i* classes (C_i), *j* dimensions (D_j) and with or without specificities (S_1 or S_0). The model parameters are as follows:

- The number of listener classes; each listener belongs to a certain class
- The number of common dimensions

- The stimulus coordinates according to the dimensions, defining its position in the common perceptual space
- The stimulus specificity value (optional), one for each stimulus
- The weight, one for each common dimension (eventually for a set of all specificities) and each class

The application of CLASCAL procedure on the studied data is described in [7]. The main results may be summarized in all five pitches as follows:

a) Two classes revealed as the best solution ($C_2D_jS_k$); division of the listeners into classes is very stable.
b) The weights of all dimensions/specificities of the first class are less then those of the second class.
c) One of the differences between classes is the measure of the exploitation of dissimilarity scale by individual listener; the mean dissimilarity value of the first class is significantly less then that of the second class.

2.2. Comparison of models

Outstanding listeners were excluded from each class before calculating individual class models to obtain more homogeneous classes. The exclusion criteria consisted in instability in class belonging with respect to models with different number of dimensions, in bootstrap instability (see Figure 2 in the next paragraph), in extreme value of the mean dissimilarity with respect to the other class members.

Euclidean and extended Euclidean models [6] applied on every clarified listener class resulted into models denoted here as $Class_iD_jS_k$. The most appropriate model for each class was established.

As a model development we will denote the sequence of models with increasing number of dimensions (number of classes and type of specificities S_0 or S_1 is not changed). Two types of model developments will be used here: $C_2D_jS_k$ = {$C_2D_1S_k$, $C_2D_2S_k$, ..., $C_2D_6S_k$} for group of all listeners and $Class_iD_jS_k$ = {$Class_iD_1S_k$, $Class_iD_2S_k$, ..., $Class_iD_6S_k$} for group consisted of listener class *i*. The Pearson correlation coefficient between stimuli coordinates of individual dimensions in two different models was used to compare them:

1. Inside model development – correlations in consecutive models exhibit stability of an individual dimension and (qualified over all dimensions) stability of the listener group model.
2. Across model developments – correlations exhibit affinity of dimensions between groups and (qualified over all dimensions) affinity of models of different groups of listeners.

3. RESULTS

3.1. Clarification of classes

The optimal model selection procedure is described in [7] where also results for all studied pitches are in details and they are summarized here in the paragraph 2.1. Two classes were found as the most appropriate in all five pitches, in the tones B3 and F#4 it was the model $C_2D_3S_0$ and in tones C5, G5 and D6 it was the model $C_2D_2S_1$.

The outstanding listeners were excluded from each class to obtain more homogeneous classes; the exclusion criteria are listed in the paragraph 2.2. For an excluded listener, several criteria were often met. Exclusion of outstanding listeners is marked in Figure 1; bootstrap instability is illustrated by a cluster tree in Figure 2. The number of listeners in clarified classes is summarized in Table 1.

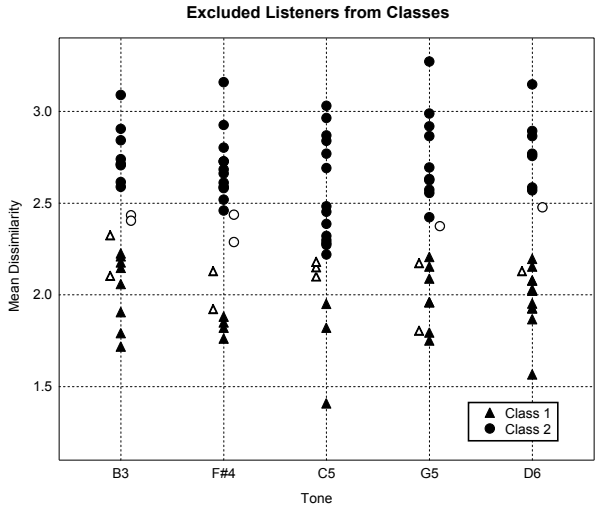


Figure 1: *Mean dissimilarity values of individual listeners belonging to class 1 or class 2. Excluded outstanding listeners are marked by not filled point patterns.*

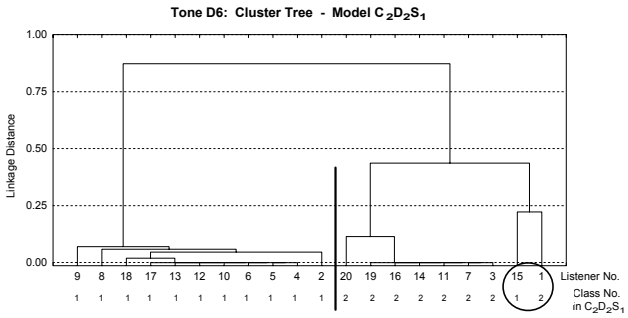


Figure 2: *Example of listeners cluster tree for tone D6. Listener distances are based on bootstrap procedure – see [5, 7]. Excluded listeners are marked by a circle.*

Tone	B3	F#4	C5	G5	D6
Class 1	**8**=10-2	**4**=6-2	**3**=6-3	**7**=9-2	**11**=12-1
Class 2	**8**=10-2	**12**=14-2	**14**=14-0	**10**=11-1	**7**=8-1

Table 1: *Number of listeners in clarified classes (bold). Original number of listeners and number of excluded listeners are on the right side of the equation.*

3.2. Class models

Euclidean models ($Class_iD_jS_0$) and extended Euclidean models ($Class_iD_jS_1$) [6] were calculated for each clarified listener class. The most appropriate models for each class both without and with specificities are summarized in Table 2 together with the

most appropriate models found previously for group of all listeners.

Tone	B3	F#4	C5	G5	D6
All (S_0)	$\mathbf{C_2D_3S_0}$	$\mathbf{C_2D_3S_0}$	$C_2D_4S_0$	$C_2D_4S_0$	$C_2D_4S_0$
All (S_1)	$C_2D_2S_1$	$C_2D_2S_1$	$\mathbf{C_2D_2S_1}$	$\mathbf{C_2D_2S_1}$	$\mathbf{C_2D_2S_1}$
$Class_1(S_0)$	D_3S_0	D_3S_0	D_3S_0	$\mathbf{D_2S_0}$	D_2S_0
$Class_1(S_1)$	$\mathbf{D_2S_1}$	$\mathbf{D_2S_1}$	$\mathbf{D_2S_1}$	D_2S_1	$\mathbf{D_2S_1}$
$Class_2(S_0)$	D_4S_0	$\mathbf{D_6S_0}$	$\mathbf{D_5S_0}$	D_5S_0	D_2S_0
$Class_2(S_1)$	$\mathbf{D_3S_1}$	D_3S_1	D_2S_1	$\mathbf{D_2S_1}$	$\mathbf{D_2S_1}$

Table 2: *The most appropriate models for group of all listeners (All) and for clarified classes; S_0: without specificities, S_1: with specificities. Optimal models irrespective to specificities are bold.*

3.3. Stability of dimensions inside model development

A stability of dimensions inside the model development was observed using the Pearson correlation coefficient between the stimuli coordinates with increasing number of dimensions. An example of a graphical view of significant correlations (α=5%) between dimensions in the model development is in Figure 3.

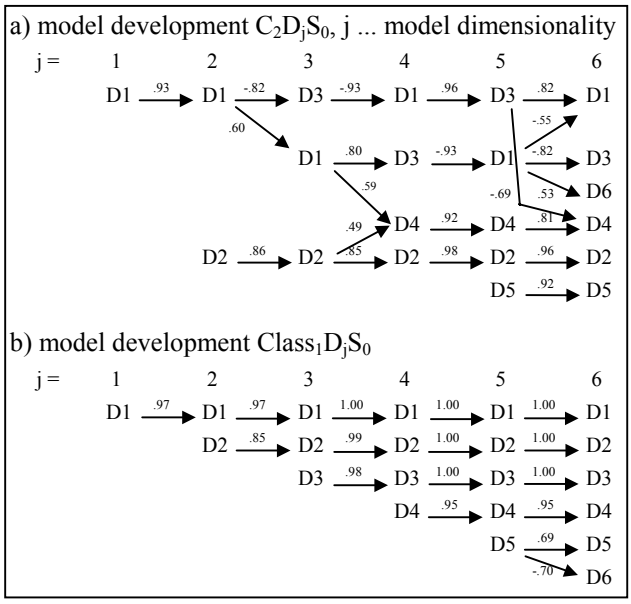


Figure 3: *Significant correlations in model development for tone F#4, a) group of all listeners, b) clarified class 1. Individual dimensions are denoted as D1, D2, ..., D6. Numbers around arrows are values of significant correlation coefficient between dimension coordinates.*

The model developments revealed similar fashion for the group of all listeners and for both classes in all five pitches.

Stability of the dimension coordinates in consecutive models is documented in Figure 4.

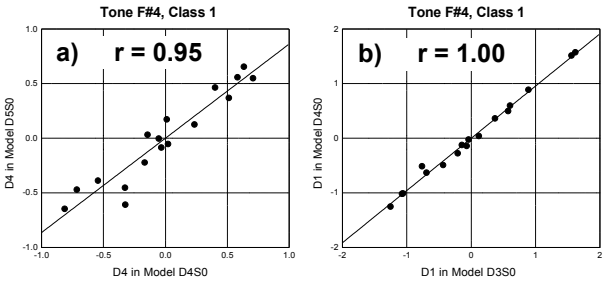


Figure 4: *Stability of dimension coordinates in model development for tone F#4, class 1: a) dimension D4 from models D_4S_0 and D_5S_0, Pearson correlation r=0.95, b) dimension D1 from models D_3S_0 and D_4S_0, r=1.00.*

3.4. Affinity of dimensions across model developments

An affinity of the dimensions across the model developments was observed using the Pearson correlation coefficient between stimuli coordinates of models of different groups (All listeners, Class1, and Class2). Dimensions from most appropriate models (see Table 2) or dimensions from other models having the same number of dimensions were correlated. The examples of significant correlations are in Figure 5.

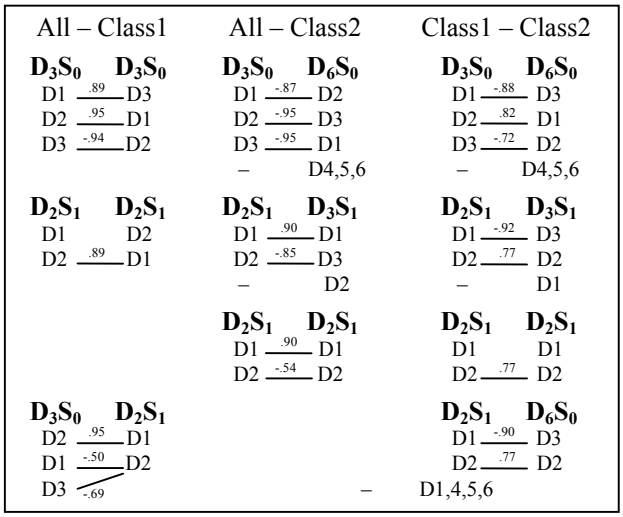


Figure 5: *Examples of significant correlations between dimensions across model developments for tone F#4. Numbers around lines are values of significant correlation coefficient between dimension coordinates.*

From the correlations across the models, it was possible to recognize the dimensions shared across groups. The dimensions were designated by letters according to their stability and stable order in class model developments. For the tone F#4, the dimensions A, B, and C are shared with the groups (this is in accordance to the CLASCAL common dimensions) but they differ in order and sign according to groups.

Shared dimensions and other dimensions used by individual groups for all five studied tones are indicated in Table 3.

Tone	Order	Class 1	Class 2	All listeners
B3	1	A	A	A
	2	B	- [(B) & (C)]	B
	3	C	**D**	C
	4		(B)	
F#4	1	- C	A	- B
	2	A	B	- C
	3	- B	C	- A
	4-6		**D**, **E**, **F**	
C5	1	- A	A	A
	2	- **C**	B	B & (-A)
	3	(**D**)	**C**	
	4	(B)	**D**	
G5	1	A	A	A
	2	B	B	B
	3	C	- C	C & D
	4	D	D	D & C
D6	1	A	A	A
	2	B	- B	B & (-A)

Table 3: *Dimensions used by groups. Extra dimensions added to CLASCAL common dimensions are bold, dimensions correlated only with significance 1%<α<5% are in parenthesis, with α<1% as normal characters.*

4. DISCUSSION

The results of analyses of the clarified CLASCAL classes exhibited some interesting properties in comparison to the usual CLASCAL approach dealing with the common dimensions.

The splitting of the dimensions and the change of their order (see example for the tone F#4 in Figure 3) is frequent in the model development in the group of all listeners. The class models showed stable dimensions (with very high significant correlations), new dimensions are added by dimension increase (dimension splitting is exceptional) and the order of the dimensions is stable.

It was shown previously [7] that the mean dissimilarity values are significantly higher in Class2 (due to use of higher marks in dissimilarity judgements) than in Class1 in all five pitches. As for most appropriate models, Class1 optimal models revealed lower number of dimensions and Class2 the higher ones than models for all listeners (Table 2). Even if there is a good concordance among common dimensions and lower dimensions of individual classes (Table 3), it is not evident whether additional dimensions in Class2 manifest a different perceptual model (use of more dimensions), or only the use of more levels in judgement decreases information "noise" allowing to discern next existing dimensions.

Usually used criteria for establishing the number of dimensions [4, 5] are information criteria (AIC, BIC) or criteria based on randomizing (Monte Carlo, bootstrap). In our approach we prefer interpretability as a criterion for the acceptation of a perceptual dimension. This is the reason we permit also models with the higher number of dimensions, to keep any potential dimension (tones B3, F#4, C5 in Table 3).

Specificity values decreased with increasing model dimensionalities, so models D_jS_0 and D_jS_1 became more similar with increasing dimensionality. The specificity feature of any signal must be judged individually by listening.

5. CONCLUSIONS

Latent class approach (CLASCAL) defines classes of listeners; clarified classes revealed better stability of perceptual dimensions and models than the group of all listeners. Affinity of dimensions and models across groups showed potential existence of additional perceptual dimensions in models of listeners using the full dissimilarity scale. The correlation with some acoustic parameters of signals will be the criterion for the acceptance of a perceptual dimension. A preliminary calculation revealed importance of spectral energy distribution (the center of gravity) and of the levels of fundamental and lower harmonics. The importance of these spectral features for violin timbre must be verified in further listening tests.

6. ACKNOWLEDGEMENTS

The research was supported partly by the Ministry of Education and Youth of the Czech Republic (Project No. 511100001) and by the Grant Agency of the Czech Republic (Project No. 202/02/1370).

The authors are gratefully acknowledged to Suzanne Winsberg for the offering of the new version of the CLASCAL program with bootstrap and for guidelines and beneficial discussions for its use.

7. REFERENCES

[1] Bismarck, G. von, "Timbre of steady sounds: A factorial investigation of its verbal attributes", *Acustica*, 30, 146-159, 1974.

[2] Plomp, R., Aspects of tone sensation, London, Academic Press, 1976.

[3] Grey, J. M., "Multidimensional perceptual scaling of musical timbres", J. Acoust. Soc. Am., 61 (5), 1270-1277, 1977.

[4] Winsberg, S., De Soete, G., "A latent class approach to fitting the weighted Euclidean model, CLASCAL", Psychometrika, 58, 315-330, 1993.

[5] Winsberg, S., De Soete, G., "A bootstrap procedure for mixture models: applied to multidimensional scaling latent class models", Applied Stochastic Models for Business and Industry, 18, 391-406, 2002.

[6] McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., Krimphoff, J., "Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes", Psychological Research, 58, 177-192, 1995.

[7] Štěpánek, J., "Latent listener classes and class models in violin timbre", DAGA 2003, Aachen, in print.

[8] Štěpánek, J., Otčenášek, Z., Melka, A., Syrový, V., "Violin Tones Spectra and their Relationship to Perceived Sound Quality" in Proceedings of the Institute of Acoustics ISMA´97, 19 (5), Edinburgh, 125-130, 1997.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

PATTERN RECOGNITION OF MUSICAL INSTRUMENTS USING HIDDEN MARKOV MODELS

Raquel Ventura-Miravet, Fionn Murtagh, Ji Ming

School of Computer Science
Queen's University Belfast
Belfast, BT7 1NN, Northern Ireland
r.ventura-miravet@qub.ac.uk

ABSTRACT

Today there is a large amount of musical recordings available from the Internet. Several useful applications emerge if users are able to automatically search this data given particular musical content e.g. different musical instruments. In order to search for a particular instrument, we extract acoustic features given a sound sample and then match these features against a database of all stored musical instruments by using pattern recognition. Whilst relatively little research has been conducted on musical instrument recognition, there already exists a large body of knowledge concerning speaker recognition. This work is concerned with applying the techniques and modelling methodologies that are used within speaker recognition to musical instrument recognition. In particular, we present a study that aims to identify musical instruments using statistical approaches that reflect the structure of the data. Given samples of each musical instrument, we perform training to generate an N-state Hidden Markov Model (HMM) that represents the spectral characteristics and time variability of musical patterns. The results show that this approach, which is widely used in speaker recognition, offers a powerful technique for feature extraction and instrument recognition given musical data. We obtain an average recognition accuracy rate of approximately 98% when discriminating between six different musical sources.

1. INTRODUCTION

Much of the work on automatic speech recognition (ASR) can be useful for identifying musical instruments. We are interested in finding out how far speech recognition techniques can be applied successfully to this purpose.

In this paper we describe the results of our work by using the hidden Markov model (HMM) [1] as a statistical signal model and different front-end analyses as an approach to identify musical instruments. We analyse front-ends based on standard processing techniques such as filter banks or linear prediction (LP). These techniques simulate properties of the peripheral auditory system or the mechanism of human articulation, being very suitable for acoustic recognition tasks. HMM and cepstral features have provided a robust model for speaker recognition over the last years.

Although relatively little research has been carried out on applying the techniques and modelling methodologies that are used within speaker recognition to musical instrument recognition, we are motivated by the optimal results of previous works [2], [3], [4] In addition, most of the studies have dealt with isolated sounds in a simple context [5]. We have extracted features from phrases collected from commercial recordings in order to create as realistic a context as possible. The database consisted of samples from 5 solo musical instruments and 1 string ensemble in order to generalize the system to polyphonic sounds, which are commonly used on the Internet. We are really interested in creating a system capable of identifying not only solo instruments but also certain sounds created by a set of instruments playing at the same time. In the rest of this paper, we will use the expressions "musical sources" or "musical sounds" interchangeably with reference to the units to be classified instead of musical instruments. Different sets for training and testing the system have been used. Our approach is an attempt to evaluate the techniques chosen in a real environment, reporting the effect of different factors on the recognition performance, such as sampling rate and feature vector sizes.

2. DATA COLLECTION

Samples from 5 monophonic musical instruments and 1 polyphonic musical sound were chosen to train and test the system. The corpus included examples from: flute (air reed wind), harp (plucked string), organ (air/mechanical reed wind), piano (struck string) and violin (bowed string). The polyphonic sound is an ensemble of stringed instruments: two violins, a viola and a cello. A random guesser would score 16.7% in the recognition task.

We have collected different compact discs to create a database as varied as possible. A number of excerpts from commercial recordings have been gathered to create training and test sets from different performers who use different types of musical instruments, playing different styles. It appears that the compact discs have been recorded under different conditions, although the differences can be insignificant.

An average of 7 different professional performers were used for each source. On average in 79% of cases, the test data was selected from different audio tracks compared to the training data. It is important to stress that there was no test data overlap with training data. It was difficult to find more than 30 minutes of monophonic sound for each source to train the system and more than 15 minutes for each set to test. The system is tested with different sample rates. All recordings were re-sampled to 44 kHz and 16 kHz with 16-bit resolution. Several aspects were considered to accomplish the development of the system, such as the size of the training and test sets on the performance. The system was tested using different lengths of excerpt, such as 0.2, 1, 5, 10 and 20 seconds, in order to study its effect on the accuracy performance. Table 1 shows the total duration used for each musical source.

Musical Source	Total Training Data	Total Testing Data	Number of Performers
Flute	2044	960	8
Harp	1978	1136	8
Organ	1876	1029	9
Piano	1818	973	7
String Quartet	1818	945	3
Violin	1994	970	9

Table 1: *Description of the commercial recordings used. The last column represents number of different performers as an indicator of variability. The size of total data is given in seconds.*

3. FEATURE EXTRACTION

Spectral envelope is successfully used in automatic speech recognition (ASR). Short-term spectrum carries information about the voice source. Motivated by this, representative and widely used signal analysis methods for speech recognition were chosen, including linear prediction coding (LPC), linear prediction derived cepstral coefficients (LPCC), mel frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP). Linear prediction is based on modelling the vocal tract while MFCC and PLP emulate properties of human perception, making them promising for front-end analysis.

Most of the experiments were performed in 20 ms length Hamming windows with 50% overlap. The length of the transform was 882 points corresponding to 44.1 kHz sample rate. For the MFCC calculations, 24 channels were used. LP order is 12. Preemphasis of 0.97 was applied to all types of spectral analysis described above. Segments of 10 seconds have been used in the test set.

We have extracted features from phrases of various lengths taken from commercial recordings. Different features were used for each front-end analysis in order to study how the energy and derivatives components influence the performance. First, we used a feature vector of 12 components. Second, the component of the energy was added to this set. Third, the velocity (delta) parameters were added to the feature space in order to take advantage of spectral dynamics. Lastly, acceleration components were joined, changing the size of the feature vector to 39 components.

4. EXPERIMENTS

The aim of this work is to evaluate methods and front-end analyses that have been successfully used for speech recognition in order to classify musical sources, reporting the effect of different factors on the recognition performance, such as sampling rate and feature vector sizes. The system was developed using Gaussian mixtures for modelling the described features above in a Hidden Markov Model (HMM) classifier. The hidden Markov model toolkit (HTK) [6] was used for this purpose. Different numbers of mixture components were tested to represent any variability successfully. We have supposed that our unit to be identified does not need to be modelled as a sequence of patterns that must occur in the same order. There is a high probability that this case is similar to speaker recognition where just one state is needed. Several numbers of states were evaluated to demonstrate that one state is able to model the features adequately. An ergodic topology was used in which any transition from any state is allowed.

5. RESULTS

Several experiments were carried out by using different Hamming window sizes, from 10 ms to 60 ms, applied at 10 ms intervals. Results show that the different overlaps between adjacent windows do not affect significantly the performance. Another factor that does not alter the performance in a significant manner is the frame rate. The following studies were carried out with a 20 ms Hamming window (884 samples at a 44 kHz rate) and 24 filter banks. Modifying the number of filter banks from 12 to 48 showed a small variation on the results, less than 0.3%.

We also tested the performance using different training set sizes. MFCC plus energy, delta and acceleration components were used as features. As we expected, increasing the training data set size from 1 minute to 30, improves the performance from 88.9% to 98.1%. Results are shown in Figure 1.

Experiments have shown that the use of acceleration components with MFCC does not add relevant information and leads to similar results as those obtained by using velocity parameters. However, PLP shows an improvement of 1.1% whilst LPCC and LPC decrease their performance by a factor of 1.3% and 5.4% respectively. One state and 5 component Gaussian mixture was used to classify 10-second excerpts. Results are reported in Table 2.

Different orders of the LPC and LPCC analysis were used to study how this factor influences the performance. The number of coefficients coincides with the analysis order. The results are shown in Figure 2. LPCC shows a more regular behaviour than LPC while the analysis order increases. The best accuracy using LPC with energy and spectral dynamics was obtained by using order 6 (88.9%) and the worst performance was with 18^{th} order (64.6%). The order affects flute, harp and organ performance in a more visible manner than the rest of sound sources. So we can find a negative variation in harp recognition of 58% when the order is changed from 6 to 20. The order 6 gives the best performance to LPCC with energy and dynamic components (94.7%) as well. The worst accuracy corresponds to 14^{th} order (87.2%). The variation is smaller than in LPC case. For MFCC and PLP with energy component and their derivatives, the recognition rates are approximately identical, about 98% with a small variation.

		Energy/Derivatives		
		_e	_e_d	_e_d_a
Lpc	66.7	67.4	78.0	72.6
Lpcc	81.8	83.1	86.6	85.3
Mfcc	97.0	97.6	97.9	98.1
Plp	97.6	97.0	97.9	99.0

Table 2: *Recognition accuracy for different features. The letters "_e_d_a" in the labels refers to addition of energy, velocity and acceleration parameters respectively. Entries are expressed as percentages.*

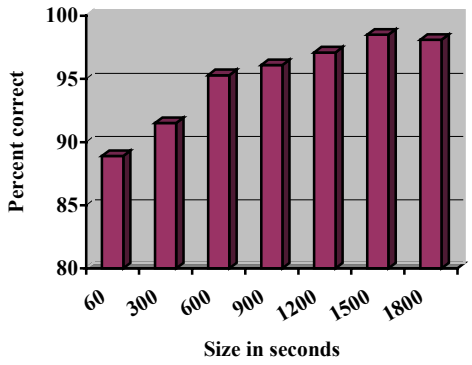


Figure 1: *Recognition accuracy as a function of the training data size used for each musical source. The size is given in seconds.*

We have carried out an experiment to study how increasing the number of Gaussian mixture components affects the performance of the system. A 39 dimensional feature vector with energy, first and second differentials added represents each frame. It was found that 5 mixtures per state were optimal in the LPCC-MFCC-PLP framework. Above this number, computation time increases without modifying the performance considerably. LPC analysis is shown to be very sensitive to the number of mixture components. Its performance improves by a factor of 5.5% increasing the number of mixtures from 5 to 10. Results are listed in Table 3.

Features \ Mixture	1	5	10	16	32	64
Lpc_e_d_a	55.0	72.6	78.1	79.5	82.3	84.4
Lpcc_e_d_a	77.6	85.3	85.3	85.1	84.9	84.5
Mfcc_e_d_a	94.3	98.1	98.4	98.5	99.0	99.5
Plp_e_d_a	93.5	99.0	98.9	98.8	99.2	99.0

Table 3: *Recognition accuracy for different features in different Gaussian densities. . The letters "_e_d_a" in the labels refers to addition of energy, velocity and acceleration parameters respectively. Entries are expressed as percentages.*

Results in Table 4 show the performance of different HMM topologies using mel frequency cepstral coefficients with energy and their first and second differentials. Each state was characterized by 5 components mixture Gaussian. Increasing the number of states from 3 to 64 does not vary the performance importantly.

States	**1**	**3**	**8**	**12**	**16**	**32**	**64**
Correct	98.1	99.2	99.2	99.2	98.9	98.9	99.2

Table 4: *Recognition accuracy as a function of number of states for Mfcc_e_d_a features. Entries are expressed as percentages.*

In order to evaluate the system in a more complex context, we changed the sample rate of training data and test data. Table 5 shows the performance of a single model system by using different sample rates. The best performances occur when the test and training sample rates are the same. Increasing the number of mixtures to 10 and training the system with two data sets, one sampled to 44.1 kHz and the other to 16 kHz, the results are optimal, although this complicates the system and uses more computation time. Table 6 shows the performance of a multi model system. Two models have been created for each musical source based on the different frequency rate. That is, each model represents a musical source sampled at a certain frequency rate. Results indicate that the performance does not depend on the sample rate used.

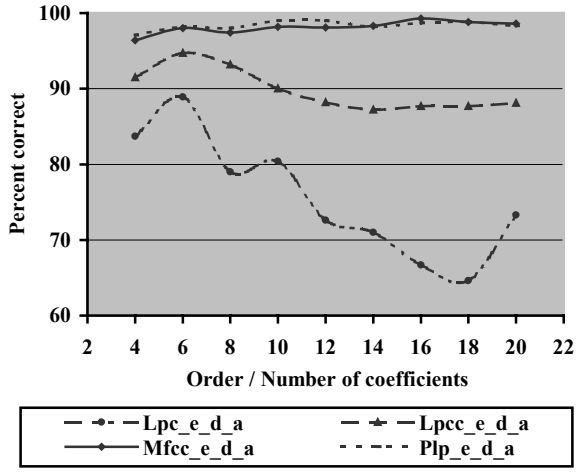


Figure 2: *Recognition accuracy as a function of different feature vector sizes. The number of components generated is the same as the order for LPC, LPCC and PLP. The letters "_e_d_a" in the labels refers to addition of energy, velocity and acceleration parameters respectively.*

		Test	
		16 kHz	**44 kHz**
Train	**16 kHz**	98.2	59.9
	44 kHz	60.7	98.1
	16 kHz + 44 kHz	94.3	98.1

Table 5: *Recognition accuracy using different HMM single models depending on the sample rate. The last row shows a model trained with 16 kHz and 44.1 kHz sample rate sounds by using 10 mixtures. Entries are expressed as percentages.*

Test	
16 kHz	**44 kHz**
98.0	98.0

Table 6: *Recognition accuracy for a HMM multi model representation. 12 models were used for this purpose: 6 for 16 kHz and 6 for 44 kHz. Entries are expressed as percentages.*

The recognition accuracy depends on the test segment length to be recognized. As we expected, results improve with long segments. The results for classifying different phrase lengths by using different features are shown in Figure 3. A total of 39 coefficients were used.

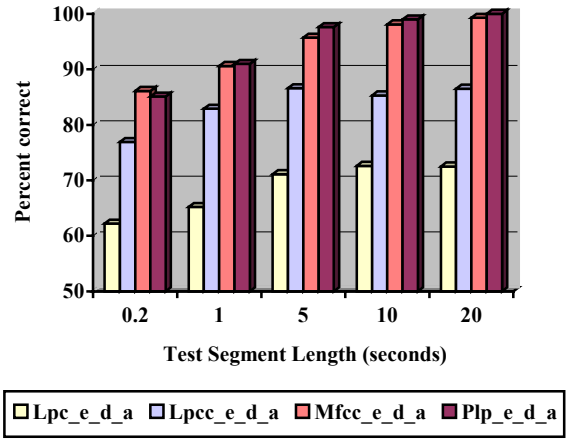


Figure 3: *Classification performance with different test segment lengths given in seconds for different features. The letters "_e_d_a" in the labels refers to addition of energy, velocity and acceleration parameters respectively.*

Table 7 illustrates the performance of the musical sources used in this study given 4 front-end analyses. Experiments were carried out with 1 state and 5 mixtures formed based on 39 features. Order 12 was chosen for LPC, LPCC and PLP. Segments of 10 seconds were used in the test set. LPC provides a bad performance for flute, harp, organ and violin whereas gives good results for recognizing piano sounds (100%). It is shown that LPCC is not very useful for classifying successfully harp and organ sounds (40.2% and 86.1 respectively).

Musical Source	**Lpc_e_d_a**	**Lpcc_e_d_a**	**Mfcc_e_d_a**	**Plp_e_d_a**
Flute	54.2	97.9	99.0	99.0
Harp	32.1	40.2	95.5	100
Organ	68.3	86.1	98.0	97.0
Piano	100	100	97.9	97.9
String Q.	98.9	100	97.9	100
Violin	82.1	100	100	100

Table 7: *Recognition accuracy by musical sources with different parameters. The letters "_e_d_a" in the labels refers to addition of energy, velocity and acceleration parameters respectively. Results are reported as percentages.*

6. CONCLUSIONS

The results obtained in this work show that MFCC and PLP analysis in the HMM framework leads to a good recognition performance. It was shown that both techniques give similar performances, as in speech recognition.

Expectedly, LPC front-end processing gives a worse performance than a recognizer based on filter bank front-ends since it analyses the signal based on the mechanism of human articulation. LP cepstral coefficients (LPCC) have been shown to be a more robust, reliable feature set than the LP analysis (LPC). These results coincide with speech recognition theory where the addition of the cepstral transformation improves recognition performance. LPC provides an unpredictable model for musical instrument recognition purposes. Results vary in a significant manner depending on the order of the model chosen, especially for flute, harp and organ sounds.

All front-end analyses gave good results classifying string quartet excerpts. Results show that the system is capable of distinguishing successfully between string quartet sounds and violin sounds. Taking into account that string quartet refers to a polyphonic sound composed of two violins, a viola and a cello, and the other independent musical sound to be identified was the violin, we can say that the task was relatively more complex in these cases.

As we expected, the accomplishment improves by increasing the length of the excerpts used as test set. From 5 seconds, the differences are not quite marked. Results show that the accuracy of the system does not depend on the sample rate used provided both sample rates for testing and training sets coincide, as speech recognition theory affirms. We have verified that a multi model representation gives the best solution to recognize samples with different sampling frequency as the training set sampling frequency.

7. REFERENCES

[1] Rabiner, L. R., A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. of the IEEE, 77, pp. 257-286, 1989.

[2] Eronen, A. and Klapuri, A. "Musical instrument recognition using cepstral coefficients and temporal features". Proc. of the ICASSP, Vol. 2, pp. 63-71, 2000.

[3] Logan, B., "Mel frequency cepstral coefficients for music modeling", Proc. International Symposium on Music Information Retrieval (ISMIR), 2000.

[4] Marques, J. and Moreno, P., "A study of musical instrument classification using Gaussian mixture models and Support Vector Machine", Tech. Report 99-4, Cambridge Research Lab, 1999

[5] Herrera P., Amatriain X., Batlle E. and Serra X. "Towards instrument segmentation for music content description: a critical review of instrument classification techniques". Proc. International Symposium on Music Information Retrieval (ISMIR), 2000.

[6] Young, S., Evermann, G., Kershaw D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P., "The HTK book", Microsoft Corporation, December, 2001.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

INFLUENCE OF RHYTMIC, MELODIC, AND SEMANTIC VIOLATIONS IN LANGUAGE AND MUSIC ON THE ELECTRICAL ACTIVITY IN THE BRAIN

S Ystad[1], C Magne[2], S Farner[1], G Pallone[1], C Astesano[2], V Pasdeloup[3], R Kronland-Martinet[1], M Besson[2]

[1]CNRS, Laboratoire de Mécanique et d'Acoustique, Marseille, France; [2]CNRS, Physiological and Cognitive Neurosciences Institute, Marseille, France; [3]University of Rennes II, Department of linguistics, Rennes, France

{ystad,farner,pallone,kronland}@lma.cnrs-mrs.fr

{magne,astesano,mbesson}@lnf.cnrs-mrs.fr

ABSTRACT

The work presented here is part of a larger multidisciplinary project associating audio signal processing, linguistics, and cognitive neurosciences. It aims at comparing and better understanding how music and language are processed in the brain. From a music and speech synthesis point of view, this is important when striving for naturalness and expressiveness in synthesized music and language.

As a first experiment towards this goal we have manipulated the rhythm in music and language as well as harmony and semantics, respectively. Since we wanted to work with natural speech, we developed and used a method to extend a given part of an audio signal without altering the timbre. This allows manipulations of the syllable lengths in the language part. In the music part we used piano tones from a sampler as a first approach. The note duration and melody of the musical sequences were digitally modified by altering the MIDI codes.

In the language part of the experiment, participants were presented sentences where the final word was either semantically congruous or incongruous (e.g., "I take coffee with sugar/dog"). In the musical sequences, the final part of the melody was either in or out of tonality. Moreover, the penultimate (second last) syllable or note was either of natural duration or increased in duration, in order to produce rhythmic incongruities. Thus, the two factors rhythm and semantics/tonality were independently manipulated. Changes in brain electrical activity were measured from 28 electrodes on the scalp using an Event-Related Potential method. Preliminary results show that similar reactions can be observed in language and music, at least for rhythmic violations.

1. INTRODUCTION

One of the main motivations behind this study is to get a better understanding of how the brain processes language and music. While the brain's processing of music seem to rely on general cognitive principles, it is not clear whether the processing of language relies on principles specific to the language or on general cognitive principles like in music. By comparing the brain's reactions to violations of rules in language and music, we hope to contribute to a clarification of how these two functions are implemented in the brain and whether similar reactions can be observed both in language and in music [1].

One well known result related to language and semantics was discovered by Kutas & Hillyard [2]. Using the Event-Related brain Potential (ERP) method they showed that a semantically incongruous word in a phrase (i.e. I take coffee with dog) elicited a negative drop in potential 400ms after the word was presented, thus called an N400 component.

The aim of the present experiment was to compare semantic effects with another aspect of language processing, namely the rhythm used to pronounce words and their corresponding syllables. The rhythm is related to the duration and is therefore part of the language's prosody which comprises three main acoustic parameters, namely fundamental frequency, intensity and duration. Previous linguistic speech analyses have showed that the final syllable of the final word of a sentence is generally of longer duration than the preceding syllables [3].

In order to introduce a rhythmic violation in language, we lengthened the penultimate syllable as an attempt to displace the final syllable lengthening effect. In this way we wanted to find out whether such a rhythmic violation would elicit a specific ERP component and whether this effect would be independent of the N400 component related to a semantic violation. We further wanted to find out if rhythmic changes in music elicit an ERP component similar to the ERP component elicited by rhythmic changes in language. In order to modify the syllable duration in language, we developed a special algorithm allowing to stretch part of a word (a syllable) without altering the timbre and the pitch of the voice. This ensures that the effects are only related to the stretching and not to a change in the voice. In the music part of the experiment, we generated the sequences with piano tones from a sampler piloted by a MAX/MSP program. The rhythmic changes were produced by changing the MIDI codes. We will first describe the experimental procedure, then the time-stretching algorithm and the construction of the musical stimuli before summarizing our preliminary results.

2. EXPERIMENTAL PROCEDURE

Different sets of stimuli were made in this experiment. In the language part sentences were manipulated either semantically or rhythmically or both. In the music part melodies comprising a set

of triplets were manipulated either harmonically or rhythmically or both. The ERP method was used to detect the brain's reaction to these manipulations.

2.1 Language part

The aim was to study the relations between semantic and rhythmic processing. Each recorded sentence was different and the last word consisted of three syllables with certain requirements to the phonetic structure. The word should not be a name and never be repeated. A semantic incongruity was obtained by replacing the last word by a word that does not make sense in the context. The rhythmic incongruity was provided by increasing the duration of the penultimate syllable using the time-stretching algorithm described in section 3 on the corresponding vowel. Thus, four different experimental conditions were used;

Semantic manipulation:

- last word fitting in context (S+)
- last word in the sentence out of context (S-)

Prosodic manipulation:

- natural pronunciation (D+)
- artificial lengthening (of 90%) the penultimate syllable of the last word (D-)

The choice of the dilation factor was constrained by the necessity of being ecological (i.e. to keep the naturalness of speech). At the same time, the effect had to be large enough to be audible in order to give the listener the impression of a change in the rhythm from the last to the penultimate syllable. The ratio of the last to the penultimate syllable in French is found to vary considerably with factors such as dialect and speech style, from 1.7 to 2.5 in a study by Astesano [4]. The average ratio for the recorded sentences used in this study is 1.45 with a standard deviation of 0.47. We compromised the above information and chose a dilation factor of 1.9.

In order to obtain a good signal/noise ratio with the ERP method, each participant listened to 32 sentences within each experimental condition, meaning that 128 different sentences (D+ S+, D-S+, D+S- and D-S-) are played to each participant. In one part of the experiment the participant was asked to focus on the semantics of the sentence to determine whether the last word is meaningful or not. She/he should respond by pressing the corresponding of two buttons as fast as possible. Similarly, in the prosodic part of the experiment the participant was asked to focus the attention on the rhythm of the final word and tell whether it sounded natural or not by pressing the corresponding button. Four different experimental lists were made in order to present each sentence in each experimental condition with no repetition. Thus a total of 512 sentences were built for the experiment: 128*2 original recordings and 128*2 that were manipulated with the dilation algorithm.

2.2 Music part

A similar experiment was conducted with musical stimuli. In this case we wanted to study the relations between harmonic and rhythmic incongruity. Small melodies were constructed by five to none triplets, each forming a simple major or minor chord that was natural in the tonality of the melody. Triplets were chosen in for comparability with the language experiment since the last word of each sentence always was tri-syllabic.
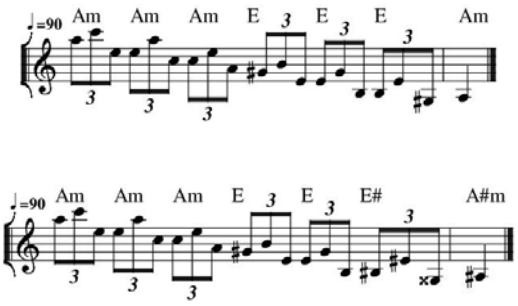


Figure 2 Examples of a melody and its corresponding harmonically modified version. To be coherent with typical musical structure, each melody ends with a fourth note and each triplet consists in an arpeggio of chords.

A MAX/MSP program [5] was made to facilitate the composition of the musical pieces. The program made it possible to choose each chord, whether it is a major or a minor chord and the order of three notes. Rhythmic incongruity could be selected and the dilation factor changed. The harmonic incongruity was made by transposing the last triplet and last note. Thus, the last four notes did always make a harmonically congruous unit, representing an analogy to the last word in the language part of the experiment. Four different beats were chosen (70, 80, 90 and 100 triplets/minute) to allow some variability in the natural tempo as for the spoken sentences. Actually, a simple statistical study of the syllable length in the language part showed that an average number of around 120 tri-syllabic words per minute were pronounced, but such a tempo was considered too fast for the musical part. The beat was fixed for each melody. Four different experimental conditions were constructed

Rhythmic manipulation:

- natural length of the penultimate note of the last triplet (D+)
- lengthening (of 70%) of the penultimate note of the last triplet (D-)

Harmonic manipulation:

- the final triplet consists in a chord which is congruent in the melodic context (S+)
- the final triplet consists in a chord which is not congruent in the melodic context (S-)

As for the language part, each participant listens to 32 melodies within each experimental condition, in total 128 different melodies (D+S+, D-S+, D+S- and D-S) and was asked to pay attention to either rhythm or harmony. Four different experimental lists were made in order to present each melody in each experimental condition. Thus, a total of 512 melodies were built for this part of the experiment.

2.3 Event-Related brain Potential (ERPs)

The ERPs offers an excellent temporal resolution of the brain electrical activity of the order of a millisecond. Changes in the brain electrical activity were recorded for 2200 ms starting 200 ms before the onset of the final word/note, from 28 scalp electrodes, mounted on an elastic cap, and located according to the International 10/20 system. These electrodes, plus an electrode placed on the right mastoid, were referenced to the left mastoid. The data were then re-referenced off-line to the algebraic average of the left and right mastoid. To detect blinks and eye movements, the horizontal electrooculogram (EOG) was recorded from electrodes placed 1cm to the left and right of the external canthi, and the vertical EOG was recorded from an electrode beneath the right eye, referenced to the left mastoid. Trials containing ocular artefacts, movement artefacts or amplifier saturation were excluded from the average ERP waveforms. The EEG and EOG were amplified by a SA Instrumentation amplifier with a bandpass of 0.01-30 Hz, and were digitized at 250 Hz by a PC compatible microcomputer.

3. TIME-STRETCHING ALGORITHM

The time-stretching algorithm used here is based on a method initially developed to adapt the speed of the sound track of a film when converting it from cinema format (24 images per second) to video format (25 images per second) or vice versa [6]. It makes it possible to change the duration of speech and musical signals without timbre and frequency modifications. In this experiment, we optimized the processing for vowels in speech signals to make it possible to use time-stretching ratios α higher than 1.5. This method is an improvement of time-domain methods described in [7-10]. For this specific experiment we could have used a PSOLA method [8]. However our method was directly available and in addition more straightforward. It gives high quality results and is compatible with a wide variety of sounds since we do not need to make any hypothesis on the signal. The principle of the method is illustrated in figure 3.

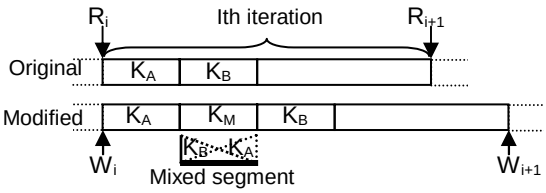


Figure 3 Principle of the time-stretching algorithm.

It relies on the local dilation of a time-domain signal *s(n)* by inserting a short mixed segment K_M, and reiterating this operation as often as needed to achieve the desired ratio.

The dilation factor alpha is then given by

$$(W_{i+1} - W_i) = \alpha(R_{i+1} - R_i) \quad (1)$$

where W_i and R_i stands for "write" and "read" marks.

The inserted segment K_M is obtained by mixing (linear weighting) two successive segments K_A and K_B in order to preserve the continuity of the signal. These segments of identical duration *K* must be as similar as possible. The duration value *K* is obtained by maximizing the following normalized autocorrelation function:

$$C(k) = \frac{\sum_{n=0}^{N-1} s(n+R_i)s(n+R_i+k)}{\sqrt{\sum_{n=0}^{N-1} s^2(n+R_i)\sum_{n=0}^{N-1} s^2(n+R_i+k)}} \quad (2)$$

where *N* is the length of the segment in which the similarity is estimated. The duration value *K* generally corresponds to a multiple of the fundamental period of the signal. It is necessary to limit *K* to values corresponding to fundamental frequencies from 75 to 450 Hz.

The next write mark W_{i+1} is deduced from the former one thanks to equation (1) and by observing that

$$(W_{i+1} - W_i) = (R_{i+1} - R_i) + K \quad (3)$$

which leads to

$$W_{i+1} = W_i - K(\frac{\alpha}{\alpha - 1}) \quad (4)$$

The next iteration will use the modified signal, setting $R_{i+1} = W_{i+1}$.

4. PRELIMINARY RESULTS

As of now, since we have tested 3 subjects, the following results are to be considered as very preliminary results.

Language part

The results showed that semantically incongruent words were associated with a negative component (N400) both in the rhythmically congruous (D+S-) and incongruent (D-S-) conditions, and did not depend on whether the participants focused on semantics or rhythm. The double incongruity condition (D-S-) also elicited a positive component (P600) following the negative one independently of focus. More interestingly, in the rhythmically incongruous and semantically congruous condition (D-S+), the result depended on whether the participants were focusing on semantics or rhythm. When focusing on the rhythm, a positive component was observed,

while different negative components were observed when focusing on the semantics.

Musical part

When the last triplet of the melody was harmonically incongruent, but rhythmically congruous (D+S-) it was associated with a negative component. This was also the case in the double incongruity condition (D-S-) both for the harmonic and rhythmic tasks. In addition the first notes of the last triplet in the double incongruity condition elicited a positive component (P600) in both the harmonic and the rhythmic tasks like in the language part of the experiment. Then, like in the language part, the second note of the last triplet was associated with a positive component in the rhythmically incongruous condition (D-S+) when the participants were asked to focus their attention on rhythm, while they were associated with a negative component when participants were asked to focus on harmony.

5. CONCLUSION

Preliminary results show that rhythmic violations in both language and music seem to be interpreted differently depending on whether a person is asked to focus on semantics (harmony in music) or rhythm. A positive ERP component was observed when the participants were asked to focus on the rhythm, while negative components were observed when participants were asked to focus on the meaning. The ERP components related to the rhythmic violations were similar in language and music. It is important to note that these results are based on an insufficient number of subjects and need to be confirmed by a larger number of participants. The final results of this study will be presented at the conference.

6. REFERENCES

[1] Besson, M. & Schön, D.,"Comparison between language and music". In The biological foundations of music, R. Zatorre & I. Peretz, Eds., Annals of the New York Academy of Sciences, New York, Vol. 930, 2001, pp. 232-258.

[2] Kutas, M. & Hillyard, S. A., "Reading senseless sentences: Brain potentials reflect semantic incongruity". Science, 207, 1980, pp. 203-205.

[3] Pasdeloup, V., "Organisation de l'énoncé en phrases temporelles : analyse d'un corpus de phrases réitérées", in proceedings XVIIIèmes Journées d'Etudes sur la Parole Montréal (Québec), Canada, 28-31 mai 1990.

[4] Astesano, C., "Rythme et accentuation en français – Invariance et variabilité stylistique". Collection Langue et Parole, editions Harmattan, 2001.

[5] Puckette, M. & Zicarelli, D. "Max - an Interactive Graphic Programming Environnement", Opcode Systems, 1985.

[6] G. Pallone, P. Boussard, L. Daudet, P. Guillemain, R. Kronland-Martinet: "A wavelet based method for audio-video synchronization in broadcasting applications". In proc of the DAFX'99, Trondheim, Norway, Dec 9-11 1999.

[7] Verhelst, W. and Roelands, M., "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech", IEEE Proceedings of ICASSP-93, Minneapolis, vol. II, Apr. 1993, pp 554-557.

[8] Hejna, D.J. and Musicus, B.R. and Crowe, A.S., "Method for time-scale modification of signals", U.S. Patent N 5,175,769, Dec. 29, 1992.

[9] Laroche, J., "Autocorrelation method for high quality time/pitch scaling", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, 1993.

[10] Moulines, E. and Charpentier, F., "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones", SC,Vol 9, N 5/6, 1990, pp 453-467.

INSTRUMENT CONTROL

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

REMUTUALIZING THE INSTRUMENT
CO-DESIGN OF SYNTHESIS ALGORITHMS AND CONTROLLERS

Perry R. Cook

Computer Science Dept. (also jointly in Music)
Princeton University, Princeton, NJ, USA
prc@cs.princeton.edu

ABSTRACT

From the advent of electronic music, and even from early organ consoles and other remote manipulated instruments, much of the design and research of new musical interfaces has focused on abstracting the "controller" from the "synthesizer" and then investigating how to best interface those two classes of hardware with each other and the player. Yet, many of the striking lessons from our history of intimate expressive musical instruments lie in the blurred boundaries between player, controller, and sound producing object. Bowed strings, winds, and certainly the human voice all blur these boundaries, both in the design and construction of the "instrument" and in the resulting controls and expressions.

This paper looks at some of the issues involved with creating new expressive electronic musical instruments, and presents a number of recent projects in the co-design of musical controllers and computer sound synthesis algorithms. Specific cases are described where the traditional engineering approach of building a controller (a box), and connecting it to a synthesizer (another box) would never have yielded the final product that resulted from the tightly coupled development of a complete musical system all at the same time. Examples are given of where a discovery from synthesis algorithm development suggested a new control metaphor, and where a control component suggested a new aspect of synthesis.

1. INTRODUCTION

On first reflection, one might assume that the notion of a musical "controller" has a relatively recent history, dating perhaps only since the advent of MIDI and digital controllers and synthesizers. But the controller/generator metaphor has a longer history than might be expected. By necessity of it's size and other factors, the organ console is likely the earliest of common instruments to have separated the controller (the console) from the tone-generator (the pipes). The earliest specimen of a hydralis (water powered) organ dates to before 300 BC [1]. By the late 1700's, organs were growing in size, were finding they way into more churches, and builders were devising new means to physically separate the console and pipes. Tracker mechanisms, pneumatically assisted valves, and eventually electronic switches and valves all increased the distance between player and sound production [2].

In the industrial era of the early 1900's, various player, reproducing, and recording pianos possessed features such as speed and volume controls, and some even had remote controls for those functions (a flexible cable with handle that, when turned, changed a motor speed, or a handle/cable that opened and closed shutter doors for muting the sound). Many Victrolas and cylinder players also possessed remote speed and volume controls.

Of course the notion of transforming music listening into an afferent act, with a human issuing commands to an instrument (playing machine), the machine responding with a "perfect" performance, and with the auditory channel being the only mechanism of feedback, has a distinct bias. The notion that everyone in the world wants to be an "armchair conductor" is a well-held tenet of modern times, but it has little to do with the construction, playing, and personal enjoyment of most musical instruments.

The major flaw in the controller/synthesizer paradigm is the loss of intimacy between human player and instrument. I pose three primary reasons (lacks) for this intimacy loss:

- Lack of haptic feedback from the controller/instrument to the player. Haptic (combined senses of touch, including skin vibration and pressure, and the muscle senses of motion, position, and force) feedback has been increasingly addressed in musical interface research projects. Commercially, the most successful haptic systems are electronic keyboards that copy (passively, through weights and levers) the feel of piano keys.

- Lack of fidelity in the connections from the controller/sensor to the generator, primarily delays and distortions in response to gestures. "Distortion" here refers to any response that doesn't meet some usual, learnable, or repeatable expectation.

- Lack of any sense that sound comes from the instrument (the controller) itself. More generally, this is a subset of a larger feeling that no meaningful physics goes on in the controller. Trends toward larger concert venues, greater amplification, and larger loudspeakers have consistently worked to diminish the importance of the actual acoustical instrument sound. The aesthetic toll of this has been great, most importantly for the player and composer, and shows profoundly in the musical results.

Figure 1 shows a block diagram of a typical electronic music controller system. Obvious potential flaws in a system of this type include the one-way flow (lack of feedback to the player) from controller to synthesizer to speaker, and the potential latencies and other distortions that might result from the sensors, quantization, transmission, etc.

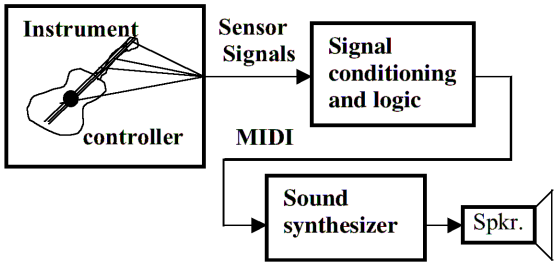


Figure 1: *Standard controller/synthesizer system*

Figure 2 shows a block diagram of a human performer connected to a musical instrument, specifically noting the multiple auditory and haptic feedback paths. Degradation in the playing "feel," (and in any resulting music) can result from distortion of any path in either system of Figure 1 or 2.

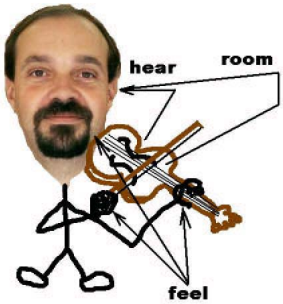


Figure 2: *Traditional acoustic player/instrument system.*

Most work in musical interfaces, protocols, and systems has worked to improve the fidelity and flexibility of the forward control channels (Figure 1), by improving the sensors, connection speeds, signal processing algorithms, computer hardware, and synthesizer quality. Much less work has been done on the feedback (Figure 2) channels. The next sections will describe a number of recent projects that attempt to address the "lacks" listed earlier, by completing and improving both the feedforward and feedback paths.

2. CLOSED-LOOP MULTI-MODAL SYNTHESIS: THE HAPTIC MARACA

Based on the Physically-Informed Stochastic Event synthesis Model (PhISEM) [3], a number of controllers containing accelerometers, force-sensing resistors, and switches were developed to translate the gestures of shaking and scraping into parameters for the particle synthesis model. Early controllers (Figure 3) were selected for their relation to the origins of the computational model. Some controllers were constructed to give a natural "feel" of the sound/instrument being controlled. For example, a ratchet device with multi-position rotary switch can feel like the ratchet sound it controls. This is an example of "passive haptics," where no explicit forces are computed and fed back to motors in the controller, but rather the basic feel of the controller is matched as closely as possible to the sound (or the sound is matched to the feel). This is the type of haptic interface provided in commercial "weighted" electronic pianos.

In constructing the early controllers shown in Figure 3, it was noted that the synthesis algorithm automatically provides useful information about the sound production physics. In PhISEM, each sound producing particle collision is computed. The Haptic Maraca was an early attempt at mutualizing synthesis and control, by feeding the PhISEM particle collision impulses to a small solenoid motor (mass glued to a speaker coil) located in the maraca gourd. Figure 4 shows the PhISEM haptic maraca controller, with the Analog Devices 2181 DSP board used to compute the model. Using the DSP board afforded essentially no delays between control gestures and audio/haptic responses. Even though the system had wires attached, testers reported an uncanny feeling of connection between gesture, sound, and feel.

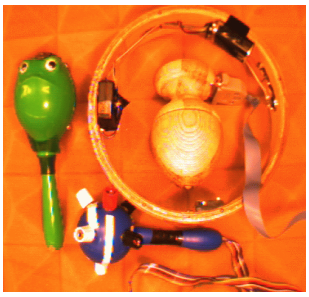


Figure 3: *Early PhISEM synthesis controllers.*

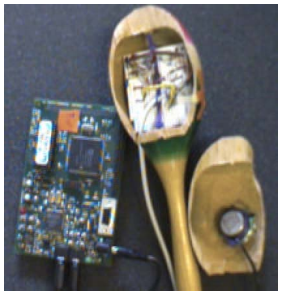


Figure 4: *The PhISEM haptic maraca controller.*

3. A CRAZY IDEA: SOUND FROM THE ACTUAL INSTRUMENT

3.1 The Bowed Sensor Speaker Array (BoSSA)

The Nbody project [4] was inspired by a desire to make new musical instruments that engage the performers, the performance space, and the audience more effectively than the forward-facing stage speakers most often used in computer music concerts. It was also desired to extend the work of Weinreich [5], Causse [6], and others in the directional sound radiation of acoustic instruments, but to collect and freely distribute a large number of radiation transfer functions for a variety of instruments. Multi-directional (12–72) impulse responses were collected from six different stringed instruments. These were studied and used to construct various simulated and virtual performance instruments.

The Bowed Sensor Speaker Array (BoSSA) [7] was the first project to fully integrate a spherical radiating speaker body and a complete electronic performance interface into one musical artifact (Figure 5). Inspired by the violin, BoSSA consists of a set of four bowed sponges (the Bonge) to emulate the strings, a bow with force and acceleration sensors (the Rbow), a violin fingerboard with linear force and tilt sensors (the Fangerbored), all mounted to a spherical speaker with 12 drivers (the Critter). This instrument allows a new kind of intimacy for computer music performance, with the sound coming from the actual controller itself.



Figure 5: *Dan Trueman's BoSSA.*

3.2 SqueezeVoxen: "Singing" Accordions

As might be expected, attempts to create controllers for computer voice models consistently point up the same problems. From a technical standpoint, the sheer number of parameters that need to be controlled in an expressive voice model present daunting issues of sensors, bandwidth, systems, and mappings. From a musical, linguistic, and perceptual standpoint, all humans posses a voice, have years of experience "playing" it (not necessarily musically, but expressively) and attending to the voices of others. Thus humans are extremely critical of synthesized voices.

Even given the problems, a further attempt at creating a meaningful and expressive (or at least fun) interface for controlling computer voices was undertaken with Colby Leider in 2000. The SqueezeVox project [8] recognized that to control a vocal model, independent controls are needed for pitch, breathing, and phoneme articulation (spectral features). The accordion is an instrument with components that map somewhat naturally to these requirements. Melody pitch is controlled with the right hand keyboard, "breathing" is provided naturally in the bellows mechanism (though the accordion "sings" when breathing both in and out), and the left hand provides an array of buttons (10 or so in a concertina or 12-bass instrument, 100+ in a full sized instrument).

The SqueezeVox project exploited these features of the accordion to control a variety of voice models (formant models, acoustic tube models, FOF synthesizers, orchestras of chanting monks, and more). Air pressure sensors, a linear FSR located next to the keyboard, left-hand buttons and four nearby bend sensors (addressing/controlling phonemes, articulator positions, or formant positions), internal speakers, tilt sensors, and other sensors were incorporated into a variety of instruments. Bart, Lisa (shown in Figure 6), Maggie (a concertina) and Santa's Little Helper (a child's toy accordion) make up the complete fleet of SqueezeVoxen.



Figure 6: *SqueezeVox (Lisa) vocal synthesis controller.*

4. FULLY INTEGRATED PHYSICAL MODEL SYNTHESIS & CONTROL: THE NUKELELE

Research on physical models, and work at Interval Research on control projects such as the Virtual Haptic Maraca and "The Stick" [9] led to a question of just how tightly one could bind a virtual instrument to it's form, algorithm, and playing technique. The result was the Digital Koto, or the "Nukelele" (dubbed this name by Michael Brook). The main idea of the Nukelele was to attempt to create a true virtual stringed instrument, with responsiveness to all the subtleties of rubs, damps, and plucks of many types. Further, as with most projects in virtual instrument design, it was desired to extend the capabilities beyond what an actual stringed instrument can do.

The Nukelele uses a simple Karplus-Strong plucked string model, controlled by linear force sensors, to simulate the interactions of a virtual string. The left hand FSR behaves as a fretless fingerboard, and can also sense hammer-on and hammer-off gestures. The right hand "string" is pressure sensitive (to sense damping), position sensitive (for pluck position), and also sends an audio signal into the plucked string model in response to striking and rubbing. This audio signal is the key to the intimacy of the interface, where even the slightest scratching or rubbing causes the virtual string to oscillate. Figure 7 shows a block diagram of the Nukelele synthesis system, and Figure 8 shows the original Nukelele.

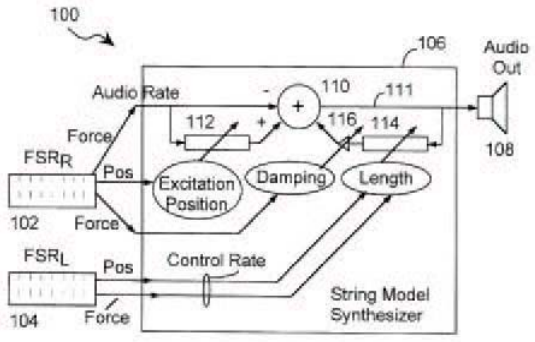


Figure 7: *The Nukelele integrates the components of a physical stringed instrument. The key is in the audio-rate excitation force signal fed into the string, allowing plucking, striking, and rubbing, in addition to position and damping.*



Figure 8: *The original Nukelele instrument.*

A 2nd generation Nukelele, called the Nukelele'elua (Figure 9) was constructed, adding a longer fingerboard FSR, and two speakers in the body of a 1/2 size guitar. The right hand FSR is situated in a "strummer" orientation, allowing multiple virtual strings to be plucked. One speaker faces outward, and the other back toward the player. Different filter functions are used to drive these two speakers, based on impulse responses collected in the NBody Project [4].



Figure 9: *The Nukelele'elua (Nukelele-2).*

5. CONCLUSIONS

With some effort, new electronic instruments can be constructed that extend our notions of acoustical instruments, but keep a sense of intimacy, connectedness, and embodiment for the player and audience. Other experiments with self-contained sensor/speaker systems [10] have yielded provocative and rewarding artistic results, and the author looks forward to building more of such systems in the future.

6. ACKNOWLEDGEMENTS

Thanks to Bob Adams and the Interval Research Expressions group for many inspiring and creative "work-jams." Thanks to Interval Research, specifically Bill Verplank and Geoff Smith, for making the artifacts of my work there available to me after Interval's demise. Thanks to Dan Trueman for great collaborations on Nbody, Rbow, BoSSA, and for good discussions on "that most perfect of musical instruments." Thanks to Colby Leider for cooking up and carrying out the SqueezeVox project with me.

3. REFERENCES

[1] www.organisten.de/oberlinger/agonline/200207.htm, The Hydralis of Dion, D. Panternalis, AGO Online, July 2002.

[2] Snyder, K. ed. "The Organ as a Mirror of Its Time: North European Reflections, 1610-2000," Oxford University Press, 2002.

[3] Cook, P. "Physically Informed Sonic Modeling (PhISM): Synthesis of Percussive Sounds," Computer Music Journal, 21:3, 1997.

[4] Cook, P. and Trueman, D. "Spherical Radiation from Stringed Instruments: Measured, Modeled, and Reproduced," Journal of the Catgut Acoustical Society, November 1999.

[5] Weinreich, G. "Sound Radiation from the Violin – As We Know It Today," Journal of the Catgut Acoustical Society, May, 2002.

[6] Caussé, R., Bresciani, J.F., and Warusfel, O. "Radiation of Musical Instruments and Control of Reproduction with Loudspeakers," Proceedings of the International Symposium of Music Acoustics, August 1992.

[7] Trueman, D. and Cook, P. "BoSSA: The Deconstructed Violin Reconstructed," Journal of New Music Research, Fall, 2000.

[8] Cook, P. and Leider, C. "SqueezeVox: A New Controller for Vocal Synthesis Models," Proceedings of the International Computer Music Conference, 2000.

[9] Levitin, D., McAdams, S. and Adams, R. "Control Parameters for Musical Instruments: A Foundation for New Mappings of Gesture to Sound," Organised Sound, 7(2), pp. 171-189, 2002.

[10] Trueman, D., Bahn, C. and Cook, P. "Alternative Voices for Electronic Sound: Spherical Speakers and Sensor-Speaker Arrays (SenSAs)," Proceedings of the International Computer Music Conference, August 2000.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

INTERACTIVE MUSIC SYSTEMS FOR EVERYONE: EXPLORING VISUAL FEEDBACK AS A WAY FOR CREATING MORE INTUITIVE, EFFICIENT AND LEARNABLE INSTRUMENTS

Sergi Jordà
Music Technology Group
Universitat Pompeu Fabra
Ocata 1, 08003 Barcelona, Spain
sergi.jorda@iua.upf.es

ABSTRACT

New digital musical instruments designed for professional and trained musicians can be quite complex and challenging, offering as a counterpart a great amount of creative freedom and control possibilities to their players. On the other hand, instruments designed for amateur musicians or for audiences in interactive sound installations, tend to be quite simple, often trying to bring the illusion of control and interaction to their users, while still producing 'satisfactory' outputs. Logically, these two classes of instruments are often mutually exclusive. But wouldn't it be possible to design instruments that can appeal to both sectors? In this paper we will show, with two projects developed in our research group, how visual feedback can highly increase the intuitiveness of an interactive music system, making complex principles understandable.

1. INTRODUCTION

1.1. Music Instruments: Questions

What is a *good* music instrument? Are there instruments better than others? These are tricky questions. Obviously, some instruments are more powerful, flexible or versatile than others, but there are so many dimensions by which a musical instrument can be evaluated, that comparisons between most instruments do not make sense, unless we clearly specified the parameters involved in the evaluation. Parameters such as degrees of freedom of control, correlation between these controls, difficulty of use, apprenticeship and learning curve scalability,… are parameters that deal with the use of an instrument, that is with the relation between the instrument and its player. Sonic richness and variability, both at a macro level (e.g. dynamic range, tessitura…) and nuances at a micro level (e.g. pitch "resolution", portamento, vibrato and other modulation capabilities…), deal more directly with the output possibilities of the instrument.

Other measurable dimensions could be the instrument's temporal precision, its responsiveness, or its control reproducibility and variability. For example: how similar can two performances be? How predictable are the outputs of small control changes? Some instruments behave definitely less linear than others. How could we evaluate the potential expressiveness of an instrument? Can expressiveness only be studied from a cultural viewpoint or is there an absolute way to evaluate it? Would it possibly be, in that case, related to non-linearity?

1.2. Music Instruments: Possible Taxonomies

Too many dimensions and far too many questions to be seriously discussed within the scope of these pages. Many authors have addressed classifications and taxonomies, both for traditional and new music instruments, based on several of their properties. Joseph Paradiso's exhaustive classification of new electronic instruments [1] can hardly be surpassed, but his article is more written from the *techno-luthier*'s viewpoint that from the player´s or from the listener's one. Gabriele Boschi's instruments classification [2] (which considers both traditional and electronic instruments) is focused on the kind of control parameters instruments allow (quantity, discrete vs. continuous, etc.) and how they apply to sound (pitch, dynamics, spectral content, articulation, transition between notes, phrasing, rhythm, etc.). Many more instruments classifications and surveys could be added, although I do not know of any of them which systematically studies concepts such as playability or apprenticeship scalability, not even for traditional instruments[1].

1.3. Music Instruments: Apprenticeship and Playability

Acoustic musical instruments as we know them now are the fruit of centuries or even millennia of evolution; they have settled into canonical forms. But that does not necessarily imply that these instruments should always excel at whatever parameter we evaluate. Let's consider the learning curve, for instance. Many traditional instruments are quite frustrating for the beginner. A violin, to mention only one, can hardly be taken as a toy (a piano could). Do not misunderstand me: I am not suggesting that the violin could be improved or that it should be made more toylike; the violin excels in many other dimensions (e.g. sound nuances)

[1] An essential selection of recent authors that have studied from different perspectives "how instruments are used", trying to infer from that "how they could be better designed", should at least include Perry Cook [3], Andy Hunt [4], Axel Mulder [5], Joel Ryan [6], Marcelo Wanderley [7] or David Wessel [8], among many others. The CD-ROM *Trends in Gestural Control of Music* [9] constitutes also a perfect introduction to this topic.

that are directly related with its initial difficulties. What I am affirming is that there is a lot to be studied from the perspectives of ergonomics and playability, and that this knowledge can only improve the design of new instruments. As Axel Mulder points, "we know how to make musical instruments that sound great, but can we make them easier to learn, better adapted to each individual's gestural expression preferences and movement capabilities?" [5].

At the opening of the NIME 2002 conference in Dublin, keynote speaker Tod Machover launched several questions of which I here retain two: "How do we create controls and interactions that feel *inevitable* to expert and amateur users?", "how do we create interactive situations that stimulate rather than placate, leading the participant beyond the surface and into thoughtful consideration of rich, expressive, meaningful experiences?". According to him, the last two decades have seen successful designs of controllers capable of virtuosity and subtlety, and also of controllers that "hook" novice users, but in this last case, very few systems have been nourishing as well, capable of encouraging deeper exploration and continued discovery and creativity [10].

2. INTUITIVE YET SOPHISTICATED MUSIC INSTRUMENTS

2.1. Playability and Efficiency of a Music Instrument

Let's write a simple equation (with quite fuzzy variables!) that will allow us to evaluate the playability *efficiency* of a music instrument according to the following ratio:

$$Efficiency = \frac{MusicalOutputComplexity}{InstrumentInputComplexity} \quad (1)$$

More complex instruments are usually capable of more complex music (e.g. the piano vs. the kazoo), but this equation is clearly misleading. What happens if we consider for example the CD player[2]? It is an instrument very simple to use, yet capable of all the imaginable music complexity. This tricky illusion is used indeed in many of the interactive sound installations which want to guarantee a complex musical output: they do not give to their interactors more than a couple of bits to play with. So let's modify our equation with an additional term that we will call *freedom*, vaguely dependent on the degrees of freedom accessible to the player, and the range of each of these degrees.

$$Efficiency = \frac{MusicalOutputComplexity \times PlayerFreedom}{InstrumentInputComplexity} \quad (2)$$

The goal is settled and it is an ambitious one: let us design and build instruments that can appeal to professional musicians as well as to complete novices; efficient instruments that like many traditional ones can offer a *low entry fee with no ceiling on virtuosity* [8]; systems in which basic principles of operation are easy to deduce, while, at the same time, sophisticated expressions are possible and mastery is progressively attainable.

2.2. Reintroducing FMOL

My first attempt at building interactive systems that tried to satisfy these conditions came with FMOL, a project I started in 1997. FMOL has evolved since its debut and several articles have already been written ([12-14]). In this paper I want to mention only the peculiar aspects brought by FMOL's unique user interface, which presents a closed feedback loop between the sound and the graphics: in FMOL, the same GUI works both as the input for sound control and as an output that intuitively displays all the sound and music activity.

FMOL has been used in many different contexts: as a tool for collaborative music on the net, as a system for introducing newcomers into experimental electronic music making (I have given many workshops to non-musicians and even to 5 to 8-year old kids!) and as an instrument for professionals as the concerts and CDs of the FMOL Trio may prove, and I'm convinced that FMOL's unique visual feedback has been a fundamental component for its success as a powerful and at the same time intuitive and enjoyable instrument. Its mouse-controlled GUI is so tightly related to the synthesis engine architecture, that almost every feature of the synthesizer is reflected in a symbolic, dynamic and non-technical way in the interface, that works both as an input devices (i.e. a controller) that can be picked and dragged with the mouse, and as an output device that gives dynamic visual feedback, and as anything in the screen behaves simultaneously as an output and as an input, no indirection is needed to modify any of these parameters

3. THE REACTABLE*

3.1. Mouse Limitations and New Intentions

Looking at the way people have used FMOL, and using it myself for improvisation in different contexts, raised ideas for new features and modifications. But I also felt that this control complexity could not be permanently increased: for availability reasons, the original FMOL had to be a mouse-driven software and there are limits to what can be efficiently achieved in real-time by means of a mouse and a computer keyboard. These and other considerations took us to a completely new path, which should profit the knowledge gained during these years and bring it to a much more ambitious project.

The *reacTable** aims at the creation of a state-of-the-art interactive music instrument, which should be collaborative (off and on-line), intuitive (zero manual, zero instructions), sonically challenging and interesting, learnable, suitable for complete novices (in installations), suitable for advanced electronic musicians (in concerts) and totally controllable. The *reacTable** uses no mouse, no keyboard, no cables, no wearables. It allows a flexible number of users, and these will be able to enter or leave the instrument-installation without previous announcements. The technology involved should be, in one word, completely transparent [15].

2 Meaning regular CD-players and regular discs! Not the manipulated ones like used by Yasunao Tone or Nicolas Collins [11].

4.2. Computer Vision and Tangible Objects

As the Tangible Media Group directed by Professor Hiroshi Ishii at the MIT Media Lab states, "People have developed sophisticated skills for sensing and manipulating our physical environments. However, most of these skills are not employed by traditional GUI…. The goal is to change the *painted bits* of GUIs to *tangible bits*, taking advantage of the richness of multimodal human senses and skills developed through our lifetime of interaction with the physical world" [16]. Several tangible systems have been constructed based on this philosophy. Some for musical applications, like *SmallFish* [17], the *Jam-O-Drum* [18], the *Musical Trinkets* [19], *Augmented Groove* [20] or the *Audiopad* [21], but we believe that no one attempts the level of integration, power and flexibility we propose: a table-based collaborative music instrument that uses computer vision and tangible user interfaces technologies, within a MAX-like architecture and scheduler, and with FMOL-inspired HCI models and visual feedback.

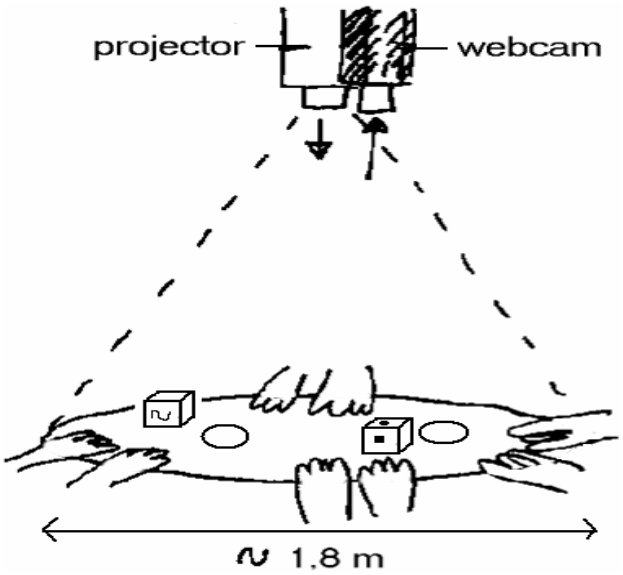


Figure 1: The *reacTable** simplified scheme.

The *reacTable** is a musical instrument based on a round table, which has no sensors, no cables, no graphics or drawings. A video camera permanently analyses the surface of the table, tracking the hand movements over the table, and detecting the nature, position and orientation of the objects that are distributed on its surface, while a projector draws a dynamic and interactive interface on it. These objects are mostly passive and made out of plastic or wood of different shapes. Users interact with them by moving them, changing their orientation on the table plane or changing their faces (in the case of volumetric objects). More complex objects include (but are not limited to) flexible plastic tubes for continuous multiparametric control, little wooden dummy 1-octave keyboards, combs (for comb-filters), or other everyday objects. Figure 1 shows a *reacTable** scheme.

4.3. Visuals

The projection follows the objects on the table wrapping them with auras or drawing figures on top of them, and covers also the whole table surface with dynamic and abstract elements that reflect all the system's activity, the objects' types and positions, and the relations between them all. The projection never shows buttons, sliders or widgets of any kind.

Like MAX and all of its cousins, the *reacTable** distinguishes between control and sound objects, and between control and sound connections. When a control flow is established between two objects, a thick straight line is drawn between them, showing by means of dynamic animations, the flux direction, its rate and its intensity. Visual feedback also guarantees that LFOs and other macrotemporal are perceived as blinking animations projected on top of the related objects, showing frequency and shape (e.g. square vs. sinusoidal).

While control flow lines are straight and simple, audio flow lines are *organic* and complex, as shown in figure 3. Their dynamic shapes show the macrotemporal audio variations (vibratos, tremolos, tempo and rhythms…) and their interior (colors, intensities…) depend on their spectral audio content. Users will also be able to control, modify or fork audio flows without using additional objects, but just by waving their hands, as if they were digging water channels in the beach sand.

4.4. Avoid user's frustration at any cost

To avoid frustrations, a system does not necessarily have to be completely understandable, but it has to be coherent and responsible. Unlike MAX, the *reacTable** has to work "by default" and any gesture has to produce audible results. Here are some of its laws: objects are not active until they are touched; active objects have a dynamic visual *aura;* objects are interconnected by proximity; moving an object on the table can change the relations with the other objects.

4.5. Future Work

The *reacTable** project has started in December 2002 coinciding with the foundation of the Interactivity Team within the Music Technology Group. We are currently working and researching all the main threads in parallel (computer vision and objects recognition, sound engine architecture, interactivity logic, sound visualization, etc.). Meanwhile we are designing the core and the integration of all its branches, and a virtual software-only version (shown in Figure 2) is already available. The *reacTable** is an ambitious project. Unlike many new designed instruments, its origin does not come from approaching its creation by exploring the possibilities of a specific technology, nor from the perspective of mimicking a known instrumental model. The *reacTable** comes from our experience designing instruments, making music with them, and listening and watching the way others have played them. Needless to say, we have deposited a great hope and expectation on it. We plan to have the first integrated by autumn 2003 and a first full working version by spring 2004.

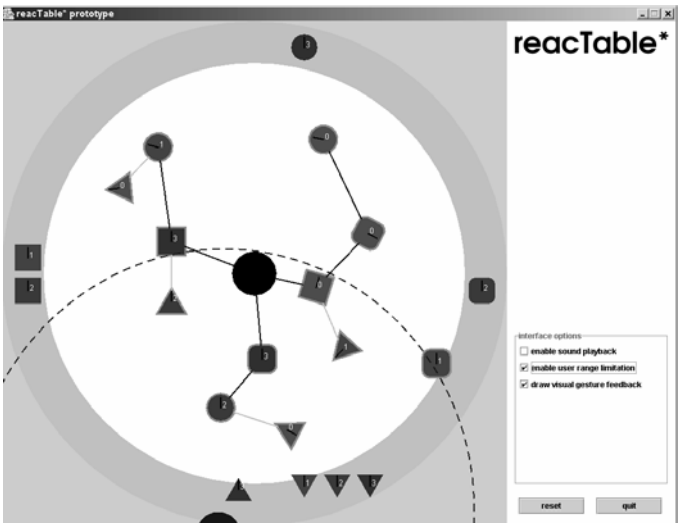


Figure 2: *reacTable** simulator snapshot

5. CONCLUSION

The field of new instruments and interactive music systems design seems to be lately in good shape. A yearly international conference (New Interfaces for Musical Expression, NIME [22]) is being held since 2001, a COST action (Con-Gas, for Gestured Controlled Audio Systems) has just started this year, and specialized journals are devoting more issues to this topic. Personal computers are already capable of complex real-time audio operations that demand for new interactive control interfaces. Yet the field is still in its infancy.

Not many recent electronic instruments have even attained the reduced popularity of the Theremin, created in 1920. The next public or standard electronic instruments after the Theremin may have not arrived yet. Or they have in fact (like the electric guitar and the turntable), but they are still not digital ones. Are we supposed to see a digital instrument standardization? Do we really need the new musical controller? Probably not. But on the other side, highly idiosyncratic instruments which are only used by their respective creators may not be the best sign or strategy for a serious evolution in this field.

New music instruments or controllers convey also new possibilities and paradigms to the act of making music. In this paper we have only scratched a few, like the chance to bring the joy of real-time music creation to non-trained musicians.

6. REFERENCES

[1] Paradiso, J., "Electronic Music: New ways to play." IEEE Spectrum, 34(12): 18-30, 1997.

[2] Boschi, G., “A Classification Of Controllers For Parameters Highlighting”, MOSART Report on Control and Virtualisation of Instruments, 2002. Available at http://www.diku.dk/forskning/musinf/mosart/midterm/T3/Boschi-controllers.pdf

[3] Cook. P., "Principles for Designing Computer Music Controllers", NIME Workshop – CHI, 2001.

[4] Hunt, A. 1999. “Radical User Interfaces for Real-time Musical Control”, PhD Thesis, University of York UK.

[5] Mulder, A., “Virtual Musical Instruments: Accessing the Sound Synthesis Universe as a Performer”, Proceedings of the first Brazilian Symposium on Computer Music, 1994.

[6] Ryan, J., “Some Remarks on Musical Instrument Design at STEIM”, Contemporary music review, 6(1), pp. 3-17, 1991.

[7] Wanderley, M., “Performer-Instrument Interaction: Applications to Gestural Control of Music”. PhD Thesis. Paris, France: University Pierre et Marie Curie - Paris VI, 2001.

[8] Wessel, D. and M. Wright, "Problems and Prospects for Intimate Musical Control of Computers." NIME Workshop – CHI, 2001.

[9] Wanderley, M. and M. Battier, eds. CD-ROM, “Trends in Gestural Control of Music”, Ircam-Centre Pompidou, 2000.

[10] Machover, T. “Instruments, Interactivity, and Inevitability”, Proceedings of the NIME International Conference, 2002.

[11] Archives of Silence, 1996:Q2 Music for CD http://www.newalbion.com/artists/cagej/silence/html/1996q2/0098.html

[12] Jordà, S., “Faust music On Line: An Approach to Real-Time Collective Composition on the Internet.” Leonardo Music Journal, Vol. 9, pp. 5-12, 1999.

[13] Jordà, S., Wüst, O., “Architectural Overview of a System for Collaborative Music Composition Over the Web.” Proceedings of the 2001 International Computer Music Conference. International Computer Music Association, 2001.

[14] Jordà, S., “FMOL: Toward User-Friendly, Sophisticated New Musical Instruments.” Computer Music Journal, Vol. 26, No.3. pp 23-39, 2002.

[15] Jordà, S., “Sonigraphical Instruments: From FMOL to the reacTable*”, NIME Proceedings, 2003.

[16] Ishii, H and B. Ullmer, “Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms”, Proceedings of CHI 97 Conference on Human Factors in Computing systems, Atlanta, Georgia USA, 22-27 March 1997.

[17] SmallFish homepage:
http://hosting.zkm.de/wmuench/small_fish

[18] Blaine, T. and T. Perkis, "Jam-O-Drum, A Study in Interaction Design," Proceedings of the ACM DIS 2000 Conference, ACM Press, NY, August 2000.

[19] Paradiso, J. & Hsiao, K., ”Musical Trinkets: New Pieces to Play”, SIGGRAPH 2000 Conference Abstracts and Applications, ACM Press, NY, July 2000, p. 90.

[20] Poupyrev, I, Augmented Groove: Collaborative Jamming in Augmented Reality, ACM SIGGRAPH 2000 Conference Abstracts and Applications, p. 77.

[21] Audiopad homepage:http://tangible.media.mit.edu/projects/

[22] NIME: http://www.nime.org

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

CONTROLLING THE VIRTUAL BODHRAN - THE VODHRAN

Roberto Bresin[1], *Sofia Dahl*[1], *Matthias Rath*[2], *Mark Marshall*[3], *Breege Moynihan*[3]

[1]Department of Speech Music and Hearing – KTH Royal Institute of Technology, Stockholm
{roberto, sofia}@speech.kth.se

[2]Dipartimento di Informatica – Università di Verona
rath@sci.univr.it

[3]Interaction Design Centre – University of Limerick
{Mark.T.Marshall, Bridget.Moynihan}@ul.ie

ABSTRACT

In order to have a successful interface it is preferred to employ a metaphor that the end user of the artefact is familiar with. The application described in this paper aims to provide users with an expressive virtual musical instrument, based on the traditional Bodhran. This is not designed entirely to simulate the Bodhran, but to provide an instrument which can be played in a similar way, and which creates a recognisable Bodhran-like sound; a virtual Bodhran, the Vodhran. This instrument is to be an extension of the original, allowing for additional playing techniques and styles, which could not be accomplished with the real instrument. In two experiments the same sound model implementing the Bodhran was used. In the first experiment the sound model was controlled both with data obtained from measurements of gestures of drum player, and with a traditional playing approach by using two different controllers. The Radio Baton and the ddrum (a drum pad) were used as interfaces for this experiment. Both interfaces were controlled with a traditional bodhran double beater. In the second experiment a sensor-based controller was used to translate player's gestures into sound model controls. The sound generation mechanism for the Vodhran is based on the modal description of the drum and a robust numerical solution of a nonlinear stick-membrane interaction model. The model includes different forms of a player's interference.

1. INTRODUCTION

In order to have a successful interface it is widely known that it is preferred to employ a metaphor that the end user of the artefact is familiar with. In the following pages we illustrate an application that aims to provide users with an expressive virtual musical instrument, based on the traditional *Bodhran*, the *Vodhran*. This is not designed entirely to simulate the *Bodhran*, but to provide an instrument which can be played in a similar way, and which creates a recognisable Bodhran-like sound. This instrument is to be an extension of the original, allowing for additional playing techniques and styles, which could not be accomplished with the real instrument.

In the *Bodhran*, sound is emitted in response to a stick (beater) beating the skin, generally executed by the right hand. The sound is modulated/dampened by the use of pressure placed on the inside of the *Bodhran* by the left hand. The beater is held loosely in the hand and is moved/controlled primarily by wrist action (rotation). The contact with the *Bodhran* is made with alternative ends of the beater in rapid succession. The damping factor is discrete application of pressure, but more often than not a dynamic/colourful range in pitch can be achieved by continuous control over the damping achieved by its application in counter direction to the beater. There are a variety of different applications of the damping factor employed by the musicians e.g. fingers only, side of hand only (see Figure 1).

In the following sections four control modalities are discussed. In all of them the same sound model implementing the *Bodhran* was used.



Figure 1: Traditional *Bodhran*.

The sound generation mechanism for the *Vodhran* is based on the (approximate and simplified) modal description of the drum and a robust numerical solution of a nonlinear stick-membrane interaction model [1]. The model is described in detail by Matthias Rath in [2].

2. CONTROLLING THE VODHRAN

The sound model was controlled with four different modalities. These involve both software and hardware controllers (see Figure 2) , and are presented in the following.

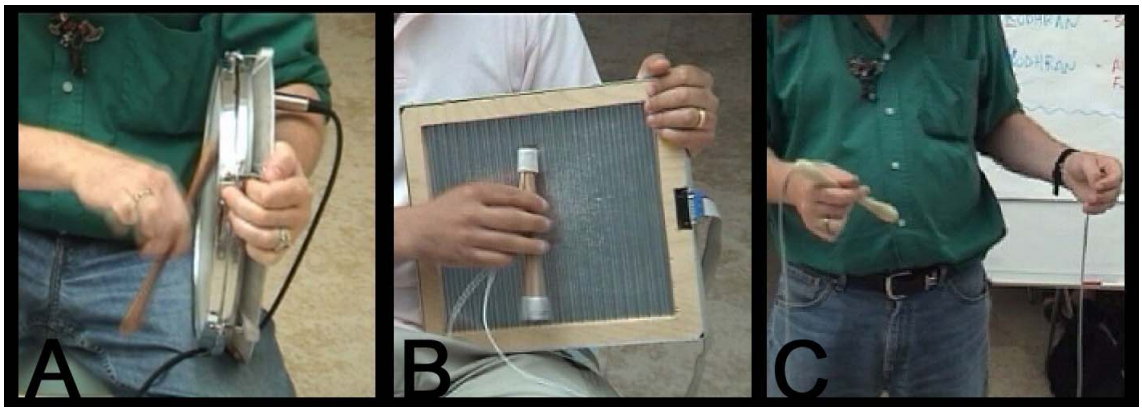


Figure 2: The three hardware controllers used to implements the Virtual *Bodhran*, the *Vodhran*. (A) The ddrum, (B) the Radio Baton, and (C) the Polhemus Fastrack.

2.1. Control by modelling real players

A `pd` patch implementing the model of professional percussionists was used for controlling the sound model of the *Bodhran*. This control model was obtained by the measurements described [3]. The measured striking velocities of two players were used as excitation velocity for the sound model. The `pd` patch, a "metronome rhythm generator", plays metronome beats with a striking velocity that is affected by dynamic level (*pp*, *mf*, or *ff*), chosen playing style (symphonic or drumset) and tempo. Any time an accent can be generated, and depending on the current tempo, dynamic level and playing style the striking velocity exciting the model will be changed accordingly. The shift between the two playing styles can also be generated in real-time.

To enable a swift change between the different playing styles and dynamic levels these control parameters were connected to an external custom made switch board. The switches were connected to the PC through a device to take analogue signals through the parallel port, the `pico` [4].

The control for tempo was mapped to the Korg KAOSS pad, a MIDI controller with a touch pad. The x-axis on the touch pad controls the tempo between 100 and 250 beats per minute. The y-axis controls the fundamental frequency of the physical model. By using the touch sensitivity, from the touch pad it is possible to start the metronome by touching the pad and stopping it on release. In this manner the onset and offset of beats becomes immediate and very flexible, resulting in less resemblance to a metronome. The rhythm generator was played in a live concert at Centro Candiani in Mestre-Venezia, Italy, on the 20th of June 2002 (see the Sounding Object website for more information about this event, including movies and sound examples: http://www.soundobject.org).

2.2. The Radio Baton as control device

In the same concert, the metronome rhythm generator was used together with a more direct control device for playing, Max Mathew's Radio Baton [5] (see Figure 2 B).

The set-up used in the concert is seen in Figure 3. The set-up used the Radio Baton with the two radio transmitters at each end of a *Bodhran* stick, and the antenna was played with the stick as a "normal" *Bodhran* by the player. The position of each end of the stick versus the antenna controlled the playing position of the sound model, i.e. the distance from the frame of the drum. During the concert the dampening and the fundamental frequency of the model was also controlled through devices such as the KAOSS pad and Doepfer pocket dial knobs. During the concert this was not done by the player herself, although this could be remedied by using foot switches that would leave the hands free for playing.

2.3. The ddrum as control device

In addition to the control devices described above, the ddrum [6] was used to play the sound model of the *Bodhran* (see Figure 2 A). The ddrum is a commercial available electronic drumpad and the MIDI velocity out from the control unit was used to excite the sound model. For the pad used there was also MIDI poly-aftertouch, which was used to control the dampening of the model. The ddrum is a nice interface to play the model because of it's tactile feedback to the player and the lack of cables to the sticks used for playing.

2.4. Control by tracking gestures in real time

A fourth modality of control was based on tracking users body gestures. The process of tracking users movements requires a means of capturing gestures in real time, and extracting the relevant features of the gesture. This requires some form of input device which can take a gesture as input and extract the relevant characteristics of this gesture.

Traditionally, the use of gesture as input involves the use of computer vision techniques, such as recording of gesture with a camera, and the tracking of the movement of a person over time. These systems have a number of limitations, for instance they can be too slow for real-time use, and do not cope well with tracking more than one user. Also, they might not be accurate enough for present requirements.

So, it was decided to use a Polhemus Fastrak [1] device in order to track the users movement (see Figure 2 C). This system tracks the position of up to four small receivers as the move through 3D space, relative to a fixed electromagnetic transmitter. Each sensor returns full six degree-of-freedom measurements of position (X, Y, and Z Cartesian coordinates) and orientation (azimuth, elevation,

[1] Polhemus Fastrak `http://www.polhemus.com/fastrak.htm`

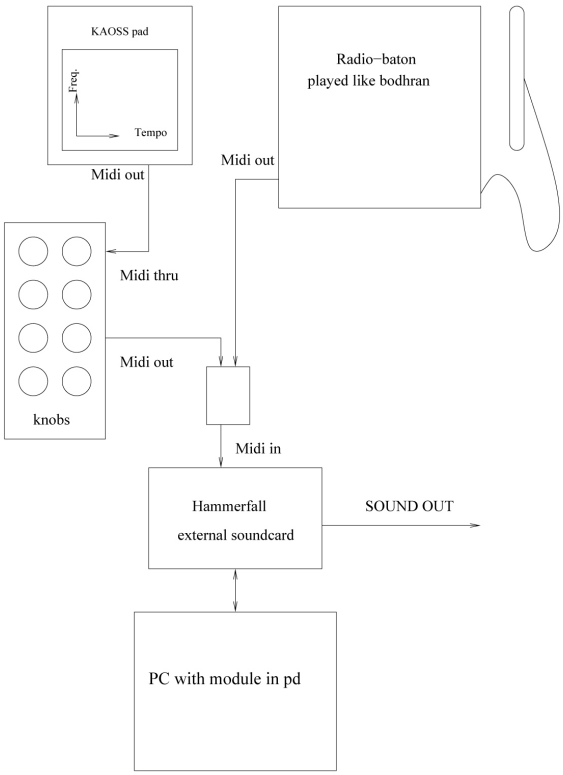


Figure 3: Set-up used for the concert at Centro Candiani in Mestre-Venezia, Italy, on the 20th of June 2002. A player used a *Bodhran* stick, modified with cables to transmit radio signals, and played the antenna like a *Bodhran*. While the position across the antenna controlled the striking position of the model, the fundamental frequency and dampening was manipulated of another person through the Korg KAOSS pad and a set of knobs. The same person also played the "metronome rhythm generator" and then used the x-axis of the KAOSS pad to control the metronome tempo.

and roll). The device connects to a computer through its RS-232 port, and operates at a speed of up to 115.2 K baud, with an accuracy of 0.03" (0.08 cm) RMS for the X, Y, or Z receiver position, and 0.15 RMS for receiver orientation, and a resolution of 0.0002 inches/inch of range (0.0005 cms/cm of range), and 0.025. For a more detailed description of the system see [2].

2.4.1. System design

In order to establish the requirements for the control system, in terms of usability, methods of interaction and sound quality, a day-long workshop was held. Three expert *Bodhran* players, each with their own distinct and different styles and techniques took part. The players also had varying amounts of experience in the use of technology in performance.

The players were asked to perform different traditional rhythms with the sensor attached to the beater, the results were recorded in order to further analyse the gestural patterns involved in *Bodhran* playing. Video data was also gathered for further analysis.

The results of the analysis of this data are being used to determine any common movements, gestures, or techniques that are used by the players, so that the control parameters of the model may be extended in order to allow more interaction for the players.

By examining the data in this way, and continuing to extend the model, we ensure that the overall goal of the project is met, in that a virtual instrument is created, which can be played like a *Bodhran*, but is not limited to the physical characteristics of a *Bodhran*.

The analysis of the workshop has led to a number of additional features in the design of the system. The nuances of how a *Bodhran* is played are very complex, involving both the left and right hands, with the right hand beating the skin, while the left hand can be used to dampen certain modes of the skin.

Damping. This left-handed damping was used by all of the players, and in some cases was used to produce very complex tone changes, even to provide a melody from the rhythmic beating of the skin. This damping effect is a major part of how the *Bodhran* is played, and as such will have to be incorporated fully into the system. Currently the sound model does contain a facility to damp the skin, but only at a single point at any given time. Also, the damping portion of the model would have to be coupled to the Fastrak hardware, and a Fastrak sensor attached to the left hand of a player, to allow them to damp the virtual Bodhran in a similar way to the real instrument.

Tactile feedback. During the course of the workshop, when the players were being asked to use the beater from the virtual *Bodhran* in order to create a simple beat by striking a virtual plane, it was noticed that some players require more tactile feedback from the system than others. While some of players were able to hold the beat, using just the beater and listening to the generated sound, one in particular found this difficult. The addition of a physical plane of reference, which matches the virtual one, was found to alleviate this problem. This opens to some further investigation, to determine whether or not a physical reference is required, and if so, the form of this reference.

Frame of reference. Another point that was raised by this workshop was that of a frame of reference for the instrument. Currently, the system uses a fixed reference point, which does not move with the player. In order for any virtual instrument to be internalised there needs to responsive in a non-arbitrary way and the modification was made for an extension to expressivity and also to allow deliberate musical control on the part of the musician in terms of control sensitivity and control selection. To meet this requirement, the system must allow the players to move naturally, as they would while playing the actual instrument. This would allow players to add to their movement range, without infringing on their standard movements. To enable this, the frame of reference for the system needs to move with the player, so that should they turn or lean as they play (which most players seem to do), the system will continue to function normally, in its new frame of reference.

3. CONCLUSIONS

The control systems that were described in this paper cover three different kinds of interaction with a sound model based on different levels of abstraction; (1) algorithmic control, based on measurements of real players gestures, (2) interaction by mean of percussive instruments, that allow players using traditional gestures

with haptic feedback, (3) use of a gesture tracking system providing continuos controls to the sound model. In particular, the system based on the Fastrack is a novel gesture-based interface to a sound model, which is used as a virtual musical instrument. By basing the design of the system on the real instrument, and by involving players in the analysis and design of the system, it is hoped to create an instrument, which captures the intrinsic characteristics of the real instrument.

However, by not restricting the system to just the characteristics of the real instrument, and by not necessarily tying the instrument to any physical presence, an instrument which allows the players to expand their playing vocabulary can be created.

4. ACKNOWLEDGMENTS

Research presented in this paper was supported by the *Sounding Object* (SOb) EU project IST-2000-25287, `http://www.soundobject.org`.

5. REFERENCES

[1] F. Avanzini and D. Rocchesso, “Modeling Collision Sounds: Non-linear Contact Force”, In *Proc. Conf. on Digital Audio Effects*, 61-66, Limerick, December 2001

[2] R. Bresin, S. Dahl, M. Rath, M. Marshall, and B. Moynihan, “Devices for manipulation and control of sounding objects: the *Vodhran* and the *InvisiBall*”, In D. Rocchesso and F. Fontana,“TheSoundingObject”, Edizioni di Mondo Estremo, Florence, Italy, 235-256, 2003

[3] R. Bresin and S. Dahl, “Experiments on gestures: walking, running, and hitting”, In D. Rocchesso and F. Fontana,“TheSoundingObject”, Edizioni di Mondo Estremo, Florence, Italy, 95-117, 2003

[4] Pico Technology. The ADC-11/10 multi channel data acquisition unit, `http://www.picotech.com/dataacquisition.html`

[5] R. Boulanger and M. Mathews, “The 1997 Mathews radio-baton and improvisation modes”, In *Proc. Int. Computer Music Conference*, 395-398, 1997

[6] ddrum4. Registered trademark of Clavia Digital Musical Instruments AB, `http://www.clavia.se/ddrum.htm`

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

THE RADIO BATON AS CONFIGURABLE MUSICAL INSTRUMENT AND CONTROLLER

Roberto Bresin, Kjetil Falkenberg Hansen, Sofia Dahl

Department of Speech Music and Hearing (TMH) – KTH Royal Institute of Technology, Stockholm
{roberto, hansen, sofia}@speech.kth.se

ABSTRACT

The Max Mathews' Radio Baton (RB) has usually been applied as an orchestra conducting system, as interactive music composition controller using typical percussionist gestures, and as a controller for sound synthesis models. In the framework of the Sounding Object EU founded project, the RB has found new applications scenarios. Three applications were based on this controller. This was achieved by changing the gesture controls. Instead of the default batons, a new radio transmitter based on a thimble that fits the fingertips was developed. This new transmitter allows musicians' interaction based on hand gestures and it can also fit different devices. The RB and the thimble-based transmitter are used for the control of `apd` model implementing DJ scratching techniques. This new configuration of the RB allows DJs a direct control of sampled sounds maintaining hand gestures similar to those used on vinyl. In another application the sound model of a bodhran was controlled with a traditional playing approach. In this case the RB can be controlled with a traditional bodhran double beater with one thimble radio transmitter fitted at each end. This allows detection of the beater position on the RB surface, the surface corresponding to the membrane of the bodhran in the sound model. In a third application the thimble controller was used to move a virtual ball rolling along the elastic surface of a box placed over the receiving antennae of the RB.

1. INTRODUCTION

The Max Mathews' Radio Baton (RB) consists of an array of receiving antennae and two low-frequency radio transmitters working at two different frequencies. The receivers are placed in a way that allows the detection of the three-dimensional position of each transmitter. Position coordinates are converted into MIDI poly-aftertouch values which are accessible via a MIDI-output port placed on the RB control unit. The system can measure the position of each transmitter with a time resolution of 1 ms, and therefore it is possible to track movements in real time. This makes the RB a gesture controller suitable for musical applications. The RB has been produced in about 40 pieces until today. It has usually been applied as an orchestra conducting system, as interactive music composition controller using typical percussionist gestures, and as a controller for sound synthesis models [1, 2, 3].

Because of its versatility, the RB has found new applications scenarios in the framework of the Sounding Object project[1]. In this project three applications were based on this controller. This was achieved by changing the gesture controllers. Instead of the default batons, a new radio transmitter based on a rubber thimble that fits the fingertips was built. This new transmitter allows musicians' interaction based on hand gestures and it can also fit different devices.

[1]The Sounding Object project: http://www.soundobject.org

2. THREE APPLICATIONS OF THE RADIO BATON

In the following sections three applications of the RB are presented; (1) a DJ console, (2) a percussion instrument, and (3) a navigation system for sounding objects.

2.1. DJ scratching control

The Max Mathews' RB has been implemented as gestural controller for Skipproof, a `pd` patch for the simulation of typical DJ scratching techniques [4]. The RB original drumstick-like batons has been substituted by a newly developed radio transmitter fitting the fingertips and based on a rubber thimble.

The thimble controller allows the manipulation of the three-dimensional signal received by the antennae. The surface of the RB receiving antennae (XY-plane) has been divided into ten sectors. One sector, long as the width of the surface (visible on the left side in the right picture in Figure 1), is used for the detection of free hand movements over the surface. The remaining nine sectors are smaller and are used for mapping a pre-recorded scratch technique each. The vertical distance from the surface (Z coordinate) determines the repetition rate of a particular technique, the closer the distance the higher the rate.

This new transmitter allows control of sampled sounds based on hand gestures similar to those used on vinyl (see Figure 1). The user can add her/his own choice of sounds to be scratched. Skipproof can also be applied to the control of sound models, such as friction sound models.

The DJ console was played by a professional DJ in two public concerts at Centro Candiani in Mestre-Venice and at Kulturhuset in Stockholm, in year 2002.

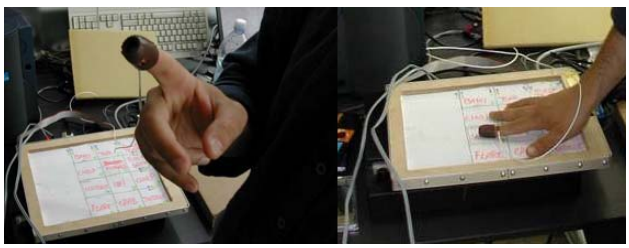


Figure 1: The thimble-based gestural controller for the Max Mathews' Radio Baton. It sends a radio signal from the fingertip to the receiving antenna.

2.2. Percussion instrument control

In another application, the RB was used as a traditional percussion instrument. The sound model of a *Bodhran* (see [5] for a detailed description of the model) was controlled with a traditional playing approach. The RB was controlled with a traditional *Bodhran* double beater with one thimble radio sender fitted at each end. This allowed detection of the beater position on the RB surface, the surface corresponding to the membrane in the sound model.

The surface with the receiving antennae was played with the beater as a "normal" *Bodhran* by the player. The position of each end of the stick versus the antennae controlled the playing position of the sound model, i.e. the distance from the frame of the drum (see Figure 2). The dampening and the fundamental frequency of the model was also controlled through other devices. It would be possible to control these parameters with the transmitters; for example the same low-frequency could be associated to the double beater ends and the other frequency could be used for damping. Also, one could imagine to play the beater at different heights from the receiver and associate the Z-coordinate to the fundamental frequency of the sound model.

The "radio bodhran" was played by a professional *Bodhran* player, Sandra Joyce, in a concert at Centro Candiani in Mestre-Venice, in year 2002 (see Figure 3).

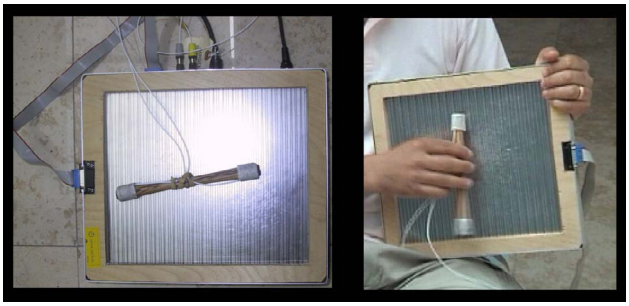


Figure 2: The Radio Baton transformed into a "radio bodhran". The thimble transmitters, shown also in Figure 1, are fitted to the two ends of a traditional double beater.



Figure 3: Sandra Joyce playing the Radio Baton transformed into a "radio bodhran" (see Figure 2).

2.3. Control of a sounding object

In a third application the thimble controller was fitted to a finger tip. The controller was used to move an "invisible" ball rolling along the elastic surface of a box placed over the receiving antennae of the RB, see Figure 4.

The interface is made of stretching material and by pressing the finger on the surface the user determines the target position for the "invisible" ball. This position, in terms of the X, Y, and Z coordinates, is feed to an impact sound model controlled with the mechanics equation of a rolling ball [6]. The ball will roll towards the position of the finger . With this interface, the user has both haptic and acoustic feedback when controlling the target position of the ball.

A visual feedback of the position of the ball, as a projection on the XY plane, i.e. as seen from above, is also possible and it is presented on the computer screen. The position is represented as a colored disk assuming colors in the red-range at high speed (hot ball) and blue-range at low speed (cold ball).

This particular setup was designed for an pilot experiment for testing the relevance of different feedback channels in the manipulation of sounding objects. This new interface allows users an interaction by using three different types of feedback:

- Acoustic; the sound model of the rolling ball
- Haptic; control of the position of the ball by pressing the elastic membrane with a finger
- Visual; graphical projection of position and speed of the ball

The main goal of the experiment was to investigate the degree of reality of the rolling-ball sound when controlled with this interface. Since the interface allows three different feedback, acoustic, haptic and visual, three different experiments were conducted. The three experiments were run in parallel at the same time so that 3 subjects at a time could listen to exactly the same sound stimuli, while receiving feedback on different perception channels. Main result of the pilot experiment was that subjects rated as more realistic those sounds corresponding to balls of smaller size and with lower damping. They also found the acoustic and visual feedback



Figure 4: Haptic interface for the control of the Radio Baton. The interface is placed over the receiving antenna. Finger position in the 3-dimensional space is detected in real-time.

easier to use compared to the haptic one. More details about this experiment can be found in [5].

3. CONCLUSIONS

The easy of use and versatility of the RB has been the driving force in the applications presented in this paper. The RB has revealed to be a useful sensor in very different applications. Both the DJ and the *Bodhran* player found the systems based on the RB of easy control and they both used them in live concert situations without extensive rehearsal training (less than one hour).

Many are the possible applications involving the RB, even in a non musical context, imagination been the only limit. For example it would be possible to use the RB as a navigation system based on gesture control; the radio trasnmitter, such as the thimble one, would be the pointing device and the antenna's surface the navigation area.

In the web repository of the Sounding Object project, it is possible to download videos excerpts of concerts in which the RB was used both as a DJ console and as a percussion instrument.

4. ACKNOWLEDGMENTS

Research presented in this paper was supported by the *Sounding Object* (SOb) EU project IST-2000-25287, `http://www.soundobject.org`.

5. REFERENCES

[1] M. Mathews, "The Radio Drum as a synthesizer controller", In *Proc. Int. Computer Music Conference*, 42-45, 1989

[2] M. Mathews, "The Radio Baton and Conductor program, or: Pith, the most important and least expressive part of music", Computer Music Journal, 15:4, 37-46, 1991

[3] R. Boulanger and M. Mathews, "The 1997 Mathews radio-baton and improvisation modes", In *Proc. Int. Computer Music Conference*, 395-398, 1997

[4] K.F. Hansen and R. Bresin, "Complex gestural audio control: the case of scratching", In D. Rocchesso and F. Fontana,"TheSoundingObject", Edizioni di Mondo Estremo, Florence, Italy, 191-234, 2003

[5] R. Bresin, S. Dahl, M. Rath, M. Marshall, and B. Moynihan, "Devices for manipulation and control of sounding objects: the *Vodhran* and the *InvisiBall*", In D. Rocchesso and F. Fontana,"TheSoundingObject", Edizioni di Mondo Estremo, Florence, Italy, 235-256, 2003

[6] M. Rath, "Sound design around a real-time impact model". In *Models and algorithms for sounding objects*, 32–45. SOb project, 2001. Available at http://www.soundobject.org.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

DJ SCRATCHING PERFORMANCE TECHNIQUES: ANALYSIS AND SYNTHESIS

Kjetil Falkenberg Hansen and Roberto Bresin

Department of Speech Music and Hearing (TMH)
Royal Institute of Technology (KTH)
{hansen,bresin}@speech.kth.se

ABSTRACT

Scratching is a popular way of making music, turning the DJ into a musician. Normally scratching is done using a vinyl record, a turntable and a mixer. Vinyl manipulation is built up by a number of specialized techniques that have been analysed in a previous study. The present study has two main objectives. First is to better understand and model turntable scratching as performed by DJs. Second is to design a gesture controller for physical sound models, i.e. models of friction sounds. We attached sensors to a DJ equipment set-up. Then a DJ was asked to perform typical scratch gestures both isolated and in a musical context, i.e. as in a real performance. He also was asked to play with different emotions: sad, angry, happy and fearful. A model of the techniques used by the DJ was built based on the analysis of the collected data. The implementation of the model has been done in pd. The Radio Baton, with specially adapted gesture controllers, has been used for controlling the model. The system has been played by professional DJs in concerts.

1. INTRODUCTION

Music performed with the turntable and mixer, scratching, has become surprisingly popular, both among the musicians and the audience. Surprisingly of several reasons: Not many would guess that vinyl records would get a second chance after CDs arrived, and as instruments turntables are big and clumsy with a lot of necessary equipment (records, pick-ups, slipmats, audio mixers and steady tables to put it on). Not least, the sound of scratching is incomparable to that of any other instrument, as it lacks both explicit melodic or rhythmical qualities, but rather combines the two into a fast, fragmentary sounding instrument. This does not have to be a disadvantage, but to some it may sound too unfamiliar.

Most of what is played in scratching can be traced down to basic movements, as DJs control two parameters: 1) moving the record forward and backward and 2) cutting the sound on and off with the crossfader.

It seems to be a simple task to model scratching and replace the turntable, with all its drawbacks, but attempts from the producers of DJ equipment has not been all that successful. We need to know more about how scratching works both musically and as an instrument. Also it is interesting to look further than manipulating only the recorded sounds, and extend the scratch music vocabulary to controlling physics-based sound models.

1.1. Background

In earlier studies of scratching, it was found that DJs build their music around a set of learned gestural patterns, and these patterns are recognized as different "techniques". From the growing number of techniques, there are about 50 that can be called fundamental or even necessary to know as a scratch DJ [1]. Still, the techniques are seldom performed in a simple one-after-the-other fashion, but rather intertwined and in complex structures [2]. This is probably varying in degree from one DJ to another, but not yet investigated.

1.2. Method and set-up

A professional DJ from Sweden, *DJ 1210 Jazz*[1], volunteered for the experiment and was the only subject. DJ 1210 Jazz has no formal musical training, but has for almost 15 years been considered to be among the best turntablists in Sweden and Europe.

In the recording sessions eight performances were executed, all of which without a backing drum track. Since DJ 1210 Jazz is an experienced performer, the lack of backing track was not considered a restrictive or unnatural condition even though scratching often is performed to a looped beat.

The DJ was asked to play in a normal way, as he would do in an ordinary improvisation. The performances from that session were by all means representative examples of improvised solo scratching with a clearly identifiable rhythmic structure. All sounds were originated from the popular "ahhh" sound from "Change the beat" [3]. This sampled part is found on the DJ 1210 Jazz [4] record we used. The recordings were done at KTH during 2001, and the equipment used for the experiment is summarized in Table 1.

Table 1: Equipment used for the experiment.

Equipment	Description	
Turntable	Technics SL-1210 Mk2 with felt slipmat	
Cartridge	Shure M44-7	
DJ-mixer	Vestax PMC-06 Pro	
Faders	Vestax PMC-05 Pro	
Record	DJ 1210 Jazz, "Book of Five Scratches: Book 2." [4]	
Potentiometer	Bourns 3856A-282-103A 10K	
DAT-recorder	Teac RD-200T Multichannel	Ch.1, 20 kHz, Pot.meter
		Ch.2, 10 kHz, Crossfader
		Ch.3, 10 kHz, Audio
Wave analysis software	Soundswell Signal Workstation [5]	
	Wavesurfer [6]	

The analysis was done on the basis of three signals; the crossfader, the record movement and a waveform of the recorded

[1] His real name is Alexander Danielsson.

sound. The right audio channel was output to a multichannel DAT recorder, while the left channel was output to a headphone mixer so the DJ could hear himself.

A potentiometer was used to track the vinyl movement, and it was mounted to the vinyl with the help of a stand. No effect could be noticed in the performance and friction on the vinyl when it was attached, and the DJ felt comfortable with the set-up. The output was recorded by the DAT.

Two cables connected from the mixer's circuit board to the DAT tracked the crossfader slider movement. The crossfader run is 45 mm, but the crucial part from silence to full volume (the so-called cut-in point), spans only a distance of 2-3 millimeters, a few millimeters away from the (right) end of the slider run. Positioned to the right, the crossfader completely muted all sound, and it let through all sound when moved a few millimeters (to the left).

Only the right channel of the stereo sound output signal was recorded to the DAT, but that was sufficient for evaluating the record movement output against the sound output. The original sound from the record had no significant stereo effects, and both right and left channel appear similar.

2. ANALYSIS

Considering the analysis from earlier experiments [2], a scratch simulator must include a volume on/off function, as almost none of the scratches are performed with the volume constantly on. There is no need to be able to control bigger scratch areas than 360° while 180° should be easily controlled.

From the analysis and data a model of scratching was built. The readings of the potentiometer and the crossfader recorded were used to control an audio file. By first using the output from the potentiometer to change the sample-rate of the audio file that was played back, and then using the output from the crossfader circuit board to change the playback volume level, we successfully resynthesized the few techniques we tested on. Three techniques involved record movement only; *baby*, *tear* and *scribble*, while six techniques, *chop*, *roll-tear*, *crab*, *flare*, *uzi*, *chirps* and *twiddle*, also involved crossfader movement.

2.1. Models of scratch techniques

Analysis of the data relative to one *baby scratch* cycle shows that the DJ moves the record forward and backward to its starting position in about 260 ms. Knowing the sounds' position on the record, it was possible to calculate the distance travelled by the record and the velocity of the record itself. The velocity of this movement has the typical shape of target approaching tasks [7]. In the DJ pulling action, the velocity reaches its maximum when the record has travelled a little over half of the final distance, then velocity decreases to 0 value when the DJ starts to push the record forward. During the pushing action, the record increases in velocity in a shorter time than in the pulling phase. It thus reaches maximum velocity before having travelled through half the distance covered in the pushing action.

The same observations and measurements can be done for the other scratching techniques taken in consideration. *Chop*, *crab*, *flare*, *chirps* and *twiddle* scratch models use *baby* scratch as the basic record movement as do most scratches where the crossfader is the main feature. Still the record movement varies in each case from the simpler *baby*.

To simulate the record playing, the sample to scratch should be looped. When the sample is running in a loop, a mouse may be used for dragging the "record" forward and backward. It will not feel much like scratching for real, however, as you have to press the mouse button on the right place on the screen and move the mouse simultaneously. Even if the ability to do this smoothly and efficiently can be trained, there are better devices and sensors for this purpose.

3. SYNTHESIS

Results from measurements and analyses were used for the design of a model of scratching. `Pd` [8] was chosen as software environment for this model. Recorded sounds can be controlled by the scratch software in the same manner as by turntables, but the model can also control physics-based sound models of i.e. friction sounds [9]. The `pd` patch, *Skipproof*, is open and customizable to be controlled by various types of input devices. We have tested the patch with both recorded sounds and physics-based models, and we have controlled the model by use of computer mice, keyboards, MIDI devices, the Radio Baton[2], and various sensors connected to a Pico AD converter[3].

3.1. Skipproof – a `pd` patch

Skipproof[4] has three main functions. It is an interface for manipulating the playback tempo of a sound-file by using a computer input device, and it is an interface for triggering single scratching techniques. With these two functionalities, Skipproof acts as both a virtual turntable and a scratch sampler/synthesizer. In addition, the output volume can be controlled manually or by the modelled scratch techniques. With this third functionality, Skipproof also simulates the scratch audio mixer.

Skipproof uses GrIPD [10], a GUI front-end program for `pd`. `Pd` processes all sounds, and GrIPD offers control over the different options programmed in `pd`.

The sounds manipulated in Skipproof are 44.1 kHz or 88.2 kHz, 16 bit mono wave-files, but other formats can easily be supported. Sounds, or 'samples' in analogy to DJ-terms, are meant to be 1.6 s long in order to imitate a real skip-proof record, yet there is no restriction to the length.

Apart from direct manual "scratch control" of a sound-file, it can be accessed via recorded scratch movements. These recordings originate from the measurements analysed in the earlier reported experiments.

Skipproof can be controlled by different sensors and hardware, and is easy to adjust to new input devices. Sensors, MIDI input and computer inputs (keyboard and mouse) are used both to trigger the scratch models and to manipulate the virtual turntable and audio mixer, and their parameters. The intention of Skipproof is that it should be controlled by hand gestures, and that the graphical

[2]The Radio Baton is a controller for musical performances that tracks the motion of a sender in a 3-dimensional space. It is developed by Max Mathews.

[3]Pico Technology. The ADC-11/10 multi channel data acquisition unit, http://www.picotech.com/data-acquisition.html

[4]The name Skipproof is taken from a feature found on DJ-tools records called a skip-proof section, where a sound is exactly one rotation long and repeated for a couple of minutes. If the needle should happen to skip during a performance, chances are quite good that it will land on the same spot on the sound, but in a different groove. The audience will probably not register this jump.

interface is just for visual feedback. Computer mouse is not the controller of choice, but is a tool most users have, so we did not want to exclude it.

3.1.1. Implementation

In the following, the `pd` patch will be commented, explaining briefly how Skipproof is designed.

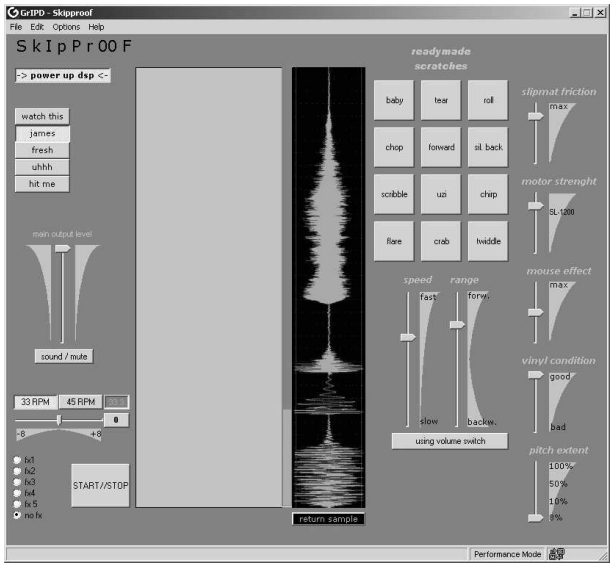


Figure 1: Graphical interface for Pure Data, GrIPD with turntable and audio mixer controls.

Figure 1 shows the performance window in Skipproof. One main focus designing the graphical interface was to some extent copy the functionalities of a turntable and a mixer. There are also a number of other buttons and sliders not found on the standard hardware, enabling the DJ to change parameters of, amongst others, motor strength.

The large light grey rectangle in Figure 1 is the part that registers mouse action (when left mouse button is held down). The meter next to the grey area displays the sound progression in relation to the wave plot. Around this 'vinyl record part' all the standard turntable buttons are collected; start/stop, two buttons for toggling 33 and 45 RPM, and a pitch adjustment slider. On a turntable there is also a power switch that lets the platter gradually stop rotating by its own low friction. When stopping the turntable with the stop-button it is the motor that forcefully breaks the rotation speed. The power-switch is sometimes used to produce a slow stop, but is omitted as a controller here.

Only two controllers are chosen from the audio mixer's many possibilities. The up-fader is replaced by a logarithmic master-volume slider. Normally the crossfader is far more utilized than the up-fader, but a slider is not an advantageous way to control volume when the mouse is occupied with record speed. Under the slider is a push-button which shuts out the sound (or turns on the sound) when activated. This button mixes the functions of the line/phono switch and the crossfader and can be assigned to keys or other inputs.

Under the heading "readymade scratches" (up right in Figure 1) there are several buttons for triggering the recorded techniques. Below, two sliders define the speed and the depth these techniques will have. The speed range is a bit broader than what is normally performed. The depth slider goes from long backward movements to long forward movements, also making exaggerated performances possible.

Both horizontal (x) and vertical (y) mouse activity is measured. Vertical mouse movement values are used for adjusting the playback sample rate. Horizontal movement can be used for several purposes, i.e. different audio effects, or simulating the hand placement on the vinyl as the effective record movement vary depending on where the record is held.

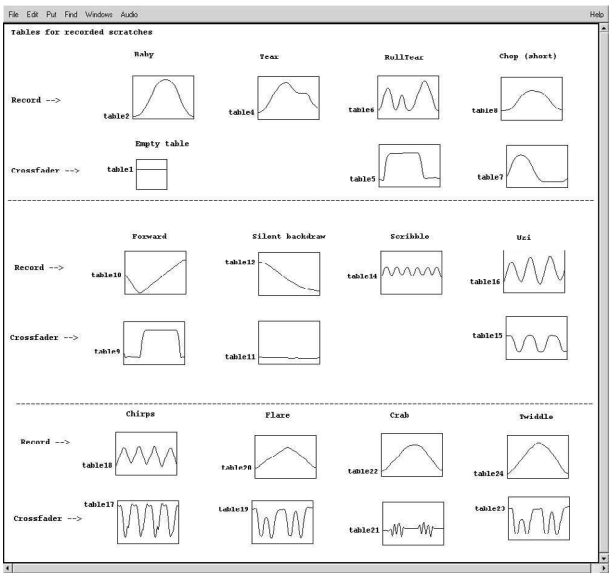


Figure 2: `Pd` tables: The recordings of 12 scratch techniques in tables. The upper row in each section shows record movement, and the lower row shows crossfader movement.

All the recorded scratches used for synthesizing single techniques are collected in "`pd` tables" for simplicity, see Figure 2. The empty "table1" is for techniques where the crossfader is not utilized, *baby*, *tear* and *scribble*. In this way all techniques can follow the same procedure in the model.

Signals from crossfader and record movements are low-pass filtered at 30-50 Hz before implemented in Skipproof. Each of the techniques is picked out from a series of repeatedly performed single techniques, and so represent an idealized model. Apart from those techniques that consists of many small movements on the record, as in *chirps* and *scribble*, only one stroke forward and backward is included for all scratches. The scratches vary in length.

To simulate the wearing of the vinyl, as explained in an earlier experiment [2], a simple noise model was implemented. Following several notions, it generates low-level white noise, narrow-band noise and vinyl defects as cracks and hisses. All the noise signals are recorded to a 1.6 s long table, so the vinyl defects always occur at the same spot on the record when it is looped and responds accordingly to scratching.

Sounds in Skipproof are sampled sounds. The user can add her own choice of sounds to be scratched. Skipproof can also be applied to the control of sound models.

3.1.2. Controllers

In this section some alternatives to standard MIDI and computer input controllers are presented. An analogue-digital converter from Pico sends the signals from the control devices to the computer. The voltage output is then read in `pd`, controlling the described parameters.

The crossfader on modern scratch mixers is becoming easier and easier to move; some models have friction-free faders. Still it takes a lot of training to accurately jump the fader over the critical cut-in point. To make it easier to accomplish fast and precise clicks, a light sensor replaces the crossfader.

The Radio Baton was used as gestural controller for Skipproof. The drumstick-like batons were substituted by a newly developed radio sender that fits the fingertips using a thimble. This new radio sender allows users' interaction based on hand gestures (see Figure 3), as the three-dimensional position of a finger in the volume over the antennae is detected.

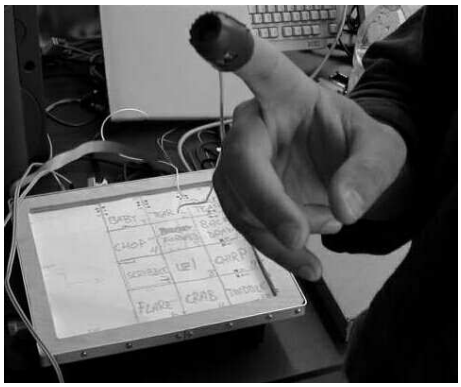


Figure 3: Thimble gestural controller for the Radio Baton and the antennae with sections drawn in.

The Radio Baton's receiving antennae is divided into one area for scratching and 12 smaller sectors, each sector hosting a prerecorded scratch technique (see Figure 3). Each modelled technique will be played according to the movements towards the antennae within the constraints of a sector. A fast dive with the hand down to the surface will produce a fast scratch performed with long record movements (high pitch), while a slow diving gesture in the air will achieve the opposite effect, a scratch performed slowly over a small area.

While scratching on a turntable only allows movement in the plane, the Radio Baton offers a third dimension, height, for controlling the instrument. So far we have tried to control the physical parameters, i.e. friction and movement amplitude, and also added functionalities not possible with a turntable, i.e. digital audio effects. Even the position on the "vinyl" in the horizontal axis can be mapped to new parameters.

4. CONCLUSIONS

Skipproof have been tested a few times, both by professionals and novices. Two times it have been played in concert situations before an audience[5]. Professional DJs tend to compare it with their usual equipment immediately and find problems in accuracy and controllability. It can be explained in different ways, but the Radio Baton has some clear disadvantages, among other factors the radio sender must be wired to the antennae, and the detectable area is more spherical than cubical. This makes the Radio Baton and Skipproof somewhat hard to play, yet both professional and novice users find it not only amusing but useful in a musical situation.

Especially the DJs like the extra dimensions the Radio Baton provides. All seem to have different ideas how to use them. It will require some training to be able to play Skipproof, which is the case for most musical instruments.

To simulate every aspect of the turntable, the vinyl, the needle and aspects like wearing, will probably turn out to be the only suitable option for making an acceptable replacement for today's instrument set-up. Scratching with physics-based sound models has only been tried once in concert, but the DJ seemed intrigued by the possibilities. An approach built on physics-based modelling technique seems therefore appropriate and worth to experiment further with in the future [11].

Arguably, the most characteristic quality in scratching is the big range of recognizable and universally agreed-upon playing techniques. Future research can reveal interesting issues regarding these. Also, technology aiming to replace the turntable should take into consideration the role and practises of scratch techniques. The techniques and characteristics of the hand movements associated with different types of scratches will be examined in future investigations.

[5]Online video clip from DJ 1210 Jazz' performance with Skipproof: http://amarone.sci.univr.it/pub/bscw.cgi/d16000/RadioBaton-scratch

5. REFERENCES

[1] K. F. Hansen, "The Basics of Scratching", J. of New Music Research, Vol. 31, No. 4: 357–367, 2002.

[2] K. F. Hansen and R. Bresin, "Complex gestural audio control: the case of scratching.", *In*: D. Rocchesso and F. Fontana, "The Sounding Object", Edizioni di Mondo Estremo, Florence, Italy, 191–234, 2003.

[3] Fab Five Freddie, "Change the beat", Celluloid Records CEL 156, 1982.

[4] DJ 1210 Jazz, "Book of Five Scratches: Book 2.", Snickars Rec., SR1206, 2001.

[5] Hitech Development, Soundswell signal workstation, http://www.hitech.se/development/.

[6] K. Sjölander and J. Beskow, Wavesurfer, http://www.speech.kth.se/wavesurfer/.

[7] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model.", The Journal of Neuroscience, 5(7):1688–1703, 1985.

[8] M. Puckette, Pd, Pure data – a real-time computer music software package, http://www.pure-data.org.

[9] F. Avanzini, "Low-level models: resonators, interactions, surface textures." *In*: D. Rocchesso and F. Fontana, "The Sounding Object", Edizioni di Mondo Estremo, Florence, Italy, 119–136, 2003.

[10] J. A. Sarlo, GrIPD, Graphical Interface for Pure Data, http://crca.ucsd.edu/˜jsarlo/gripd/.

[11] D. Rocchesso, R. Bresin and M. Fernström, "Sounding objects", IEEE MultiMedia Magazine, April/June, (10)2: 42–52, 2003.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

GESTURE-TACTILE CONTROLLED PHYSICAL MODELLING MUSIC SYNTHESIS

David M Howard and Stuart Rimell

Media Engineering Research Group
Department of Electronics, University of York, Heslington, York, YO10 5DD, UK
dh@ohm.york.ac.uk

ABSTRACT

Despite enjoying the resulting freedom from physical constraints that can result from the incorporation of computer technology in electronic musical instruments as musicians are enabled to push back the creative boundaries within which they work, musicians are still searching for virtual instruments that come closer to their physical counterparts in terms of their playing experience. This paper describes a physical modelling music synthesis system known that enables 'virtual instruments' to be controlled in real-time via a force-feedback joystick and a force-feedback mouse. These serve to provide the user with gestural controllers incorporating tactile feedback. Virtual instruments are set up via a graphical user interface in a manner that is highly intuitive, since users design these virtual instruments by interacting directly with their physical shape and structure in terms of the physical properties of basic objects such as strings, membranes and solids which can be interconnected to form complex structures. Acoustic excitation can be applied at any point mass via virtual bowing, plucking, striking, or from an external sound source. Virtual microphones can be placed at any point masses to deliver the acoustic output. The organic nature of the resulting acoustic output is illustrated.

1. INTRODUCTION

The acoustic output from electronic musical instruments is often described as 'cold' or 'lifeless' by players and audience alike. Indeed, many report that such sounds become less interesting with increasing exposure. The acoustic output from acoustic musical instruments may, on the other hand, be described as 'warm', 'intimate' or 'organic'.

This difference can to some degree be addressed by making use of the following two means in electronic instruments:

1. by using physical modelling for sound synthesis to create organic sounds that more closely resemble those of physical entities
2. by creating user interfaces that enable musicians to interact with the electronic instrument in ways that are both more intuitive and intimate.

Both of these means have been incorporated into the new electronic musical instrument described in this paper, known as 'Cymatic' [1-3]. Cymatic has been designed to provide its players with an immersive, organic and tactile musical experience that is more commonly associated with acoustic instruments but rarely found with computer-based instruments. The audio output from Cymatic is derived from a physical modelling synthesis engine which has its origins in TAO [4]. It shares some common approaches with other physical modeling sound synthesis environments such as Mosaic [5] and CORDIS-ANIMA [6].

The system is controlled via tactile and gestural interfaces with which users both design and play their virtual Cymatic instruments. This approach ensures that users an interact with the computer in a manner that is both intuitive and intimate. Although the player is physically detached from the virtual structures that are responsible for sound production, s/he interacts with it indirectly via interface peripherals such as a mouse, MIDI keyboard or musical instrument controller.

Cymatic provides haptic feedback in addition to the acoustic output to provide users with an immersive and tactile playing experience. It is the haptic senses rather than the visual senses that provide the second most important means (after the audio output) by which users observe and interact with the behaviour of musical instruments [7]. Complex and realistic musical expression can only result when both tactile (vibrational and textural) and proprioceptive cues are available in combination with aural feedback [8, 9].

In order to give Cymatic the potential for having a wide appeal, it has been implemented on a Windows PC and it makes use of widely-available PC gaming devices for real-time gestural control and haptic feedback via a force feedback joystick and a tactile feedback mouse

2. THE CYMATIC SYSTEM

Cymatic makes use of a mass-spring paradigm of physical modelling to synthesise resonating structures in real time. Users design virtual instruments by placing graphically individual elements, such as string, sheets and solids using the interface illustrated in figure 1. These elements can be interconnected by linking individual masses, and the

edges can be locked as desired. The tension and mass parameters of each element can be user defined either to be fixed or to be dynamically changed during synthesis using either the mouse or joystick. Individual masses can be deleted to enable arbitrary shapes to be realised (see figure 1). Excitation can be placed on any mass within the virtual instrument, selected from the following:

- Pluck
- Bow
- Random
- Sine wave
- Square wave
- Triangular wave
- Noise
- Live audio.

Audio output is derived from one or more virtual microphone(s) which can be placed on any individual mass(es), and a selection of standard sampling rates from 48kHz to 8kHz is available.

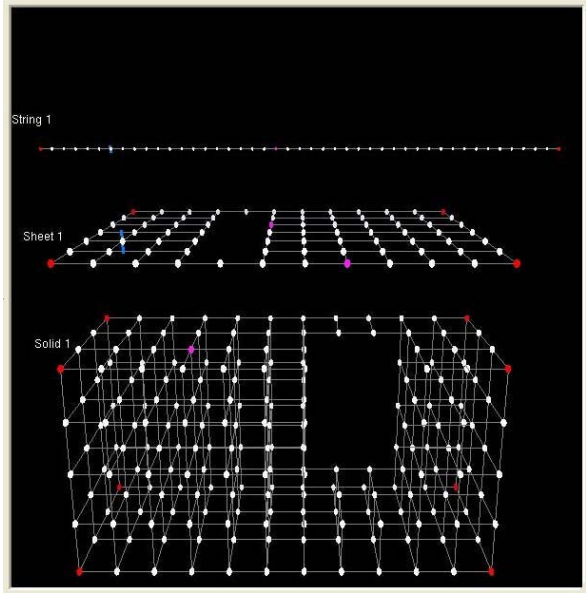


Figure 1: *Screenshot of an example Cymatic virtual instrument consisting of a sheet and solid with individual masses removed as well as a string.*

2.1. Cymatic's synthesis engine

Synthesis in Cymatic is carried out by solving for the mechanical interaction between the masses that make up the virtual instrument. Feynman's numerical discrete time approximation to the solution of the differential equation for a harmonic oscillator is employed. This method calculates the mass velocity half a time step ahead of its position. This results in a more stable model than an implementation of the Euler approximation

Acceleration of cell at time *t* is described by:

$$a = F/m \qquad (1)$$

where: a = acceleration, F = total force acting on the cell, m = mass of the cell.

The total force on the cell is the sum of three forces:-

$$F_{total} = F_{spring} + F_{damping} + F_{external} \qquad (2)$$

Where: F_{spring} = the force from springs connected to neighbouring cells, $F_{damping}$ = the frictional damping force on the cell due to the viscosity of the medium, $F_{external}$ = the force on the cell from any external excitations (pluck, bow etc.).

F_{spring} is calculated by summing the force on the cell from the springs connecting it to its neighbours. This force can be calculated for each neighbour by Hooke's law:

$$F_{spring} = k(p_n - p_0) \qquad (3)$$

Where: p_n = the position of the nth neighbour, p_0 = the position of the current cell.

$F_{damping}$ is the frictional force on the cell caused by the viscosity of the medium in which the cell is contained. It is proportional to the velocity of the moving cell, where the constant of proportionality is the damping parameter of the cell. The damping force is therefore given by:

$$F_{damping} = -\rho v(t) \qquad (4)$$

Where = the viscosity as given by the damping parameter of the cell, $v(t)$ = the velocity of the cell at time t.

Combining these forces with equation 3, we can find the acceleration, and hence the velocity and position, of a particular cell at any instant:

$$a(t) = (1/m)(k\sum (p_n - p_0) - \rho v(t) + F_{external}) \qquad (5)$$

3. HAPTIC CONTROL OF CYMATIC

Haptic control is realised in Cymatic via PC gaming interface devices that incorporate force feedback and interface directly with Windows. Currently, a Microsoft Sidewinder Force Feedback Pro Joystick and a Logitech iFeel mouse [10] are used. The Sidewinder joystick offers four degrees of freedom (x, y, z-twist movement and a rotary "throttle" controller) as well as eight buttons. Each of these can be user customised to control Cymatic parameters such as the mass or tension of any element, or excitation parameters such as bowing pressure, force, or velocity. The joystick offers the potential to output six forces simultaneously. The Logitech iFeel mouse has two degrees of freedom and three buttons. It is an optical mouse and it contains a vibrotactile device to produce tactile feedback over a range of frequencies and amplitudes.

4. EXAMPLE CYMATIC OUTPUTS

In order to illustrate the organic nature of the acoustic output from Cymatic visually, spectrograms are used.

Figure 2 shows a spectrogram for an acoustic anechoically recorded double bass (upper plot), and that for a Cymatic bowed string whose fundamental frequency has been set such that it matches that of the double bass (lower plot). The Cymatic output has little constancy as it exhibits change throughout the sound, and it can be seen that this is a feature of the acoustic double bass. In addition, it is clear that there is similarity between these spectrograms in the nature of the note onset and note offset in each case. Every individual note played on a Cymatic virtual instrument will be acoustically subtly different since each will have its origins in its own unique physical gesture. This illustrates to some degree the organic nature of the Cymatic output; one of the key cited advantages of using physical modelling.

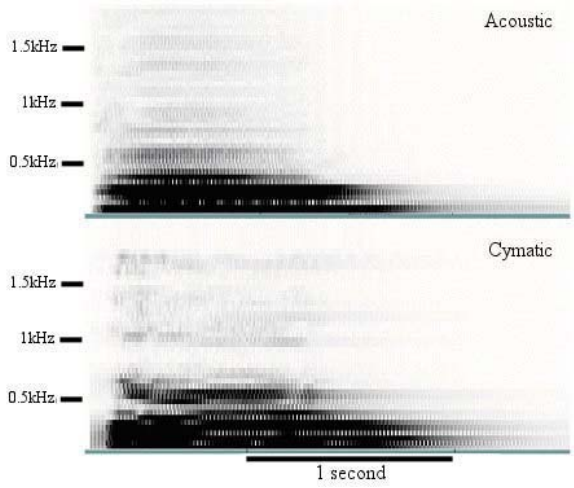


Figure 2: ***Spectrograms of the outputs from an acoustic double bass (upper) and a Cymatic bowed virtual instrument tuned to the same pitch***.

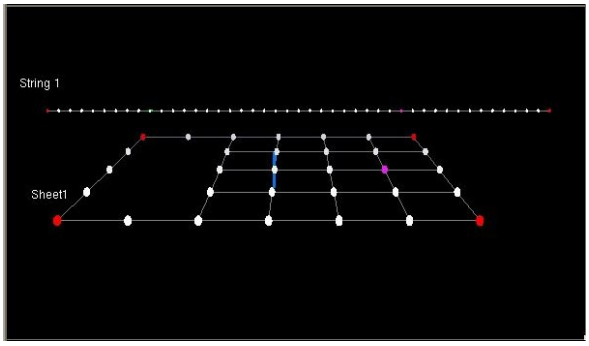


Figure 3: *Cymatic virtual instrument consisting of a string and modified sheet. They are joined together between cell 30 (from the left) on the string to point (6, 3) on the sheet. A random excitation is applied at point 10 of the string and the virtual microphone is located at point (4, 3) of the sheet*

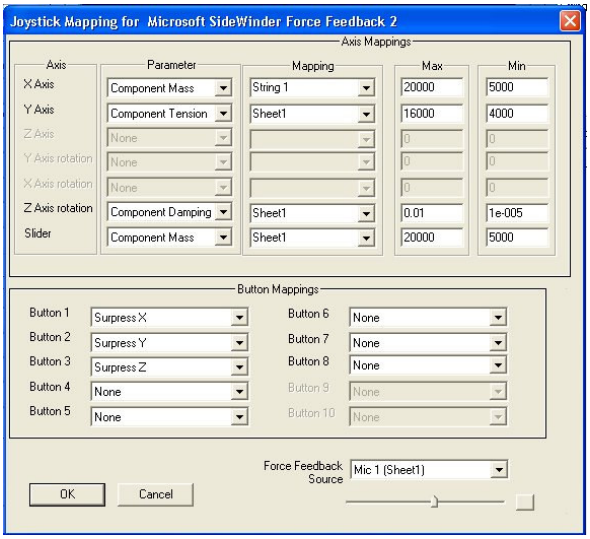


Figure 4: *Cymatic joystick controller mapping dialogue for the virtual instrument shown in figure 4.*

Figure 3 shows a virtual Cymatic instrument consisting of a string and a modified sheet which are joined together between cell 30 (from the left) on the string to point (6, 3) on the sheet. A random excitation is applied at point 10 of the string and the virtual microphone is located at point (4, 3) of the sheet. Figure 4 shows the force feedback joystick settings used to control the virtual instrument, and it can be seen that the component mass of the string, the component tension, mass and damping of the sheet are controlled by the X, Y, Z and slider (throttle) functions of the joystick. Three of the buttons have been set to suppress X, Y and Z; a feature which enables a new setting to be jumped to as desired, for example by pressing button 1, moving the joystick in the X axis and then releasing button 1. Force feedback is applied based on the output amplitude level from the microphone (mic 1).

Figure 5 shows a spectrogram of the output from mic 1 of the instrument shown in figure 3. The tonality is entirely due to the resonant properties of the string and sheet since the input excitation is random. Variations in the tonality are rendered through gestural control of the joystick, and the step change is a result of using one of the 'suppress' buttons.

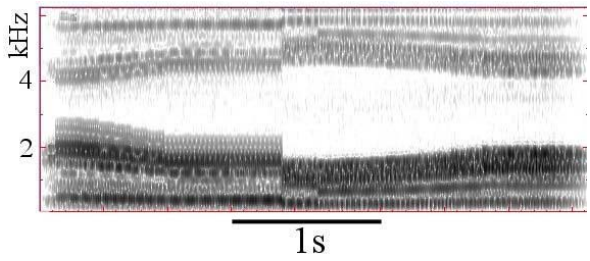


Figure 5: Spectrogram of output from the Cymatic virtual instrument consisting of a string and modified sheet shown in figure 3.

5. DISCUSSION AND CONCLUSIONS

Cymatic was not designed to enable the accurate modelling of physical acoustic musical instruments as virtual entities. Rather, it arose from a desire to develop a new musical instrument that had the potential to enable musicians to work with original and inspiring new timbral palates in a manner that is intuitive. The intuitive side is ensured through the use of a physical modelling paradigm, which enables users to make sense of what they are doing through reference to the likely behaviour of physical resonant structure such as strings, sheets and solids.

Cymatic has the further potential in the future (as processing speed increases further) to move well away from the real physical world, while maintaining the link with this intuition, since the spatial dimensionality of the virtual instruments can in principle be extended well beyond the three of the physical world. In addition, other dynamic variations can be employed that are not possible in the real world such as changing the mass and tension of the cells of one element (string, sheet solid etc.), or changing the shape of individual elements by cutting with virtual scissors.

The intuitive nature of Cymatic is further supported by the increased sense of immersion created by the use of gesture control and haptic as well as audio feedback to the player. This is particularly useful when developing performance skills because it reinforces the visual and aural feedback cues and helps the player internalise models of how the instrument reacts to gesture. The haptic feedback also has the potential to prove invaluable in group performance, where traditionally computer instruments have placed an over-reliance on visual feedback, thereby detracting from the player's visual attention which should be directed elsewhere in a group situation, for example towards a conductor.

Cymatic made its public live debut during the Christmas Carol Concert given by the Beningbrough Singers in York in December 2002, conducted by David Howard. A piece "the child is born" was specially composed by Stuart Rimell for a capella choir and Cymatic. The composer performed the Cymatic part, which made use of three cymbal like structures controlled by the mouse and joystick. There were over 200 in the audience, and the piece was widely acclaimed; indeed, many commented that they had been highly skeptical about the use of an electronic instrument during a carol concert. An extract from the piece is linked to from the following web page [11].

6. ACKNOWLEDGEMENTS

The authors acknowledge the support of the Engineering and Physical Sciences Research Council UK under grant number GR/M94137.

7. REFERENCES

[1] Rimell, S., Howard, D.M., Hunt, A.D., Kirk, P.R., and Tyrrell, A.M. (2002). The development of computer-based, physically modelled musical instrument with tactile feedback, for the performance and composition of electroacoustic music, *Proceedings of the 10th Anniversary Conference of the European Society for the Cognitive Sciences of Music*, University of Liege, CD-ROM.

[2] Howard, D.M., Rimell, S., Hunt, A.D., Kirk, P.R., and Tyrrell, A.M. (2002). Tactile feedback in the control of a physical modelling music synthesiser, In: *Proceedings of the 7th International Conference on Music Perception and Cognition*, Stevens, C., Burnham, D., McPherson, G., Schubert, E., and Renwick, J. (Eds.), Adelaide: Casual Publications, 224-227.

[3] Rimell, S., Howard, D.M., Tyrrell, A.M., Kirk, P.R., and Hunt, A.D. (2002). Cymatic: restoring the physical manifestation of digital sound using haptic interfaces to control a new computer based musical instrument, *Proceedings of the International Computer Music Conference*, Sweden: Göteborg, 114-117.

[4] Pearson, M.D., and Howard, D.M. (1996). Recent developments with TAO physical modelling system, *Proceedings of the International Computer Music Conference*, **ICMC-96**, 97-99.

[5] Morrison, J.D. and Adrien, J.M. (1993). MOSAIC: A Framework for Modal Synthesis, *Computer Music Journal,* **17**, (1), 45-56.

[6] Cadoz, C., Luciani, A., and Florens, J.L. (1993). CORDIS-ANIMA: A Modelling system for sound and image synthesis, the general formalism, *Computer Music Journal* **17**, (1), 19-29.

[7] Cook, P.R. (1999). Music, Cognition and Computerised Sound: An Introduction to Psychoacoustics, London: MIT Press, pp229.

[8] MacLean, K.E. (2000). Designing With Haptic-Feedback, www.cs.ubc.ca/~maclean/publics/icra00-DesignWithHaptic-reprint.PDF

[9] Howard, D.M., Rimell, S., Hunt, A.D., Kirk, P.R., and Tyrrell, A.M. (2002). Tactile feedback in the control of a physical modelling music synthesiser, In: Proceedings of the 7th International Conference on Music Perception and Cognition, Stevens, C., Burnham, D., McPherson, G., Schubert, E., and Renwick, J. (Eds.), Adelaide: Casual Publications, 224-227.

[10] http://www.immersion.com

[11] http://www-users.york.ac.uk/~dmh8

MISCELLANEA

including

OPTICAL AND ACOUSTICAL MEASUREMENTS

and

TEACHING MUSIC ACOUSTICS TO MUSICIANS

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

VISUALIZATION OF SOUND FIELDS WITH APPLICATION TO MUSICAL INSTRUMENTS

Nils-Erik Molin

Division of Experimental Mechanics, Luleå University of Technology,

SE-97187, Luleå, SWEDEN

nem@mt.luth.se

ABSTRACT

Modern optical measuring methods have the advantage of being contact less, non-disturbing and whole field methods. They often are quantitative and they present ”pictures”. In the following some modern optical, holographic, speckle and interferometric methods that visualize and measure vibrations and sound fields of musical instruments are presented and discussed.

1. INTRODUCTION

Pictures that demonstrate physical phenomena are important in science. *Atlas de phénomènes d`optique* (Atlas of optical phenomena) [1] is a collection of photographs of interference, diffraction and polarization phenomena. Phase contrast methods that visualize transparent objects are pictured. In *An Album of Fluid Motion* [2] different kinds of fluid flows are visualized. Classical optical methods like shadowgraph, schlieren and interferometry are used to illustrate phenomena in turbulence, convection, subsonic and supersonic flow, shock waves etc. Some of them like shock waves certainly also generate sound fields.

During the last decades, new measurement techniques have been presented for the study of sound fields. Holographic Interferometry [3] (HI), Speckle Interferometry [4] (SI), Speckle photography [3-5] (SP) and Laser Doppler Velocimetry (LDV) [6] allow diffusely reflecting objects and do not need “plane or well defined waveforms” in the setup. They are “digital and computer aided” and have reached high enough sensitivity to record sound fields.

Lökberg [7] measured harmonic sound fields in air. Phase modulated TV holography (also called Digital or Electronic Speckle Pattern Interferometry, DSPI or ESPI) was used. Wåhlin et al [8] used pulsed holographic interferomety to measure transient sound fields outside impacted plates and beams. This pulsed technique was earlier applied to wave propagation in solid objects, see figure 1. It was for instance used to measure Operational Deflection Shapes (ODS) of a bowed violin at about 2000 Hz [9]. Tomographic reconstruction of sound fields [10-11] was reported. Runnemalm [12] recorded aerial standing waves inside a rectangular box and inside the sound box of a violin and a guitar. Zipser et al [6] used scanning Laser Doppler Vibrometry (LDV) to measure sound and flow fields.

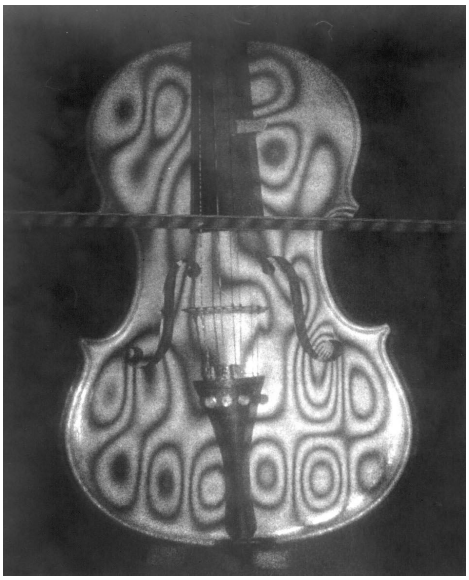


Figure 1: *Operational deflection shape at steadily hand bowing of the E-string stopped at about 1/3 (~2000 Hz). Double pulsed interferogram with 0,2 ms between the pulses, the first one randomly fired, [9].*

2. RECIPROCITY COMBINED WITH TV HOLOGRAPHY MEASURES RADIATIVITY

An indirect point wise method to measure the sound distribution from an instrument is reciprocity combined with TV holography, [13]. A movable loudspeaker at constant (unit) driving conditions excites a violin from different positions outside the violin. The corresponding velocity component at the bridge foot is measured by TV holography for the specific excited ODS. Then, it can be argued using reciprocity, that if a constant (unit) driving force instead is acting at the bridge foot it will create a sound pressure of the same magnitude as the measured velocity component at the same place as the loudspeaker was placed. So moving the loudspeaker around when measuring the vibration amplitude at the bridge foot estimated the radiativity.

Experiences from the experiments were:
- Radiativity is measured in an ordinary laboratory.
- Mode shape is simultaneously recorded.
- Violin radiativity is almost spherical up to the 500 Hz.
- Complicated radiativity patterns result at higher modes.

Important conditions for dynamic reciprocity measurements are that the system is linear, passive and that the vibration has stable equilibrium positions.

3. VISUALIZATION OF SOUND FIELDS USING INTERFEROMETRY

A more direct way to measure sound fields is interferometry. A reference wave is now added to the object, measuring wave. Such methods record the change in optical phase, $\Delta\Omega$, which is the change of the product of refractive index n, the geometrical path length L and the wave number k (assumed to be constant). That is, the change in optical phase is:

$$\Delta\Omega=\Delta(knL)=k(n\Delta L+L\Delta n) \qquad (1)$$

The wave number equals $k= 2\pi/\lambda$ where λ is the laser wavelength. The first part in (1), $n\,\Delta L$, equals zero if L is kept constant. Then only the $L\Delta n$-part remains. Since an increased air density or air pressure gives an increase to the refractive index, Δn, the "trick to see the sound" is simply to let the index of refraction vary along L in a measuring volume. Figure 2 illustrates the measuring principle. Laser light, illumination and object waves, travel through the measuring volume to and from the rigid reflector along almost the same path to the measuring unit of the interferometer.

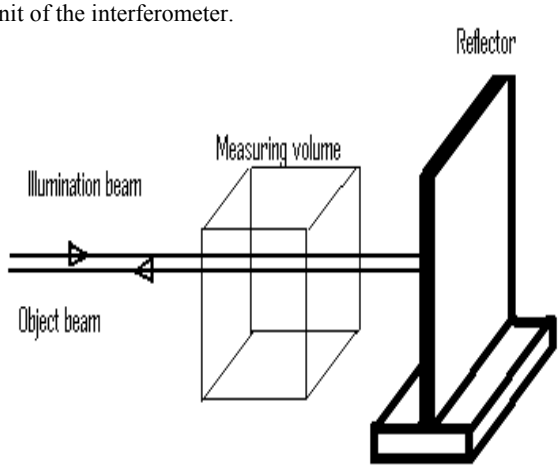


Figure 2: *Part of an interferometer that measures sound fields. The reflector is a rigid steel plate.*

Since $\Delta\Omega$ can be measured, the integrated product of $L\Delta n$ is known. The refractive index is related to the air density by the Gladstone-Dale equation. If adiabatic conditions are assumed, the relations between air density and pressure are also known. By combining these equations, a quantitative measure of the sound pressure p in all points in the field of view is thus possible to obtain.

Figure 3 shows standing waves inside a transparent guitar measured by phase modulated TV holography [13]. In figure 4 the transient sound field outside an impacted cantilever beam is depicted [8] recorded by pulsed holographic interferometry. It is interesting to see that bending waves in the beam acts as supersonic sound sources creating a pattern looking like a "bow wave" behind a ship. It differs however in that this one has opposite phases on each side of the beam.

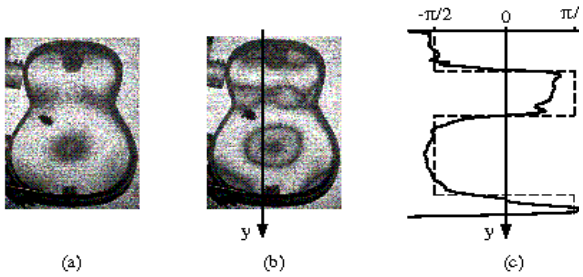


Figure 3: *Standing waves inside a guitar cavity, with and without cover of the sound hole measured by phase modulated TV holography. The measured phase distribution along the y-axis is shown.*

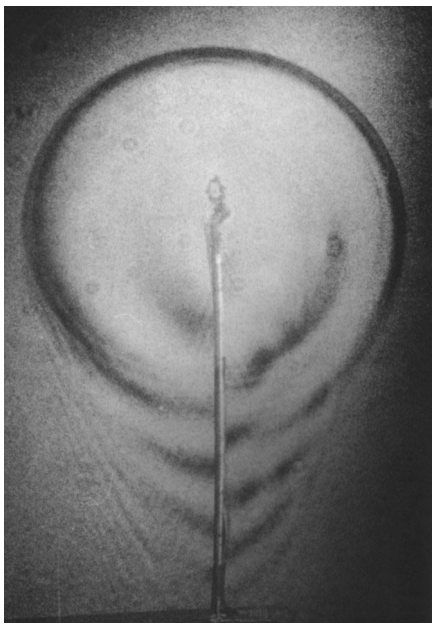


Figure 4: *A cantilever steel beam 1*30*190 mm, clamped at the bottom is impacted from the top right by an air gun bullet. The interferogram is obtained 230 μs after impact start by pulsed holographic interferometry. Supersonic bending waves travel down the beam and create "bow waves" moving faster than the spherical wave centered at the impact point.*

4. A SURVEY OF SOME MODERN OPTICAL METHODS TO VISUALIZE SOUND FIELDS

In table 1, four modern, optical, all-electronic methods that visualizes vibrations and sound fields are listed, they are:

1 **TV holography,** also called Digital or Electronic Speckle Pattern Interferometry (DSPI or ESPI) is used for "real-time" observation and time average recordings of quasi-static and harmonic events. Continuous lasers are used to illuminate the object. The technique is a modern, all-electronic, version of holographic interferometry and it allows quantitative field measurements. It needs vibration-isolated tables. Phase modulation of the reference beam is of great help in the study (of the phase) of vibrations. Rather small vibration amplitudes, down to about 0.05 μm are measured. Larger amplitudes than about 25 fringes (which corresponds to about 2.5 μm) across the CCD-censor can get hard to resolve. It is fast and effective method to use.

2 **Phase modulated TV holography** is used for the measurement of small (< 0.5 μm) vibration amplitude and phase distributions, also used for amplitude and phase measurements of sound fields (transparent "phase" objects). It is about 50 times more sensitive than ordinary TV holography. It is not working "in real-time". It is also sensitive to environmental disturbances. Both amplitude and phase maps are measured see for instance figure 3.

3 **Pulsed TV holography** (also called pulsed holographic interferometry) makes use of double pulsed or multiple pulsed lasers for object illumination. It is used to measure transients and other fast events. The recording time is so short, in the order of 10 ns that most mechanical events can be looked upon "as if " stationary. Vibration isolation of the setup is not necessary. It is an "all electronic method" but it does not work in "real-time". Pulsed lasers are quite expensive and they emit light with very high peak powers. Therefore protective goggles must be worn for safety reasons.

4 Scanning **L**aser **D**oppler **V**ibrometry (**LDV**) also called a **Scanning Vibrometry** measures velocity fields and scans a measuring laser point across the object. A signal, for instance from a microphone or accelerometer at the object, is necessary to get the relative phase between successive measuring points. A restriction is that the sequences recorded must be repeatable. Its high sensitivity allows the measurements of sound fields. Disturbances can be filtered electronically, which is a great advantage. Small vibration amplitudes down to the sub-nm-range, smaller than with phase modulated TV holography, can be measured. Special equipment like vibration-isolated tables is not necessary to use since it is the velocity signal that is measured and unwanted and slowly varying disturbances can be filtered out. Another advantage is that ordinary modal analysis tools can be applied to the signals, as the well-developed techniques with Fourier analysis of accelerometer data. Figure 5 shows a scanning vibrometry of a 2D- ultrasound acoustic field.

Methods **1-4** above measure the integrated sound pressure field (integrated optical phase change along the probing light rays) of the sound distribution. This is often presented as a two-dimensional interferogram. To obtain full volume information a tomographic reconstruction can be made.

Sound fields	Harmonic variations	Repetitive, multi-frequency sequences	Transient, fast, dynamic events
Standing waves	**1, 2, (3), 4**	**-**	**-**
Propagating waves	**1, 2, (3), 4**	**3, 4**	**3**

Table 1: *Optical methods to visualize sound fields*

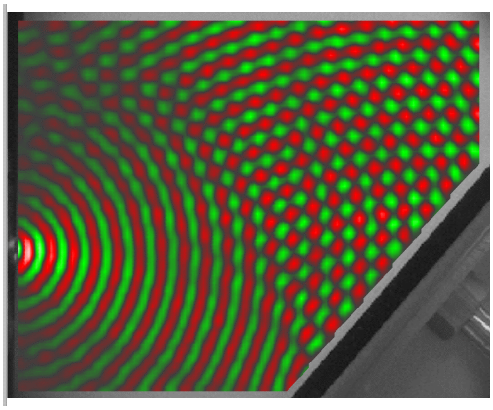


Figure 5: *Interference pattern between direct ultrasound from a "point" source (to the left) and the reflected ultrasound via a mirror (to the right), recorded by scanning Vibrometry (LDV).*

5. SPECKLE PHOTOGRAPHY AND CORRELATION METHODS.

The equipment needed in methods **1-4** above, is still rather expensive and sometimes also quite complicated to use. Simpler and cheaper methods are looked for. A less expensive but also less sensitive method is digital speckle photography (DSP), today also named speckle correlation [4]. A random pattern that "acts as if it is glued" to the object surface is imaged twice with a CCD-camera on different frames, before and after some load is applied to the object that changes the position of the random pattern. So far this reminds of TV holography (or Digital Speckle Pattern Interferometry, DSPI) but this time no reference wave is added.

Two types of random patterns are in use: First laser speckles; this random (speckle) pattern is generated by the

imaged microstructure of the laser illuminated diffuse object surface itself (laser speckles in DSP). In the second type (White light DSP) there may already exist a random pattern at the object surface as "random" fibers in paper or a random microstructure in a wood sample, which is imaged. Otherwise a random pattern may be painted or sprayed onto the surface. Ordinary white light is used for such measurements.

A digital correlation technique is then used to determine how much small sub areas in the first recording have to be moved to correlate as exactly as possible, with the second pattern. Performing this for all sub-areas, an in-plane deformation map is formed. Figure 6 shows an example of a defocused laser DSP-recording of a transparent He-gas jet, which is entering into air. From such recordings it is possible to calculate where the gas jet is situated and how strong the refractive index field of the gas is [14]. This technique, in its focused version, reminds of PIV (Particle Image Velocity) and determines velocity fields in turbulent flows for instance at the labium of a blown organ pipe [see the proc of the present conference].

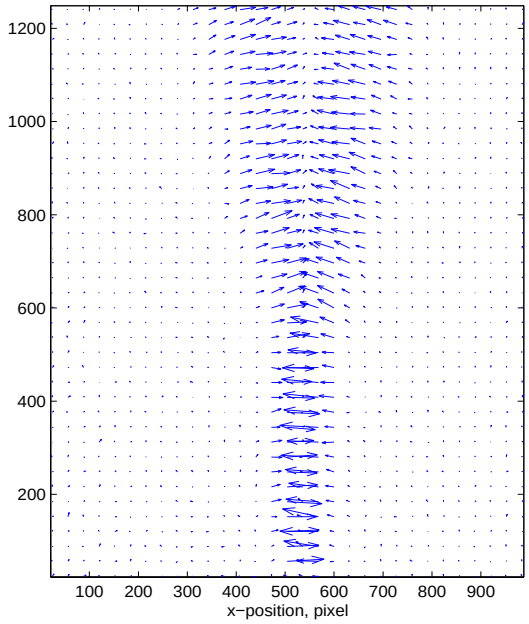


Figure 6: *A DSP recording of defocused laser speckles motion when penetrating a transparent He-gas jet.*

6. CONCLUSIONS

Optical metrology continues to grow in importance and applicability to scientists and engineers in mechanics, acoustics and fluid mechanics. Modern optical techniques like phase modulated TV holography, pulsed TV holography and scanning vibrometry-methods measure phase shifts in transparent, phase objects like sound fields in air and water. They present "pictures" and they produce quantitative measures. Reciprocity methods can also be of great help in such experiments. Optical measuring methods have the advantage of being contact less, non-disturbing and whole field methods. The future for optical metrology in acoustics looks bright.

7. REFERENCES

[1] Cagnet, M., Francon M. and Thrierr, J. C., Atlas de phénomènes d`optique (Atlas of optical phenomena). Springer Verlag, Berlin, Göttingen, Heidelberg. 1962.

[2] Van Dyke, M., An Album of Fluid Motion. The Parabolic Press. Post office box 3032. Stanford, California 94305-0030, 1982.

[3] Kreis, T., Holographic Interferometry Principles and Methods, Academie Verlag GmBH, Berlin, 1996.

[4] Rastogi, P. M. (ed). Digital Speckle Pattern Interferometry and related Techniques, John Wiley & Sons LtD, Chichester, England, 2001.

[5] Fomin, N. A., Speckle Photography for Fluid Mechanics Measurements, Springer-Verlag Berlin Heidelberg, 1998.

[6] Zipser, L. and Lindner, S., "Visualisation of Vortexes and Acoustic Sound Waves", Proc. 17th Int. Congress on Acoustics, Volume I, Physical Acoustics part B, Ultrasonics, quantum acoustics and physical effect of sound: 24-25, Rome September 2-7, 2001.

[7] Lökberg, O. J., "Sound in flight: measurement of sound fields by use of TV holography". Appl.Opt. 33(13) 2574-2584, 1994.

[8] Wåhlin, A. O. Gren, P. O. and Molin, N-E., "On structure-borne sound: Experiments showing the initial transient acoustic wave field generated by an impacted plate." J. Acoust. Soc. Am. 96 (5): 2791-2797, 1994.

[9] Molin, N-E., Wåhlin, A.O and Jansson, E.V., "Transient wave response of the violin body revisited." Journal of Acoustical Society America. 90, (4), 2192-2195. Oct 1991.

[10] Lökberg, O. J., Espeland, M. and Pedersen, H. M., "Tomographic reconstruction of sound fields using TV holography", Appl.Opt. 34(10), 1640-1645, 1995.

[11] Gren, P., Schedin, S and Li, X., "Tomographic reconstruction of transient acoustic fields recorded by pulsed TV holography", Applied Optics Vol 37, No 5: 834-840, 1998.

[12] Runnemalm, A. and Molin, N-E., "Operating deflection shapes of the plates and standing aerial waves in a violin and a guitar model", ACUSTICA united with acta Acustica, Vol 86, No 5:883-890, 2000.

[13] Saldner, H. O., Molin, N-E. and Jansson, E. V., "Sound distribution from forced vibration modes of a violin measured by reciprocity and TV holography", CAS Journal, Vol 3, no 4 (Series II):10-16, 1997.

[14] Johansson, E.L., Benckert, L. and Sjödahl, M.,"Phase object data obtained from defocused laser speckle displacement". Submitted for publication 2003

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

BREATHING MODES OF ORGAN-PIPE BODIES: EXPERIMENTAL DETECTION WITH RING-SHAPED PIEZOELETRIC-POLYMER SENSORS

M. Wegener[1], R. Gerhard-Multhaupt[1], W. Wirges[1], A. Bergner[2], S. Bergweiler[2]

University of Potsdam, Department of Physics, ACMP, Potsdam, Germany (1),
UP Transfer GmbH at the University of Potsdam, Potsdam, Germany (2)

mwegener@rz.uni-potsdam.de

ABSTRACT

For the operation of wind instruments, the breathing of their bodies plays only a secondary role. But the change of the cross-sectional area may influence the sound (1) by radiation of "secondary" sound waves and (2) by changing the boundary conditions for the sound field inside the cylindrical resonator, which will in turn affect the sound radiated from the instrument.

Here we present measuring techniques based on piezoelectric polymer sensor films for detecting wall oscillations of cylindrical wind instrument parts. Piezoelectric sensor films with various electrode patterns were slightly tensioned around e.g. an organ pipe and used to measure the circumferential oscillations of its body. A possible disadvantage of this technique as compared with interferometry or acoustical measurements may be a slight restriction of the body-wall breathing by the tensioned piezoelectric film, but the advantage is a complete recording of the breathing for the full circumference of the cylindrical body if a sensor film with non-patterned electrodes is employed. With patterned electrodes on the piezoelectric film, it was also possible to directly detect the ellipsoidal quadrupole oscillation at the upper end of an organ pipe, which is already known from interferometric measurements [2]. Furthermore, a piezoelectric sensor film was used as ring-shaped microphone. The advantage of this technique is that the breathing of the circumference of the pipe body can be detected without influencing the pipe body through any mechanical contact. However, a disadvantage of the microphone technique is that any loud sound radiated near the microphone will also be detected. We discuss the details of the new sensor technique as well as typical measurements in order to demonstrate the applicability. In general, the proposed sensor can be employed to characterize body-wall breathing oscillations of wind instruments independent of tube material and surface smoothness.

1. INTRODUCTION

Air-column oscillations in wind instruments are always exited by air jets. This is the case for brass instruments, for reed instruments and for flutes and organ pipes. The primary oscillator part of these groups may take various forms, but the resonator that controls the spectral envelopes is always a tubular cavity. According to the dimensions and the shape of this tube, the air column inside the instrument exhibits resonances, amplifying certain harmonic frequencies. The resulting sound is radiated from the openings. These mechanisms are well investigated for most wind instruments. Reasonable agreement between the physical models and the observed behavior is found [1].

On the other hand, it is often not known whether the resonator body of a wind instrument influences the audible sound in any other way but through its geometry. Usually, the tubes are considered to be perfectly stiff and motionless. But if this assumption were true, then why is the choice of material so important for wind instruments? Recently, organ pipes were investigated with respect to wall oscillations of their resonator bodies. By means of optical interferometry, a planar quadrupole oscillation was found at sections perpendicular to the pipe axis [2]. However, the possible influence of these oscillations on the sound of the pipe has not been studied yet.

Because of the pressure distribution of the standing waves inside an organ pipe or other wind instruments, it is conceivable that a “breathing” motion occurs, with the same phase in all directions around the pipe circumference. This “breathing” wall motion may be superimposed onto the previously observed quadrupole oscillation and will then influence the sound radiation. In general, sound is radiated if some area oscillates and if an air volume displacement takes places. With wind instruments such as organ pipes or flutes, sound is mainly radiated from the stationary pressure wave inside the pipe resonator. The tube cross section of a wind instrument may oscillate in several modes. But its diameter and circumference are very small in comparison with the wavelength of the relevant sound waves. Thus, only monopole-like “breathing” of the cross section may significantly contribute to the sound of a wind instrument.

Here we present techniques that are useful for measuring the “breathing” of cylindrical resonators. In order to investigate the influence of an attached sensor on the resonator wall, two sensors, a non-attached ring-shaped microphone and a directly attached sensor film, were developed. These techniques may also be useful for the investigation of other wind instruments or tube systems.

2. EXPERIMENTAL TECHNIQUES

In order to investigate the in-phase wall motion of the cylindrical part of a pipe, we employ films which have for instance a length of at least 14.3 cm in order to surround a c (2’) labial organ pipe completely. With continuous electrodes on its front and rear surfaces, such a film integrates over the whole sensing area around one circumference so that it is only sensitive to monopole (“breathing”) oscillations, but not to dipole,

quadrupole, octupole, etc. contributions. Suitable sensors are e.g. piezoelectrically active polymers.

Piezoelectric polymers exhibit a thickness or length change, when an electrical voltage is applied to their electrodes (inverse 33 and 31 effect, respectively) or they generate an electric signal when mechanical force is applied to the thickness or length of the film (direct 33 and 31 effect, respectively). The resulting piezoelectric coefficient d for the direct and the inverse effect may be written as (1).

$$d = \left(\frac{\partial Q}{\partial F}\right)_V = \left(\frac{\partial x}{\partial V}\right)_F \qquad (1)$$

Here Q is the charge on the electrodes, F the applied mechanical force, x the sample thickness and V the applied voltage. The electromechanical properties are intrinsic material properties of piezoelectric or quasi-piezoelectric polymers. Suitable electromechanically active polymers are polar non-cellular (and often ferroelectric) polymers such as polyvinylidene fluoride and some of its copolymers [3] as well as the new class of the usually non-polar cellular space-charge electrets [4]. For the necessary symmetry breaking, an electric poling process leads to an orientation of molecular dipoles within domains or to micro-discharges and charge trapping at internal void surfaces within the polar non-cellular or non-polar cellular polymers, respectively. After suitable preparation and poling, the materials are metallised on both outer film surfaces and may then be used as piezoelectric sensors with arbitrary areas and various shapes. In comparison to piezoelectric sensors, an electret microphone requires a well defined air gap between one electrode and the electrically charged polymer film. This reduces the possibility of unusual microphone shapes and sizes.

We developed two different measuring techniques for investigating the "breathing" of pipe resonators: (1) The piezoelectric polymer film is used as a microphone for airborne sound. To this end, the sensor film is attached to the inside of a metal ring, which serves as acoustically hard backing and as electrode (Figure 1). (2) For the second technique, the piezoelectric film is directly attached to the pipe resonator (schematic picture and photograph in Figures 2 and 3, respectively). The sensor film is pre-stressed: a force of 0.74 N is applied to one of the sensor ends in order to ensure optimal mechanical contact to the organ-pipe body. The other end of the sensor film is fixed to the organ-pipe body with adhesive tape. We also investigated other attachment schemes such as free film ends and pre-stressing at both film ends as well as fixing of both sensor ends to the pipe body, but the above-mentioned arrangement was found to be the most suitable for wall-motion detection. In order to study the expected "breathing" as well as the reported quadrupole oscillation, sensor films with various electrode patterns were prepared (Figure 4).

The sensor with the continuous electrode (upper half of Figure 4) detects the volume displacement. Any higher-order multipole modes of the surface oscillations are only taken into account if the integral of their elongation includes some volume displacement. The technology thus selects only the monopole part of the surface motion. The metallized electrode has an area of 1.5 cm by 11.5 cm, so that 290° of the pipe-body circumference are covered. Further investigations with sensors covering the whole circumference are in preparation.

The selective electrode areas shown in the lower half of Figure 4 are arranged in such a way that they are located at 0, 90, 180 and 270° of the pipe-body circumference. The sensor with patterned electrodes is useful for local measurements of wall motions at selected areas of the organ-pipe circumference.

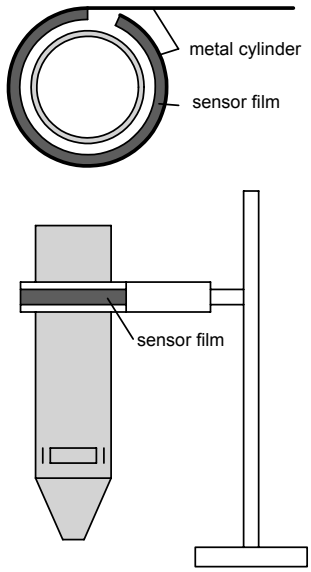


Figure 1: *First measuring scheme: piezoelectric sensor film used in ring-shaped microphone surrounding the organ pipe.*

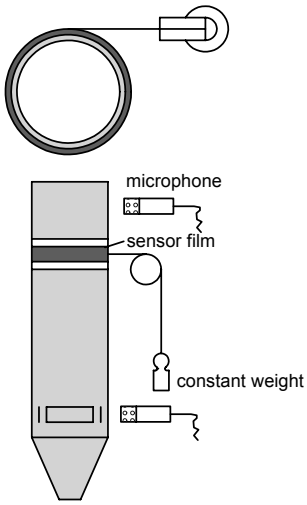


Figure 2: *Second measuring scheme: piezoelectric sensor film for direct wall-motion measurements, attached around the organ pipe.*

The piezoelectric polymer used for all sensors is polyvinylidene fluoride – trifluoroethylene (PVDF-TrFE) with thicknesses of 50 µm and 25 µm for the ring-shaped microphone and the attached sensor, respectively. The polymer films were electrically poled by means of a point-to-plane corona discharge at –22 kV. This voltage was suitable for generating a homogenous polarisation across the sample thickness and significant piezoelectricity. After the poling process, the sensor films were metallized first with chrome and then with aluminium

on both surfaces. The chrome layer was necessary in order to enhance the stability of the electrode during movements of the sensor films.

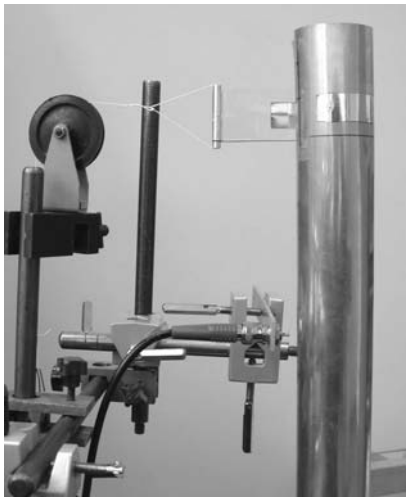


Figure 3: *Photograph: piezoelectric sensor film for direct wall-motion measurements, attached around the organ pipe.*

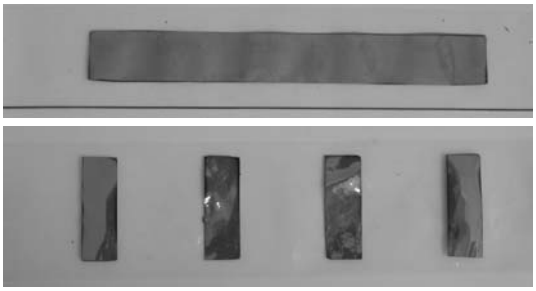


Figure 4: *Piezoelectric sensor films with continuous (upper half) and patterned (lower half) electrodes.*

In order to contact the sensor elements, thin silver wire was glued to the sensor areas. The electrical signal was recorded by means of a Spectrum Analyzer SR 770 FFT. With both sensors (1) and (2), the wall motions of the pipe body were investigated along the resonator length in steps of 1 cm and 2 cm in case of the ring-shaped microphone and the non-patterned sensor film, respectively. With the patterned sensor film, the ellipsoidal quadrupole oscillation was investigated with the sensor electrodes rotated around the circumference of the pipe body in steps of 22.5°. The influence of the attached sensor stripe to the radiated sound was investigated by means of microphone measurements (B&K 2669), 1 cm away from both openings of the c (2') principal organ pipes. The time-dependent signal was recorded with a sampling frequency of 44.1 kHz and 32 bit A/D conversion by means of a M-Audio 2496 measuring system.

3. TYPICAL RESULTS AND DISCUSSION

3.1. Polymer film in ring-shaped microphone

If the piezoelectric sensor film is used as a non-contacting ring-shaped microphone, there should be no influence on the oscillation of the organ-pipe wall or on the sound radiation of the organ pipe. However, this technique has the disadvantage of large signals at the openings of the organ pipe, so that small variations may be masked.

3.2. Attached sensor film – Investigation by means of microphone measurements

In order to study the influence of the sensor on the acoustical behavior of the pipe, we did measurements on a pipe without and with reduced wall thickness. The latter type of organ pipe features a very small (0.5 cm) wall thickness at its open end. The piezoelectric sensor film was tensioned at a position of 2cm from the open end. The transmitted sound was recorded by means of the condenser microphone near the upper and lower openings of the pipe body. Each measurement was repeated 6 times with the sensor attached and 6 times without the sensor film. Only a close look at the sound pressure levels (SPL) of the harmonics possibly shows some influence. For the measurements with the microphone near the open end, the mean, maximum and minimum sound-pressure levels at the first five partials are summarized in Table 1. These measurements show that the attached sensor does not significantly influence the sound radiation of the organ pipe. In principle, the same result was found for the microphone measurements near the lower opening of the organ pipe. The SPL differences as shown in Table 1 are essentially within the experimental error margins. Fluctuations of between 0.4 dB for the first harmonic and 1 dB for the fifth harmonic are found on pipes with constant wall thickness. Pipes with a wall thickness that decreases from 1 mm at the mouth to 0.5 mm at the open end exhibit SPL changes of up to 2 dB for the first harmonic [6].

3.3. Attached sensor film – Investigation of wall motion

The piezoelectric sensor film with the patterned electrode was attached at the center of the pipe body. Two sensor elements with an angular relationship of 90° were connected to the signal-recording system. Figure 5 shows the simultaneously measured voltage signals of both sensors as a function of time at a fundamental frequency of 263 Hertz, i.e. the first harmonic. For the 90° shifted sensor elements, signals of opposite polarity were found. This result demonstrates an ellipsoidal quadrupole oscillation around the circumference of the pipe body at the sensor location.

Figure 6 presents the measurements with the two sensor elements at different angular positions, but at the same distance from the open end of the pipe. The observed values are the relative levels of the first harmonic, taken from simultaneous time-dependent measurements of both sensors. For better readability, the curve from the 90° shifted sensor was shifted back by 90° so that the phase of both sensors should be the same. The observed angular dependence shows the well-known quadrupole oscillations and thus demonstrates the applicability of our sensor technique.

Measurements with piezoelectric sensor films with continuous electrodes allow the investigation of the "breathing" of organ pipes. As expected, the "breathing" was observed and investigated on different types of organ pipes as further described in [5].

order of harmonic	SPL without sensor (dB)			SPL with sensor (dB)			SPL difference (dB)
	mean	maximum	Minimum	mean	maximum	Minimum	SPL without sensor – SPL with sensor
1	-9.24	-9.13	-9.40	-8.88	-8.79	-8.96	-0.36
2	-11.90	-11.60	-12.01	-13.23	-13.13	-13.37	1.33
3	-16.90	-16.60	-17.19	-17.38	-17.15	-17.55	0.48
4	-23.48	-23.18	-24.10	-24.55	-24.22	-24.67	1.06
5	-34.36	-34.22	-34.61	-34.24	-34.10	-34.43	-0.12

Table 1: *Measured sound-pressure levels of a thick-walled organ pipe without and with sensor attached.*

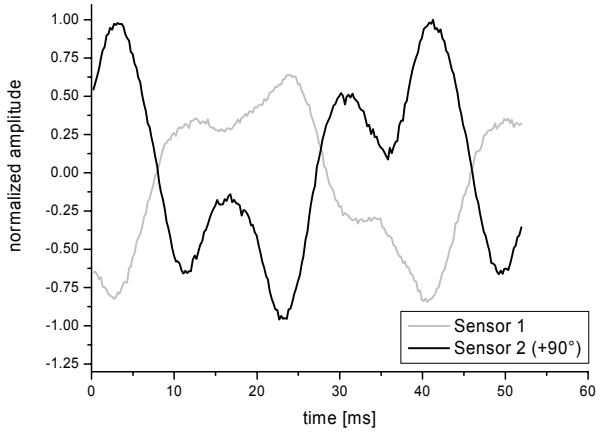


Figure 5: *Signals from two electrodes of the patterned sensor film attached at the center of the pipe body.*

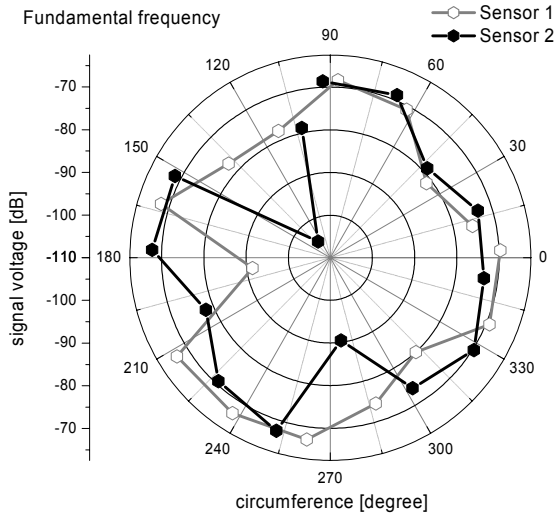


Figure 6: *Signals from two active elements at different positions around the circumference of the organ pipe (angle of the second sensor shifted by 90°).*

4. CONCLUSIONS

Measurement techniques for the investigation of wall motions of tubular acoustic resonators were developed on the basis of thin piezoelectric polymer films. The sensors were used to detect the wall motions without and with mechanical contact to the pipe body. By means of non-patterned sensors, the „breathing“ of organ-pipe resonators was demonstrated. With suitably patterned sensors, the ellipsoidal quadrupole oscillations were detected. In general, the proposed techniques are suitable for vibration measurements not only on organ pipes, but also on other tubular acoustic resonators.

5. ACKNOWLEDGEMENTS

We kindly acknowledge Schuke Orgelbau Potsdam, for making the organ pipes and the laboratory-scale wind supply system available. We are indebted to Thomas Görne and Dr. Detlef Zscherpel for many stimulating discussions. We also would like to thank the German Federal Ministry for Economics and Technology (InnoNet Project 16IN0001 and ProInno Project 0203902KBV0) for financial support.

6. REFERENCES

[1] Rossing, T., Fletcher, N., “The physics of musical instruments”, Springer, New York, 1998.

[2] Runnemalm, A., Zipser, L., Franke, H., “Structural vibration modes of a blown organ pipe”, Acta acoustica Vol. 85, Nr. 6, p. 877-882, November 1999.

[3] H.S., Nalwa (Ed.), “Ferroelectric Polymers”, Marcel Dekker Inc., New York 1995.

[4] Gerhard-Multhaupt, R., “Less can be more: Holes in polymers lead to a new paradigm of piezoelectric materials for electret transducers”, IEEE Transactions on Dielectric and Electrical Insulation, Vol. 9, No. 5, 850-859 (2002).

[5] Bergweiler, S., Bergner, A., Görne, T., Wegener, M. and Gerhard-Multhaupt, R., “Breathing modes and sound radiation of metallic organ pipes”, Proceedings SMAC 03, 2003, Stockholm

[6] Görne, T., Zscherpel, D., Zierenberg, M., Bergweiler, S., and Gerhard-Multhaupt, R., „Klangliche Unterschiede zwischen nominell baugleichen Labialorgelpfeifen aus parallelwandigen oder ausgedünnten Metallplatten“, (“Sound differences between otherwise identical labial organ pipes made from plane-parallel and from wegde-shaped metal sheets”), Proceedings 29. DAGA, 2003, to be published.

LEAK DETECTION IN MUSICAL WIND INSTRUMENTS USING ACOUSTIC PULSE REFLECTOMETRY

V. Chilekwa, D.B. Sharp and T.J.W. Hill

Acoustics Group, Department of Environmental and Mechanical Engineering, Open University, Milton Keynes, MK7 6AA, United Kingdom
V.Chilekwa@open.ac.uk, D.Sharp@open.ac.uk

ABSTRACT

In this paper, the technique of acoustic pulse reflectometry is applied to the problem of leak detection in tubular systems, such as musical wind instruments. Acoustic pulse reflectometry is a method for measuring the input impulse response of a duct. Suitable analysis of the input impulse response enables both the input impedance of the duct and changes in area along the duct to be calculated. The presence of a leak results in a spurious increase in the calculated cross-sectional area at the position of the leak. This is illustrated through measurements made on historical wind instruments. It is then shown that information about the size of a leak can be obtained using both time-domain and frequency-domain approaches.

1. INTRODUCTION

The detection of leaks in tubular systems has applications ranging from water pipe networks to industrial components to musical instruments. In the latter case the presence of leaks can have serious effects on the tone-quality and playability of the instrument. Leaks can occur at keys, valves, joints or other parts of instruments as a result of damage, wear or poor craftsmanship.

Acoustic pulse reflectometry, which has been widely applied to musical wind instruments, provides a useful method by which the location and size of leaks can be identified. The technique involves injecting a pulse of sound into a tubular object and digitally recording the resultant reflections. Suitable analysis of these reflections yields the input impulse response of the object from which the input impedance and bore profile can be calculated [1, 2, 3]. The procedure is non-invasive and affords the additional benefit that instruments can be measured without having to be sealed and pressurised

In this paper, we describe how acoustic reflectometry can be applied to simple systems and musical instruments in order that both time and frequency-domain information can be used to determine the size and location of a leak. In Section 2, reflectometry data from historical wind instruments are presented that illustrate the effect that a leak has on the reconstructed bore profile. Following on from this, in Section 3, the size of a leak is calculated from frequency-domain data by means of a lumped impedance model. Finally, the potential for a new time-domain approach to calculating the size of a leak is explored.

2. DETECTING THE PRESENCE OF LEAKS IN WOODWIND AND BRASS INSTRUMENTS

The technique of reflectometry involves measuring the time sequence of reflections produced when a sound pulse is injected into a duct of varying cross sectional area. With this information and knowledge of the input pulse waveform, application of a layer-peeling algorithm [4] enables the change in impedance along the duct and hence the bore profile to be calculated.

A leak, or hole, in the duct wall presents a significant change in impedance to the probing sound pulse. The relative size of this change in impedance combined with the layer peeling approach results in a systematic error in the bore profile reconstruction that starts at the location of the leak. This can be seen quite clearly through examination of the solid line in Figure 1, which shows the bore reconstruction of a Rudall Carte 'Conical bore model' cornet measured with all three valves operated. At a distance of about 1.2 m from the mouthpiece end of the instrument there is a sudden expansion. Beyond this, the predicted radius deviates considerably from that measured manually using calipers, suggesting the presence of a leak at this location. Visual inspection of the cornet confirmed this to be the case with the discovery of a small crack in the third-valve tuning slide. The dotted line in Figure 1 shows the result obtained when the crack was sealed. The reconstruction and the direct measurement are now in good agreement.

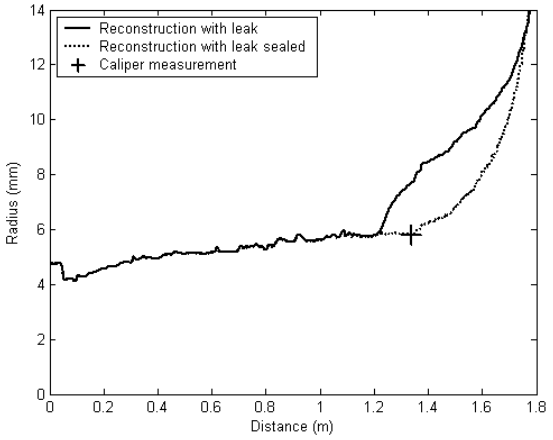


Figure 1: *Reconstructed bore profile of Rudall Carte cornet with all three valves operated.*

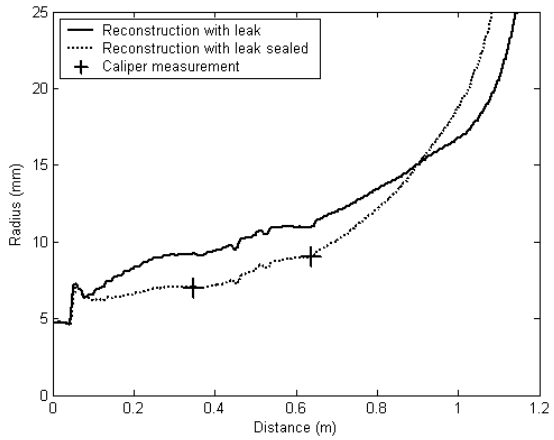


Figure 2: *Reconstructed bore profile of King trombone bell.*

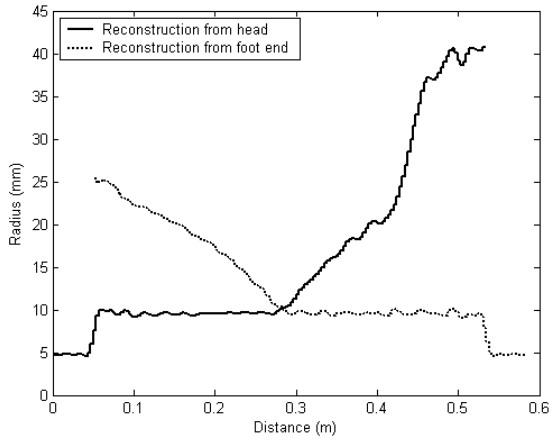


Figure 3*: Reconstructed bore profiles of Blessing flute with a leaking pad.*

Similarly, Figure 2 shows bore reconstructions of the bell section of a King trombone. The solid line deviates significantly from the two manual measurements, indicated by crosses at 350 mm and 640 mm, suggesting the presence of a leak at the beginning of the bell section. Again, visual inspection confirmed this to be true. With the leak sealed the bore reconstruction is in close agreement with both manual measurements.

Unless the entire bore profile of an instrument is known in advance, an expansion in the reconstruction due to a leak may be indistinguishable from a real widening of the bore. Hence, although it may be possible to infer that a leak exists in a particular region of the instrument, the precise location of the leak may be more difficult to determine.

One method by which the precise location of a leak can be deduced is to perform one reflectometry measurement from each end of the instrument. Plotting the resultant bore reconstructions on the same axes gives 'good' reconstructions from each end of the instrument up to the region containing the leak. The precise location of the leak is identified as the position where the two reconstructions cross. This approach is clearly demonstrated in Figure 3 which shows bore profiles calculated for a Blessing flute, which has a leaky keypad halfway along its length. This method is more readily applied to cylindrical ducts or ducts with gradual tapers (e.g. woodwinds), as coupling the reflectometer to wider flares (e.g. brass instruments) is less practical.

3. DETERMINATION OF THE LEAK SIZE

As well as being able to identify the presence of a leak using reflectometry measurements it is also possible to determine its size. This can be approached from either the frequency or time-domain.

3.1. Frequency-domain approach

A frequency-domain approach has been developed [5] which enables the size of the leak to be determined from complex input impedance measurements made with a reflectometer. However, this method has, so far, only been applied to a cylindrical pipe and requires prior knowledge of its geometry and of the location of the leak.

When determining the size of the leak in a cylindrical geometry the system can be modelled as two cylindrical sections, of length l_1 and l_2, either side of the leak The leak is treated as a cylindrical side branch of radius r_h and length l_h, equal to the wall thickness of the duct (Figure 4).

Using a lumped impedance approach the impedance of the leak, z_h, can be expressed in terms of the two impedances z_1 and z_2, as

$$z_h = \frac{z_1 z_2}{z_2 - z_1}. \qquad (1)$$

The impedance, z_1, is related to the measurable input impedance of the first cylindrical section, z_0, by

$$z_1 = \frac{z_0 - j\dfrac{\rho\omega}{k\pi r^2}\tan kl_1}{1 - j\dfrac{z_0\pi r^2}{\rho\omega}\tan kl_1}. \qquad (2)$$

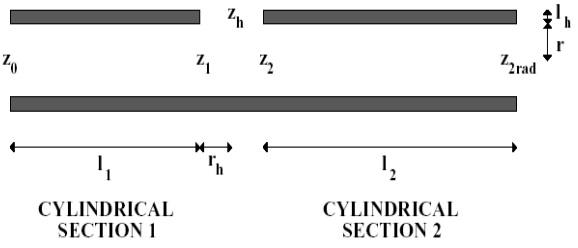


Figure 4: *Schematic diagram of leaking cylinder.*

Similarly, the input impedance of the second cylindrical section, z_2, is related to the radiation impedance of its open end, z_{2rad}, by

$$z_2 = \frac{\rho\omega}{k\pi r^2}\left(\frac{\frac{z_{2rad}k\pi r^2}{\rho\omega} + j\tan kl_1}{1 + j\frac{z_{2rad}k\pi r^2}{\rho\omega}\tan kl_2}\right). \quad (3)$$

In this case, as the open end is unflanged the radiation impedance is given by [6]

$$z_{2rad} = 0.25\frac{\rho\omega}{k\pi r^2}k^2r^2 + j0.6\frac{\rho\omega}{k\pi r^2}kr. \quad (4)$$

So, z_2 is given by the expression

$$z_2 = \frac{\rho\omega}{k\pi r^2}\left(\frac{0.25k^2r^2 + j(0.6kr + \tan kl_2)}{(1 - 0.6kr\tan kl_2) + j0.25k^2r^2\tan kl_2}\right). \quad (5)$$

Substituting equations (2) and (5) into equation (1) gives the complex impedance of the hole, z_h, in terms of the measured input impedance of the leaking cylinder, z_0.

By then substituting this value for z_h into the following cubic equation (Equation 6, derived theoretically in [5]) and solving, the radius of the hole is determined.

$$r_h^3 + \left[\frac{1.724\pi r^2\Im(z_h)}{\rho\omega}\right]r_h^2 + \left[\frac{0.431r^2}{c}\left(\sqrt{\frac{\eta\omega}{2\rho}} + (\gamma - 1)\sqrt{\frac{\kappa\omega}{2\rho C_p}}\right) - 2.75r^2\right]r_h - 1.724l_hr^2 = 0 \quad (6)$$

where c is the speed of sound in air, γ is the ratio of the specific heats of air, C_p is the specific heat at constant pressure, η is the coefficient of shear viscosity of air, and κ is the thermal conductivity of air.

3.1.1 Practical predictions of leak size

The complex input impedance was calculated from the input impulse response (measured with a reflectometer) for a cylindrical pipe (of length 501 mm, internal radius 5.0 mm and wall thickness 1.0 mm) with a hole of 1 mm radius bored in the pipe wall 200 mm from its start. The radius of the hole was predicted using Equation 6 for frequency components in the range 500 Hz to 2 kHz. When plotted as a function of frequency and compared with the actual size of the hole (Figure 5) it is apparent that the predicted values are erroneous when a pressure node is coincident with the leak, i.e. at frequencies of integer multiples of 507 Hz for this geometry. To avoid such frequencies and obtain a best prediction for the effective radius of the hole, values of r_h were averaged over the frequency ranges 712.5 to 997.5 Hz and 1282.5 to 1567.5 Hz. Further measurements and predictions were made for holes of radius 1.5 mm and 2.5 mm. The results are shown in Figure 5 and Table 1. For each of the holes the predicted radius agrees with that measured directly within experimental error.

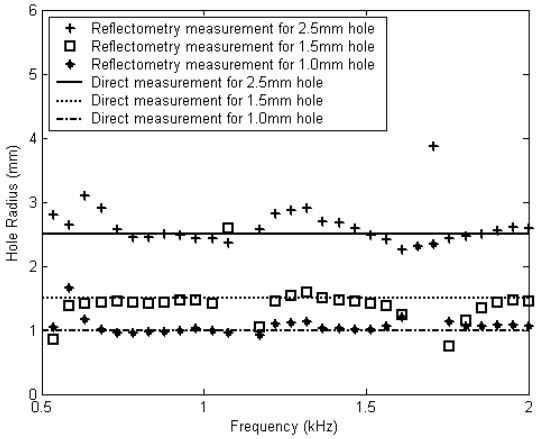


Figure 5: *Comparison of predicted and measured hole radii over a range frequencies.*

Actual radius (mm)	Average radius predicted from frequency range 712.5 to 997.5 Hz (mm)	Average radius predicted from frequency range 1282.5 to 1567.5 Hz (mm)
1.0 ± 0.05	0.979 ± 0.02	1.044 ± 0.04
1.5 ± 0.05	1.440 ± 0.05	1.463 ± 0.05
2.5 ± 0.05	2.479 ± 0.05	2.627 ± 0.05

Table 1: *Measured hole radii and average predicted radii calculated from two frequency ranges.*

3.2 Time-domain approach

It should also be possible to devise a time-domain method for determining the size of leaks in pipes based on reflectometry bore reconstructions. This approach is illustrated in this section with bore reconstructions of identical sections of pipe that have had different size holes bored in the pipe wall at the same location. Figure 6 shows that the bore reconstructions expand rapidly from the location of the leak but at different rates according to the size of the hole. The largest hole (4 mm radius) exhibits the fastest expansion and the smallest (1 mm), the slowest. With this time domain approach the location of the leak is clearly indicated by the onset of the expansion. The size of the leak obviously correlates with the gradient of the erroneous expansion, however it is not clear, at this stage, exactly how the gradient should be determined and the correlation made.

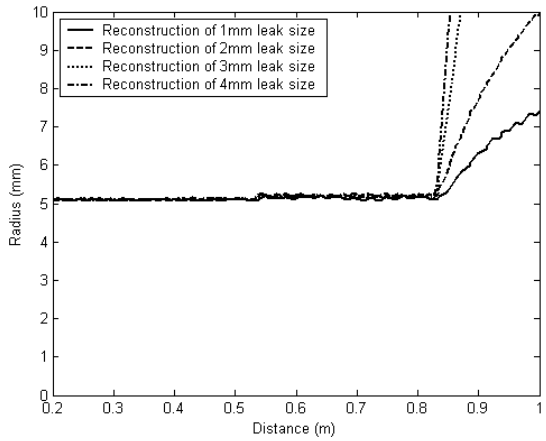


Figure 6: *Bore reconstructions of different signal leaks showing different expansion rates.*

4. CONCLUSIONS

Acoustic pulse reflectometry has been demonstrated to be a useful method by which to obtain information which can be processed in either the time or frequency-domain to confirm the existence, location and size of a leak in a tubular duct. Both time and frequency-domain approaches to determining the axial location and radius of a leak in cylindrical duct have been presented and the limitations of each technique have been indicated. Acoustic pulse reflectometry methods have been applied to musical instruments and have been shown to be capable of detecting and locating small leaks (cracks and leaking keypads) that are difficult to locate by visual inspection and may not even be noticeable by a musician initially.

5. REFERENCES

[1] Smith, R.A., "It's all in the bore!", J.Int.Trumpet Guild, 1988, 12(4).

[2] Watson, A.P. and Bowsher, J.M. “Impulse measurements on brass musical instruments”, Acustica, 1988, 66, 170-173.

[3] Sharp, D.B. "Acoustic pulse reflectometry for the measurement of musical wind instruments." PhD Thesis, 1996, University of Edinburgh.

[4] Amir, N., Rosenhouse, G. and Shimony, U. “A discrete model for tubular acoustic systems with varying cross section - the direct and inverse problems. Parts 1 & 2: Theory and Experiments”, Acustica, 1995, 81, 450-474.

[5] Sharp, D. B., and Campbell, D. M. "Leak detection in pipes using acoustic pulse reflectometry," Acustica, 1997, 83, 560-566.

[6] Kinsler, L.E., Frey, A.R., Coppens, A.B. and Sanders, J.V., “Fundamentals of acoustics”, 4th ed., John Wiley and Sons Inc, New York, 2000.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

AN EDS MODELLING TOOL FOR TRACKING AND MODIFYING MUSICAL SIGNALS

Bertrand DAVID, Gaël RICHARD and Roland BADEAU

ENST, Department of Signal and Image processing
École Nationale Supérieure des Télécommunications,
46, rue Barrault,
75634 PARIS cedex 13, FRANCE
bedavid,grichard,rbadeau@tsi.enst.fr

ABSTRACT

An analysis/synthesis scheme for musical signals is introduced in this paper. It is based on an adaptive subspace analysis and the Exponentially Damped Sinusoids model. This method leads to a new representation, called the HR-ogram, where the signal components are represented as points in the time-frequency plane. These points are gathered according to their frequency, phase and amplitude proximity from an analysis time-instant to the following one. This leads to an accurate deterministic/stochastic decomposition using a projection onto the noise subspace. The whole technique allows a separate processing for both components.

1. INTRODUCTION

Most of the analysis/synthesis schemes designed for musical sounds found in the literature are based on either a frequency-domain or a time-domain approach. Both families have evolved into a broad variety of algorithms from their very ancestors: the so-called phase-vocoder [1] and the OLA (overlap-add) method. These tools and their derivatives (Quatieri and Serra techniques [2, 3], synchronized OLA methods SOLA [4], PSOLA [5]) are widely used in the context of audio signal processing [6, 7].

The technique described in this paper is mostly related to the first class of methods: it relies on an Exponentially Damped Sinusoids (EDS) model and takes advantage of the signal decomposition into a deterministic part and a noise component. The EDS modelling allows an accurate representation of each signal frame in terms of the amplitudes, phases, damping factors and frequencies of the component sine waves while the modifications can be processed separately on both parts (deterministic and stochastic) of the signal decomposition.

This work follows earlier ones designed for estimating, tracking and modifying musical sounds [8, 9] and rely on the high resolution properties of the subspace analysis [10]. The main drawback of this approach is the computational cost of such algorithms but the newest versions have become adaptive and overcome the constraint of computing a Singular Value Decomposition at each time step [11].

The theoretical background presented in section 2 mainly recovers from [9], highlighting the relationship with the well-known quasi-stationary models [2] and [3] and the applicative context. Section 3 shows analysis results and the modification and synthesis techniques are discussed in section 4.

2. THEORETICAL BACKGROUND

2.1. Definitions and model

The discrete signal to be analysed and modified is assumed real valued and denoted $s(t)$. It is segmented in overlapping frames $x(t,u) = s(t + t_a(u))w_a(t)$ where $t_a(u)$ are the analysis marks indexed by the non-negative integer u and w_a is the analysis window assumed of finite length L_a. The time-instants $t_a(u)$ are usually regularly spaced, *i.e.* $t_a(u) = u\Delta_a,\ u \in \mathbb{N}$, where the interval Δ_a is a fixed increment such as $\Delta_a \leq L_a$. In this paper, w_a will always be the rectangular window of length L_a.

For each frame, an Exponentially Damped Sinusoids model is used:

$$x(t,u) = \sum_{k=1}^{M} (b_k z_k^t + b_k^* z_k^{*t}) \quad (1)$$

where $b_k = \frac{A_k}{2}\exp(j\phi_k)$ is the complex amplitude of the k^{th} component (A_k is the amplitude of the corresponding real component and ϕ_k its initial phase), and $z_k = \exp(-\alpha_k + j2\pi f_k)$ is its complex pole. α_k and f_k denote its damping factor and frequency, ranging in $\mathbb{R}$ for the first one and in $[0\ 1/2]$ for the other. These parameters are implicit functions of $t_a(u)$, not reported in (1) for lightening purposes. Like the well-known sinusoidal decompositions found in [2, 3], the parameters are assumed to vary slowly, *i.e.* they are considered constant over the window duration. Anyway, it should be noted that, in contrast, the definition of the EDS model includes a possible variation of the components envelopes leading to a more accurate representation for each frame, yielding to a residual of lower energy than the one obtained by an undamped modelling for the same order M and window length L_a.

2.2. Analysis stage

The estimation of the $4M$ parameters for each frame $x(t)$ (for simplicity the u-dependency is removed) is performed using a subspace-based technique for the frequencies f_k and damping factors α_k. The b_k are jointly estimated by a least squares method. The subspace analysis takes into account the particular mathematical structure of the model, leading to a high resolution (HR) estimation: when the signal is noise-free the limit for separating two close components is only restrained by the limited computational capacity.

2.2.1. HR method outlines

Using the $L_a = 2N - 1$ samples of the frame, the square Hankel structured data matrix H is defined as

$$\mathbf{H} = \begin{pmatrix} x(0) & x(1) & \dots & x(N-1) \\ x(1) & x(2) & \dots & x(N) \\ \vdots & \vdots & \ddots & \vdots \\ x(N-1) & x(N) & \dots & x(L_a-1) \end{pmatrix}. \quad (2)$$

Assuming $2M \leq N$, the symmetric real matrix $\mathbf{H}$ is rank-deficient of dimension $2M$ [9]. Its eigendecomposition yields

$$\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \quad (3)$$

where $\mathbf{U}$ is an $N \times 2M$ orthonormal real matrix. In presence of an additive white noise, $\mathbf{H}$ becomes full rank and the columns of $\mathbf{U}$ are defined as the $2M$ dominant eigenvectors, corresponding to the $2M$ eigenvalues of highest magnitude.

The signal poles $\{z_k, z_k^*\}_{k=1,\dots,M}$ are estimated by taking into account the rotational invariance property of the signal subspace, which is expressed in terms of a real $2M \times 2M$ matrix $\mathbf{\Phi}$ whose eigenvalues are the signal poles:

$$\mathbf{U}_\uparrow = \mathbf{U}_\downarrow \mathbf{\Phi} \quad (4)$$

where $\mathbf{U}_\uparrow$ (*resp.* $\mathbf{U}_\downarrow$) is obtained by deleting the first (*resp.* the last) row of $\mathbf{U}$.

2.2.2. Estimation of the complex amplitudes

This estimation of the complex amplitudes $\{b_k, b_k^*\}_{k=1,\dots,M}$ is performed for each frame using a Least Squares (LS) method. The $N \times 2M$ Vandermonde matrix $\mathbf{Z}$ is defined by

$$\mathbf{Z} = \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 \\ z_1 & \dots & z_M & z_1^* & \dots & z_M^* \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ z_1^{N-1} & \dots & z_M^{N-1} & z_1^{*N-1} & \dots & z_M^{*N-1} \end{pmatrix}. \quad (5)$$

When the frame is noise-free, the column vector $\mathbf{b} = [b_1, \dots, b_M, b_1^*, \dots, b_M^*]^T$ satisfies the relation

$$\mathbf{Z}\mathbf{b} = \mathbf{x} \quad (6)$$

where $\mathbf{x} = [x(0), \dots, x(N-1)]^T$. The solution of (6) in the least squares sense is $\mathbf{b} = \mathbf{Z}^+\mathbf{b}$ where $\mathbf{Z}^+$ denotes the pseudo-inverse of the matrix $\mathbf{Z}$.

2.2.3. Subspace tracking

Since the HR-method relies on the eigenvalue decomposition (EVD) of the data matrix $\mathbf{H}$, without further modification the estimation of the frequencies and damping factors would require an EVD at each time step and thus would lead to a high computational cost (*cf.* [9] for more details). However this cost can be reduced by using an adaptive algorithm which avoids the computation of the EVD. It is based on an iterative algorithm called *Orthogonal Iteration* [12] and uses a two steps procedure which yields the matrix $\mathbf{U}$ when the convergence is reached. For our tracking purpose, this algorithm is applied in a sequential way, assuming that it converges much faster than the variations of the signal subspace. As shown in table 1, it involves two auxiliary matrices $\mathbf{A}(u)$ and $\mathbf{R}(u)$ and an economy size QR factorization since $\mathbf{A}(u)$ is of dimension $N \times 2M$.

Table 1: Sequential iteration for subspace tracking

Initialization of the signal subspace matrix

$$\mathbf{U}(0) = \begin{pmatrix} \mathbf{I}_{2M} \\ \mathbf{0}_{(N-2M)\times 2M} \end{pmatrix}$$

For each analysis time-instant $t_a(u), u = 1, 2, \dots$ iterate:
matrix product $\mathbf{A}(u) = \mathbf{H}(u)\mathbf{U}(u-1)$
skinny QR factorization $\mathbf{A}(u) = \mathbf{U}(u)\mathbf{R}(u)$

3. ANALYSIS RESULTS: HR-OGRAM

In this section, the capabilities of the algorithm are demonstrated for both synthetic and real data (singing voice). A time-frequency representation is introduced, called the HR-ogram[1] which serves the same goals as the spectrogram in Fourier analysis. The HR-ogram represents the components as points $(t_a(u), f_k(u))$ in the time-frequency plane for each analysis time-instant $t_a(u)$. The energy ϵ_k of the k^{th} component is represented in decibels using gray levels and defined as

$$\epsilon_k = A_k^2 \frac{1 - \exp(-2\alpha_k L_a)}{1 - \exp(-2\alpha_k)}. \quad (7)$$

This definition includes the effect of the damping factor in the graph. This avoids the overvaluing of spurious poles, often related to noise, which can be highly damped and estimated at very high magnitude values, resulting in a weak component.

3.1. Simulation example

The graphs of figure 1 show the analysis results for a three components signal with an additive white noise corresponding to a 27dB signal to noise ratio (SNR). All the components are undamped.

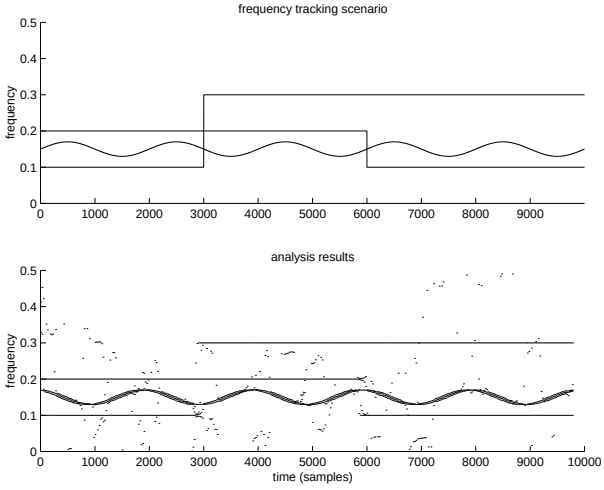


Figure 1: Analysis results for a synthetic signal

[1] standing for High Resolution Spectrogram

Two of them present a frequency jump at distinct time-instants while the third one is sinusoidally modulated at the period of 2000 samples and with a 0.05 frequency deviation.

The analysis parameters are set as follows: the window length is $L_a = 201$, the number of components is overestimated to $M = 6$ and the analysis is performed every $\Delta_a = 25$ samples.

The results show clearly the good tracking of the three components. The convergence of the algorithm is fast enough to handle the frequency jumps and when they occur for one component the estimation of the other ones remains stable and accurate. The modulated component is often represented by a few (2 or 3) poles. The signal model (1) is indeed not fully respected by the frequency modulated component. The spurious poles corresponding to the additive noise are easily identifiable for they do not aggregate in a specific shape and are widely spread in the whole frequency range.

3.2. Singing voice subspace tracking

This example is a female soprano singing voice, who realizes an up and down glissando between C5 and E5 . The signal is recorded at the 44100 Hz sampling rate. The analysis provides the HR-ogram shown in figure 2, which highlights the spectral lines associated to the harmonic structure but also the poles related to noise, gathered in formant-like shapes.

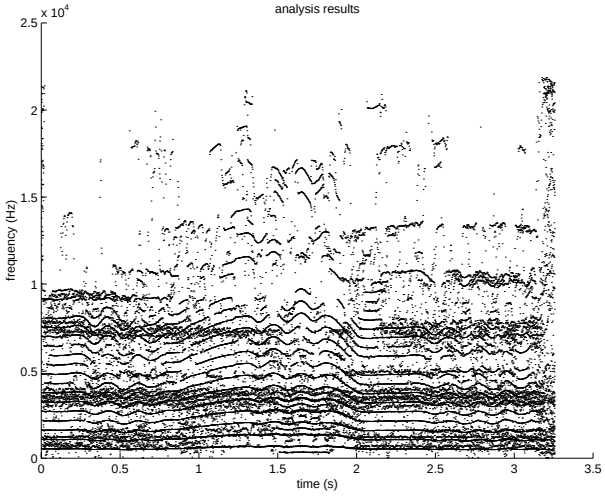


Figure 2: HR-ogram of a soprano singer

4. MODIFICATION AND SYNTHESIS

Each point of the HR-ogram is well localized both in frequency and time domains and can be modified individually. However, in order to achieve a high quality analysis/modification/synthesis the poles related to the sinusoidal components and those related to noise must be processed separately. Moreover, the EDS model does not represent accurately the stochastic part of the signal.

4.1. Deterministic/stochastic decomposition

4.1.1. Method

As in most of the analysis schemes, a noise component $w(t,u)$ is added to the model (1) leading to:

$$x(t,u) = \sum_{k=1}^{M} (b_k z_k^t + b_k^* z_k^{*t}) + w(t,u). \quad (8)$$

This component is often expressed as a time-varying filtering of a white stochastic process [6] and will be referred to as the stochastic component of the signal while the noise-free EDS model will be referred to as the deterministic component.

A common technique to derive the noise component consists in subtracting to $x(t,u)$ the deterministic part, after its estimation. But this can lead to a significative amount of sinusoidal components introduced in $w(t,u)$. In order to avoid this effect, $w(t,u)$ is obtained by projecting the signal onto the noise subspace. For the noisy model (8), the matrix $\mathbf{H}$ is full rank. The signal subspace is the space spanned by the eigenvectors associated to the $N-2M$ smallest eigenvalues.

The number M of sinusoids is chosen accordingly to the stability of the spectral lines of their associated poles. M is first overestimated and the signal poles $z_k(u)_{k=1,\ldots,M}$ and corresponding complex amplitudes $b_k(u)$ are estimated at the analysis time-instant $t_a(u)$. $z_m(u+1)$ and $b_m(u+1)$ are estimated at the time-instant $t_a(u+1)$ and distances are computed to measure how these poles are close in terms of frequency, amplitude and phase. The corresponding distances d_f, d_A and d_ϕ are thus defined as:

$$d_f(k,m) = 4(f_k(u) - f_m(u+1))^2 \quad (9)$$

$$d_A(k,m) = (A_k(u) - A_m(u+1))^2 \quad (10)$$

$$d_\phi(k,m) = \frac{(\lfloor \phi_k(u) + 2\pi f_k(u)\Delta_a \rfloor - \lfloor \phi_m(u+1) \rfloor)^2}{4\pi^2} \quad (11)$$

where $\lfloor \phi \rfloor$ denotes the principal determination of ϕ and the signal is normalized to a maximum magnitude equal to one. The spectral lines are formed according to the following steps:

1. for each k, $m_k = \arg_m \min d(k,m)$ is computed where $d(k,m) = d_A(k,m) + d_f(k,m) + d_\phi(k,m)$,
2. for each k, the poles $z_k(u)$ and $z_{m_k}(u+1)$ are connected if $|f_k(u) - f_{m_k}(u+1)|/f_k(u) < 1\%$ and $d_A < -20$dB and $d_\phi < -30$dB.

The number of components of the deterministic part $\hat{M} < M$ is defined as the number of poles which have been connected between $t_a(u)$ and $t_a(u+1)$. The stochastic component is obtained by the projection:

$$\mathbf{w} = (\mathbf{I}_{2\hat{M}} - \mathbf{U}_{2\hat{M}}^H \mathbf{U}_{2\hat{M}})\mathbf{x} \quad (12)$$

where $\mathbf{U}_{2\hat{M}}$ is the subspace matrix whose columns are the $2\hat{M}$ dominant eigenvectors, $\mathbf{w} = [w(0,u), \ldots, w(L_a-1,u)]^T$ and $\mathbf{x} = [x(0,u), \ldots, x(L_a-1,u)]^T$.

4.1.2. Results

Figure 3 shows the spectral lines tracked by the technique described in section 4.1.1, leading to a time-frequency representation of the deterministic component of the preceding singing voice signal for the time indices ranging between 1s and 1.25s (*cf.* the

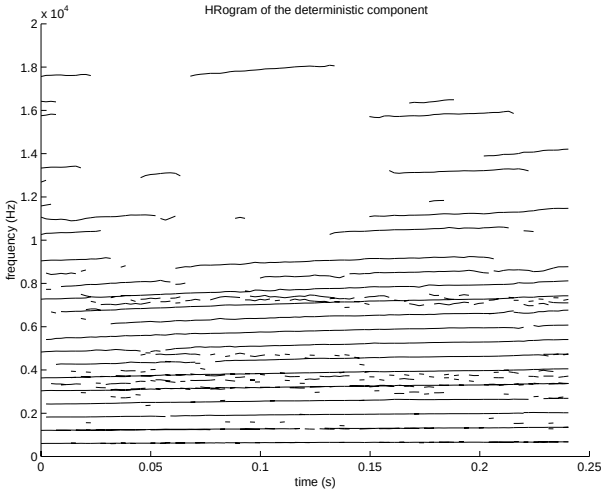


Figure 3: HR-ogram of the deterministic part

HR-ogram of figure 2). Besides a few lines related to noise components and easily identifiable because of their short time duration, the harmonic structure is highlighted and shows clearly the frequency modulation due to the *glissando* produced by the soprano singer. The short-term spectra (10ms) of the signal and its stochastic part are represented in figure 4. The sinusoidal part has been mostly removed by the projection while the formant-like reinforcements around 4000 Hz and 7500 Hz are emphasized.

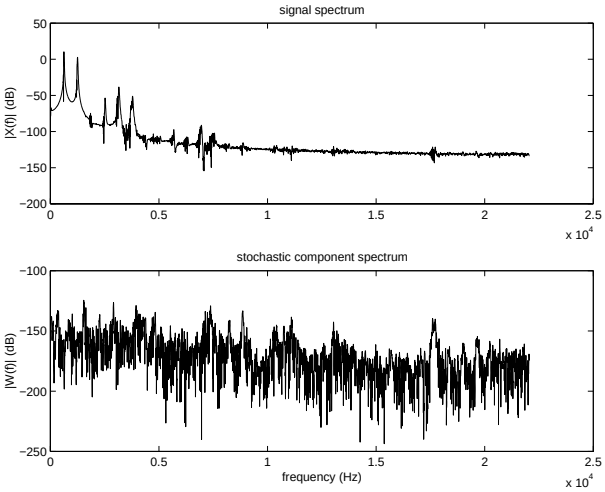


Figure 4: Short-term spectra of $x(t,u)$ and $w(t,u)$

4.2. Modifications

Coming along with the deterministic/stochastic decomposition, many audio effects can be processed. For example, pitch-shifting can be applied only on the deterministic part, eventually taking into account the spectral envelope. An interesting by-product of this decomposition is the capability of processing each spectral line to add or remove vibrato and tremolo, to adjust the pitch or the duration or to modify the ratio between the voiced and the unvoiced part of the sound.

5. CONCLUSIONS

In this paper an analysis/synthesis scheme has been proposed. It uses a High Resolution adaptive method which overcomes the Fourier resolution limit and achieves an accurate estimation of the sinusoidal components of the signal . The stochastic part is then derived by projecting the signal onto the noise subspace. The deterministic and stochastic parts can thus be processed separately, leading to high quality audio effects.

Acknowledgements. The authors are very grateful to Nathalie Henrich for having provided the musical signals.

6. REFERENCES

[1] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 3, pp. 243–248, Jun 1976.

[2] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 4, pp. 744–754, Aug 1986.

[3] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music J.*, vol. 14, no. 4, pp. 12–24, Winter 1990.

[4] S. Roucos and A. M. Wilgus, "High quality time-scale modification of speech," *Proc. IEEE ICASSP-85, Tampa*, pp. 493–496, Apr 1985.

[5] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5/6, pp. 453–467, Dec 1990.

[6] M. Kahrs and K. Brandenbourg, *Applications of Digital Signal Processing to Audio and Acoustics*. Dortrecht, Netherland: Kluwer Academic Press, 1998.

[7] E. Moulines and J. Laroche, "Non parametric techniques for pitch-scale and time-scale modification of speech." *Speech Communication*, vol. 16, pp. 175–205, Feb 1995.

[8] B. David, R. Badeau, and G. Richard, "Sintrack analysis for tracking components of musical signals," in *Forum Acusticum*, Sevilla, Spain, Sep 2002.

[9] R. Badeau, R. Boyer, and B. David, "EDS parametric modeling and tracking of audio signals," in *5th Int. Conf. on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, Sep 2002, pp. 139–144.

[10] Y. Hua and T. K. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-38, no. 5, pp. 814–824, May 1990.

[11] R. Badeau, G. Richard, and B. David, "Sliding window adaptive SVD algorithms," *IEEE Trans. Signal Processing*, to be published.

[12] G. Golub and C. V. Loan, *Matrix computations*, 3rd ed. Baltimore and London: Johns Hopkins University Press, 1996.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

DIDACTIC RESOURCES FOR THE TEACHING AND POPULARISATION OF THE PHYSICS OF THE MUSIC

Elejalde-García, M.J.; Macho-Stadler, E.; Franco-García, A.; J. Janariz-Larumbe

Dpto. Física Aplicada 1. Escuela Superior de Ingenieros.
Universidad del País Vasco
wupelgam@bi.ehu.es

ABSTRACT

The physical knowledge about the nature of the sound is shown desirable and necessary for students, players and the enthusiastic to the music in general. This extensive population implies the creation of didactic materials with easy access and coherent, simple and multiple forms of presentation of the following themes:

The physical and musical characteristics of the sound; the qualities of the musical groups and of the auditoria; the fundamentals and acoustical possibilities of the instruments; the peculiarities of the voice and the human ear; the relation among the music and the new technologies; the musical peculiarities of the folk tradition.

The didactic units include book, web page or CD formats and have served of support for university free-election subjects about musical physics. Students have mainly proceeded of the engineering field and other didactic resources as the laboratory, the videos and the tutorials have been utilized. The student's opinion about the methodology and materials has been in favor of the simplicity and scientific rigor without an extensive use of the mathematical language. This has encouraged us towards the same informational effort, directed to people that are not familiars with mathematics and physics.

1. INTRODUCTION

Music exits in all human societies and musical performance and appreciation are generally considered exquisite products of human culture. Over the past century, we have witnessed an explosion in research activities on music physical characterization, music perception and performance, music production and dissemination. The objective of this work is to contribute to the potential dynamism and richness of the multimedia in order to teach and make accessible the basic physical knowledge about the physics of music.

In this work we present some didactic materials about the physics of music, using multiple representations of physical processes. Some of the materials take advantage of the interactivity characteristic of the computers. The general topics are developed in a series of more specific ones that we list below. They contain verbal and mathematical descriptions, examples and or simulations, proposed exercises and references.

1.1. Physical and musical characteristics of the sound
Simple harmonic damped and forced oscillators.
Coupled oscillators: waves.
Longitudinal and transversal waves.
Wave superposition.
Acoustical phenomena.
Sound characteristics and musical sounds.

1.2. Qualities of the musical groups and of the auditoria
Orchestra.
Architectural acoustics.
Outdoor and indoor acoustics.
Reverberation time.
Room acoustical qualities.
Acoustical isolation and conditioning.

1.3. Fundamentals and acoustical possibilities of the instruments
String instruments.
Wind instruments.
Percussion instruments.
Electric and electronic instruments.

1.4. Peculiarities of the voice and the human ear
The human voice.
The human hearing system.
Intensity sensation.
Masking.
Directional characteristics.
Tone perception and combination.

1.5. Relation among the music and the new technologies
Music and new technologies.
Transducers.
Microphones and speakers.
Amplifiers.
Equalization.
Sound recording and reproducing systems.

1.6. Musical peculiarities of the folk tradition
Basque musical instruments.

2. MULTIPLE REPRESENTATIONS OF PHYSICAL PROCESSES

A physical process can be described using some different representations, as it can be seen in figure 1. The pictorial and physical representations are often called qualitative representations, in contrast to the quantitative mathematical representation. Experiments in cognitive science and physics education indicate that experts often apply qualitative representations (pictures, graphs, diagrams) to help themselves understand problems before they use equations to solve them quantitatively. Studies in physics education also have found that students problem-solving achievement improves when greater

emphasis is placed on qualitative representations of physical processes.

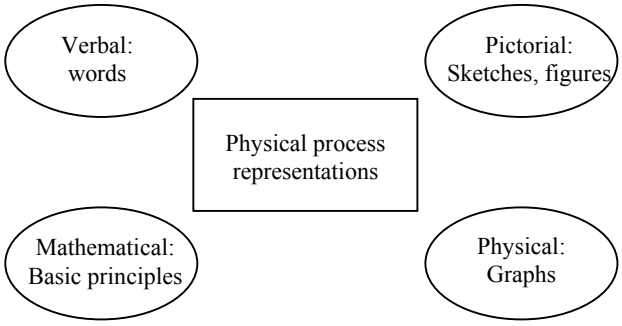


Figure 1: Different representations of a physical process.

Qualitative representations foster understanding of the process since they enhance human perceptual reasoning. Moreover, they build a bridge between the verbal and the mathematical representations: students move in smaller and easier steps from words to equations. As a consequence they help students develop images, which give the mathematical symbols meaning.

The use of verbal, pictorial, physical and mathematical representations is called multiple-representation [1].

In figure 2 we show a multiple representation scheme for the study of simple harmonic motion. In this case we can see examples of the four different representations of the same phenomenon. Verbal and pictorial representations provide a simple description of the system. The mathematical representation shows the second Newton's law for the oscillator, the value of x (mass displacement from equilibrium position) and v (mass velocity) as functions of time, and the values of kinetic, potential and total energies. The physical representation shows a graph that includes kinetic, potential and total energies of the oscillator as a function of its displacement from equilibrium position. The spring restoring force is also showed. This graph corresponds to an interactive simulation included in our on-line course [2].

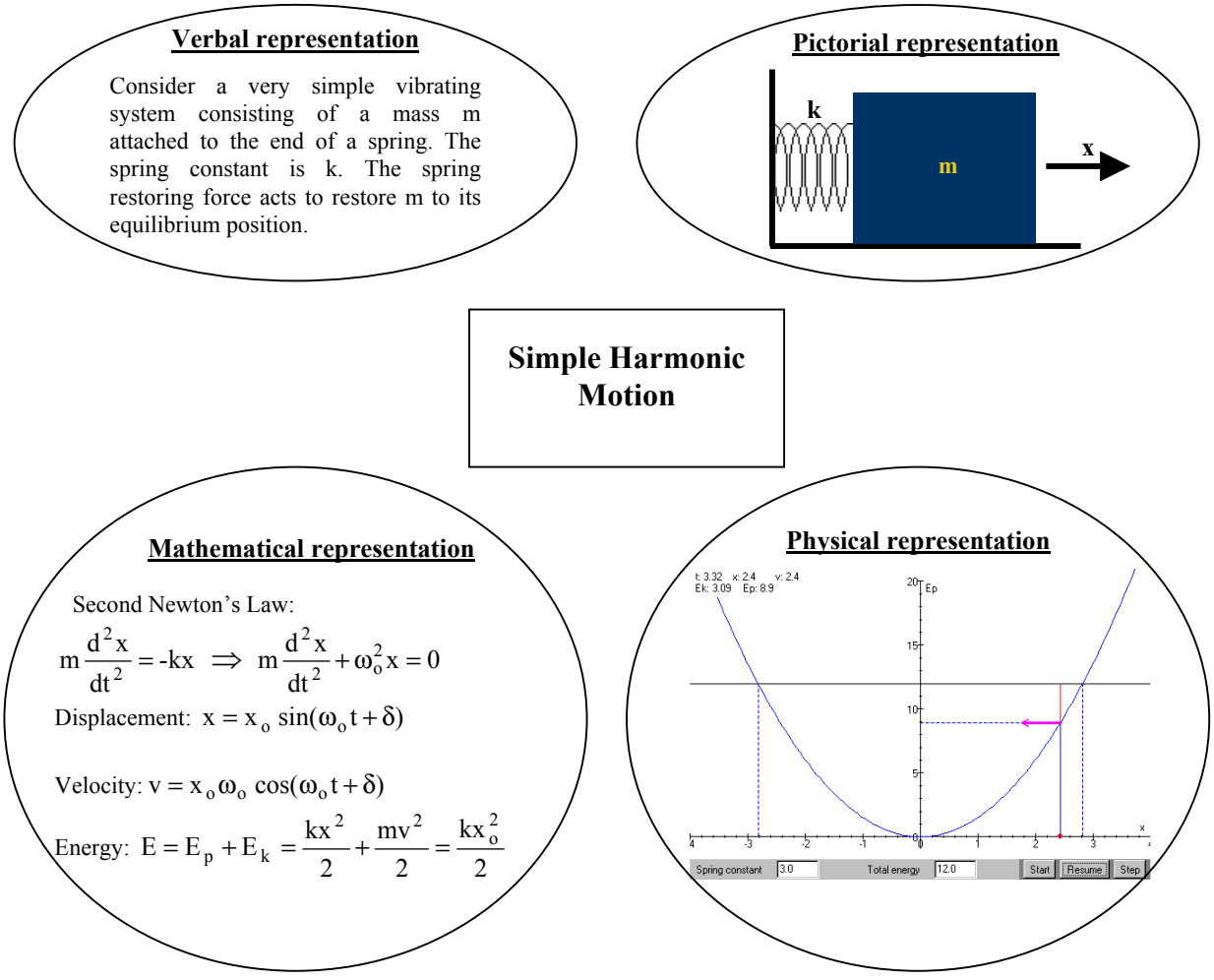


Figure 2: *An example of multiple-representation scheme for simple harmonic motion.*

Sometimes, we do not use some of the representations or we combine them in the same one. As an example, Figure 3 shows the multiple representation schemes for Diffraction. In this case, the mathematical description of the phenomenon is difficult and it is not included into the didactic materials. So, we emphasize on the pictorial and physical representations. Figure 3 shows a picture of the interactive simulation corresponding to the single slit diffraction. In this simulation we can choose the wavelength, the slit width and the number of sources.

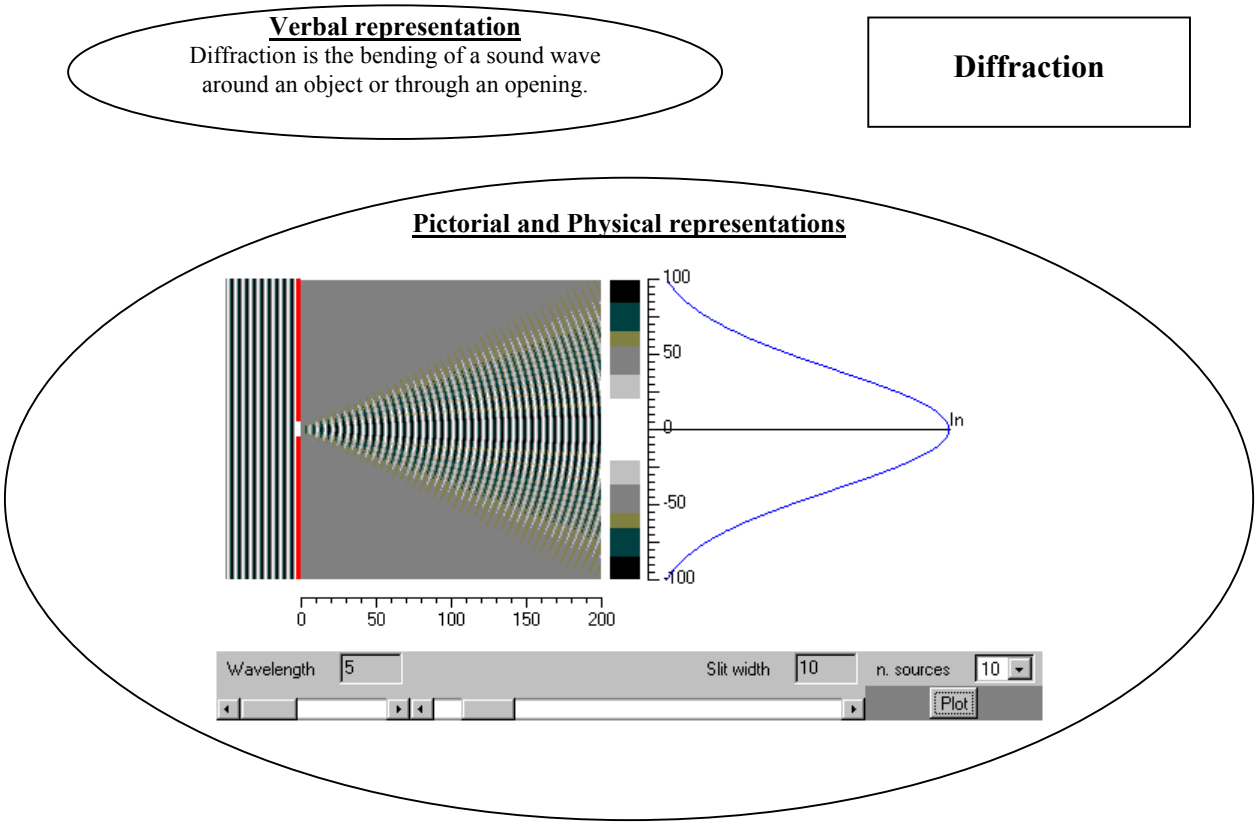


Figure 3: *An example of multiple-representation scheme for diffraction.*

3. INTERACTIVITY AND APPLETS

Interactivity is one of the most relevant characteristics of computers. By means of our on-line course we intend to take advantage of it. So, one of our tasks has been to produce coherent sets of applets, written in Java. They are simulations, computer experiments, interactive problems, etc., that are embedded into HTML documents.

With these applets we create a rich group of experiences in order that the students can get an intuition of the different physical situations programmed in the computer. By means of the interactive dialogue between the student and the program we encourage the student to be an active participant in her/his learning process, instead of a passive spectator.

The applets allow the observation of the system behavior and facilitate the step-by-step comprehension of the involved physics concepts. They are especially powerful for a better understanding of "hard" concepts, virtual labs and to learn scientific strategies for problems solving.

For example, in studying vibrations and waves, it is advisable to begin with the simplest type, a one-dimensional sinusoidal vibration that has only a single frequency component: the simple harmonic oscillator. However, when a real body is set into oscillation, dissipative forces arise. These forces will always result in a damping of the oscillations, a decrease in the amplitude of the free oscillations with time. When the oscillator is driven by an externally applied force it is called forced oscillator.

In our course, these three oscillators are studied in a similar way, and the applets produced for all of them allow the visualization of the differences between their motion, the displacement value versus time, the velocity value versus displacement and the time history of the total energy. Figures 4, 5 and 6 show an example of these applets for simple harmonic oscillator, damped oscillator and forced oscillator respectively.

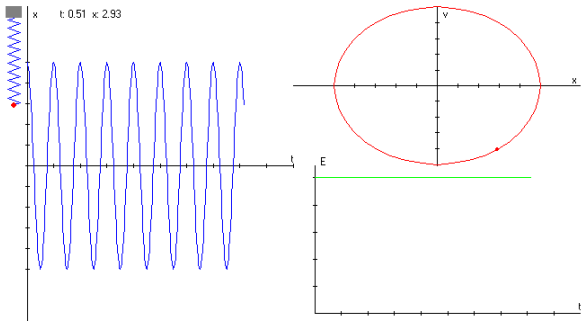


Figure 4: *Applet for simple harmonic oscillator.*

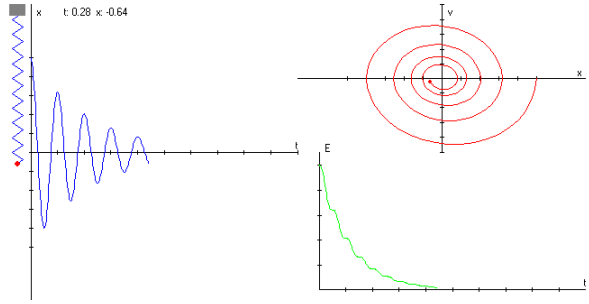


Figure 5: *Applet for damped oscillator.*

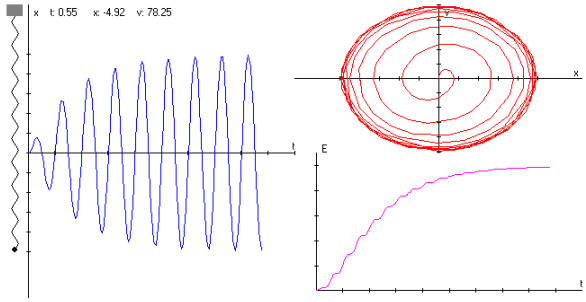


Figure 6: *Applet for forced oscillator.*

4. CONCLUSIONS

We have developed didactic materials related to Musical Acoustics. They are easy access and include book and web page or CD formats. At the same time, they are focused on simplicity, and also on scientific rigor and coherence without an extensive use of the mathematical language. In order to describe physical processes we use multiple forms of representation: verbal, mathematical, pictorial and physical.

The applets embedded into web pages with text, diagrams and formula around are learning objects that allow the observation of the system behavior and facilitate the comprehension of the involved physics concepts.

The didactic materials should be useful for students and teachers and also for other people that are not familiars with mathematics and physics, such as players and devotees of music.

5. REFERENCES

[1] Van Heuleven, A. and Zou, X., “Multiple representations of work-energy processes”, Am. J. Phys., 69(2):184-194, 2001.

[2] Elejalde-García, M. J., Franco-García, A., Janariz-Larumbe, J. and Macho-Stadler, E., “Curso de Acústica, (Acoustics Course)”, www.ehu.es/acustica, 2003.

6. ACKNOWLEDGEMENTS

This work has been partially supported by the project 1/UPV 00057.345-E-14783/2002 of the Universidad del País Vasco/ Euskal Herriko Unibertsitatea.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

QUANTITATIVE ASSESSMENT OF AIR FLOW FROM PROFESSIONAL BASS REFLEX SYSTEMS PORTS BY PARTICLE IMAGE VELOCIMETRY AND LASER DOPPLER ANEMOMETRY

E Esposito, M Marassi

Department of Mechanics
Polytechnic University of Marche – Ancona (I)
espo@mehp1.univpm.it

ABSTRACT

Particle Image Velocimetry is a well-known laser based measurement technique that un-intrusively determines the instantaneous full field velocities of a fluid flow. It uses two laser pulses to produce two images of a flow seeded by small particles; specialized software packages cross-correlate these images and process them to extract vector maps.

Also Laser Doppler Anemometry employs laser light to remotely determine fluid flow velocities, but to do so, it exploits laser beam Doppler scattering from seeding particles. In this work we will present a series of experiments conducted on different bass reflex ports designs to assess their performance in terms of flow turbulence and sound level compression at high input power levels. These issues are quite important in professional sound systems, where increasing power levels and sound clarity ask for exponentially growing cost and weight. For these reasons it is vital to optimize port design and our tests have been conducted on a well-known P.A. system normally used in medium sized venues and on a custom built similar box.

1. INTRODUCTION

Bass reflex or vented loudspeaker systems (BRS) are analogous to 24 dB/octave cutoff high pass filter and are generally well regarded when compared to closed box systems for a series of reasons [1], [2], [3]:

1. Greater maximum acoustic output (up to 5 dB for a programme signal)
2. Higher efficiency (max + 3 dB)
3. Lower cutoff frequency
4. Reduced distortion at lower frequencies

Low frequency added extension comes from exploiting the Helmholtz resonator created by the compliance of the air inside the enclosure and inertance of the air in the port. Acoustic transformer created this way has its own resonant frequency, Fb, at which most (or all, if there was no losses) of the system acoustic output comes from the port. A fairly obvious implication of this is significant velocity associated with the air flow through the port. The first patent describing BRS was granted to A.C. Thuras in 1932, followed by the seminal papers of Locanthi, Beranek, Van Leeuwen, De Boer, Lyon and Nowak in the 50's. The turnkey arrived with A. N. Thiele's paper of 1961 (re-published in 1971 [4]), and twelve years later contributions by R. Small [5]. We must also mention the contribution of J. E. Benson [6], who was the examiner of the Ph.D. thesis of R. Small. Thanks to the work of these researchers, modern design practice of BRS was born and brought to a more practical implementation including design tables by R. Bullock in 1981 [7]. Vented enclosures however present also two major drawbacks when compared to other types of loudspeaker systems:

1. Design procedure is more complicated than usual, and design errors, "misalignments", have serious consequences on sound quality.
2. Airflow in the port(s) may cause all sorts of non-musical noises to be generated by the port and also distortion and acoustic compression.

Thanks to the work of above mentioned researchers, point 1 has been resolved and nowadays there are many different software packages to help loudspeaker systems designers and also WEB calculators [8] or ready to print on-line guides [9]. Preceding studies [2], [3], [10], [11], [12], [13], have tackled point 2. but to the knowledge of the authors there not exists an accurate, non-intrusive experimental full field study of air flows coming out of reflex ports in operating conditions. This work will present a set of PIV and LDA measurements made on a sample professional BRS driven at different power levels. Measurements have been conducted using a last-generation digital PIV (DPIV) system that allows tracing vector and pictorial maps of all the flux parameters: stream lines, flux velocity, and vorticity. This system has also been compared with a custom built BRS equipped with standard cylindrical tubes instead of the almost rectangular ones used in the professional system.

2. PIV AND LDA THEORY

2.1. PIV basic theory

Particle Image Velocimetry (PIV) is a non-intrusive measurement technique used to simultaneously determine the velocities at many points in a fluid flow. The technique involves seeding the flow, illuminating the region under investigation and capturing two images of that region in rapid succession. From the displacement of the tracer particles, provided that the time interval between image captures is known, a velocity vector map can be calculated.

The technique of PIV can be considered as consisting of two stages: image acquisition and image evaluation (Figure 1).

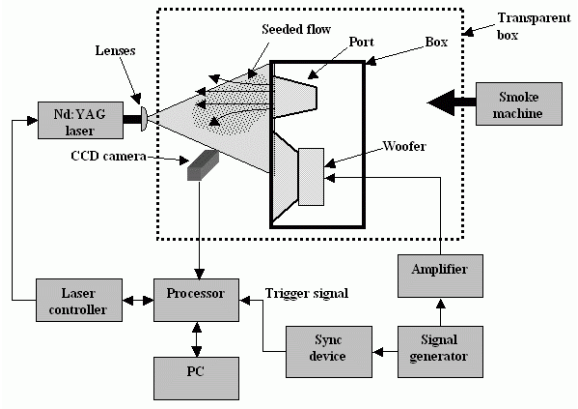


Figure 1: *A scheme of the employed PIV set-up.*

Image acquisition

Figure 1 shows a typical experimental arrangement for carrying out PIV measurements. Tracer particles added to the flow under investigation are illuminated by a laser sheet and images of the illuminated flow field are captured and stored for later analysis. The seeding particles must faithfully follow the fluid flow. In the case of turbulent or high speed gas flows, seeding particle diameters of about 1 μm or smaller are necessary [14]. As the contrast of the captured images is directly proportional to the scattered light power, the tracer particles must have strong scattering characteristics, and to efficiently illuminate the seeding particles a high intensity light source, normally a laser, must be used.

Capturing images can be done photographically or digitally. With recent improvements in spatial resolution and the numerous possibilities of data treatment, not forgetting the advantage of immediate image availability, the digital approach is becoming more and more used over photo-chemical processing. Indeed, Charge Couple Device (CCD) arrays are now widely used for PIV recordings.

Image evaluation

Since the introduction of the first PIV image evaluation methods, alternative analysis algorithms have been developed as well as error correction and post-processing procedures designed to improve speed and accuracy. However, the classical PIV analysis method is still the most frequently used and forms the basis of many other algorithms. The heart of the PIV analysis is the correlation of regions of the input images (known as "interrogation areas") with each other to determine the displacement vector of the flow in that part. Knowing the time interval between the recorded images enables a velocity vector to be calculated from the displacement vector. Using standard photographic equipment, the correlation technique is used for a single frame multiply-exposed (auto-correlation); resorting to a CCD camera, multiple frames singly-exposed (cross-correlation) is the common solution. After interrogating the images in this way and generating the vector map, post-processing is carried out to validate the data and to improve the vector map resolution and accuracy.

2.2. LDA basic theory

The LDA technique has been invented by Yeh and Cummins in 1964 and allows the determination of the point velocity of a seeded flow. It employs the Doppler effect, by which a laser illuminated particle that is moving will reflect light at a frequency linearly dependent on its velocity.

To obtain this, the laser beam is split into two parts and then these two rays are focused in a common point (the so called "measurement volume", see Figure 2) where the measured fluid will pass. Scattered light will impinge on a photo multiplier that will convert light intensity into a current signal that is then passed to a signal analyzer; knowing the laser beam wavelength (λ) and the angle of intersection of the two beams, we may write the conversion factor between the Doppler frequency shift (f_D) and the measured fluid velocity component (U_x).

$$f_D = \frac{2U_x}{\lambda}\sin\,\theta/2 \qquad (1)$$

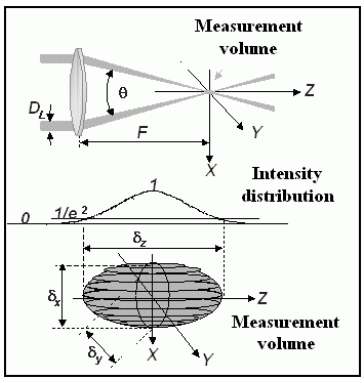


Figure 2: *Basic definitions of LDA.*

There also exists a different approach to explain LDA, based on the particle crossing the interference pattern created by the two intersecting laser beams, but we will skip it and address the reader to a well-known reference [15]. Finally, to measure not only velocity amplitude but also direction, a frequency shifter is inserted into the transmission optics, usually a Bragg cell.

3. EXPERIMENTAL SET UPS AND MEASUREMENTS RESULTS

3.1. PIV and LDA test benches

To perform our PIV and LDA measurement sessions, a particular ambient had to be prepared, due to the specific requirements of these techniques:

1. seeding had to remain inside the ambient for a time long enough to allow measurements to be done;
2. no external influence should disturb the flow;
3. flow should be optically accessible;
4. due to employed lasers high power, safety measures had to be implemented, especially to preserve operator eyes.

Following these indications, the structure of Figure 3 has been realized, consisting of an aluminum frame closed by Plexiglas and light plastic panels; LDA instrumentation was mounted on top of it, while the PIV system was positioned on tripods in front of it.

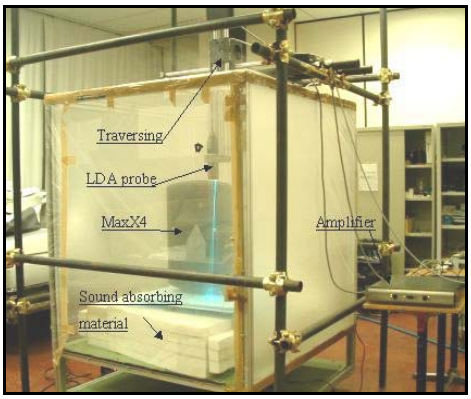


Figure 3: *LDA and PIV measurement set-up.*

In Figure 4 we have two photos of studied samples, namely a MaxX4 and our custom built reflex box, employing the same loudspeaker, and with the same volume (about 40 liters) and tuning frequency of the former system.



Figure 4: *FBT MaxX4, left, and custom built bass reflex system, right.*

Both boxes have been tested at ten power levels, starting from about 1.5 W up to power amplifier clipping, in the vicinity of 185 W. Excitation signal has been fixed to a 62.5 Hz sinusoid, coinciding with the BRS tuning frequency

3.2. Measurement results

In Figure 5 we show a cumulative graph of air velocity vs. input power for the MaxX4 at the center of the vent, considering four different time intervals during the 16 ms input signal period. Maximum value of flow velocity is 3 m/s, which is reached at 130 W: after this value, fluid flow saturates and we cannot observe any increase.

The spectrum of the air flow at 185 W is presented in Figure 6, and we may note a well behaved flux, with very little distortion.

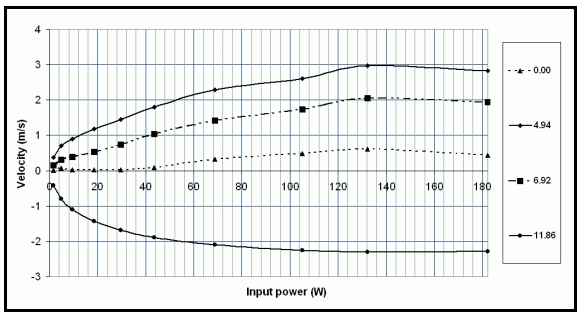


Figure 5: *LDA measurements of MaxX4 at different power levels.*

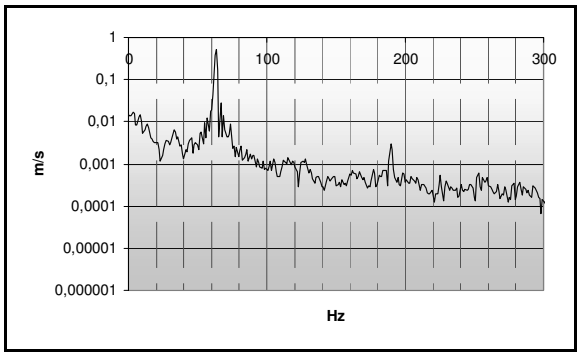


Figure 6: *LDA spectral analysis of MaxX4 at 185 W.*

Same analyses have been conducted for the custom system and results concerning velocity amplitudes are shown in Figure 7. In this case, maximum velocity value is higher, reaching more than 5 m/s, but it is rather evident a great distortion from the spectral analysis (not shown), and we may also observe that the flux saturates at just 25 W. Last two pictures show the results obtained by the PIV looking at four different planes of the vents, respectively for the MaxX4 and our system. In the latter case, the flow presents a greater vorticity and also it tends to become a jet, with an average positive velocity.

4. CONCLUSIONS

In this work we have presented an application of PIV and LDA to the measurements of airflow through the ports of bass reflex systems. We compared an industrial product with the cabinet produced by gas injection moulding, with a more traditional approach based on a wood cabinet and glued circular plastic tubes.

Although the latter system showed greater maximum velocities, its flux appears quite distorted, also in terms of spectral analysis. In both cases we found out limiting input powers that gave origin to flow saturation, and also in this respect the MaxX4 demonstrated a superior performance

reaching 130 W instead of the 20 W of the other box. We see that modern technologies offer superior performance but also require a careful design, because each mistake will cost a lot due to the necessity of modifying the cabinet mould. Our work will thus continue with the use of simulation software to predict flow behaviour prior to any prototype realization.

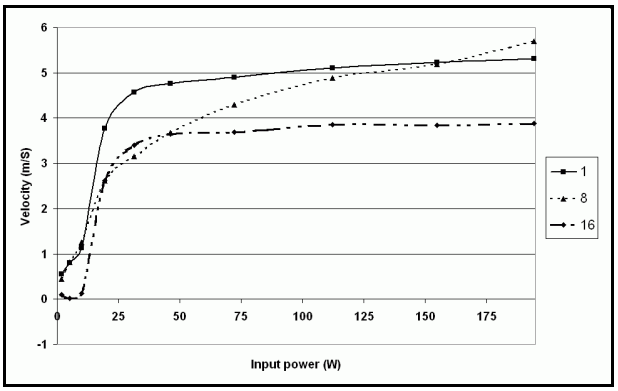


Figure 7: *LDA measurements of custom box at different power levels.*

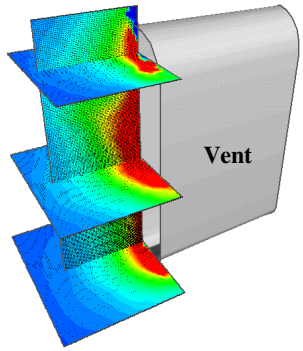


Figure 8: *PIV measurements of MaxX4 at185 W.*

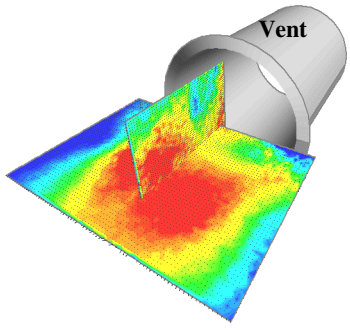


Figure 9: *PIV measurements of custom box at185 W.*

5. ACKNOWLEDGEMENTS

We wish to express all our gratitude to prof. Nicola Paone for his continuous support of this work, and to Massimiliano Rossi, who greatly contributed with his thesis to this paper. Also a big thanks to FBT Elettronica Spa for supplying the MaxX4.

6. REFERENCES

[1] Colloms, M., High Performance Loudspeakers, Pentech Press, London, 1991.

[2] Dickanson, V., The Loudspeaker Design Cookbook, Audio Amateur Press, Peterborough, NH, 1997.

[3] Raczynski, B., "How good is your port?", audioXpress, September 2001.

[4] Thiele, N., "Loudspeakers in Vented Boxes", Journal of the Acoustical Engineering Society (JAES), May-June 1971. Also available in Loudspeakers Vol.1, edited by Raymond E. Cooke, AES book store.

[5] Small, R., "Vented-Box Loudspeakers Systems", JAES, June to October issues, 1973. Also available in Loudspeakers Vol.1, edited by Raymond E. Cooke, AES book store.

[6] Benson, J.E., "The Theory and Design of Loudspeakers Enclosures", originally published by Amalgamated Wireless Australia Technical Review in three parts in 1968, 1971, and 1972. Published also as a book by Prompt Publications, Indianapolis, IN, 1996.

[7] Bullock III, R., "Thiele, Small and Vented Loudspeaker Design", Speaker Builder, 3/81.

[8] http://www.kbapps.com/audiodesign/calculators/

[9] www.ciare.it

[10] Backman, J., "The Nonlinear Behaviour of Reflex Ports", Proc. 98th AES Convention, Paris, France, preprint #3999, 1995.

[11] Pedersen, J. A., and Vanderkooy, J., " Near-Field Acoustic Measurements at High Amplitudes", Proc. 104th AES Convention, Amsterdam, NL, preprint #4683, 1998.

[12] Vanderkooy, J., "Nonlinearities in Loudspeaker Ports", Proc. 104th AES Convention, Amsterdam, NL, preprint #4748, 1998.

[13] Roozen, N.B., Veal, J.E.M., and Nieuwendijk, J.A.M, "Reduction of Bass-Reflex Port Nonlinearities by Optimizing the Port Geometry", Proc. 104th AES Convention, Amsterdam, NL, preprint #4661, 1998.

[14] Tonddast-Navaei, Ali, and Sharp, David B., "The Use Of Particle Image Velocimetry In The Measurement Of Sound Fields", Proc. International Symposium on Musical Acoustics, Perugia, Italy, pp.379-382, 2001.

[15] Durst F., Principles and Practice of Laser-Doppler Anemometry, Academic Press, 1981.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

COMPUTER-ANIMATED ILLUSTRATIONS OF VIBRATIONS AND WAVES

Donald E. Hall

Department of Physics and Astronomy
California State University, Sacramento

hallde@csus.edu

ABSTRACT

Students in Musical Acoustics courses often do not have much experience either with concepts in physics or with graphs as a representation of physical systems. We do well to provide as much as possible in the way of demonstrations to illustrate these concepts, and computer animations can be a valuable supplement to physical demonstrations. The author has written a set of eleven programs in QBasic 4.5 for the PC, which are complementary to a similar set done by Bruce Richards in C++ for Macintosh users.

The programs include depictions of basic physics (Doppler effect and normal modes), motion of plucked and bowed strings, motion of struck bars and drumheads, and air motion in various classes of pipes. Many of these are "movie versions" of illustrations in the author's textbook, and are available upon request.

1. INTRODUCTION

Most students in Musical Acoustics courses have minimal background in science and math. In particular, they often have had little experience either with concepts in physics or with graphs as a representation of physical systems. We like to introduce graphs to represent transverse wave motion, where the graph itself is a replica of the shape of the physical system But some of our physical concepts require graphs that are more than just literal pictures, with longitudinal displacement and velocity especially requiring a higher degree of abstraction. It is helpful if we can provide as much as possible in the way of demonstrations to illustrate these concepts, and computer animations can be a valuable supplement to physical demonstrations.

Under this same title, Bruce Richards presented at the 143rd meeting of the Acoustical Society of America [1] an admirable set of class demonstrations implemented on a Macintosh computer with C++ and the Mac Toolbox. This has inspired the author to prepare a similar package of animations written in BASIC for the PC. Several samples of these programs will be demonstrated during the session, and a brief summary of their coverage is given here. Many of them are "movie versions" of illustrations in the author's textbook [2]. Brief descriptions of each program follow, and the Appendix lists the connection of each program with the corresponding figure in that text.

2. BASIC PHYSICS

2.1. Doppler Effect

DOPPLER is the shortest and simplest of these programs. It shows a series of circular wavefronts spreading out from a moving source, each wave crest remaining centered at the point where it was emitted, and the resultant crowding or spreading of wavelength for observers on opposite sides.

2.2. Normal Modes

PENDMODE illustrates the motion of a pair of coupled pendulums, in order to illustrate what it means to "add" the simple motions of the two modes to obtain the more complex combined motion and its graphs. For additional examples, the initial positions can be varied, as can also the spring strength and consequent ratio of the two mode frequencies.

3. VIBRATION OF SOLID OBJECTS

3.1. Plucked String

STPLUCK shows the sequence of shapes in the idealized motion following the release of a plucked string, as well as the pair of infinite traveling waves whose sum reproduces those shapes. Addition of successive normal modes to produce increasingly better approximations to the same shape is also shown. A "second page" displays the motion again along with graphs of displacement and velocity as functions of time at some point along the string, as well as the waveform of the force upon the bridge. The plucking point and pickup locations can both be changed, and resulting pickup spectra correlated with predictions.

3.2. Bowed String

STBOW shows the sequence of shapes during idealized Helmholtz motion resulting from the bowing of a string, along with graphs of displacement and velocity at an observation point and force at the bridge, along with their spectra. The bowing location and observation point can both be changed at will, and the bowing direction can be reversed to observe how the stick-slip cycle is related to the direction of motion of the kink in the string shape.

3.3. Rectangular Bar

BARVIBR shows in succession the first eight normal modes of a simple thin bar in a one-dimensional approximation. For each mode, there is both the side view of the actual bent shape and the top view with color coding to represent the displacement, along with the frequency of each mode. Then the top view is used to show the idealized motion after the bar is struck at each of several points, and how it is approximated by the sum of those eight modes

3.4. Circular Membrane

DRUMHEAD shows in succession the first ten normal modes of a simple drumhead, both in a side view of a cut along a diameter and in the top view with color coding to represent the displacement, along with the frequency of each mode. Then the top view is used to show the motion after the drum is struck at each of several points, and how it is approximated by the sum of those ten modes. It is interesting (here as well as in BARVIBR and STPLUCK) to see how the addition of several standing waves can suggest a traveling-wave disturbance moving out from the excitation point and reflecting from the boundaries.

4. AIR MOTION IN PIPES

4.1. Infinite Cylindrical Pipe

PIPEINF shows sinusoidal waves in a pipe without the complication of boundaries. Along with a literal representation of moving air packets are six graphs. The first three give a moving picture of air velocity, displacement, and pressure as functions of position, while the last three give the same three variables at a single observation point as functions of time. The author's experience is that students need all the help they can get to see how these representations relate to each other (for instance, as in answer to Exercise 12.2 in Ref. 2).

The observation point can be moved, and the user can toggle back and forth between pure traveling and pure standing waves. A "second page" lets the pipe have one end so that the reflection of a single wave pulse can be watched. Graphs show the variables vs. time both at the reflecting end and at another observing point. Again the user can toggle between open and closed ends.

4.2. Finite Pipes

PIPEOPEN uses the same format as PIPEINF to show the first five normal modes of a cylindrical air column open at both ends. The observing point, initially in the mouth of the pipe, can be moved to any interior point at will. PIPECLSD does the same for a pipe closed at one end, and PIPECONE the same for a narrow conical pipe. These programs are all one-dimensional idealizations which do not attempt to deal with end corrections.

4.3. Pipe with Hole

TONEHOLE shows air motion, along with graphs of velocity, displacement, and pressure vs. position, for the first six modes of a cylindrical pipe open at both ends with a single tone hole located 1/5 of its length from one end. The hole diameter can be increased in steps from 0.1 to 1.1 times the pipe diameter, and the effect of its size upon both fundamental frequency and inharmonicity can be seen. There is opportunity to see that the hole affects some modes much more than others, and that the air motion immediately to one side of the hole may be in phase with that on the other side in some cases, but out of phase for other modes, as well as having quite different amplitudes.

5. CONCLUDING REMARKS

I would mention also two other (and much more complicated) programs that were written years ago in the course of research and have since been used with classes. FLEXSTRG models the striking process of a piano hammer hitting a perfectly flexible string, and STIFSTRG does the same for a string with some finite stiffness. Each of them has several input parameters to allow representation of different hammer and string properties. Copies of these programs, and the others described above, can be obtained upon request from the author at the email address given above.

It must be admitted that the use of QuickBasic 4.5 represents an earlier generation of programming formats, and was the path of least resistance for an author who has already used that language extensively. Knowing that some instructors would prefer Windows-based programs with more convenient parameter controls, the author would be happy to see someone rewrite them in a more advanced language that presents a prettier user interface.

6. APPENDIX

This list connects each simulation program with the number of the corresponding illustration in the textbook by Hall [2].

Program	Figure Numbers
DOPPLER	4.12
PENDMODE	9.6, 9.7
STPLUCK	10.4, 10.5
STBOW	11.7, 11.8, 11.9
BARVIBR	9.14
DRUMHEAD	9.16
PIPEINF	12.1
PIPEOPEN	12.2
PIPECLSD	12.3
PIPECONE	13.6
TONEHOLE	12.19

7. REFERENCES

[1] Richards, C.B., "Computer-animated illustrations of vibrations and waves", J. Acoust. Soc. Am., Vol. 111, 2002, p. 2394.

[2] Hall, D. E., Musical Acoustics, 3rd edition, Brooks/Cole Pub. Co., Pacific Grove, CA, 2002.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

MUSICAL ACOUSTICS COURSES IN THE EDUCATION OF MUSIC STUDENTS IN THE U.S.A.

Donald E. Hall

Department of Physics and Astronomy
California State University, Sacramento

hallde@csus.edu

ABSTRACT

Musical Acoustics courses in American universities and colleges, and the textbooks written for them, have been developed largely for the sake of General Education courses whose purpose is to provide token exposure to science for students majoring in other subjects. The author is not aware of any substantial previous data showing whether these courses play a significant role in the training of music majors. A survey questionnaire was sent to the heads of all music programs accredited with the National Association of Schools of Music. The results provide new information about the number of schools that have Musical Acoustics courses available, and the proportion in which this course has some integral role as a required or recommended part of the student's program.

1. INTRODUCTION

Education in Musical Acoustics in the United States was profoundly influenced around 30 to 40 years ago by two people. John Backus and Arthur Benade both carried out significant research that advanced our knowledge of the workings of musical instruments. Both also were very influential in awakening interest in the subject among other people, and in particular did so for this author. Backus wrote an excellent introductory textbook [1], still in print in its second edition over 30 years later, which encouraged many teachers to develop course offerings as well as inspiring several of us to write alternative textbooks for those courses. Benade was an inspirational guest lecturer in numerous meetings, as well as writing a book that is a gold mine for us as instructors though very challenging for students [2].

These influences worked together with growing enrollments and developing philosophies about general education to encourage the introduction of non-specialist courses in Musical Acoustics in many American colleges and universities in the 1970s and 1980s. In some few cases they were specifically aimed at serving Music students, but more commonly their existence was justified by their serving to fulfill "distribution" or "general education" requirements in the Science area for students whose own specialties were anywhere outside the sciences. But that function was always in some peril because of the interdisciplinary nature of the subject: since it was not a traditional and easily recognizable "Introduction to Physics", Musical Acoustics also sometimes lost its standing in the General Education program, which quickly translated into losing its enrollment and then its very existence. Another reason why Musical Acoustics courses have not only appeared but also disappeared is their dependence on the availability of instructors with the competence and confidence to cross the traditional boundaries between physics and music, not to mention physiology and psychology.

I would like to focus here on the more specific question of what role these courses have played in the training of Music majors. One possibility would be that Music schools or departments might regard such a course as useful enough to include among either required or recommended courses in the major. Another is that they might especially encourage their students to take advantage of Musical Acoustics as a way of fulfilling science requirements, even while not counting it as part of the major. Working against these possibilities would be two drawbacks: first, Music vies with Engineering for being one of the majors which is so packed with requirements already (history, literature, performance, ensembles, etc.) that those in charge would say "sorry, there's just no room to consider any more units of additional requirements." Second, there is the delicate question of whether the courses, when taught by physicists, will actually be perceived as helpful by the musicians anyway.

2. THE SURVEY

Following a suggestion by Professor Johan Sundberg, I thought it would be interesting to obtain some actual data about how many music programs (a) have a Musical Acoustics course available, and (b) either require or encourage their students to take that course. Through the kind assistance of Ernie Hills, chair of Music at my institution, I obtained the mailing list of the department heads in all those departments with membership in the National Association of Schools of Music. This included 596 departments in the U.S., nearly all of which are four-year colleges or universities (perhaps as many as half of which grant graduate degrees), together with a handful of conservatories, two-year colleges, and secondary schools. I sent a short questionnaire to all the schools on this list in late February, 2003.

The initial question asked whether the respondent's school either has (or had in the past) a Musical Acoustics course available to their students. If so, further inquiry was made as to which department is responsible for teaching the course, which students are its primary intended audience, and whether it is required, recommended, or purely elective for the Music students. Comments were solicited about reasons why the

course is or is not required, and in cases where a course has died as to why.

3. THE RESULTS

Replies were received from 225 schools. The return rate of over 37% is quite good for such a survey. There is roughly a two-to-one ratio between what I guess to be "small" and "large" Music programs among these responses, and fewer than ten that I would characterize as being among the most prestigious schools or conservatories.

117 responses (52%) indicated that there neither is nor has been a Musical Acoustics course; not surprisingly, the smaller schools are somewhat disproportionately represented in this group. Interestingly, for two major schools which gave this response, I know personally that there was a course in the past, so allowances must be made that institutional memories are not necessarily reliable. 20 of the 117 indicated that there is a significant block of material on Musical Acoustics included in some other course, such as Music Psychology or Music Technology.

Another 31 schools (14%) report that they do not presently have a Musical Acoustics course available but did in the past. The reason most often mentioned for the demise of the course was that the instructor retired, died, or left the institution; we may infer that this course is not routinely staffed as many others would be but is quite dependent on the presence of a particular instructor with strong interest in the subject. Many other reasons were mentioned, including difficulty in retaining approval for General Education credit, crowded program or shortage of staff, and difficulty in achieving agreement with the Physics department on suitable course content.

This leaves approximately one-third of the respondents indicating that they presently have a Musical Acoustics course available. Approximately 75% of these courses are offered through Physics departments, 15% through Music, and the rest either jointly or in a few cases "other". Not surprisingly, Music sponsorship is more common in larger schools (and of course in conservatories). The primary audience for the course is identified as Music students in a little over half the schools, and as General Education in most of the remainder. This is somewhat surprising to the author, who expected GE to dominate.

For undergraduate Music majors, responses were divided very nearly 1/3 each way as to whether the Acoustics course was required (at least for some of the majors), or recommended, or purely elective. (This 1/3 of 1/3 means that around 10% of schools have a Musical Acoustics requirement. But since that often applies only to some versions of the major, there are probably no more than 2 or 3% of all students who have such a requirement. Then again, there are probably several times that many who take the course for other reasons.) For graduate majors it is nearly always elective, presumably because they are expected to be more concerned with their own specialization at that level. Again in those schools which do have Musical Acoustics, approximately 20% of respondents indicated that some significant material on acoustics is included in some other course.

Only 39 respondents knew offhand which textbook was used, representing slightly over 1/3 of the schools which have or had a Musical Acoustics course. Interestingly, Backus [1] leads this list with 12 mentions, followed by Hall [3] with 11, Rossing [4] with 7, and Berg and Stork [5] with 5. But one would not want to draw any strong conclusions from these small numbers.

4. CONCLUDING REMARKS

Approximately 1/3 of American schools of music have a Musical Acoustics course available to their students. Allowing for the difference in enrollment levels, we might estimate that somewhere from 1/3 to ½ of the students have this opportunity. There is no clear consensus on whether this course should be an integral part of the major or only an elective. General Education requirements do provide an important motivation in many cases but by no means all.

Numerous comments confirm the idea that Music curricula are crowded with requirements, and that many more schools would like to include such a course among the requirements but feel that other pressures do not allow it. In a much smaller number of cases, individuals indicated that they personally had an Acoustics course and valued it highly but have been unable to convince their Music colleagues to agree.

Physics departments are providing a service in many of these schools which is valued by the musicians. At the same time, there are at least some places where the Physics offering is not perceived by the musicians as being sufficiently relevant to their interests. The situation is not without difficulty, because the interdisciplinary nature of the material demands a special interest in the subject, extra study, extra work, and willingness to reach out to the other side for help if the course is to reach its full potential. My own experience in doing so has been positive, and I would strongly encourage Physics chairs and professors to get together with their Music counterparts and talk about how they can work together.

5. REFERENCES

[1] Backus, J.G., The Acoustical Foundations of Music, W.W. Norton, New York, 1969.

[2] Benade, A.H., Fundamentals of Musical Acoustics, Oxford Univ. Press, New York, 1976.

[3] Hall, D.E., Musical Acoustics, 3rd Ed., Brooks/Cole Pub. Co., Pacific Grove CA, 2001.

[4] Rossing, T.D., Moore, F.R., and Wheeler, P.A., The Science of Sound, 3rd Ed., Prentice-Hall, New York, 2001.

[5] Berg, R.E. and Stork, D.G., The Physics of Sound, 2nd Ed., Prentice-Hall, New York, 1994.

Proceedings of the Stockholm Music Acoustics Conference, August 6–9, 2003 (SMAC 03), Stockholm, Sweden

UNUSUAL MOTIONS OF A NONLINEAR ASYMMETRICAL VIBRATING STRING

*Roger J. Hanson, H. Kent Macomber, and Andrew B. Morrison**

Department of Physics
University of Northern Iowa
Cedar Falls, IA, 50614-0150, USA
(* now at Department of Physics, Northern Illinois University, DeKalb, IL, 60115, USA)

roger.hanson@cfu.net, macomber@cfu.net, morrison@physics.niu.edu

ABSTRACT

A rich variety of unusual motions of a sinusoidally driven vibrating string results from nonlinearities due to varying tension and longitudinal motion not included in simple linear theory and to intrinsic asymmetries in the string. In a brass harpsichord string (wire), the asymmetries cause a small splitting of each natural frequency of free vibration into two closely spaced frequencies (relative separation of the order of 0.2% to 2%). The two frequency components are associated with transverse motion along two orthogonal "characteristic" wire axes. These asymmetric properties have a profound effect on the nonlinear behavior. Generation of higher harmonics and of motion perpendicular to the driving direction result in elaborate motions which can suddenly change from one state to another. Examples of such unusual motions of a point on the wire are exhibited on videotape at the poster presentation and as a limited series of still frames in this paper.

1. INTRODUCTION

Earlier work on the following nonlinear phenomena in a sinusoidally driven harpsichord string (wire) has been reported [1]: nonlinear generation of transverse motion perpendicular to the driving force, sudden amplitude jumps, and hysteresis effects. Theoretical and experimental work prior to 1994 was also summarized. The nonlinear effects are a result of tensional changes and longitudinal motions which are ignored in the simple linear theory. The inclusion of those factors yields coupled nonlinear differential equations for motion parallel to and perpendicular to the driving force. An extension of that work [2] shows that the situation is made much more complicated by the splitting of each natural frequency of free vibration into two closely spaced frequencies with relative separation $\Delta f/f$ of the order of 0.2% to 2%. The magnitude of the frequency splitting increases rather sharply as the tension decreases. Associated with each of these two frequencies is a "characteristic" transverse axis orthogonal to the other one. Some consequences include: (1) Dependence of nonlinear effects, including generation of motion perpendicular to the driving direction, on the relative directions of the driving force and the characteristic axes, and on the magnitude of the splitting. (2) *Linear* generation of motion perpendicular to the driving direction masquerading, under some conditions, as *nonlinear* generation. (3) Dependence of nonlinear generation of higher harmonics on the relative orientations of characteristic axes for the different natural modes (nearly harmonic). More complete details of all of these effects are being reported elsewhere [3].

2. EXPERIMENTAL APPARATUS

The experimental arrangement is similar to the earlier one. (See Figs. 1 and 2 of reference [1].) A 70-cm long brass harpsichord wire is clamped at its ends and driven electromagnetically by an AC current in the wire passing through the magnetic field of a small permanent magnet. The initially established equilibrium tension remains constant except for the effect of ambient temperature changes. The motions in two transverse directions of a very small segment of the wire are sensed with orthogonal optical-interruptor detectors [4], the voltage outputs of which are fed to a dual-channel FFT frequency analyzer and an X-Y oscilloscope.

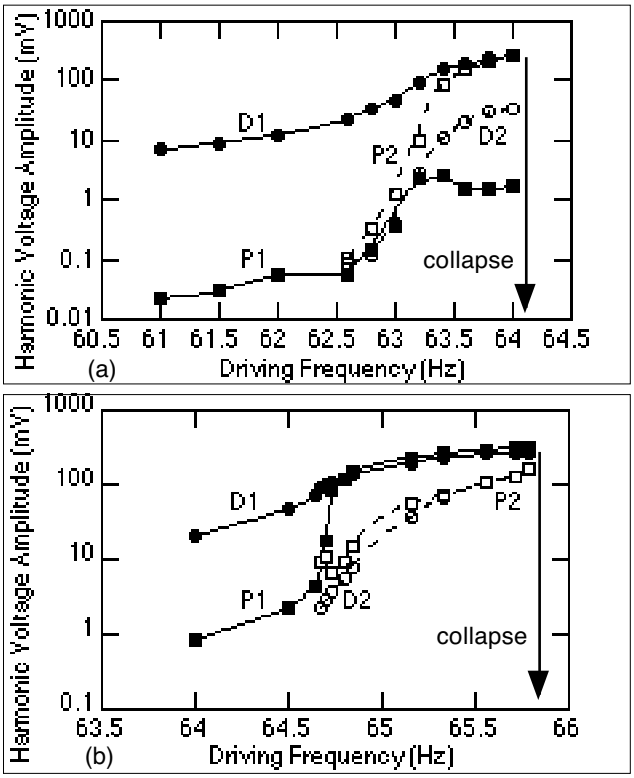
3. WIRE ASYMMETRY

Reference [2] discusses the existence of a "characteristic" axis for each of the two closely spaced frequencies resulting from wire asymmetry. (This asymmetry is probably due to the wire being differently stiff in different directions.) In free vibration, a detector aligned with one such axis yields only the frequency associated with that axis. General motion may be regarded as a superposition of such characteristic motions. The measured angle of $90^\circ \pm 0.2^\circ$ between the two axes is consistent with a simple model that predicts their orthogonality. An important consequence of the existence of frequency splitting and associated characteristic axes is that motion perpendicular to the driving direction can be

produced by purely *linear* processes. This is pronounced if the driving force is in a direction quite different from either of the two axes so that it has a significant component along *both* axes. The response of the wire along one of the characteristic axes then depends upon how closely the driving frequency matches the natural frequency for that axis. As the driving frequency is increased from one of the natural frequencies to the other, the direction of motion will change from that of the one characteristic axis to the other. The conditions for insignificant *nonlinear* generation of perpendicular motion are discussed later in this section.

For each of the three partials above the first (approximately the 2nd, 3rd, and 4th harmonics), the splitting bears no simple numerical relation to that for the first partial. As with the first, there are observed orthogonal characteristic axes associated with each closely spaced frequency pair for the higher partials. There is no apparent simple relationship of the orientation of the characteristic axes from one partial to the next. For example, it is observed that the axis for the *higher* of the two second-partial frequencies corresponds within about 10º to the axis for the *lower* of the two first-partial or fundamental frequencies. On the other hand, the axis for the *lower* of the two third-partial frequencies nearly coincides with the axis for the *lower* of the two fundamental frequencies.

The nonlinear effect of generation of motion perpendicular to the driving direction is most pronounced if the splitting is small, facilitating coupling between the two directions. Figs. 1(a, b) for driving along the high-frequency axis show that for a moderate first partial splitting of 0.16 Hz, the generation of perpendicular motion depends upon the length of time that the driving frequency is in the vicinity of the low-frequency resonance. Several prominent harmonics (at least six) are observed under these conditions, but only the first and second are exhibited in the figures. D1 denotes the 1st harmonic in the driving direction, P1 denotes the 1st harmonic in the direction perpendicular to the driving direction, and similarly for higher harmonics. In Fig. 1(a), where the driving frequency is increased in a normal way, there is just a hint of a peak in P1 in the low-frequency resonance region, perhaps from linear effects because the driving direction is not exactly perpendicular to the low-frequency axis. In Fig. 1(b), the frequency is increased very, very slowly in the vicinity of the low resonance (even sometimes adjusted back and forth to follow small resonance shifts due to ambient temperature changes), resulting in a very large P1 response. Once P1 is "nursed up" to a large value, the system is robust, and P1 is not very sensitive to further increases in driving frequency but remains large. It should be noted that the value of P2 has a sharp dependence on whether the P1 value is "nursed up" to its high value. This reciprocal relation between P1 and P2 was observed in many trials and is also very evident on the oscilloscope patterns of the motion of a point on the wire. It should also be noted that the response curves are not symmetrical resonant peaks for D1 because as the wire amplitude increases to the order of 2 mm, the increased average tension increases the resonant frequency so that the resonance peak is "chased" but never reached before collapse occurs.



Figures 1(a,b): *Wire Responses Under Different Conditions of Frequency Change*

Driving is approximately along the high-frequency characteristic axis and approximately perpendicular to the low-frequency axis. In (a), the frequency is increased normally slowly, whereas in (b), it is increased very, very slowly and thus held nearly constant in the region of the low-frequency resonance where P1 increases sharply.

The example illustrated in Figs. 1(a,b) indicates two stable modes of vibration with the same driving frequency, depending upon the history of frequency changes. Similarly, there are situations where there are seemingly "spontaneous" transitions from one mode to another, sometimes with very different amplitudes of a number of harmonics in different directions. In the videotape at the poster presentation and also at http://www.physics.uni.edu/Hanson.html there are shown several examples of these apparently "spontaneous"

transitions, resulting in very unusual patterns of the motion of a point on the wire. Frames taken from those tapes are exhibited in the next section.

4. UNUSUAL MOTIONS

In Figs. 2 and 3 some dramatic transitions are shown in the series of frames grabbed from video images of the trace on an X-Y oscilloscope to which the signals from the two optical detectors are fed. The trace shows the path of the transverse motion of a point on the wire at which the orthogonal optical detectors are sensing motion. Both series in Figs. 2 and 3 are obtained during pauses in the slow increase of the driving frequency through the resonance regions. The driving force is essentially in the high-frequency characteristic axis direction.

The transition illustrated in Fig. 2 occurs in the frequency region a little above the low resonance where a rather large P1 decreases suddenly, accompanied by significant increases in the higher harmonics. The driving frequency is held constant during these sharp changes. The transition in Fig. 3 occurs at a driving frequency which had been increased about 0.5 Hz above that in Fig. 2. The change in pattern from the end of the series in Fig. 2 to the beginning of the series in Fig. 3 occurs gradually and continuously during the slow change of driving frequency. The transition in Fig. 3 occurs during no change in driving frequency. At this point in the response curve, the large D2 and P2 drop almost a factor of ten, accompanied by a sharp drop in D4 while P4 remains constant and very significant as is easily seen in the pattern.

The series shown in Fig. 4 appears chaotic but is actually periodic with a period of about nine seconds. The driving frequency is unchanged during the exhibited sequence, and the patterns actually continue for many minutes with no obvious change in periodicity.

In Fig. 5, the driving frequency is that of the third partial and exhibited is a rather sudden transition (while the driving frequency remains constant) to a mode where there are strong peaks at all of the harmonics of *1/3 of the driving frequency*. Similar behavior is observed when driving at a frequency close to the second partial.



Figure 2: *Sudden Transition of the Pattern of the Path of a Point on the Wire*

The wire is being driven along the high-frequency characteristic axis. The driving frequency is held constant at about the low-frequency resonance resulting in dramatic pattern changes.

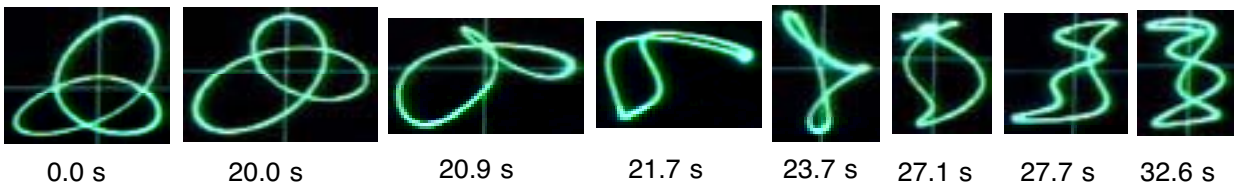


Figure 3: *A Second Transition at a Higher Driving Frequency than in Fig. 2*

Shown is a transition under the same conditions as in Fig. 2 except that the driving frequency is approximately 0.5 Hz higher than there, but held constant during the sequence shown.

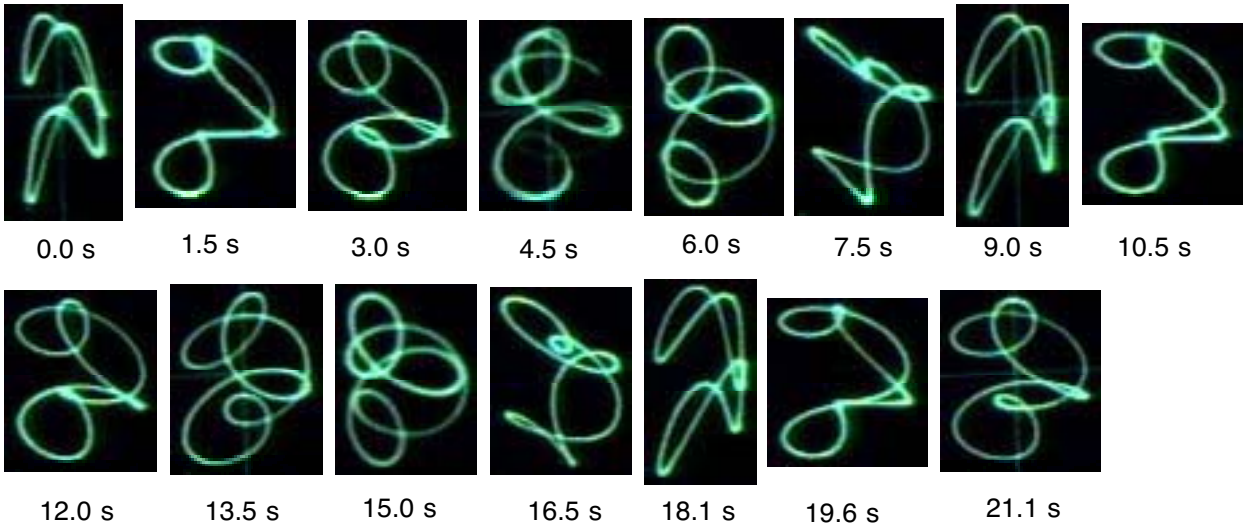


Figure 4: *Changing Pattern of the Path of a Point on the Wire with a Pattern Periodicity*

The wire is being driven in the direction of the high-frequency axis near the resonant frequency. The pattern, which appears very unstable, repeats itself with a period of about nine seconds.

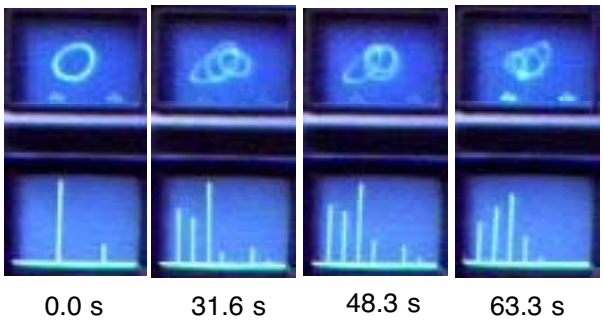


Figure 5: *Generation of Subharmonics*

The wire is driven at the third partial indicated by the prominent line in the first FFT spectrum. As the driving frequency is increased slightly, harmonics of 1/3 of the driving frequency are suddenly generated.

5. ACKNOWLEDGMENTS

Undergraduate students Mike Kassakatis, Matt Boucher, Cary Pint, and Cheryl Claassen provided invaluable assistance in data acquisition and analysis. Physics Technician Larry Dirkes constructed critical equipment. Multimedia Production Specialist Charles Sengstock digitized and captured still frames from video tape. The Graduate College provided essential financial support for student assistance.

6. REFERENCES

[1] Hanson, R.J., Anderson, J.M., and Macomber, H.K., "Measurements of nonlinear effects in a driven vibrating wire," J. Acoust. Soc. Amer., 96, pp. 1549-1556, 1994.

[2] Hanson, R.J., Macomber, H.K., and Morrison, A.B., "Nonlinear effects in an asymmetrical vibrating string," Proc. of the International Symposium on Musical Acoustics (ISMA 01, Perugia, Italy), pp. 157-161. (Copies are available from the authors.)

[3] A manuscript is being submitted for consideration for publication in J. Acoust. Soc. Amer. Preprints will be available at the poster presentation.

[4] Two GE Model H21B1 Optointerruptors cut in half and mounted in an orthogonal arrangement.

[5] Hanson, R.J., Macomber, H.K., and Kassakatis, M.P., "Nonlinear effects in a vibrating string (wire)," Proc. International Congress on Acoustics and the Acoustical Society of America (ICA/ASA '98), Seattle, WA, USA, pp. 1107-1108, 1998. (Copies are available from the authors.)

REDUCING THE SOURCE TUBE TO IMPROVE THE BANDWIDTH OF ACOUSTIC PULSE REFLECTOMETRY

Aijun Li and David B. Sharp

Department of Environmental and Mechanical Engineering
Open University, Milton Keynes MK7 6AA, UK
A.li@open.ac.uk, D.Sharp@open.ac.uk

ABSTRACT

Acoustic pulse reflectometry has become established as a useful non-invasive method of measuring the input impulse response, input impedance and internal dimensions of ducts. The technique involves injecting a sound pulse via a source tube into the duct under test. The reflections returning from the duct are recorded by a microphone embedded in the source tube wall and then analysed. The source tube is necessary to separate the forward and backward travelling waves. Although this provides an effective means of isolating the input pulse and the duct reflections, the attenuation of the signal while travelling in the source tube can be large, resulting in poor bandwidth input impulse response measurements. In this paper, a reflectometer with a shorter source tube is presented. The reduction in source tube length ensures less attenuation of the measured signals, leading to an improvement in the bandwidth of the experimentally determined input impulse response and more accurate bore profile reconstruction.

1. INTRODUCTION

Acoustic pulse reflectometry is a useful technique for measuring the input impulse response of tubular objects, from which their internal dimensions can be calculated. Applications include the monitoring of human and animal airways, quality control in musical wind instrument manufacture, non-invasive low-contact analysis of historical instruments, and the detection of leaks in pipes [1, 2, 3,4].

The accuracy of bore profile reconstructions made using the reflectometry technique depends on the bandwidth of the signal used to probe the duct under investigation. The greater the high frequency content of the input signal, the higher the resolution of the reconstruction. In this paper, modifications to the basic acoustic pulse reflectometry technique are described which result in an improved bandwidth input pulse, leading to more accurate bore reconstructions.

2. BASIC REFLECTOMETRY TECHNIQUE

A typical pulse reflectometer comprises a loudspeaker, a coil of copper tubing (referred to as the source tube) and a microphone which is embedded in the wall of the copper tubing partway along its length. A PC with data acquisition facilities is used to send signals to, and receive signals from, the reflectometer.

2.1. Experimental Procedure

Figure 1 shows a schematic diagram of a typical reflectometer. The duct under investigation is coupled to the far end of the source tube by a 0.5 m long cylindrical connector of the same radius as the source tube. The loudspeaker produces a sound pulse which travels through the source tube and cylindrical connector into the duct under investigation. The reflections from the duct are captured by the microphone, amplified and stored on the PC.

The source tube is necessary to separate forward and backward travelling signals. The section l_2 = 3 m ensures that the input pulse has fully passed the microphone before the first of the returning duct reflections reaches it. The section l_1 = 7 m ensures that the duct reflections are separated from any further reflections from the loudspeaker. Once the duct reflections reach the microphone, they can be recorded for up to $2l_1/c$ seconds (the time taken to travel the distance from the microphone to the loudspeaker and back, where c is the speed of sound) before the loudspeaker reflections return and contaminate the signal.

By deconvolving the duct reflections with the input pulse shape, the input impulse response of the duct (or, more precisely, the input impulse response of the cylindrical connector and the duct) can be determined. The input pulse shape is measured by rigidly terminating the source tube and recording the reflected pulse. This ensures that both the duct reflections and the input pulse have travelled the same path in the source tube and have therefore experienced the same source tube losses.

There is generally a DC offset present in the experimental input impulse response. However, it can be evaluated and removed. As there should be no signal reflected from the 0.5 m long cylindrical connector (as it contains no expansions or contractions) the first few milliseconds of the input impulse

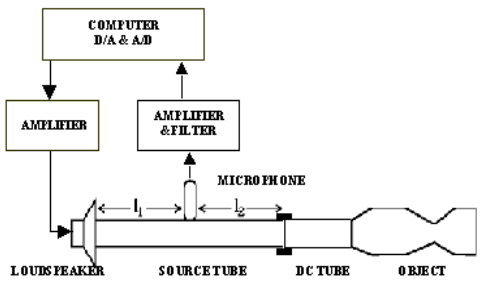


Figure 1*: Schematic diagram of reflectometer.*

response should be zero. The average value of the measured input impulse response over this range thus gives the value of the experimental DC offset. For this reason, the cylindrical connector is often referred to as a DC tube [5][6].

Once the DC offset has been removed from the input impulse response, application of a lossy layer-peeling algorithm enables the duct profile to be reconstructed [7].

2.2. Results

Using a reflectometer of the design described in Section 2.1, measurements were made of a stepped tube comprising two 0.15 m long cylindrical sections of radii 3.55 mm and 2.25 mm respectively. Figure 2 shows the input impulse response spectrum of the combination of the DC tube and the stepped tube. Above 7.5 kHz the impulse response becomes dominated by noise. This has implications for the bore reconstruction of the stepped tube (Figure 3). Although the calculated radii of the tube sections are in good agreement with direct measurement, the change in radius from 3.55 mm to 2.25 mm at the step is reconstructed as occurring over an axial distance of approximately 2 cm. (The first 50 cm of the bore reconstruction shows the DC tube.)

3. SHORT SOURCE TUBE

The design of reflectometer described in Section 2 includes a source tube to separate the forward and backward travelling signals. Although this provides an effective means of isolating the input pulse and the duct reflections, the attenuation of the signals while travelling in the source tube can be large, resulting in poor bandwidth impulse response measurements. For a duct with rapid expansions and contractions this can lead to significant errors in the bore reconstruction.

In this section, a reflectometer with a shorter source tube is presented. The reduction in source tube length ensures less attenuation of the measured signals, leading to an improvement in the bandwidth of the experimentally determined input impulse response and more accurate bore reconstruction. However, the shorter source tube also means that a new way of separating the input pulse and duct reflections is required. An alternative calibration method designed to achieve this separation is described.

3.1. Experimental procedure

Figure 4 shows a schematic diagram of the new design of pulse reflectometer incorporating a shorter source tube in which l_2 has been reduced to 0.5 m. However, l_1 is left unchanged at 7 m so the time over which the duct reflections can be sampled, and hence the maximum length of duct that can be measured, is the same as for the previous reflectometer. When using this new reflectometer, the duct under investigation is coupled directly to the source tube (a DC tube is not used).

As before, the loudspeaker produces a sound pulse which travels through the source tube into the duct. The returning duct reflections are captured by the microphone, amplified and stored on a PC. The shortening of the section l_2 results in a decrease in signal attenuation and consequently an increase in the bandwidth of the input pulse and duct reflections. However, it also means

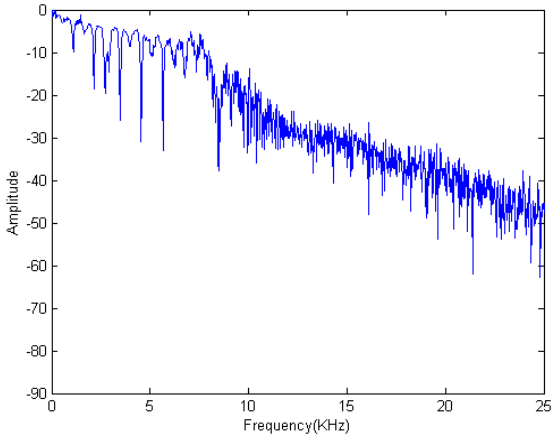


Figure 2: *Input impulse response spectrum of DC tube and stepped tube.*

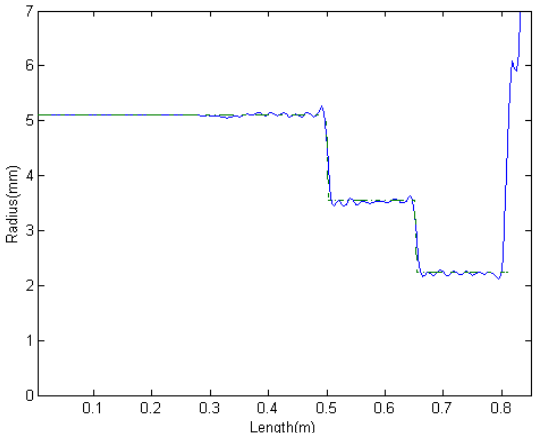


Figure 3: *Bore reconstruction of stepped tube.*

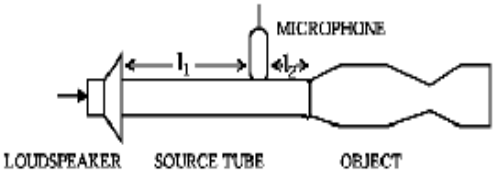


Figure 4: *Reflectometer with shorter source tube.*

that the forward travelling pulse does not completely pass the microphone before the first reflection from the duct reaches it. Figure 5 shows the signal captured by the microphone when the stepped tube described previously is measured. The forward travelling pulse and backward travelling duct reflections overlap.

To separate the duct reflections from the forward travelling pulse, a calibration measurement is performed. This involves anechoically terminating the source tube using a quasi-infinite tube of the same radius as the source tube. The forward travelling pulse can then be measured without the presence of any reflections (Figure 6).

Subtraction of the forward travelling pulse (Figure 6) from the mixture of the input pulse and duct reflections recorded by the microphone (Figure 5) separates out the duct reflections (Figure 7).

By deconvolving the separated duct reflections shown in Figure 7 with the forward travelling input pulse of Figure 6, the input impulse response of the duct (or, more precisely, of the section l_2 of source tube and the duct) is obtained. Again, the input impulse response generally contains a DC offset. The DC offset is removed in a similar manner to that described in Section 2. However, in this case the source tube section l_2 acts as the DC tube. After removal of the DC offset, a lossy layer-peeling algorithm can be applied, as before, to determine the bore profile of the duct.

3.2. Results

Figure 8 shows the input impulse response spectrum of the combination of the length l_2 of the source tube and the stepped tube, calculated from the data shown in figures 6 and 7. By shortening the source tube, the bandwidth of the impulse response has been extended from 7.5 kHz to approximately 15 kHz. Figure 9 shows the resulting bore reconstruction. The increased high frequency content has led to better reconstruction of the step. The change in radius from 3.55 mm to 2.25 mm is now reconstructed as occurring over an axial distance of 10 mm, half that of Figure 3.

4. SUMMARY

The shortening of the source tube and the subsequent modifications to the acoustic pulse reflectometry technique outlined in this paper have resulted in a one hundred percent increase in the bandwidth of experimental input impulse response measurements. This in turn has led to bore profile reconstructions with double the axial resolution.

These improvements to the acoustic pulse reflectometry technique will enable musical instruments to be measured with greater accuracy. In particular, it will be possible to identify the positions of flaws and defects, such as stray solder and developing cracks, more precisely. Similarly, when using reflectometry to detect leaks in pipelines, it will be possible to pinpoint the locations of any such holes more exactly.

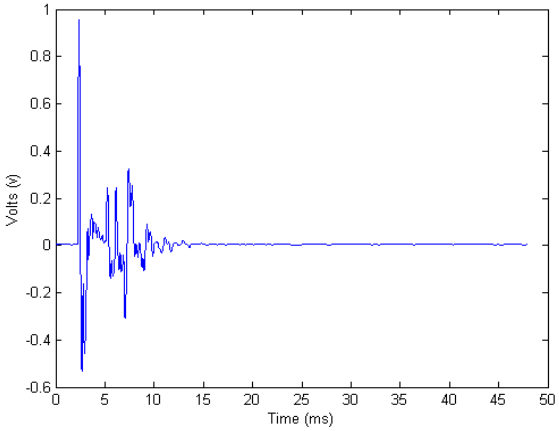


Figure 5: *Overlapping forward travelling pulse and duct reflections.*

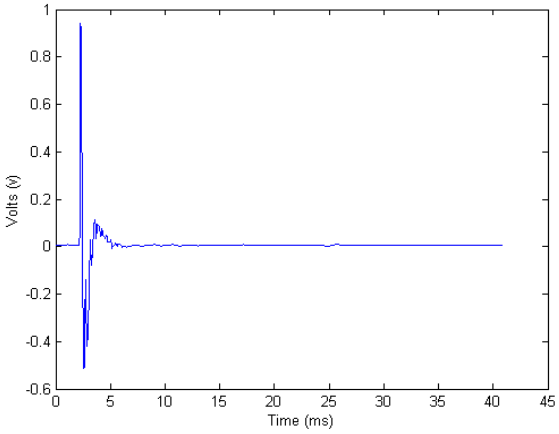


Figure 6: *Isolated forward travelling pulse.*

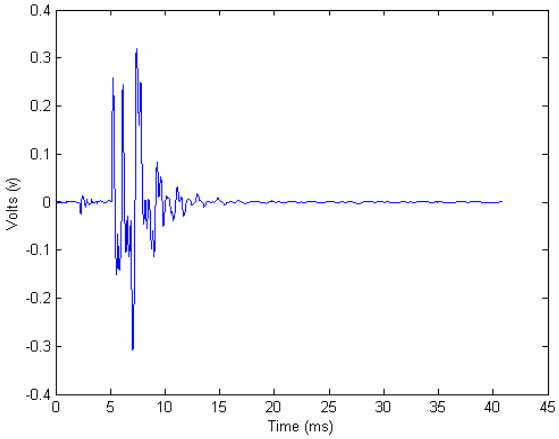


Figure 7: *Separated reflections from stepped tube.*

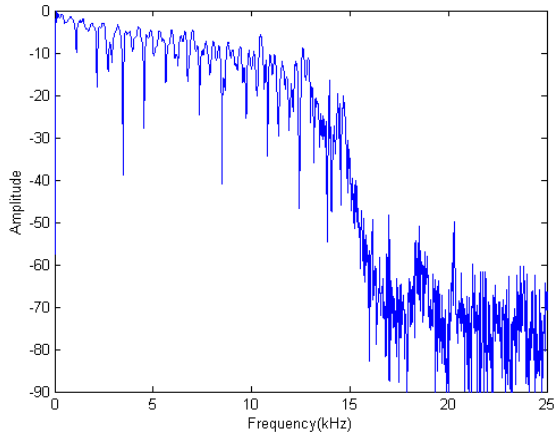


Figure 8: *Input impulse response spectrum of source tube section l_2 and stepped tube.*

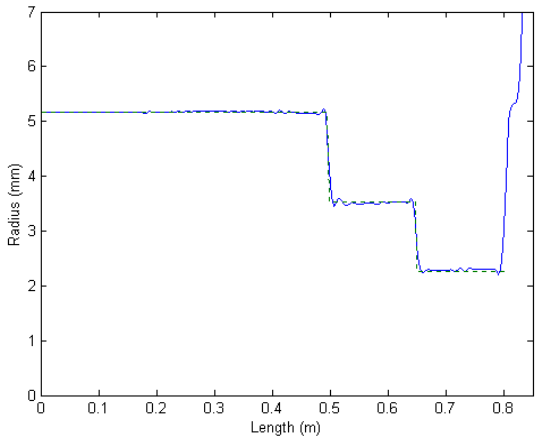


Figure 9: *Bore reconstruction of stepped tube using short source tube reflectometer*

4. ACKNOWLEDGEMENTS

This research was founded by Engineering and Physical Sciences Research Council under grant GR/N32471

5. REFERENCES

[1] Sharp, D.B., and Campbell, D.M., " Leak Detection in Pipe Using Acoustic Pulse Reflectometry" Acoustica, Vol. 83, 1997, pp. 560-566.

[2] Sharp, D.B., Myers, A., Parks, R., and Campbell, D.M., " Bore Reconstruction by Pulse Reflectometry and Its Potential for the Taxonomy of Brass Instruments", Proc. 15th International Congress on Acoustics, Trondheim, Norway, pp. 481-484, 1995

[3] Marshall, I., Rogers, M., And Drummond, G., " Acoustic Reflectometry for Airway Measurement. Principles, Limitations and Previous Work". Clin. Phys Physiol Meas. Vol. 12(2), pp. 131-141.

[4] Sharp, D.B., Myers, A., and Campbell D.M., "Using pulse reflectometry to compare the evolution of the cornet and the trumpet in the 19th and 20th centuries," Proc. of the Institute of Acoustics, 19(5), 1997, pp. 541-548.

[5] Sharp, D.B., PhD *Thesis*, University of Edinburgh, UK, 1996.

[6] Li, A., Sharp, D.B., and Forbes, B.j., "Improving the measurement of the low frequency response of the impulse response using acoustic pulse reflectometry," Proc. of Forum Acusticum 2002, Seville, Spain, 16-20 Sept 2002.

[7] Amir, N., Rosenhouse, G., and Shimony, U., "A Discrete Model for Tubular Acoustic Systems with Varying Cross Section - The Direct and Inverse Problems. Part 1 and Part 2: Theory and Experiment", Acustica, 81(5), 1995, pp.450-474.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

AN APPARATUS FOR MEASURING STRING VIBRATION USING ELECTRIC FIELD SENSING

Jyri Pakarinen and Matti Karjalainen

Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing,
P.O. Box 3000, FIN-02015 HUT, Espoo, Finland
jyri.pakarinen@acoustics.hut.fi

ABSTRACT

Vibrating string measurements by electric field sensing make use of the phenomenon where the capacitance between a metal string and an electrode plate changes as a function of the distance between them. The measuring device introduced in this study is intended for measuring the vibrational movement of a conductive string in two polarizations and possibly in multiple points. The basic idea of the construction is to apply an alternating voltage of high enough frequency to the string, and to measure the capacitively coupled voltage from a small electrode plate nearby. This AC voltage is amplitude demodulated to obtain the displacement of the string as a function of time. The simple architecture of the measuring apparatus, its relatively small size, and minor effect to string vibration make it an interesting alternative to other string vibration measurement devices. Detailed construction of the apparatus will be described in the paper, as well as analysis on its properties. These include the accuracy of the method, signal-to-noise-ratio, and system design parameters that affect these properties. Interesting cases of string vibration (particularly of the guitar) will be analyzed and depicted.

1. INTRODUCTION

Methods for measuring a vibrating string's motion can roughly be divided into three categories; optical, electromagnetic, and electric field sensing methods. This paper introduces a device for measuring the vibration of a string in both polarizations using the latter technique. Generally, electric field sensing makes use of the phenomenon where the capacitance between two electrodes changes, when the distance between them is varied. Musical applications of electric field sensing are discussed in [1].

There are two basic techniques for performing string measurements with electric field sensing. The principle of the first method is similar to the function of condenser microphones. In this approach, a conducting string is grounded and DC voltage is applied to an electrode plate nearby. The string's movement modulates the voltage between the string and the plate, and information about the string's placement is obtained. The DC voltage applied to the plate must be sufficiently high, in order to allow the changing capacitance to be measured. This means that the coating of the sensing electrode must be designed with care, to avoid breaking the measuring equipment. A condenser microphone and preamplifier setup can be used as a measuring device, if the microphone's diaphragm is replaced with the grounded string, as is done in an earlier work [2]. If commercially available microphone preamps are used, the measurement of vibration in both polarization planes becomes difficult because the preamps are relatively large. An electrostatic pickup for dual-polarization measurements of piano strings is presented by Weinreich [3].

The measuring device introduced in this paper is intended for measuring string vibration on a musical instrument using electric field sensing in a different way. The basic idea here is that an alternating voltage, the carrier signal, is applied to the string, and a small electrode is placed nearby. The voltage capacitively coupled to the electrode is then the carrier signal, amplitude-modulated by the movement of the string.

A voltage source follower is connected to the electrode plate to lower the output impedance of this measuring apparatus, and therefore to minimize the voltage drop in the cable to further analysis equipment. The signal is fed from the measuring apparatus to an AD-converter through an amplifier. The carrier signal is also fed to an AD-converter. Both signals are then read to a computer and the data concerning the distance between the string and the electrode is obtained by synchronous detection. Two electrodes are mounted on the measuring apparatus, thus allowing information of the string's placement to be obtained in both polarization planes, horizontal (i.e., parallel to the soundboard of a guitar) and vertical (i.e., perpendicular to the soundboard of a guitar).

2. THE APPARATUS

2.1. Construction of the apparatus

The measuring apparatus consists of two electrodes, an electric circuit containing two voltage source followers, and two 2-conductor microphone cables leading to the amplifier. The construction of the apparatus is illustrated in Figure 1. The circuit is enclosed by a brass casing, which is connected to the ground to make a Faraday cage, in order to cancel the effect of electromagnetic radiation to the circuit. The brass casing also provides mass to the the apparatus, which together with a thin layer of polystyrene underneath the casing constitutes the apparatus less sensitive to external vibrations, e.g., the vibration of a soundboard in a musical instrument. The casing is sealed with two plastic plates and air duct tape.

The electrodes are made of two brass plates, dimensions being $(15 \times 10 \times 0.6)$ mm, and are connected to the circuit with stiff copper wires of diameter 0.7 mm. The plates are placed at right angles relative to each other. The distance between the string and each electrode is chosen in these experiments to be approximately 5 mm when the string is at equilibrium, thus allowing vibrational amplitudes under 5 mm to occur on the string. A thin layer of insulating varnish is overlaid on the electrodes, in order to prevent galvanic coupling with the string.

The electric potential at a single point caused by an infinitely

long charged string (cylindrical charge) at distance r is of the form

$$V(r) = -\frac{1}{2\pi\epsilon_0}\lambda \ln \frac{r_{\text{string}}}{r} \quad (1)$$

where ϵ_0 is the permittivity constant of the vacuum, λ is the linear charge density of the string, and $\mathrm{r}_{\text{string}}$ is its radius. This implies that the voltage coupled to the electrode would depend on the logarithm of the distance between the electrode and string. However, if the amplitude of the string's vibrational movement is relatively small, the distance-voltage relationship can be approximated as linear, as is done in this study, or the nonlinearity can be compensated for in signal analysis.

The construction of the circuit is illustrated in Figure 2. The operating voltage for the circuit (6 V DC voltage) is produced with an external voltage source, and transferred to the circuit through cables. The cable used here is a shielded twisted pair, where the signal from the measuring apparatus and the operating voltage are in the inner conductors, while the shield is connected to the ground.

It must be noted that since the measuring device is relatively small, simple, and inexpensive, measurements at multiple points on the string can easily be conducted using several devices.

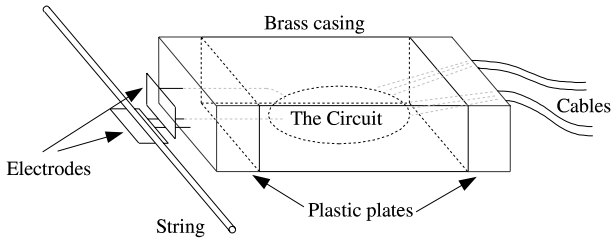


Figure 1: *The measuring apparatus is designed for obtaining data from both vibration polarizations of a conductive string.*

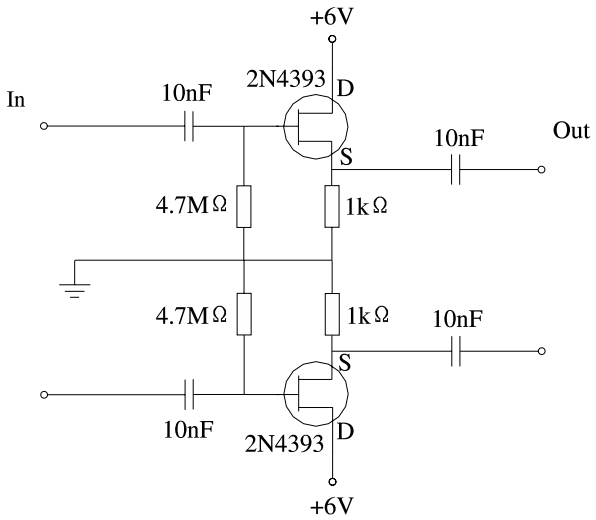


Figure 2: *The circuit used in the measuring device is a two-channel voltage source follower. The main task of the circuit is to lower the output impedance of the apparatus, thus lowering the voltage drop in the cables to further measuring equipment.*

2.2. Lowpass-filter-effect due to the longitudinal averaging of the electrode

It is important to note that because the measurements are carried out using electric fields and since the dimensions of the electrodes are not extremely small and they are not extremely close to the string, the longitudinal (in the direction of the string) spatial resolution is limited. The voltage capacitively coupled to the electrode is inversely proportional to a weighted average of its distance to a single point on the string. This phenomenon introduces a low-pass filter effect to the system, and it can be intuitively understood when considering the result of the averaging process. The wider the averaging window or higher the frequency of the displacement wave travelling in the string, the smaller displacement of mean voltage level from the equilibriun state is obtained. This low-pass effect would be reduced if the width of the electrodes could be made smaller, but this would also result in a lower signal level. Taking the electrode closer to the string would also help, but this would lead to increased nonlinearity due to Eq. 1.

3. CONSTRUCTION OF THE MEASUREMENT SETUP

The measurement setup consists of four main parts: a carrier signal source, a string vibration sensing device, an amplifier, and an analog-to-digital converter. The structure of the measurement setup is illustrated in Figure 3. The carrier signal used in the measurements is a steady 12 kHz sine-signal with peak-to-peak amplitude of 20 volts, produced with a signal generator. The signal generator also inserts some additive and multiplicative noise to the carrier signal, but this is not considered a serious problem since the amplitudes of the noise components are small, and the effect of noise is still reduced in the signal detection phase. Because the carrier signal has a relatively high signal level, it is fed to the AD-converter through a simple two-resistor voltage divider, denoted by R1 and R2 in Fig. 3.

The signal from the sensing apparatus is amplified with a low-noise mixer amplifier (denoted by B in Fig. 3), and then fed to an AD-converter. The sampling frequency used in the measurements is 48 kHz and the sample depth is 24 bits. It is possible to derive an expression for the voltage at the gate of the transistor using Figure 3. Clearly, the varying reactance of the string-electrode pair $X_{s,e}$ and the combined impedance $Z_{s\&i}$ of the stray capacitance and the source follower input resistor act as a voltage divider for the voltage V_0 generated in the signal generator, yielding for the gate voltage

$$V_{\text{gate}} = V_0 \frac{Z_{s\&i}}{X_{s,e} + Z_{s\&i}} + V_{\text{noise}} \quad (2)$$

and expressed in terms of the circuit components

$$V_{\text{gate}} = V_0 \frac{X_{\text{stray}} R_{\text{input}}}{X_{s,e} X_{\text{stray}} + X_{s,e} R_{\text{input}} + X_{\text{stray}} R_{\text{input}}} + V_{\text{noise}} \quad (3)$$

where V_{noise} is the voltage noise at the gate, discussed more thoroughly in subsection 3.2. X_{stray} and R_{input} denote the stray reactance between the ground and the gate and the input resistance of the source follower, respectively.

3.1. Signal detection

Since the carrier signal can be obtained easily from the signal generator, synchronous detection is used to obtain the data concerning the movement of the string. First, the carrier signal is phase shifted

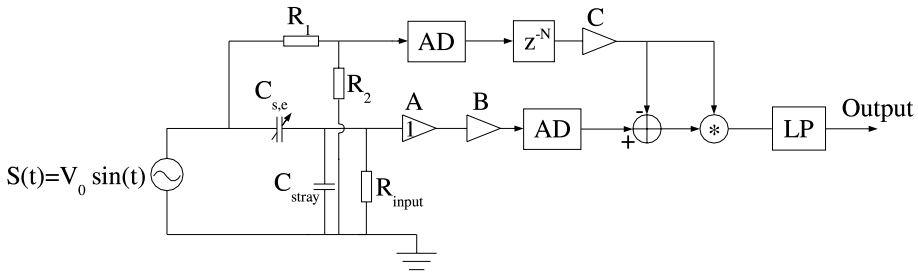


Figure 3: *The measurement setup. The varying capacitance between the string and the electrode is denoted with $C_{s,e}$, while C_{stray} symbolizes the stray capacitance between the electrode and ground. R_{input} is the input resistance of the voltage follower circuit, presented in Figure 2. R_1 and R_2 present the resistance values of the voltage divider for lowering the carrier signal level. The signal generator is denoted with $S(t)$. Signal amplifiers A and B denote the voltage follower circuit and a mixer amplifier, while the amplifier block C is only a constant scaling coefficient used in the signal detection phase. The delay element z^{-N} is used in synchronizing the carrier signal with the signal from the apparatus. The output signal is lowpass-filtered by LP to remove the remaining carrier signal components.*

in the time domain, so that it is in phase with the signal obtained from the measuring apparatus, and then normalized so that their amplitudes are approximately equal.

The carrier signal is then subtracted from the apparatus' output signal, and the resulting signal is multiplied with the phase- and amplitude-corrected carrier signal and finally lowpass-filtered at 10 kHz in order to remove the remaining high-frequency carrier. The final results can then be presented as the distance between the string and the electrode, as a function of time.

3.2. Theoretical study of SNR

The noise voltage at the gate of the transistor consists of three components: the input noise voltage of the transistor V_{nFET}, the noise voltage caused by the noise current of the transistor V_{iFET} and the noise voltage generated by the thermal noise of the input resistor V_{nI},

$$\mathrm{V_{noise}} = V_{\mathrm{nFET}} + V_{\mathrm{iFET}} + V_{\mathrm{nI}} \tag{4}$$

The equivalent input noise voltage of the JFET transistor is reported to be approximately $5\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$ at frequency of 10 kHz. If the bandwidth of the signal is considered to be 10 kHz, the equivalent input noise of the JFET would then be 500 nV. Because the voltage output difference of the measuring apparatus between the distances of 5 mm to 6 mm is about 20 mV, and if the vibrational movement with amplitude 1 mm is considered as 0 dB, the equivalent input noise voltage of the transistor would cause a signal-to-noise ratio of about 92 dB.

The equivalent noise current of the transistor is specified as $0.4\,\mathrm{fA}/\sqrt{\mathrm{Hz}}$ at 10 kHz. A 10 kHz bandwidth would therefore result in a noise current of 40 fA, and this noise current combined with the 4.7 MΩ resistance of the input resistor would finally produce a noise voltage of 0.19 μV. Using the same SNR calculation method as above, we find that the noise current of the FET results in SNR of 100 dB, and so the effect of the noise current can therefore be neglected in practice.

The thermal noise of a resistor can be calculated from equation

$$\mathrm{V_{nR}} = \sqrt{4\,k\,T\,B\,R} \tag{5}$$

where $k = 1.38 * 10^{-23}$ (Boltzmann's constant), T is the absolute temperature of the resistor, B is the bandwidth of the signal, and R is the resistance. An input resistor of 4.7 MΩ would therefore result in a noise voltage of approximately 39.3 $\mu\mathrm{V_{rms}}$.

The noise voltage V_{nI} at the gate of the FET due to the input resistance can be derived if the stray capacitance between the gate and the ground is known. The measurement data shows that the signal from the apparatus lags the signal obtained from the signal generator by approximately 30° This suggests that the capacitive reactance of the stray capacitance at the carrier frequency is

$$\mathrm{X_{stray}} = \frac{R}{\tan 30^\circ} \approx 8.14\ \mathrm{M\Omega}, \tag{6}$$

and thus the noise voltage at the gate due to the input resistance can be presented as

$$\mathrm{V_{nI}} = \frac{|\,\mathrm{j}\,X_{\mathrm{stray}}\,|}{|\,R + \mathrm{j}\,X_{\mathrm{stray}}\,|}\mathrm{V_{nR}} \approx 34\ \mu\mathrm{V}. \tag{7}$$

Assuming again that the string's movement with 1 mm amplitude would correspond to 0 dB, the thermal noise of the resistor would cause a signal-to-noise-ratio of approximately 55 dB. In real measurements however, lower SNR values were obtained. When comparing this noise level to the ones due to the transistor itself, it is obvious that the thermal noise produced in the input resistor is the main source of noise in this system, and it cannot be eliminated without dramatically changing the construction of the measuring apparatus.

If a higher carrier signal frequency would be used, the voltage noise at the gate could be made smaller, because the input resistor's contribution to the total impedance would be lower. On the other hand, if the carrier signal frequency would be very high, the signal detection would have to happen at hardware level instead of computational means after AD-conversion used here for simplicity. This would reduce the flexibility of the system and require more complex hardware. With approximative calculations we have found that the voltage noise produced in the transistor will be the dominating noise component for carrier frequencies higher than approximately 1.6 MHz.

4. MEASUREMENTS PERFORMED ON A GUITAR STRING

Signals obtained from a real measurement performed on a steel-string acoustic guitar are depicted in Figure 4. In this figure, signals from two different polarizarions are plotted separately. Figure 5 shows data from the same measurement, illustrated so that different polarization movements correspond to different axes. It must be noted that in both figures the distance units not calibrated.

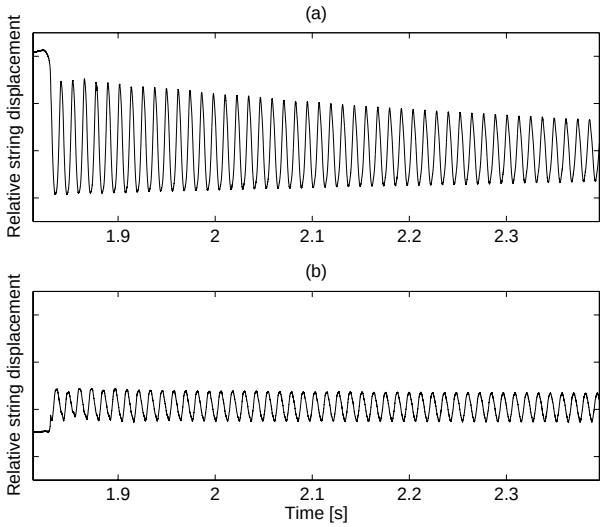


Figure 4: *Measurement data from a plucked guitar string (E_2) obtained from the apparatus. The horizontal vibration component is depicted in (a), and the vertical component in (b). The string was plucked at its midpoint primarily in the horizontal direction, and the scaling in both (a) and (b) is the same. Note that the y-axis values in both figures are roughly in millimeters but not calibrated.*

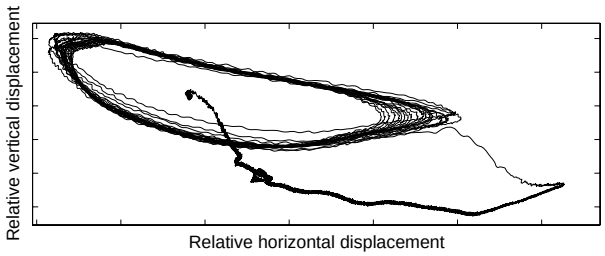


Figure 5: *Data from a plucked guitar string. Horizontal movement of the string is presented in the x-axis, while vertical movement is presented in the y-axis. In this figure the pluck and approximately 0.2 seconds of vibrational movement is presented. The string was plucked primarily in the horizontal plane and same scaling was used for the x- and y-axes. The axis values were not calibrated.*

5. DISCUSSION

There are many possible ways to improve the method proposed here to measure string vibration. Some of the possibilities are discussed in this section.

The dynamic range achieved in the prototype system, 40–50 dB, is useful for basic measurements of relatively large string displacements. The present dynamic range means 0 dB SNR for a displacement of about 3–10 μm. It could be however improved remarkably by several means. The most obvious one is to increase the frequency of the electric signal fed to the string. If the carrier frequency is above 1.6 MHz in the case described, the SNR could be up to 90 dB, which means that vibration displacements down to 0.03 μm will not be below measurement noise floor. At the same time the detection signal bandwidth can be increased since the low-pass filtering cutoff after demodulation can be made higher.

A signal processing technique improving SNR is to utilize the fact that after the attack the signal is concentrated around the modal frequencies, while measurement noise is a wide-band signal. For example high-resolution autoregressive moving average (ARMA) modeling of a measured pluck response can be applied to the signal above noise floor and let the model predict the practically periodic signal down below the noise floor [4].

Another problem of the proposed measurement method is the inherent nonlinearity of the system as a function of string displacement. This can be compensated, however, by careful calibration of a correction curve, based on measurements.

The next problem, if real playing of the instrument is desired to be measured, is to isolate the player whose finger is touching the string not to change the electric field around the electrodes. This may be realized to some degree by shielding and connecting the player to ground potential.

The problem of spatial low-pass filtering, discussed in 2.2., is difficult to avoid since the capacitance between the string and the electrode is spatially distributed. Again, to some degree it can be compensated by inverse filtering of the low-pass effect caused by spatial averaging. Another possibility is, if there are many electrodes along the string, to do spatial filtering with the channel outputs in order to counteract the effect.

6. CONCLUSIONS

We have presented a principle for measuring string vibrations in two polarizations and in multiple points along a metal string using electric field sensing. A prototype device was constructed that senses capacitively the distance of string from a small electrode using AC voltage carrier fed to the string. The signal from a sensing electrode is down-modulated and low-pass filtered to obtain an audio signal proportional to string displacement.

Theoretical SNR of the device was also derived, and two-polarization data from a real measurement on a steel-string guitar was presented. Although the prototype system has limitations in performance and functionality, such as relatively small dynamic range, the method shows potential to be improved further by proper combination of dedicated hardware and software.

7. ACKNOWLEDGEMENTS

This research was done within the ALMA project "ALgorithms for Modelling of Acoustic interactions" (IST-2001-33059); homepage at http://ftp-dsp.elet.polimi.it/alma/, and the Academy of Finland project "Technology for Audio and Speech Processing" (SA 53537).

8. REFERENCES

[1] J. Paradiso and N. Gershenfeld, "Musical Applications of Electric Field Sensing," *Computer Music Journal*, vol. 21, no. 2, 1997.

[2] X. Boutillon, "Model for piano hammers: Experimental determination and digital simulation," *Journal of the Acoustical Society of America*, vol. 83, pp. 746, 1988.

[3] G. Weinreich, "Coupled Piano Strings," *Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1474–1484, 1977.

[4] M. Karjalainen, P. Esquef, P. Antsalo, A. Mäkivirta, and V. Välimäki, "Frequency-zooming arma modeling of resonant and reverberant systems," *Journal of the Audio Engineering Society*, vol. 50, no. 12, pp. 1012, 2002.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

AN ENSEMBLE METHOD FOR AUTOMATIC ALIGNMENT OF SOUND AND ITS TRANSCRIPTION

Kåre Sjölander

Department of Speech Music and Hearing
Kungl Tekniska Högskolan
kare@speech.kth.se

ABSTRACT

The paper describes a novel method for automated segmentation and transcription of audio. A sound signal and a textual transcription of the contents are aligned automatically, using a procedure that borrows ideas from the fields of Machine Learning and Automatic Speech Recognition. The method utilises an ensemble of Hidden Markov Model (HMM) recognisers to decode the contents of an audio signal. By using an ensemble the combined output is more accurate than any single recogniser would have been. The result is a description where the signal has been divided into segments with start and end times and where each of these segments has been assigned a label from a predefined set. Label categories can for example be silence, speech, or music, but more detailed categories can also be used. For example, speech is typically segmented at a more detailed phonetic level where every phone is marked. The automatic alignment system has been implemented in a software toolkit for audio analysis and processing. It has been trained and evaluated on a manually annotated speech corpus as well as on radio programmes from the Swedish broadcasting corporation. The system is able to provide automatically aligned segments that are more than 85.5% in agreement with those provided by a trained human transcriber, given a 20ms tolerance between reference and system output. These results are competitive with other published results. Segmentations created by automatic methods have the attractive properties of being consistent and reproducible as well as much cheaper and faster to produce.

1. INTRODUCTION

Transcription of speech recordings at phone-level is a fundamental task in phonetics and speech technology research. Identification of phone segments in a speech material is the starting point for many studies. Typically this is done manually. For concatenative speech synthesis such segmentation provides the foundation for a unit inventory and the accuracy of the segmentation is crucial for the overall quality of the generated synthesised speech. However, it takes skill and considerable effort to produce such phone-level transcriptions. In comparison, orthographic transcriptions can be had at a much lower cost, especially when no time information for word boundaries is needed. Techniques borrowed from automatic speech recognition have been successfully applied to such word-level transcription in order to produce time aligned phone transcriptions automatically [1-2]. Speech recognition used in this context is usually referred to as forced alignment. The recognition grammar only allows for what actually has been said and only the segment boundaries is the output of interest, since the word output will be the same as the input.

This paper describes an automatic speech alignment system developed for Swedish, which is an extension to earlier work in this area [3]. The system takes as input speech sound files and matching text files containing word-level transcriptions of the speech. A specialised speech recognition system, based on Hidden Markov Models (HMMs), is applied and a time aligned phone-level transcription is produced.

In recent years there has been a lot of interest in using ensemble methods in the field of machine learning [4-5]. The main idea is that an ensemble of models can be more accurate than any of the individual models themselves. This idea has been used to increase the performance in the current speech alignment system. A set of HMMs is used to create one aligned transcription each. These transcriptions are combined in a second stage to give a more accurate result.

2. METHOD

The proposed automatic alignment system is based on standard techniques used in automatic speech recognition (ASR).

The sound signal is converted into feature vectors using standard ASR parameterisation methods. In this case, Mel-Frequency Cepstral Coefficients (MFCCs) are used. These provide spectral patterns of the speech at equidistant intervals. Between one hundred and a thousand vectors are created per second of sound signal.

The textual transcription is used to create a finite state network of HMMs. Silence, music and speech units (phones) are all modelled by HMMs with continuous density output distributions. The network represents the expected (and allowed) content of the sound signal. Each word in the text will result in a sequence of phone models depending on its pronunciation as shown in Figure 1 (top). If a certain word has alternative pronunciations these will generate parallel branches in the network, cf. Figure 1 (middle). Similarly, spontaneous pauses can be handled by inserting optional silence models, see Figure 1 (bottom). The Viterbi algorithm is used to find the optimal path through the network for the given sequence of parameter vectors. This optimal path is the resulting segmentation. In the implementation used each state in the network can be assigned one or more text tags. This feature has several applications. Tags can be used to assign lexical stress-level on vowels or to mark the beginnings of words, paragraphs, or even speaker change in a

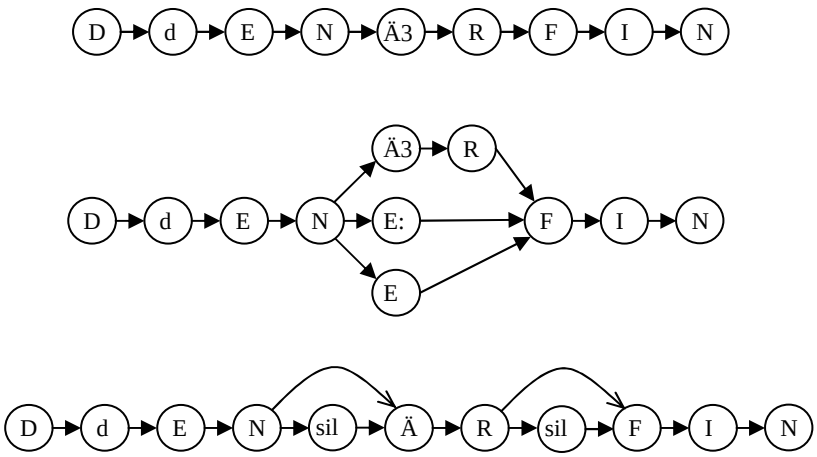


Figure 1.Example finite state networks of phone HMMs used in the automatic aligner to model the content of a speech recording being processed. Top graph shows a simple left-to-right transition network. Middle graph shows a network allowing for multiple pronunciations, one alternative per path. The path that best matches the speech will be used in the output. Bottom graph shows a network that allows for optional inter-word pauses, where "sil" is the silence model. Only pauses that match with the recorded speech will be contained in the output.

dialogue. Types of information that is not contained in the phone sequence alone can be transferred through the decoding process unaffected and later be used to structure the aligned phones for further processing.

The alignment system works in two stages. First a collection of segmentations is produced each based on HMM-set and matching feature vectors with different parameterisation settings. Typically, the interval between the features vectors and the analysis window size is varied. Seven such HMM-sets were used in the current system. In the second stage these segmentations are combined into a single segmentation, which is more accurate than any of the initial hypotheses. In order to accomplish this, the phone segments have first been clustered into broad phonetic categories, e.g. all vowels are assigned to the class Vowel. In the ensemble combination stage a segment boundary is taken from the segmentation that best models the current segment category to segment category transition.

One application of the automatic alignment system has been aligning radio recordings. These contain music sections, usually at the beginning and the end, but sometimes also interspersed with the speech. In order to handle these music sections, a special music HMM was trained.

2.1 Implementation

The system has been implemented using a two-level approach. The basic performance-critical speech technology algorithms have been put in a function library. In this way they could also be used for other applications with little effort. On top of this library a command line tool has been written, called nAlign. This tool provides a more user-oriented interface. It has several control options and takes care of the different file formats involved. In most cases it can be used with the default settings for fully automated phone-level time-alignment.

2.1.1 Library functions

The library functions were implemented as new commands in the Snack Sound Toolkit [6]. In this way it was possible to use existing signal processing routines for speech parameterisation and also to benefit from other sound handling functions during the development. Two new commands were added, one for the creation of feature vectors and one to perform the Viterbi alignment.

2.1.2 The nAlign tool

The alignment and segmentation tool nAlign takes sound files and transcription text files as input. A transcription file typically consists of an orthographic transcription of the speech contained in the corresponding sound file, but it can also be a phone-level label file if desired. Output is generated as phone label files and, optionally, as word-level files. There is a possibility to use a single transcription file archive, instead of creating individual files. Several options are available to control the tool, for example, what kind of output to generate, if inter-word modelling is to be used, and so on. Also, the tool can be controlled through tags in the transcription files, for example, using the symbols <silence>, <music>, or <garbage>, to mark pauses, music, or disturbances in the recording. The special tag <fixpoint: *time*> can be used if the alignment process fails somewhere. Using this tag, a certain position in the input text can be manually locked to the correct time position in the recording. This feature can also be used to decrease processing requirements for long recordings. Basic text normalisation functionality has been built directly into the

tool. Phone transcription from words is done either through lexicon look-up or through the KTH text-to-speech component RULSYS [7].

3. RESULTS AND DISCUSSION

The performance of the system was evaluated on a test material from a manually annotated speech corpus [8]. This material consists of 327 sentences, with 9318 segments manually marked at phone-level. The total length of the test material is about 19 minutes. The material took 6 minutes to process on a 2.5GHz Pentium IV computer. The system was able to put 85.5% of all boundary locations within 20 milliseconds of the manually segmented reference boundaries. For 10 milliseconds allowed deviation the corresponding figure is 68.4%. As a comparison the best single model from the ensemble gave 84.4% correct boundaries within 20 ms. The performance of the same model for 10 ms was only 60.0%, which shows that the ensemble was able to perform better than any single model on its own. Complete results for all the seven HMM-sets used as well as for the combined ensemble are given in Table 1 below.

For the music sections the performance is a markedly lower. Only 39.1% of the boundaries were correctly placed within 20ms. The corresponding figures were 13.0% for 10 ms allowed deviation and 56.5% for 40 ms. One problem is that music and speech usually overlap in the recordings. A speaker will typically start speaking while the music is playing and only then will the music start to fade away.

The system has been applied to a variety of other tasks, for example, to duration measurements in a study on prosodic boundaries in conversational speech [9], to segmentation used for articulatory measurements in expressive speech [10], and for measuring speaking rate variation [11]. It should be noted that manual checking and correction was applied after the automatic alignment processing.

HMM model	< 10 ms	< 20 ms	< 40 ms
10 / 25	56.6	83.3	93.9
10 / 16	60.0	84.4	94.0
5 / 25	63.2	82.5	92.9
5 / 16	66.0	83.8	93.4
5 / 8	68.4	84.2	93.4
2 / 5	66.1	80.5	90.5
1 / 5	64.2	78.7	89.2
Total ensemble	68.4	85.5	94.0

Table 1: *Performance for the individual segmentations and for the ensemble after combination. The HMM model column shows the frame interval and analysis window length used for that model. Both values are in milliseconds. The other columns show the percentage of boundaries within 10, 20, and 40 milliseconds of reference.*

3. CONCLUSIONS

The automatic alignment system does not perform as well as a trained transcriber for all purposes. But it is fast and convenient to operate and gives consistent results. One obvious use is for creating an initial starting point for further manual refinement. The input orthographic transcription needs to reflect the speech as closely as possible in order to give best results. But investigations are underway how to handle speech recording where only part of the speech has been transcribed. The current system applies text normalisation to the input text before processing, for example, every character is converted into lower case or non-letter characters are removed. In many cases it might be more appropriate to get the original input back but with added time stamps as desired. This is currently being implemented.

In the development of the automatic alignment tool the goal has been to minimise overall boundary deviation. But in some cases it might be more appropriate to try to minimise the maximal deviation. If only a sentence-level or even paragraph-level segmentation is desired it does not matter much that some phone boundaries in the middle of the segment are a few milliseconds off from their ideal positions. But the desired segment boundaries themselves should be reliable. This will be the goal of further research.

The method for model combination used in this study is very simple. Each model is weighted with either 1 or 0 in the combined output depending on the model's performance for the given broad phonetic category transition. Other weighting schemes will be the aim of future investigations.

The automatic alignment system has been used for several different purposes and on recordings of varying types with a high degree of success. Even though the resulting alignment is far from perfect the amount of manual labour saved is high. In the future its applicability and performance is expected to continue to grow.

4. ACKNOWLEDGEMENTS

The research reported here was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations. The author would like to thank Mattias Heldner for his many useful comments and suggestions in testing and using the system.

5. REFERENCES

[1] Brugnara, F., Falavigna, D. and Omologo, M., "Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models", Speech Communication, 12, 4, 357-370, 1993.

[2] Hosom, J.-P., "Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information", PhD Thesis, Oregon Graduate Institute of Science and Technology, 2000.

[3] Sjölander, K., "Automatic alignment of phonetic segments", Working papers 49: Papers from Fonetik 2001, Lund, Lund University, Dept. of Linguistics, 140-143, 2001.

[4] Dietterich, T. G., "Machine Learning Research: Four Current Directions", *AI* Magazine, 18 (4), 97-136, 1997.

[5] Cook, G. D., "Model Combination and Data Selection in Connectionist Speech Recognition", PhD Thesis, Queens' College, Cambridge, 1997.

[6] Sjölander, K. and Beskow, J., "WaveSurfer - an open source speech tool", Proceedings of the ICSLP 2000, IV, 464-467, 2000.

[7] Carlson, R., Granström, B., and Hunnicutt, S., "A multi-language text-to-speech module", Proceedings of ICASSP '82, Paris, Vol. 3, 1604-1607, 1982.

[8] Bertenstam, J., Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., Nord, L., de Serpa-Leitão, A. and Ström, N., "The WAXHOLM application database", Proceedings of Eurospeech -95, Madrid 1, 833-836, 1995.

[9] Heldner, M. and Megyesi, B., "Exploring the prosody-syntax interface in conversations", Proceedings of ICPhS 2003, Barcelona, 2003.

[10] Svanfeldt, G., Nordstrand, M., Granström, B., and House, D. "Measurements of articulatory variation in expressive speech", In Proceedings of Fonetik 2003, Umeå, Sweden, 2003.

[11] Bell, L., Gustafson, J. and Heldner, M. "Prosodic adaptation in human–computer interaction", Proceedings of ICPhS 2003, Barcelona, 2003.

Proceedings of the Stockholm Music Acoustics Conference 2003 (SMAC 03), Stockholm, Sweden, August 6-7-8-9, 2003

MEASUREMENT OF THE TERMINATION IMPEDANCE OF A TUBE USING PARTICLE IMAGE VELOCIMETRY

D.J. Skulina, D.M. Campbell, C.A. Greated

Acoustics and Fluid Dynamics Group
Univerity of Edinburgh
dskulina@ph.ed.ac.uk

ABSTRACT

The non-linear behaviour of the acoustic velocity field in the vicinity of the open end of a musical wind instrument is strongly affected by its detailed geometry. As the sound pressure level within the instrument increases phenomena such as acoustic streaming and vortex shedding contribute increasingly to the dissipation of sound energy, resulting in an increase in the real part of the terminating impedance.

To characterise quantitatively the effect of varying the geometry of a tube termination, particle image velocimetry (PIV) has been used to measure the velocity field near the open end of a tube in which a high amplitude sound field is generated by a loudspeaker. A high speed digital camera and pulsed laser were used to acquire images with high spatial and temporal resolution from which velocity maps were derived.

Three tube ends with different profiles were studied. The acoustic volume flow was obtained from the PIV maps and the sound pressure level measured using a probe microphone, allowing the terminating impedance to be calculated at a number of sound levels. In addition, the full field velocity maps obtained using PIV allow a qualitative description of the development of the acoustic flow at the tube termination as the intensity increases.

1. INTRODUCTION

At low acoustic amplitudes, the behaviour of the acoustic field near the end of a resonant tube is well understood in terms of linear acoustics and the local flow well described by a potential flow model. As the sound pressure level within the tube increases, however, non-linear propagation effects such as acoustic streaming, the development of shockwaves and localised effects such as vortex shedding increasingly affect the flow behaviour at the open end of the tube. These effects additionally result in an increase in the real part of the terminating impedance of the tube and have implications in the study and construction of woodwind instruments.

The detailed geometry of the termination of the open end of a tube plays an important role in the strength of the effects described above. It is the aim of this work to investigate experimentally, using particle image velocimetry (PIV), the outcome of changing the termination profile of the open end of a tube in which a high amplitude sound field has been generated by a loudspeaker. PIV is a powerful tool for the non-intrusive (both quantitative and qualitative) examination of instantaneous velocity fields and is used here to provide detailed information about the flow at the open end of the tube.

The acoustic volume flow is obtained from the PIV velocity maps and the sound pressure level measured using a probe microphone allowing the terminating impedance to be calculated at several sound levels. Furthermore, analysis of the velocity maps allows the examination of, for example, the development of velocity profiles and of vortex shedding over the acoustic cycle and with increasing velocity amplitude.

2. THEORY

The acoustic radiation impedance Z_r is defined as the ratio of the acoustic pressure p and the volume flow rate U and is given by:

$$Z_r = R + jX \tag{1}$$

where R is the acoustic resistance corresponding to the dissipation of energy and X is the acoustic reactance relating to the fluid inertia.

For an unflanged tube Levine & Schwinger [2] showed that in the linear regime the radiation impedance Z_{rlin} is given for low frequencies by:

$$Z_{rlin} = \left\{ \frac{(ka)^2}{4} + jk\delta_0 \right\} Z_c \tag{2}$$

where k is the wavenumber, a the tube radius, $\delta_0 = 0.6133a$ is the unflanged end correction and $Z_c = \rho_0 c / S$ is the characteristic impedance of tube of cross sectional area S.

The termination impedance Z_t is given by:

$$Z_t = \left\{ \frac{p}{U} \right\} e^{j(\phi_p - \phi_u)} \tag{3}$$

where p and U are the first harmonic components of the pressure and volume velocity amplitudes and ϕ_p and ϕ_u are the relative phases of pressure and velocity respectively. With knowledge of the phase difference between pressure and velocity at the open end of the tube it is therefore possible to calculate the terminating impedance if the pressure and volume velocity are also known.

Assuming cylindrical symmetry, the volume velocity at the open end is calculated by splitting the tube cross section into annular regions of width δr_i, each with an associated velocity u_i measured from the PIV maps using the following equation:

$$U = \sum_i \{2\pi a \delta r_i - \pi \delta r_i^2\} u_i \tag{4}$$

where i is the number of annular regions in the cross section.

The non-dimensionalised acoustic resistance R_{nd} can then be established by taking the real part of the measured terminating impedance Z_t and dividing by the characteristic impedance Z_c:

$$R_{nd} = \frac{Re(Z_t)}{Z_c} \tag{5}$$

Recent work by Atig [1] has shown using a two microphone method that the radius of curvature of the tube termination is linked to and has a significant effect on the behaviour of the real part of the terminating impedance of an open ended tube at high acoustic amplitudes. Also noted was a discontinuity in the magnitude of the terminating impedance for certain termination geometries at specific acoustic amplitudes.

3. EXPERIMENTAL METHODS

The main features of the apparatus used are shown in figure 1. A 20W copper vapour pulsed laser capable of repitition rates up to 50kHz is connected via an optical fibre to a 'fibrewire' light sheet delivery box, producing a light sheet 50mm wide and 1 mm thick. The high repitition rate of the laser allows accurate measurement of acoustic velocity amplitudes up to about 20ms^{-1}. The light sheet is aligned with the tube axis and a digital camera positioned to view an area of interest around the open end of the tube, perpendicular to the light sheet. A copper tube of radius 8mm is mounted onto the loudspeaker with an adaptor in which a pressure transducer is located, and one of three carefully machined terminations (see figure 1(ii)-(iv)) is fitted to the open end. The effective length of the tube, adaptor and termination system is 347mm giving a fundamental resonant frequency of 490Hz.

The apparatus in figure 1(i) is enclosed within a cubic aluminium box of side length 1.2m for safety reasons (to prevent extraneous laser light from escaping, and to reduce the sound level in the lab) and to contain seeding particles. Seeding particles from a fog generator are added and the loudspeaker and laser light turned on.

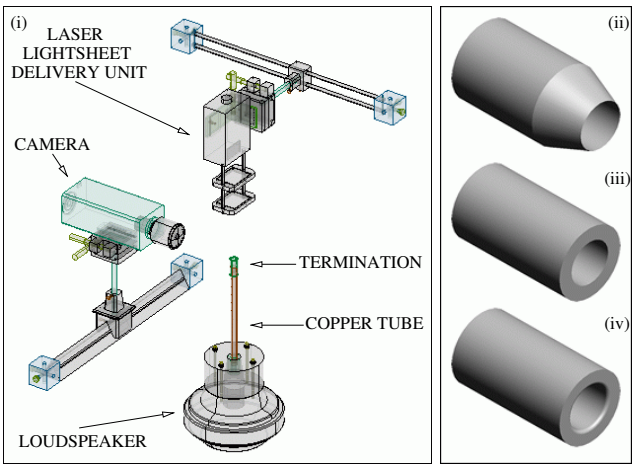


Figure 1: *(i) Experimental apparatus showing camera, laser light sheet delivery unit, loudspeaker, tube and termination. Three tube terminations used: (ii) Sharp edges, (iii) R=0mm, (iv) R=1mm.*

The system shown schematically in figure 2 was used to ensure that the phase difference between the pressure and velocity fields at a specific point in the flow was known. The pressure signal from a probe microphone placed a short (known) distance from the open end of the tube was sent to a PC along with a reference signal used to drive the loudspeaker through a data acquisition card. Once acquired, this data was analysed to give the phase difference ϕ_p between the reference and pressure signals.

A pulse synchronised with the reference signal from the frequency generator was used to trigger the synchronisation of light pulses from the laser with digital camera exposures at specific points throughout the phase of the acoustic cycle. This was accomplished by introducing a delay between the reciept of a trigger signal from the frequency generator and the firing of the laser. Given that the period of the acoustic cycle is known and that typically 10 PIV image pairs per acoustic cycle are adequate, the delay was incrementally increased by one tenth of a period for each point in the phase. At each phase step 20 PIV image pairs of separation 20μs were taken allowing averaging of vector maps and error evaluation during the analysis process.

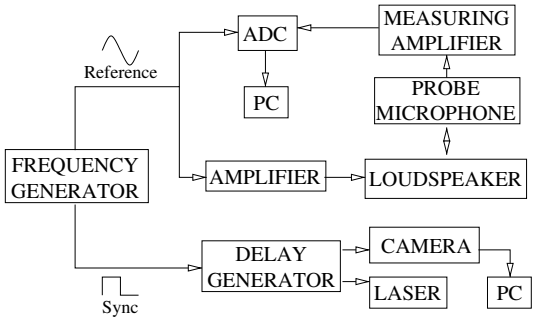


Figure 2: *(i) Schematic diagram showing the synchronisation setup.*

This process of capturing velocity maps throughout the acoustic cycle was repeated at several intensity levels, ranging from 150dB to 170dB (measured using the pressure transducer mounted on the loudspeaker adaptor) and for each of the three tube terminations.

PIV analysis of the images was carried out and the resultant PIV vector maps validated. Further averaging of the 20 vector maps at each phase step reduced the data to ten velocity maps per acoustic cycle and yielded information about the standard deviation in velocity at each point in the vector map. The volume velocity , which has the same phase angle as the acoustic velocity, was evaluated across the exit of the tube using equation 4 and the phase difference between the reference signal and the volume velocity, ϕ_u, calculated.

From this information the phase difference between pressure and velocity was gained. Fourier analysis of the pressure and volume velocity signals yielded the magnitudes of the first harmonic components and the terminating impedance calculated using equation 3. Combining equations 3 and 5 the non-dimensionalised acoustic resistance was deduced and plotted against the peak-to-peak voltage used to drive the loudspeaker.

4. RESULTS AND DISCUSSION

Example images taken using the system described above with the $R = 0mm$ tube termination and their resultant averaged vorticity contour maps with velocity vectors can be seen in figures 3 to 6. The location of the walls of tube termination are represented in the PIV maps by the solid grey blocks.

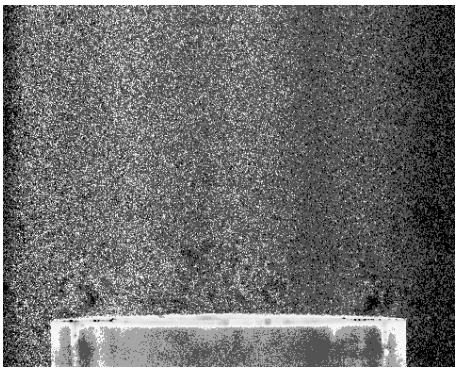


Figure 3: *Photograph used for PIV analysis taken at 160dB and 216°*

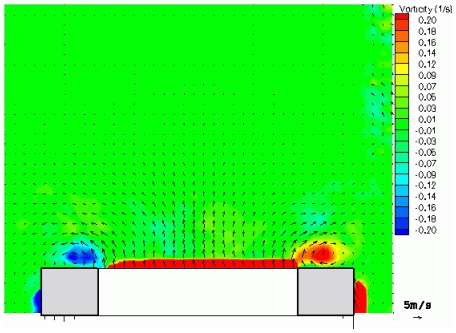


Figure 4: *PIV velocity map taken at 160dB and 216° with vorticity contours. Blue and red areas represent counter-clockwise and clockwise rotation respectively.*

Figures 3 and 4 correspond to the situation where the sound pressure level measured at a transducer located at the base of the tube is 160dB at a phase angle of 216°. The peak-to-peak driving voltage to the loudspeaker is 8V and the acoustic velocity amplitude averaged over the tube cross-sectional area at the termination is 6.7ms^{-1}. In figure 4 there are acoustic velocities up to about 8ms^{-1} and some localised vorticity can be seen close to the edges of the tube exit.

Figures 5 and 6 correspond to the same situation at 165dB with a driving voltage of 16V and an acoustic velocity amplitude at the end of the tube of 10ms^{-1}. At this level velocities range up to about 13ms^{-1} and the localised vortices have become detached from the tube edges.

Examination of PIV maps throughout the acoustic cycle confirm the presence of a change in acoustic flow behaviour for each of the tube terminations at an acoustic amplitude corresponding to that at which a discontinuity in the terminating impedance was observed by Atig [1].

The PIV maps in figures 4 and 6 demonstrate this difference in flow behaviour for the $R = 0mm$ tube termination. Up to a certain threshold acoustic amplitude vortices increase in strength and appear to cling to the edges of the tube. In the outward flowing part of the acoustic cycle these vortices can be seen just outside the tube edges, as shown in figure 4. On inflow these vortices are drawn back into the tube.

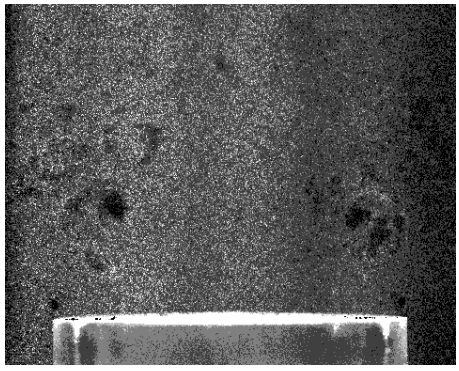


Figure 5: *Photograph used for PIV analysis taken at 165dB and 216°*

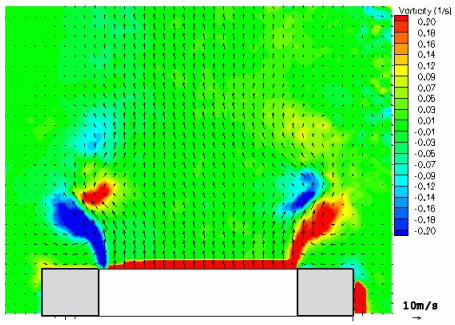


Figure 6: *PIV velocity map taken at 165dB and 216° with vorticity contours. Blue and red areas represent counter-clockwise and clockwise rotation respectively.*

Above a threshold acoustic amplitude the vortices no longer cling to the edges and are shed periodically. This situation is shown in figure 6. Here the strength of the inflow is not enough to draw the vortices back into the tube and vortex motion outside the tube can be observed throughout the acoustic cycle.

It was found that the threshold between the two types of flow behaviour occurred at a peak-to-peak driving voltage to the loudspeaker of around 11V with a measured acoustic velocity amplitude of around 8.8ms^{-1} for the $R = 0mm$ tube termination, and in a separate experiment at a driving voltage of around 17V for a termination with radius of curvature $R = 0.3mm$. A transition to the second flow regime was not observed at the acoustic velocity amplitudes studied here for the R=1mm termination. The type of vortex shedding observed in the second flow type described above was observed for all acoustic amplitudes studied when using the sharp edged tube termination.

Acoustic volume velocities calculated using equation 4 for the $R = 0mm$ and $R = 1mm$ tube terminations can be seen in figure 7. As the acoustic amplitude within the tube increases it is clear that non-linear effects become more prominent. The volume flux out of the tube gradually becomes larger than the inflow, as the acoustic streaming velocity magnitude approaches that of the acoustic particle velocity. A phase shift is observed for the $R = 0mm$ termination at driving voltages above the threshold (11V). No such shift is observed for the $R = 1mm$ termination.

Preliminary measurements of the acoustic resistance at several

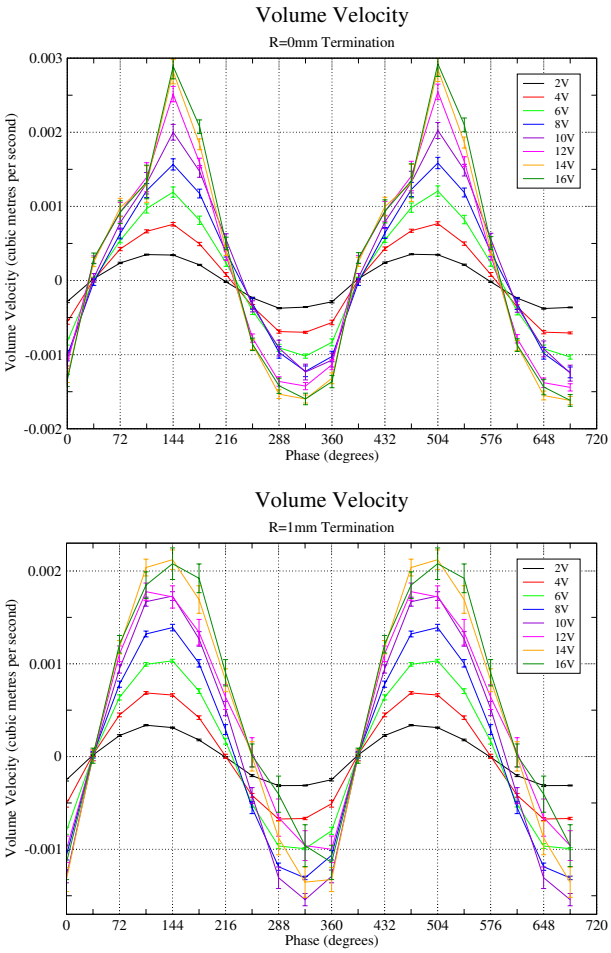


Figure 7: *Acoustic volume velocity with increasing loudspeaker voltage for two terminations. Top: R=0mm, bottom: R=1mm.*

intensity levels are shown in figure 8. Also shown for comparison are the results of Atig [1] and the linear prediction of Levine and Schwinger [2] for an unflanged tube. There is good agreement in the location of the discontinuity described previously and it appears that losses below 10V may be more significant than predicted by Atig. However, this may be due to an underestimation in the calculation of the volume velocity at the tube exit caused by problems aligning the physical location of the tube walls with the grid used in the PIV analysis. This clearly requires further review.

As the acoustic amplitude decreases, the flow behaviour becomes predominantly linear in fashion. The resistance should tend towards a value slightly higher than the result predicted for an unflanged tube by Levine and Schwinger [2] due to the finite width of the $R = 0mm$ tube termination. This is in agreement with observations.

5. CONCLUSIONS

The geometry of the termination of an open tube plays an important role in determining the strength of the real part of the terminating impedance of the tube. Two distinct flow behaviours were observed and the threshold between them appears to be dependent

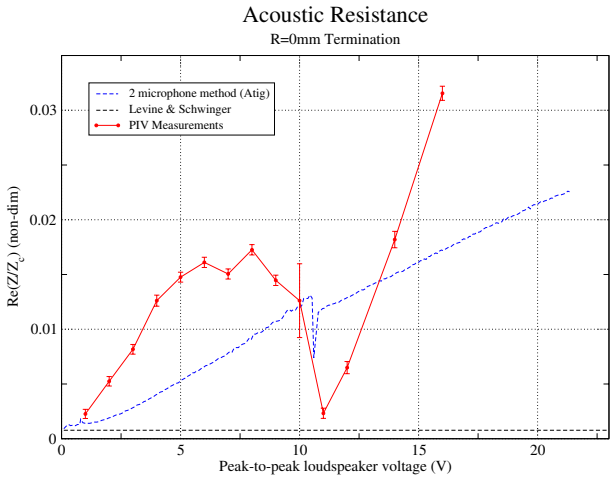


Figure 8: *Acoustic resistance with increasing loudspeaker voltage for the R=0mm termination.*

on the radius of curvature of the inner edge of the tube termination. As the radius of curvature increases, so does the threshold. It appears that both vortex shedding and acoustic streaming contribute greatly to the acoustic flow behaviour observed on either side of this threshold and hence to the acoustic energy losses at the tube termination. This has important implications for manufacturers of woodwind instruments. Sharp edged tubes begin to convert acoustical energy into rotational vortex motion at lower acoustic amplitudes than tubes with rounded edges and therefore become more difficult to play loudly.

6. REFERENCES

[1] Atig, M., Dalmont,J.-P., Gilbert, J., "Impedance of pipe ends at high sound pressure level". Private communication. Submitted to Compte Rendus de l'Academie des Sciences - Mecanique (2003).

[2] Levine,H., Schwinger,J., "On The Radiation of Sound from an Unflanged Circular Pipe", Phys. Rev 74 (1948) 383-406.

[3] Dalmont, J.-P. *et al* ,"Experimental determination of the equivalent circuit of an open side hole: linear and non-linear behaviour", Acustica, 88 (2002) 567-577.

[4] Ingard, U., Ising, H., "Acoustic nonlinearity of an orifice", J. Acoust. Soc. Am., 42(1) (1967) 6-17.

[5] Rockliff, D., "Application of particle image velocimetry to the measurement of non-linear effects generated by high intensity acoustic fields", PhD Thesis, University of Edinburgh (2002).

[6] Ingard, U., Labate, S., "Acoustic circulation effects and the nonlinear impedance of orifices", J. Acoust. Soc. Am., 22(2) (1950) 211-218.

[7] Disselhorst, J.H.M., Van Wijngaarden, L., "Flow in the exit of open pipes during acoustic resonance", J. Fluid Mech., 99(2) (1980) 293-319.

[8] Salikuddin, M., Ahuja, K.K., "Acoustic power dissipation on radiation through duct terminations: experiments", J. Sound Vib., 91(4) (1983) 479-502.

THE CONCEPT OF 3-D ACOUSTIC RESISTANCE AND ITS PRELIMINAR EXPERIMENTAL EVALUATION AT THE OPEN END OF AN ORGAN PIPE USING A MICROFLOWN® PROBE

Domenico Stanzial, Davide Bonsi

The Musical and Architectural Acoustics Laboratory
FSSG-CNR, Fondazione Cini, Isola di S. Giorgio Maggiore, I-30124 Venezia, Italia
domenico.stanzial@cini.ve.cnr.it - davide.bonsi@cini.ve.cnr.it

ABSTRACT

The concept of 3-D acoustic resistance is here introduced on the basis of energetic properties of a general sound field and its measurement by means of the USP microflown® particle-velocity probe is proposed. In the present application a preliminary set of measurements of acoustic pressure and three components of air particle velocity together with the amplitude spectrum of their complex ratio at the open end of an organ pipe is analyzed. Experimental data were collected both when the pipe was playing alone and when the same pipe was playing but coupled at the same fundamental frequency, with a second organ pipe also radiating its own sound nearby the first one. This way the so called "mitnahme" effect was acoustically monitored, and gathered data are here reported and compared. Finally the potentiality of measuring sound resistance in musical acoustics is addressed.

1. INTRODUCTION

The definition of specific acoustic resistance R_s is usually given in literature [1] as the real part of the specific acoustic impedance

$$Z_s(\omega) = R_s + iX_s = \frac{\hat{p}(\omega)}{\hat{v}_{in}(\omega)} \qquad (1)$$

where Z_s is clearly a frequency dependent complex function coming from the rightmost ratio of the complex amplitude $a+ib$ of a single frequency-component of the acoustic pressure signal

$$p = \mathrm{Re}(\hat{p}e^{i\omega t}) = P\cos(\omega t + \phi);\ P = \sqrt{a^2+b^2};\ \phi = \arctan\frac{b}{a}$$

by the complex amplitude, at the same frequency, of the inward component of acoustic velocity projected normally to a surface. In the definition of specific impedance the acoustic velocity is thus subjected to the same mathematical complex formalism as the acoustic pressure but having, from the physical point of view, a significance somewhat involved and a little bit obscure. The word "specific" stands in fact for "unit area" being Z_s measured in Pa·s/m or Kg/(m^{2}s) and then this area has to be related to any surface defined within a spatially 3-D sound field but possibly with a clear physical meaning. For instance in the case of an open organ pipe the measured value of the impedance at its top end is usually multiplied by the area of the pipe end section, so finding the actual mechanical radiation impedance which is then measured in Kg/s. Clearly in this special case the sound field into the pipe is considered a one-dimensional one and the pipe section refers to an area over a surface orthogonal to the pipe axis.

The concept of acoustic impedance is then essentially connected with an oriented area over which in the general case a surface integral could be calculated. The only spatial direction which, from the energetic point of view, has interesting local implications for the sound propagation in any general 3-D sound field is that defined at each point of the field by the active sound intensity vector. Furthermore, as shown in [2], this direction can usefully be employed for defining the pressure in-phase component of sound velocity giving rise to the instantaneous radiant intensity [3]. This interesting feature allows us accordingly to define the specific acoustic resistance in terms of active intensity as

$$R_s := \frac{\langle p^2 \rangle}{A} \qquad (2)$$

so possibly linking Eq. (1) to the energetic analysis of sound field based on the energy flux density properties. It is interesting to notice here that Eq. (2) foresees an infinite value of R_s for null values of the active intensity vector modulus A.

Anyway this paper is only concerned with some preliminary results gathered for this promising research topic in order to show the experimental feasibility of this approach using a new commercially available p-v probe by microflown®. In this respect a set of acoustic quad signals (pressure and three particle velocity components) has been measured for different acoustic systems built-up by two strongly interacting fields generated by organ pipes and a brief analysis of the behavior of significant transfer functions between pressure and velocity components is reported in the following.

2. EXPERIMENTAL SETUP AND REULTS

The experimental setup assembled and used for measurements is shown in Figure 1. There, two flue organ pipes having a fundamental frequency of about 200 Hz are shown and two different sound fields were built up by playing the left pipe alone (field "A") and the two pipes together (field "A+B"). A plastic folder was used for the exact tuning of the two pipes during the *mitnhame* effect (field "A+B"). The microflown® USP was located along the axis of the left pipe about .34 m above its open end and the acoustic quad signals for both field "A" and "A+B",

were sampled and digitized by a PCMCIA 4-channel audio board installed on a notebook PC.

Figures 2 (a and b) show respectively the measured pressure spectrum for the field A and A+B. Comparing the two spectra it can be seen that when the two pipes played together, a difference of about -5 dB at low harmonics, has been detected. Considering now the twin (a and b) spectra relatively to the x, y, z components of the velocity spectrum, practically no difference at all is present in the z component of velocity spectrum (Figures 5a and 5b). Detected differences become significatively large when components lie in the xy plane. In fact, from the comparison of velocity spectra reported respectively in Figures 3a and 3b we find an increment of about 10 dB at the fundamental frequency when passing from the field "A" to "A+B". A similar increment of about 5 dB can be observed in Figures 4a and 4b.

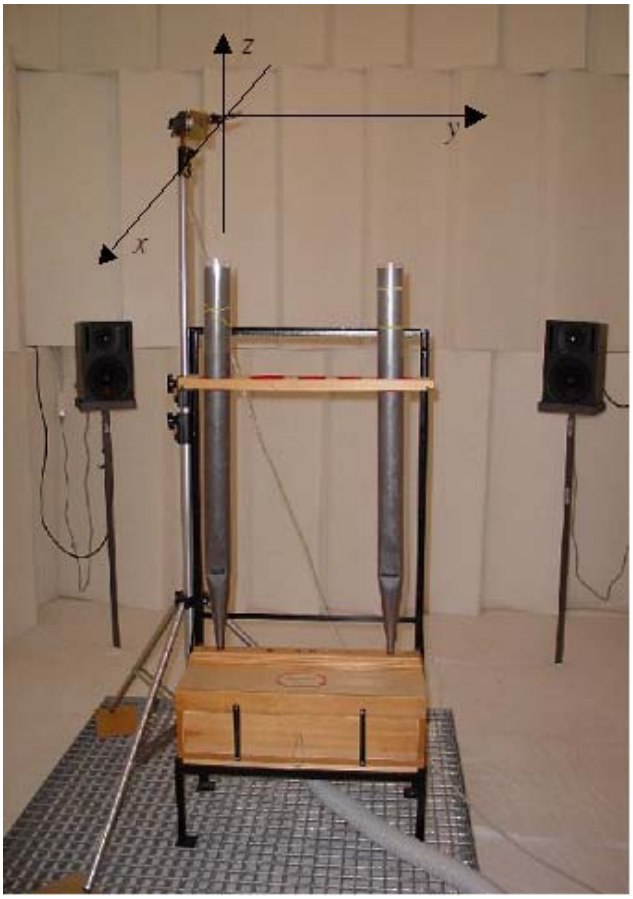


Figure 1: *Experimental setup. The origin of the Lab reference frame coincides with the position of the USP whose velocity sensors have been oriented according with x,y,z axis directions. The field called "A" in the text was generated by the left pipe playing alone while the field "A+B" is the coupled field generated by both the pipes in the "mitnhame" acoustic regime.*

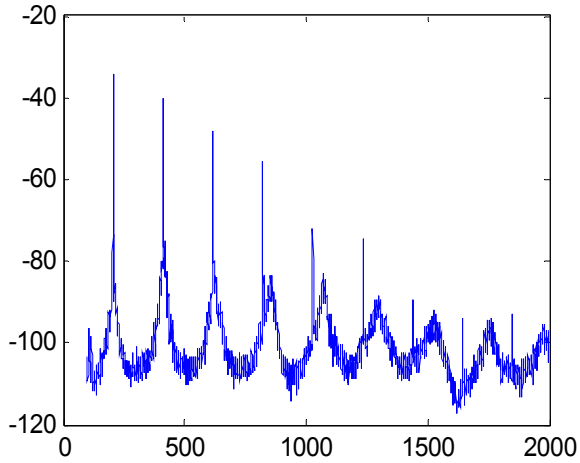


Figure 2a: *Field "A", pressure sprectrum (dB vs Hz).*

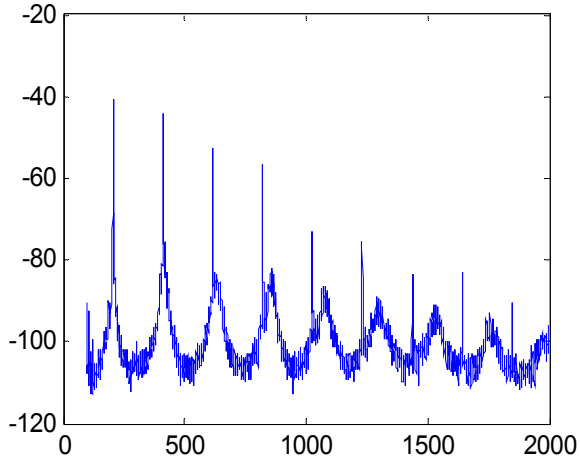


Figure 2b: *Field "A+B", pressure sprectrum (dB vs Hz).*

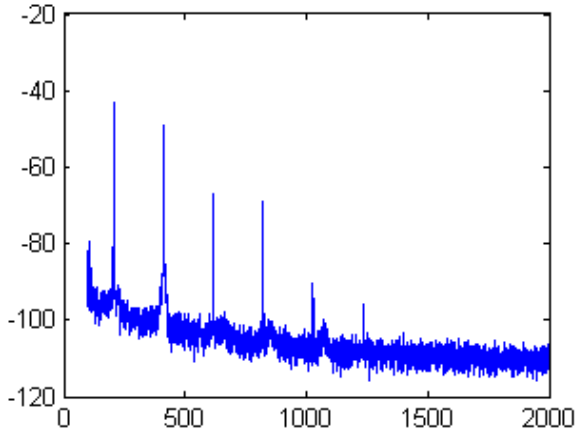


Figure 3a: *Field "A", x-velocity spectrum (dB vs Hz).*

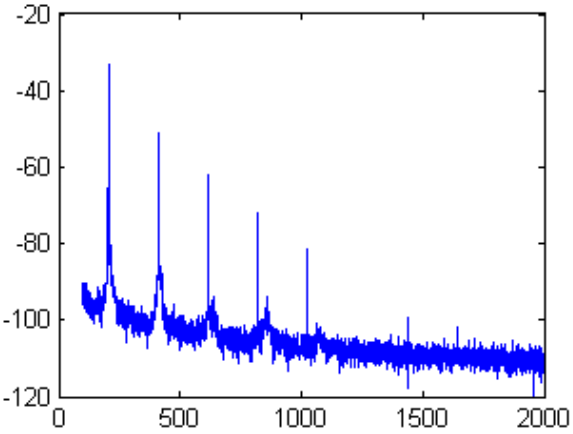


Figure 3b: *Field "A+B", x-velocity spectrum (dB vs Hz).*

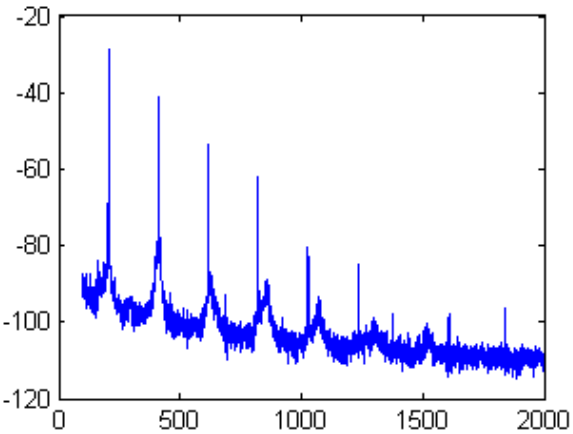


Figure 4a: *Field "A", y-velocity spectrum (dB vs Hz).*

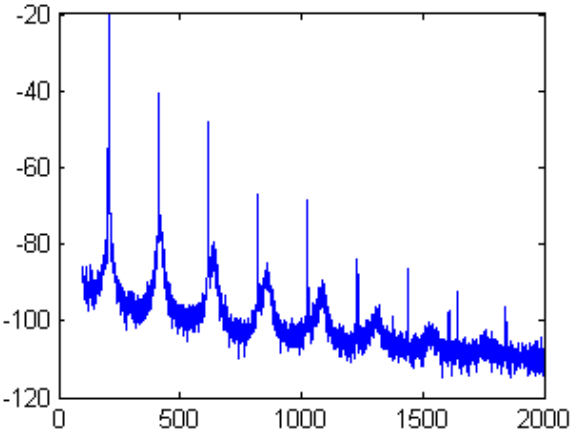


Figure 4b: *Field "A+B", y-velocity spectrum (dB vs Hz).*

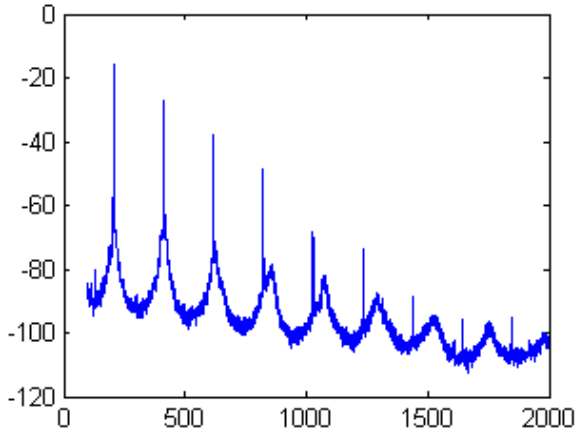


Figure 5a: *Field "A", z-velocity spectrum (dB vs Hz).*

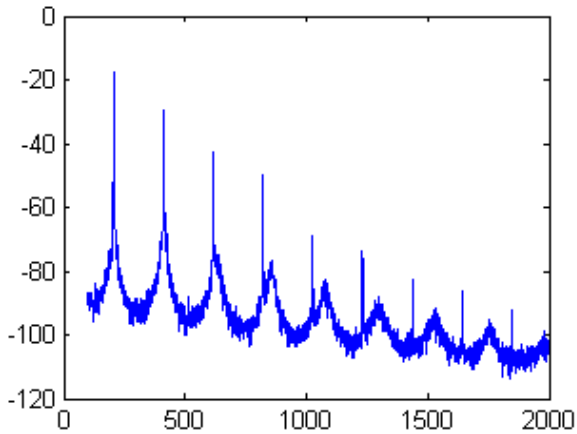


Figure 5b: *Field "A+B", z-velocity spectrum (dB vs Hz).*

It is worthwhile noting that no preliminary calibration of the probe is required for this kind of experimental data analysis as results are used here only for relative comparison and not for measurements of absolute magnitudes and phases of physical related quantities. A calibration of the microflown® USP can be anyway always be done as reported for instance in [4].

Finally the modulus and the phase of the transfer function between signals of the acoustic pressure and the y-component of the velocity for fields "A" and "A+B" are shown in Figures 6a, 6b and respectively in Figures 7a, 7b. As it is evident from the magnitudes comparison the *mitnhame* effect sharpens the peaks around the harmonics frequencies and changes considerably the relative phases between pressure and velocity (from the flat curve about 90° of Fig. 7a, to the graph of Fig. 7b showing very different values for the relative phase of each harmonic). This means obviously that significant variations have to be expected also when measuring the specific acoustic resistance, as defined in Eq. (2).

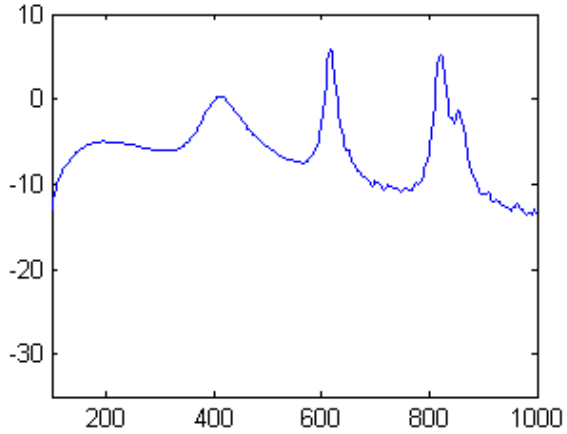


Figure 6a: *Field "A", frequency response between p and the y-velocity component (dB vs Hz).*

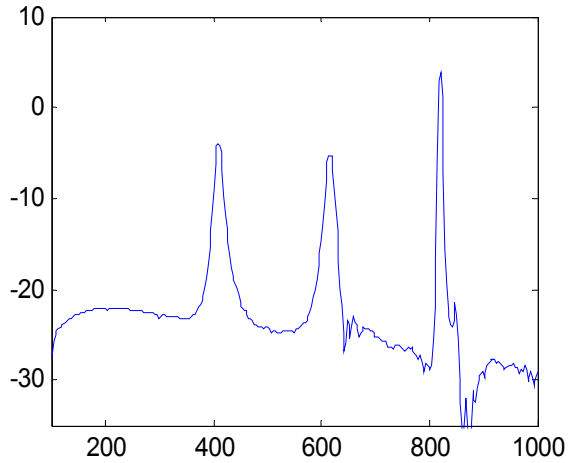


Figure 6b: *Field "A+B", frequency response between p and the y-velocity component (dB vs Hz).*

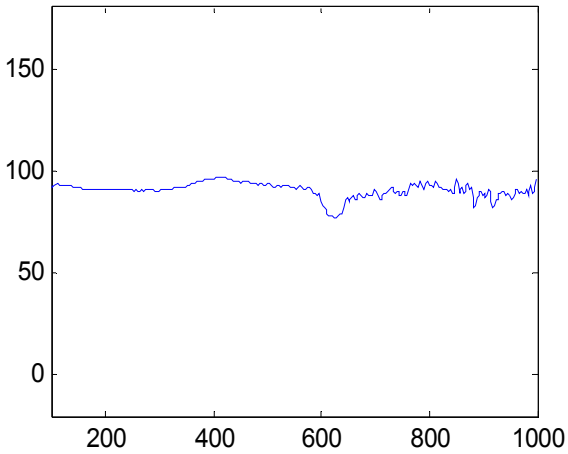


Figure 7a: *Field "A", phase response between p and the y-velocity component (degrees vs Hz).*

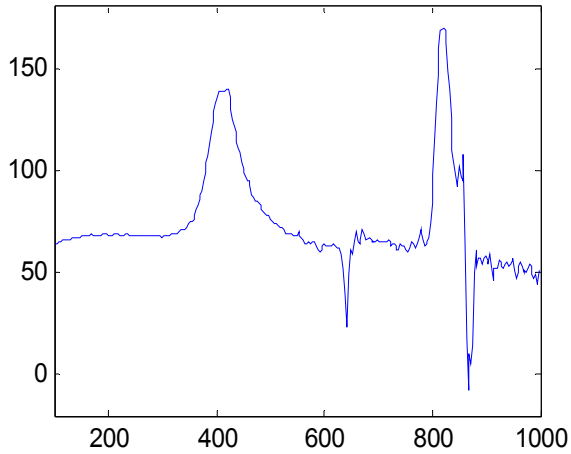


Figure 6a: *Field "A+B", phase response between p and the y-velocity component (degrees vs Hz).*

2. CONCLUSIONS

The theoretical definition of specific acoustic resistance in term of the modulus of active sound intensity has been given for general sound fields. As an immediate consequence of this definition one infer that, according with energetic properties of sound fields as given in [2] and [3], pure resonant fields (i.e. fields where A=0 everywhere) are characterized by an infinite acoustic resistance (open sound energy circuits). This might be a first significant trace aimed to establish a full general analogy of 3-D acoustics in air with electrical circuits theory. However more theoretical analysis is needed for reaching this ambitious but extremely attractive goal. As a preliminary confirmation of the excellence of this approach some measurements were carried out with a microflown® USP monitoring the mitnhame effect between two organ pipes. Measures indicate not only the effectiveness of the used probe for this kind of applications but also that specific acoustic resistance as defined by Eq. (2) (even though not yet explicitly measured in this case) is certainly a very sensible impedance-related indicator for the 3-D analysis of sound fields especially in musical acoustics.

3. REFERENCES

[1] Pierce, A.D., Acoustics, An introduction to its physical principles and applications, ASA-AIP 2nd printing, p. 107, New York 1991,.

[2] Schiffrer, G., Stanzial, D., "Energetic properties of acoustic fields", J. Acoust. Soc. Am., Vol. 96(6), 1994, Sec V, p 3649.

[3] Stanzial, D., Prodi, N., "Measurement of newly defined intensimetric quantities and their physical interpretation", J. Acoust. Soc. Am., Vol. 102(4), 1997, p 2033-2039.

[4] Stanzial, D., Bonsi D., "Calibration of the p-v microflown® probe and some considerations on the physical nature of sound impedance", Proc. Euronoise, Naples, May 19-21, 2003.

SOUND AND VIBRATION FROM A SPECIAL MUSICAL INSTRUMENT: THE MAZZACORATI THEATRE

Lamberto Tronchin[1], Enrico Esposito[2], Valerio Tarabusi[1]

[1]Department of Energetic, Nuclear and Environmental Engineering
University of Bologna
[2]Department of Mechanical Engineering
Polytechnic University of Marche

tronchin@ciarm.ing.unibo.it

ABSTRACT

The sound characterization of theatres represents a common target especially for architects and acousticians. Italian theatres are usually characterized by their special horse-shoe shape, and only acoustical measurement are performed. The "1763 Teatro of Villa Aldrovandi-Mazzacorati", in the town of Bologna, Italy, represents a little, unique baroque wooden theatre, which is quite similar to a huge violin sound-chest. Never measurements were performed in the hall and in the two balconies. Therefore, vibro-acoustical measurements were planned. Experimental acoustical values have been obtained by using of a dummy head (Sennheiser) and a 3D microphone (SoundField), whilst vibrations have been measured by using a Laser Doppler Vibrometer (LDV). Both the techniques allowed the measurements of IRs in listening positions. The results are compared and presented

1. INTRODUCTION

The sound quality perceived by performers and listeners and produced by the soundboard of Musical Instruments and singers is strongly related with the acoustics of the enclosure in which music is played. Many times in theatres the influence of ceilings, walls, and curtains could limit the performance, and very often acoustical chambers are designed in order to improve sound quality.

The 1763 Teatro Mazzacorati in Bologna is a very particular little theatre, built in XVIII Century, in which the dimension and the shape of the hall, and especially the huge quantity of wood on the ceiling, the balcony and the walls, enhance sound properties of string instruments, and is still utilized with its original wooden scene since its foundation. The thin structures of balconies, stuccoes, and the very little stage allow both measurements of room acoustical parameters and vibration of wooden parts of the theatre, in order to correlate sound generation in the stage and natural sound reinforcements of the structure of the theatre, which enhance sound generation of performers exactly like a sound-chest.

1.1. Microphones and sound source

Two different microphones have been utilized during the measurements: a dummy head and a B-format microphone. The dummy head allows the measurements of binaural parameters, like IACC, whilst the 3D microphone allows both the measurements of mono-aural parameters, like reverberation time, both spatial parameters, like LE and LF, and B-format (3D) Impulse Responses. In the first session, acoustical parameters have been measured by meaning of dummy head (Sennheiser) and B-format microphone (SoundField MkV), all connected to an 8 channel-in soundboard (Layla, by Event). A test signal (30 second logarithmic swept-sine) was fed into an omnidirectional loudspeaker (Look Line), and acoustical and vibrational IRs have been measured, by means of Aurora system.

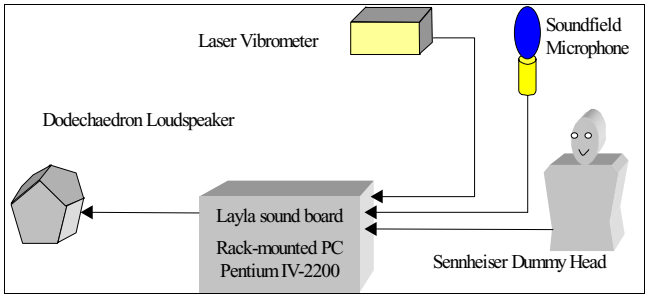


Figure 1: *set-up of instrumentation*

The sound source was put into the stage, whilst measurement points where chosen both in the main hall both on the balcony. In the same time acoustical and velocity measurements were performed, and synchronous 7-channels Impulse Responses were measured.

1.2. Laser Doppler vibrometers

In the following lines a particular type of laser based instrument will presented, i.e. the so-called laser Doppler vibrometer (LDV) [9], [10].

The SLDV may accurately measure surface velocities using interferometric techniques with a maximum frequency in the range of some tens of MHz, and with a lower limit of less than a Hertz. Full-scale highest range is typically 10 m/s with lower ranges in the order of 1 mm/s, corresponding to a displacement of some tens of nanometres.

These features make the LDV an ideal instrument in applications where it is impossible or very difficult to use standard vibration measuring devices, such as accelerometers. For example, this happens with distant structures, a common situation encountered in the musical acoustic sector. Moreover, accelerometers will load the examined structures and may even damage delicate surfaces and, to perform an accurate vibrational analysis, it would require to employ many transducers or to move one all around the tested piece; in both cases time and cost would rise considerably.



Figure 2: *The measurements in the theatre*

As its own name says, this type of vibrometer uses the Doppler phenomenon to remotely acquire vibration velocities: surface vibrations induce a Doppler frequency shift on the impinging laser beam generated by the instrument, and this shift is linearly related to the velocity component in the direction of the laser beam. In this way a linear connection between laser beam frequency variations and velocity values have been established. The obstacle facing us now is that Doppler shifts are usually very small when compared to the laser fundamental frequency, typically 1 part out of 10^8; the only way to appreciate such small quantities is to use interferometry, so that high frequency oscillations are combined and reduced to much lower values that can be dealt with by standard electronics.

A picture of a vibrometer incorporating a Mach-Zender interferometer, together with dummy head and SoundField microphone inside the theatre is presented in Figure 2. Laser emission is split in two equal power beams by beam splitter, and, while one beam is focused on the vibrating object, the "measuring beam", the other one stays in the laser head and becomes the so-called "reference beam".

After being reflected, the measuring beam re-enters the laser head and is recombined with the reference beam. Surface displacement varies the optical path difference between the two laser beams and this results into a phase lag varying with vibration velocity **v**. So a time varying phase difference is established, corresponding to an instantaneous frequency component that follows **v**. This frequency shift is equal to the Doppler shift (f_D) and easily from basic physics the relation between f_D, **v** and source wavelength (λ) is the following:

$$f_D \propto 2v/\lambda \qquad (1)$$

Finally this signal is FM demodulated and a voltage proportional to velocity is sent to the LDV output.

2. ANALYSES OF MEASUREMENTS

The acoustical and vibrational measurements were performed positioning the sound source in the proscenium and the microphones in many points in the audience and balcony. The synchronous measurements allowed comparing acoustical results with velocity frequency responses of wooden moving parts.

An example of Impulse Response measured in the audience area is presented in Figure 3, whilst the same W channel and the LDV Impulse Response are represented in Figure 6.

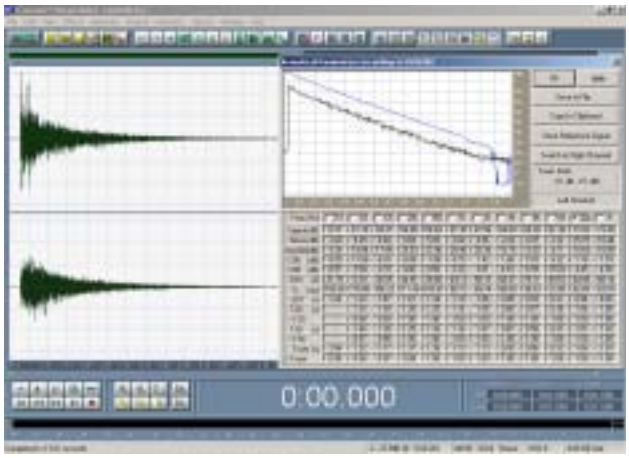


Figure 3*: Acoustical measurements (W e Y channels) in the balcony*

The calculation of all acoustical parameters is obtained post-processing IRs, therefore the measurement of IR allows to fully determine sound propagation. For example, *Lateral Efficiency*, that computes lateral energy perceived in a measuring position, could be calculated as follows:

$$LE = \frac{\int_{25ms}^{80ms} p_{\infty}^2(t)\,dt}{\int_{0}^{80ms} p_0^2(t)\,dt} \qquad (2)$$

Where $p_{\infty}^2(t)$ and $p_0^2(t)$ describe the Impulse Responses measured by means of an omni-directional microphone (W channel) and figure-eight microphone (Y channel) respectively.
The measurements were conducted in 24 positions, and the most important acoustical parameters (RT, Clarity, CT, G, LE, IACC, STI, etc.) were calculated.

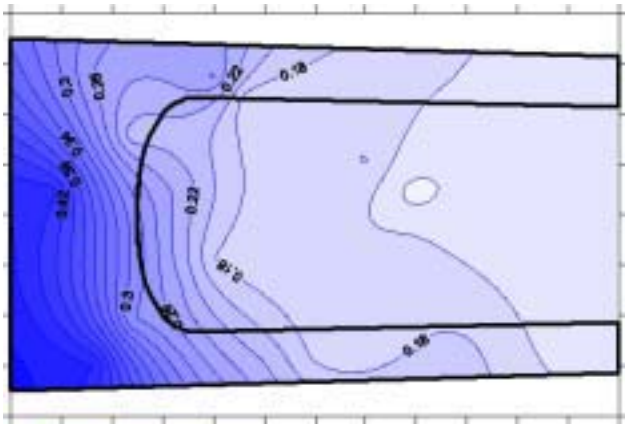


Figure 4 *Map of Lateral Efficiency measured in the Theatre*

Considering the spatial distribution of sound energy, the mapping of parameters reveals that moving away from the proscenium the spatiality enhances, due to increasing of energy reflected by the walls and the ceiling of the balcony. It should be noted that also approaching to the stage Lateral Efficiency has still good value, which means the behaviour of the panels is perceived until the first seats.
The frequency response could be evaluated analysing in the frequency domain the acoustical parameters. Considering the Early Decay Time measured in the theatres, there is evidence that initial reverberation is characterized by more low frequencies than late reverberation.

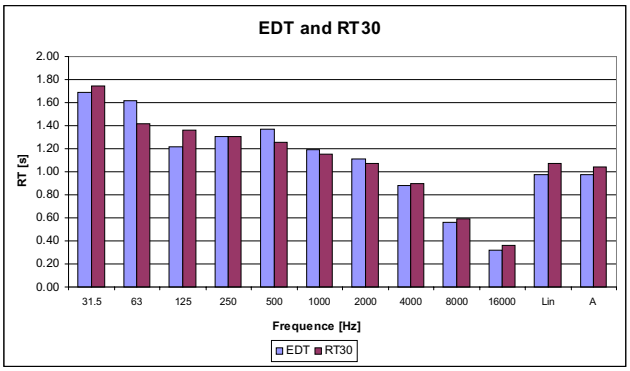


Figure 5 *EDT and RT30 measured in the Theatre*

The subjective impressions of colour and warmness of sound are more related with early reverberation at low frequencies, whilst brightness and liveness of sound with more reverberation at high frequencies. The Mazzacorati theatre seems to have both the characteristics.
Analysing the value of the most important acoustical parameters, the following considerations could be drawn
The reverberation time ranges between 1 s (EDT) to 1.1 s (RT20), i.e. not variability. The sound decay is therefore almost perfect, and the coupling effect of stage-audience that troubles many Opera Houses is here missing. Strength is almost perfectly distributed, as evidence to the very little dimension of the theatre. Clarity (C50 and C80) is settled to perfect values especially for little ensembles (ranging between 2 and 4 dB).

3. SOUND AND VIBRATION

The Mazzacorati theatre represents a unique theatre in Italy. The shape, the dimensions, and especially the materials provoke a spectacular acoustical performance.

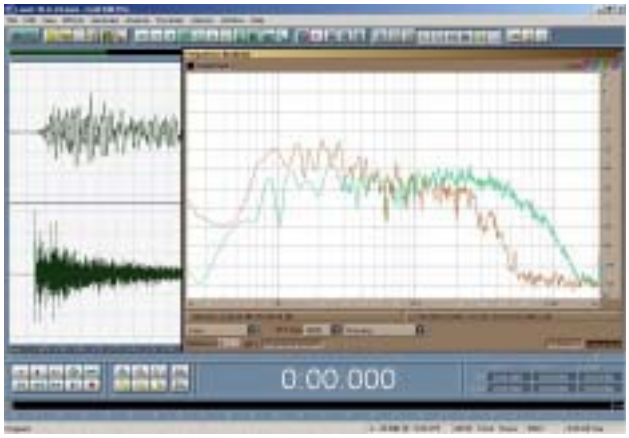


Figure 6: *SoundField (W channel) and Laser (Velocity) Impulse Responses in the balcony: Initial 150 milliseconds*

The interaction between displacement of wooden panels (in the balcony, in the ceiling, in the stage) and sound

quality was analysed measuring Impulse Responses captured with LDV.
The measurements were conducted analysing different vibrating structures, in different points. Special vibrating patterns were found in the balcony, where the floor, the ceiling and the walls together were vibrating when solicited by acoustical stimulus. It is interesting to compare frequency response of velocity IR and pressure IR (in near acoustical field), and find out how the vibrating frequencies are different from the acoustical ones. Considering the frequency response of W (SoundField) and Velocity (LDV) shown in figure 6, the vibration of the balcony is particularly significant until 2 kHz, and below 300 Hz partials are very relevant. It could suggest the balcony acts like an acoustical dipole, being a secondary sound source enforcing sound production generated in the proscenium.

4. CONCLUSIONS

The interaction between sound generation and propagation in such a particular theatre has revealed a strong influence of vibrating parts (like the ceiling and the walls) in the assessment of sound quality in the theatre. The little dimensions of the hall, the huge utilization of wood to build floors and ceilings, the proscenium and balconies, provoked a peculiar acoustical characterisation that makes Mazzacorati theatre close to the sound chest of a string Instrument. The vibrations of wooden part have revealed a consistent frequency response. Nevertheless a throughout study on point impedance in the structures might find out a closer relation between sounds coming from the proscenium and from lateral walls, which are considered very important in the subjective evaluation of sound quality. This might lead the design of new vibrating (wooden) parts in Auditoria, which could give special results on sound quality.

5. ACKNOWLEDGMENTS

The authors wish to thank Prof. Angelo Farina for his help during the setting-up of instrumentation, Prof. Alessandro Cocchi for his help during the discussion of the results, Ramona Loffredo for the precious information regarding the history of the theatre, and Roberto Fiorilli who collaborated during the calculation of physical parameters. Work partially funded by Italian Minister.

6. REFERENCES

[1] Farina, A., Langhoff, A., Tronchin, L.: “Acoustic characterisation of "virtual" musical instruments: using MLS technique on ancient violins” Journal of New Music Research, Vol. 27(4), 359-379, 1998

[2] Farina, A., Tronchin L.: “3D Impulse Response measurements on S. Maria del Fiore Church, Florence, Italy” Proc. 16th International Congress on Acoustics, Seattle, USA, 1998

[3] Ponzi, F., Cocchi, A., Tronchin, L.: **“**Objective surveys concerning the validation of the restoration processes on the pianos of the romantic period: Rossini's Pleyel and other cases” Proc. Colloque acoustique et instruments ancient, Cité de la Musique, Paris, France, 1998

[4] Cocchi, A., Massobrio, D., Tronchin, L.: “Room Acoustics and Cultural Heritage: how to preserve?” Proc. “2nd International Congress on "Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin”, Paris, France, 1999.

[5] Farina, A., Tronchin, L., “On the “virtual” reconstruction of sound quality of trumpets”, Acustica United with Acta Acustica, Vol. 86(4), 737-745, 2000.

[6] Tronchin, L., Farina, A., Pontillo, M., Tarabusi, V.: “The calculation of the impulse response in the binaural technique”, Proc. 7th International Congress on Sound and Vibration (ICSV), Garmisch, Germany, 2000

[7] Tronchin, L., Cocchi, A., Consani, C.**:** **“**Vibrational and acoustical radiation in two Italian harpsichords of XVII century”*,* Proc. International Symposium on Musical Acoustics (ISMA2001) Perugia, Italy, 2001

[8] Cocchi, A., Tronchin, L., Farina, A., Cesare-Consumi, M.: “A comparison between some measurement techniques in the Foligno auditorium” Proc. Auditorium Acoustics: Historical and contemporary design and performance, IOA, Imperial College, London, UK, 2002

[9] Esposito, E., “Non-invasive characterization of musical instruments and amplification systems by scanning Laser Doppler Vibrometry”, atti seminario AES “Fisica degli strumenti musicali”, pp.157-215, TOP AUDIO, Milano, 22 settembre 2002.

[10] Castellini, P., Revel, G. M., Tomasini, E. P. “Laser Doppler Vibrometry: a review of advances and applications”, The Shock And Vibration Digest. vol. 30-6, pp. 443-456, 1998.

Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

LARGE-BANDWIDTH MEASUREMENT OF ACOUSTIC IMPEDANCE USING TWO MICROPHONES AND FOUR CALIBRATIONS

Maarten van Walstijn and D Murray Campbell

School of Physics
University of Edinburgh
maarten@ph.ed.ac.uk

ABSTRACT

A new method for measuring the acoustic impedance of tubular objects, that gives accurate results for a wide range of frequencies, is discussed. Similar to various previous methods, the experimental setup involves a cylindrical measurement duct embedded with two microphones. The impedance is deduced from the microphone signal ratio and three coefficients that are pre-determined by means of calibration measurements with four precisely manufactured cylindrical tubes. The calibration procedure is based on the simple mathematical relationships between the impedances of the calibration tubes. Since these relationships are independent of the temperature and the the thermodynamic constants of air, the method does not require accurate knowledge of the propagation constant. Accurate results can be obtained at frequencies up to 20 kHz. The attenuation constant for plane waves in a cylindrical tube is measured to be less than 1.4% different from what is predicted by theory.

1. INTRODUCTION

Musically useful information about the internal bore of musical wind instruments can be obtained by measuring its response to an acoustic input signal. For wind instruments, the dominant resonances of the air column usually fall below 3 kHz, which is one of the reasons why most measurement systems have not been designed to measure accurately at higher frequencies. In recent years however, the experimental approach has been combined with theory with the objective to reconstruct bore profiles from measured response data (see for example [1, 2, 3, 4]). For this type of application, it is advantageous to be able to measure at higher frequencies, because the axial resolution of the reconstructed bore profile is largely determined by the frequency bandwidth of the acoustic data. In that respect, the problem of reconstruction of an internal bore profile from band-limited acoustic data is similar to the problem of reconstruction of a non-smooth time-domain signal from a truncated set of Fourier components. This limitation does not affect the reconstruction of smooth bore profile objects, but presents a significant problem when reconstructing any abrupt changes and discontinuities in diameter. An increased measurement bandwidth is also advantageous for studying small components of wind instruments, such as bassoon crooks, because their resonances often occur at frequencies much higher than the main resonance frequencies of a complete instrument.

One of the key features of a suitable experimental method is that it should be insensitive to the influences of local non-propagating modes, which are typically significant for short wavelengths. For this reason, the classic experimental set-up used in two-microphone methods (see Fig. 1) is employed, because it makes it possible to separate planar-mode information from higher-mode information, and to fully take into account the influence of non-propagating modes via calibration.

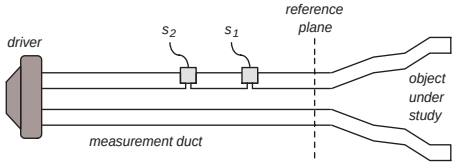


Figure 1: *Sketch of the experimental set-up used in the two-microphone approach. The air inside the measurement head is excited using a compression driver. The impedance of the object connected to the measurement head is derived from the microphone signals s_1 and s_2.*

For a full calibration of this system, three closed tubes of different lengths can be employed, which is the basic principle of the "Two-Microphone-Three-Calibration" (TMTC) method developed by Gibiat and Laloë [5]. The main disadvantage of this method is that the calculation of the impedance of closed tubes requires precise knowledge of the propagation constant, which is a frequency-dependent complex-valued parameter that depends on a number of air constants, such as the speed of sound, air density, and the viscosity coefficient. These constants are temperature-dependent and it is usually difficult to obtain the precise set of constants of the actual environment in which the experiments take place.

In this paper we propose a method that measures the planar mode input impedance at frequencies below the first cut-on of the measurement duct. It is shown that accurate measurements are possible up to 20 kHz, using a new calibration method that does not depend on accurate knowledge of the propagation constant.

2. THE METHOD

2.1. Theory

The experiments are carried out under the following conditions and assumptions. The air inside the measurement duct is excited only at frequencies below the cut-on frequency of the first higher mode. There is no significant coupling between any possible non-propagating modes at or beyond the reference plane and those excited at the microphone positions, because the distance between

the reference plane and each of the microphones is signficantly greater than the duct diameter. Furthermore, the wavelength is significantly larger than the dimensions of the microphone holes in the duct wall at all frequencies within the measurement bandwidth. Therefore the pressure is assumed to be constant over the surface of the membranes of the microphones, and the way the microphones are coupled to the acoustic field may be assumed to be independent of the impedance load at the reference plane. Finally, the oscillations in the duct are limited to the linear regime.

Under these conditions, the microphone signals s_1 and s_2 are linear functions of p and U at the reference plane:

$$s_1 = \alpha_1 p + \beta_1 Z_c U, \quad (1)$$

$$s_2 = \alpha_2 p + \beta_2 Z_c U, \quad (2)$$

where α_1, α_2, β_1, and β_2 are coefficients that depend on the manner in which waves propagates between microphones and the reference plane, as well as on the characteristics of the microphones, but are independent of the boundary condition at the reference plane. Z_c is the characteristic impedance.

For passive terminations, one may define the impedance at the reference plane, which is the impedance load due to connecting the object to the measurement duct, as:

$$Z = \frac{p}{U}. \quad (3)$$

Following the general two-microphone approach, we define a new variable y as the ratio of the two microphone frequency-domain signals. From equations 1, 2 and 3, one then obtains:

$$y = \frac{s_2}{s_1} = \frac{\alpha_2 Z + \beta_2 Z_c}{\alpha_1 Z + \beta_1 Z_c}, \quad (4)$$

or

$$\frac{Z}{Z_c} = A\left(\frac{y - B}{y - C}\right), \quad (5)$$

with

$$A = -\beta_1/\alpha_1, \quad B = \beta_2/\beta_1, \quad C = \alpha_2/\alpha_1. \quad (6)$$

Equation 5 states that the normalised impedance (the ordinary impedance divided by the characteristic impedance of the measurement duct) of the object under study can be determined from a measured value of y and three coefficients A, B, and C.

The response of the object under study may also be described as the ratio R of the backward- and forward-propagating planar mode waves at the reference plane, commonly known as the (planar mode) reflection coefficient. If no non-propagating modes are excited at the reference plane, R is related to the normalised impedance $\bar{Z} = Z/Z_c$ by:

$$R = \frac{\bar{Z} - 1}{\bar{Z} + 1}. \quad (7)$$

2.2. Calibration

In the present study, the calibration coefficients A, B and C are determined by means of four calibration measurements. In the first calibration measurement, the duct is closed off at the reference plane, and the microphone ratio signal y_0 is measured. The hard wall that closes the duct off presents a zero-admittance termination. In that case the denominator in equation 5 must be zero, i.e.

$$C = y_0. \quad (8)$$

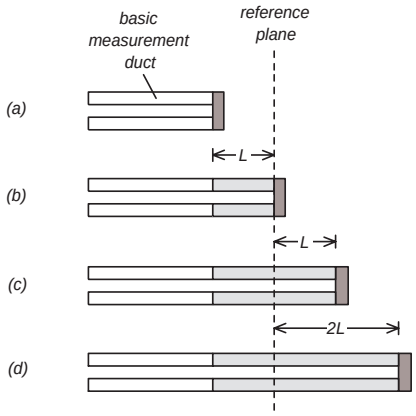


Figure 2: *Bore configurations for the calibration measurements. (a) negative-length tube, realised by closing off the basic measurement duct. (b) Hard wall at the reference plane, realised by connecting a closed tube of length L to the basic measurement duct. (c) positive-length tube, realised by connecting a closed tube of length $2L$ to the basic measurement duct. (d) double-length tube, realised by connecting a tube of length $3L$ to the basic measurement duct.*

For the determination of coefficient B, two tubes of the same cross-section as the measurement duct, one of length L and one of length $-L$ are employed. The negative length can be realised by defining the position of the reference plane at a distance L from the basic measurement duct (see Fig. 2). The microphone ratio signals measured with the positive- and negative-length tube are y_p and y_n, respectively. The normalised planar mode impedance of a closed tube of length L is $\bar{Z} = \coth(\Gamma L)$. Since

$$\coth(-\Gamma L) = -\coth(\Gamma L), \quad (9)$$

we have

$$\bar{Z}_p + \bar{Z}_n = 0, \quad (10)$$

where $\bar{Z}_p$ and $\bar{Z}_n$ are the normalised impedances of the positive- and the negative-length tube, respectively. Evaluation of the impedances in equation 10 using equation 5 gives

$$A\left[\left(\frac{y_p - B}{y_p - C}\right) + \left(\frac{y_n - B}{y_n - C}\right)\right] = 0. \quad (11)$$

After substitution of equation 8, the value of coefficient B is obtained as:

$$B = \frac{2y_p y_n - y_0\,(y_p + y_n)}{(y_p + y_n) - 2y_0}. \quad (12)$$

For the determination of the coefficient A, another closed tube is connected to the measurement duct, this time of length $2L$, and the microphone ratio y_2 is measured. The reflection coefficient of a closed tube of length L is $R = \exp(-2\Gamma L)$, thus we have

$$R_2 = R_p^2, \quad (13)$$

where R_p and R_2 are the reflection coefficients of the closed tubes of length L and $2L$, respectively. Substitution of equation 7 yields:

$$\bar{Z}_p^2 - 2\bar{Z}_p\bar{Z}_2 + 1 = 0, \quad (14)$$

where $\bar{Z}_2$ denotes the normalised impedance of the double-length tube. After evaluation of the impedances $\bar{Z}_p$ and $\bar{Z}_2$ using equation 5, and substitution of equations 8 and 12, one obtains:

$$A^2 = \left[\left(\frac{y_2 - y_0}{y_p - y_n}\right)(2y_0 - y_p - y_n)^2\right] \times [y_p y_2 + 3(y_p y_0 + y_2 y_n) \ldots - 4(y_p y_n + y_2 y_0) + y_0 y_n]^{-1}. \quad (15)$$

Energy must be conserved, thus $|R_p| < 1$, on basis of which it can be decided which of the two roots $\sqrt{A^2}$ or $-\sqrt{A^2}$ is the correct one. Note that the correctness of the three assumptions upon which the calibration procedure relies, namely that equations 8, 10, and 13 are true, is independent of the value of the propagation constant.

Finally, we note that in order to minimise errors within the measurement bandwidth, the basic calibration length should be chosen such that the first singular frequencies respecively related to the microphone distance d and the calibration length L cooincide, i.e. $L = d/2$.

2.3. Apparatus

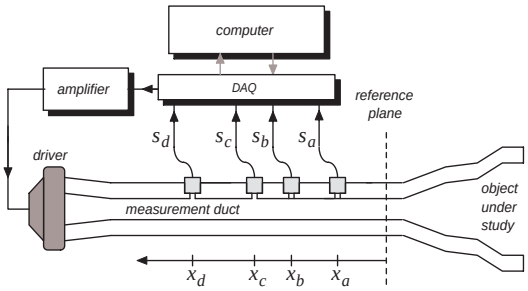


Figure 3: *Schematic drawing of the experimental apparatus.*

Fig. 3 shows a schematic drawing of the experimental apparatus. The air inside the measurement duct and the object under study is brought into vibrational motion by the compression driver, and the pressure in the duct can be measured at four different positions along the duct axis, using cylindrically-shaped microphones that are mounted in the duct wall. The position of the reference plane, and therefore the distance x_a, depends on the calibration length L. The other microphones are positioned at $x_b = x_a + 14$ mm, $x_c = x_a + 21$ mm, and $x_d = x_a + 49$ mm. With these four microphones, measurements with six different microphone distances can be carried out (see Table 1). Only two microphones are actively used per individual measurement session.

Empirically it was found that when using a microphone distance d, the effective measurement bandwidth within which accurate results can be obtained is approximately $[0.3 - 0.81]f_c$, where $f_c = c/(2d)$ is the first singular frequency. The effective bandwidths of the different possible microphone combinations of the experimental set-up are listed in Table 1. The radius of the measurement duct is 4.9 mm; the cut-on of the first asymmetric duct mode is 20.5 kHz.

3. RESULTS

For testing the accuracy of the method, one needs to measure an object of which the impedance is known with relatively high certainty. For the TMFC method, the most suitable test object is probably a cylindrical tube with closed end and of the same cross-section as the measurement duct. In our experiments we used such a tube of 128 mm length.

Table 1: Possible microphone combinations in the experimental set-up and their measuring properties. BW is the measurement bandwidth. See text and Fig. 3 for an explanation of the other quantaties.

combin.	s_1	s_2	d (mm)	f_c (kHz)	BW (kHz)
I	s_b	s_c	7	24.6	$7.4 - 20.0$
II	s_a	s_b	14	12.3	$3.7 - 10.0$
III	s_a	s_c	21	8.2	$2.4 - 6.7$
IV	s_c	s_d	28	6.1	$2.0 - 5.0$
V	s_b	s_d	35	4.9	$1.5 - 4.0$
VI	s_a	s_d	49	3.5	$1.0 - 2.9$

Fig. 4 shows the measurement results obtained with microphone combination I. From the reflection coefficient magnitude curve, it can be seen that the measurement error is primarily structural, i.e. the error exhibits an oscillatory rather than noisy pattern. The discrepancy between theory and experiment does not exceed 0.2 dB.

The method can also be employed to measure the attenuation of plane waves in a cylindrical duct. The attenuation constant is the real part of the propagation constant, and can be determined from the reflection coefficient using the formula:

$$\alpha = -\frac{log\,|R|}{2L}. \quad (16)$$

The oscillating curves in Fig. 5 show the attenuation constant as a function of frequency, as deduced directly from the reflection coefficient of a 404 mm long cylindrical tube measured at frequencies between 1 and 20 kHz. The results obtained with six different measurement sessions and microphone combinations are plotted. It is known from theory that the attenuation is a monotonically increasing function of frequency [6]. Therefore, a smoothly rising function, in a form that follows from theory, was fitted to the data (solid black line). This results in a better estimate of the attenuation constant, one that is free of the spurious oscillations due to calibration errors.

For comparison, the attenuation constant as calculated from theory is also plotted (solid grey line). To this end, we used the formula:

$$\alpha = \alpha_w + \alpha_c + \alpha_m, \quad (17)$$

where α_w represent the losses associated with visous drag and heat conduction at the tube wall, α_c indiactes the classical losses associated with plane wave propagation, and α_m is the attenuation coefficient associated to molecular relaxation processes. Pierce [6] gives formulas of each of these quantities. The discrepancy between the the fitted curve α_f and the theoretical curve is less than 1.4% at all frequencies up to 20 kHz. The losses associated with classical and molecular effects are often neglected, but as can be seen from Fig. 5, this would in our case result in a significant underprediction at the higher frequencies.

4. CONCLUSIONS

A new method for calibrating a measurement set-up with two microphones fitted in a cylindrical measurement duct has been pre-

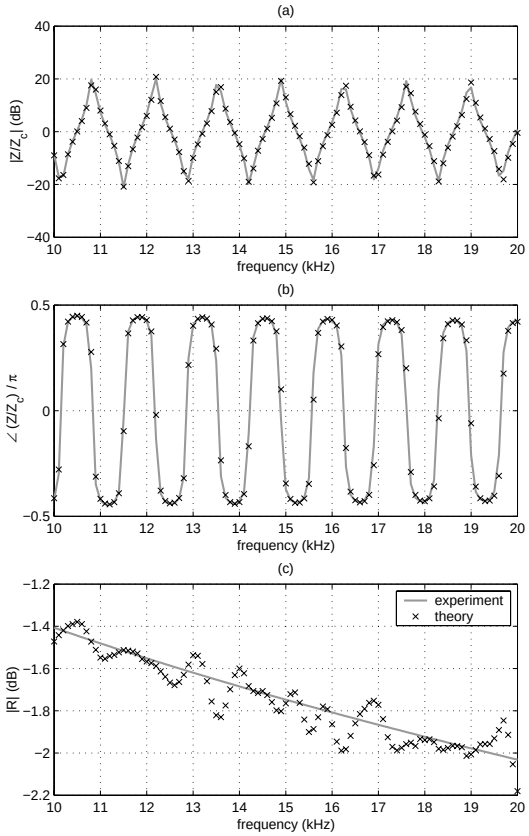


Figure 4: *Response data measured with the short cylindrical tube of equal cross-section, using microphone combination I. (a) Impedance magnitude (normalised). (b) Impedance phase. (c) Reflection coefficient magnitude.*

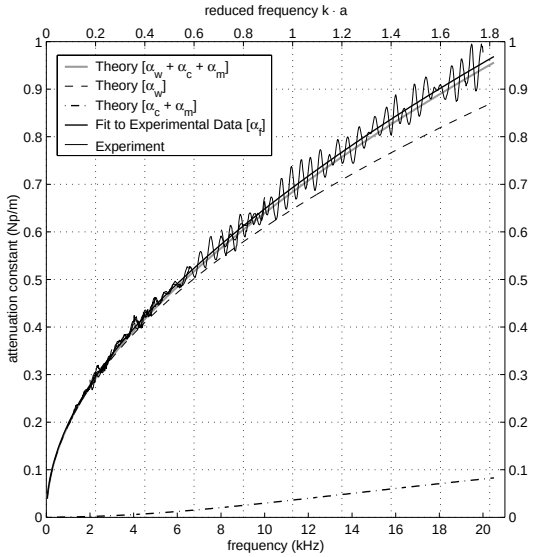


Figure 5: *The attenuation coefficient associated with plane-wave propagation in a cylindrical duct. The thin, oscillating curves indicate the experimental results obtained with six different microphone combinations.*

sented. A major advantage in comparison with many previous methods is that it does not require knowledge of the propagation constant. The calibration is completely general for a given constant temperature, and is relatively insensitive to ambient noise. Proximity effects and other influences related to the microphones are automatically calibrated out. Accurate experimental results have been obtained up to 20 kHz, which is just below the cut-on frequency of the anti-symmetric mode of the measurement duct.

The method has been applied to experimentally determine plane wave losses in a cylindrical tube. The measured attenuation constant has been shown to be within 1.4% of the coefficient predicted by theory that includes the contributions of boundary layer losses, classical plane wave losses, and molecular relaxation processes.

A more detailed description of the TMFC method, including the application to tubular objects with an entry diameter that differs from te measurement duct diameter, can be found in [7].

5. REFERENCES

[1] N. Amir, U. Shimony, and G. Rosenhouse, "Discrete model for tubular acoustic systems with varying cross section - the direct and inverse problems. Part I: Theory," *Acta Acustica*, vol. 81, pp. 450–462, 1995.

[2] D. B. Sharp, *Acoustic pulse reflectometry for the measurement of musical wind instruments*, Ph.D. thesis, Department of Physics and Astronomy, University of Edinburgh, 1996.

[3] J. A. Kemp, *Theoretical and experimental study of wave propagation in brass musical instruments*, Ph.D. thesis, University of Edinburgh, 2002.

[4] W. Kausel, "Bore reconstruction from measured acoustical input impedance; equipment, signal processing, algorithms and prerequisites," in *Proc. 2001 Int. Symposium on Musical Acoustics*, Perugia, Italy, 2001, pp. 373–378, The Musical and Architectural Acoustics Laboratory FSSG-CNR, Venezia, Italia.

[5] V. Gibiat and F. Laloë, "Acoustical impedance measurements by the two-microphone-three-calibration (tmtc) method," *J. Acoust. Soc. Am.* , vol. 88, pp. 2533–2544, 1990.

[6] A. D. Pierce, *Acoustics - An Introduction to its Physical Principles and Applications*, Acoustical Society of America, New York, 1994.

[7] M.O. van Walstijn, D. M. Campbell, J. Kemp, and D. B. Sharp, "Wideband measurement of the acoustic impedance of tubular objects," Submitted to Acta Acustica.

REFRACTOSCOPIC VISUALISATION OF SOUND IN MUSICAL INSTRUMENTS

Lothar Zipser and Heinz Franke

ZAFT, Centre of Applied Research and Technology at HTW University of Applied Sciences, Dresden, Friedrich-List-Pl. 1, D-01069 Dresden, Germany
zipser@et.htw-dresden.de

ABSTRACT

Laser interferometers prove to be multifunctional devices for the investigation of musical instruments. With interferometers the vibration of the surface of musical instruments and their influence on the tone can be measured and visualized [1], as well as the generation of air borne sound in wind instruments and the distribution of the sound inside and outside the instruments. Interferometers allow investigations of the mode of operation of musical instruments, their mechanical quality and they can indicate manufacturing errors or defects by abrasion. The latter is important for the reconstruction of valuable historical musical instruments. However, direct statements about the musical quality of instruments based on interferometric measurements are not known so far.

This paper concerns with the interferometric measurement and visualisation of the generation and propagation of sound from wind instruments, e.g. organ pipes, bugles and trumpets.

1. INTRODUCTION

The problem of sound generation and emission in musical instruments has been investigated by several researchers for a long time with different results. The non-interacting measurement and visualisation of acoustic waves in air is solved only in part so far. Acoustic waves of high intensity can be seized and visualised by Schlieren methods, holographic methods, laser-speckle methods or with refraction tomography. However, all these methods often give no satisfying results in the cases of low audio sound intensity, high demands of spatial resolution or small measuring volume, as inside of musical instruments.

A new method for interferometrically measuring and visualising acoustic sound waves and vortexes in air [2] eliminates these problems to a large extent. There, the measuring effect is based on the phase difference between two interfering laser beams, caused by a variation of the optical refractive index *n(x,y,z,t)* which is determined by acoustic sound or by vortexes in air. Possibilities for the application of this method for measuring and visualising acoustic sound inside and outside of musical instruments are demonstrated. A commercially available scanning Laser-Doppler vibrometer is used as measuring system.

2. MEASUREMENT PRINCIPLE

Mach-Zehnder Doppler interferometers (see Figure 1) can be used for measuring displacement and velocity of vibrations. The beam splitter BS1 splits the laser beam into measuring and reference beams. The reference beam reaches BS3 via a mirror. The measuring beam hits the reflector at the end of the section of measurements, is reflected there and then guided by BS2 towards BS3. There, measuring and reference beams interfere and result in a light signal with variable intensity *I*.

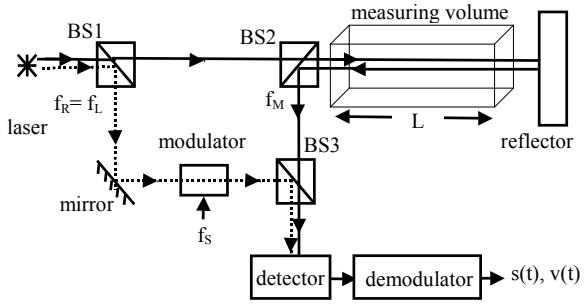


Figure 1:
Mach-Zehnder Doppler interferometer with a measuring volume influenced by noise

If the reflector is moved with a velocity *v*, then, because of the Doppler effect, the frequency f_M of the reflected measuring beam is different from the primary frequency $f_L = f_R$ of the laser. Behind BS3, the resulting signal is received by the detector as a frequency signal. For identifying the actual direction of the velocity *v*, the reference beam with the frequency f_R is sent through a modulator, where a shift frequency f_S is superposed. Behind BS3 the light intensity is with $\Delta f = f_M - f_R$ [2]

$$I(f_S, \Delta f, t) = \frac{I_0}{2}(1 + \cos(2\pi (f_S \pm \Delta f)t)) \quad (1)$$

After a demodulation, amount and direction of both velocity $v(t)$ and displacement $s(t)$ of the reflector, e.g. a violin body, are recognisable.

2.1. Vibration measurement

Modern types of interferometers according to Figure 1 are scanning Laser-Doppler vibrometers (LDV). They scan the surface of a vibrating object pointwise and present the

results as animated frames (Figure 2). This mode of operation is often used for investigating the vibration behavior of musical instruments.

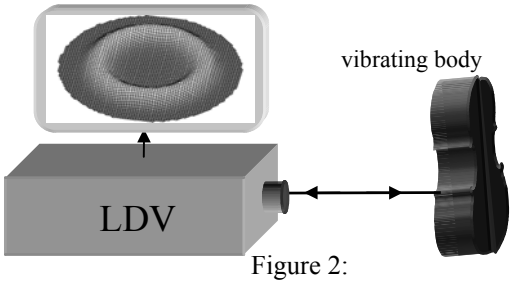


Figure 2:
Measurement of the vibrations of the body of a musical instrument

2.2. Sound measurement

For measuring and visualising temporal and spatial pressure fluctuations $p(x,y,z,t)$, as they occur in sound waves, the reflector of the set-up in Figure 1 must be absolutely rigid. The sound pressure changes $p(x,y,z,t)$ result in density fluctuations $\rho(x,y,z,t)$ which are related, in connection with the Gladstone-Dale constant K, with fluctuations of the refractive index $n(x,y,z,t)$ [3].

If the pressure varies by $\Delta p(x,y,z,t)$, then it causes proportional changes of the optical refractive index $\Delta n(x,y,z,t)$ and, consequently, changes the optical length of the measuring volume in x-direction by $\Delta l(x,t) = L\,\Delta n(x,t)$. Now, an interferometer according to Figure 1 senses, in the case of phase demodulation, a *virtual* reflector displacement, $s(x,t) = L\Delta n(x,t)$ although the reflector itself is immovable. This virtual displacement is

Equations should be placed on separate lines and numbered:

$$\breve{s}(y,z,t) = \int_L \Delta n(x,y,z,t)\,dx \qquad (2)$$

In the case of frequency demodulation, a *virtual* reflector velocity

$$\breve{v}(y,z,t) = \int_L \dot{n}(x,y,z,t)\,dx \qquad (3)$$

is measured. Figure 2 shows an arrangement for measuring and visualising acoustic waves with a scanning LDV using a rigid reflector. Symbolically, a violin is used as a sound source. Assumed is a harmonic wave that spreads in negative y-direction, i.e. orthogonal to the measuring beam. Within the measuring volume the refractive index is, in this case, a time- and space-dependent function $n_1(y,z,t)$. In real measurements the acoustic wave is not ideally shaped along the measuring volume. According to equations (2) and (3) all fluctuations of the refractive index $\Delta n(x,y,z,t)$ along the laser beam contribute to the measuring result. Therefore an optimised position of the sound source is required.

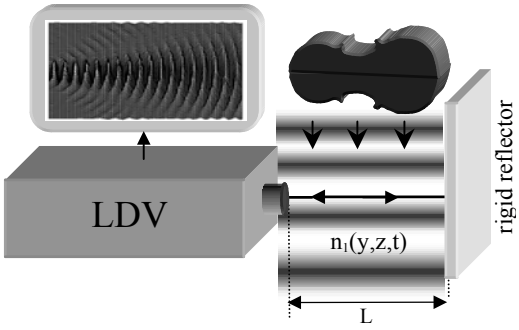


Figure 3:
Measurement of sound emitted from a musical instrument

2.3. Measurement of vortexes

Similar to sound, vortexes in air generate fluctuations of pressure $\Delta p(x,y,z,t)$ and, consequently, fluctuations of the refractive index $\Delta n(x,y,z,t)$. As it is known, vortexes in air are used to generate sound in wind instruments. These facts allow to investigate, with the same method, the generation of sound by vortexes and, simultaneously, it's propagation inside and outside the instruments. A typical example is the generation of sound in organ pipes immediately at the labium (see Figure 4).

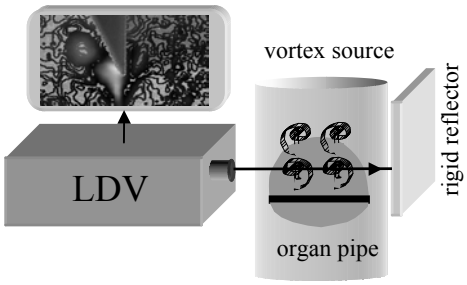


Figure 4:
Measurement of vortexes at the labium of an organ pipe

3. MEASUREMENTS ON ORGAN PIPES

With the methods described above organ pipes were investigated. Figure 5 shows, as an example [1], the unrolled vibration pattern of the body of a 100 years old thin-walled organ pipe at a frequency of 320 Hz. The investigations proved that a coupling exists between the vibration modes of the air column inside the organ pipe and the vibrations of the pipe body. Both vibrating components contribute to the resulting acoustic wave.

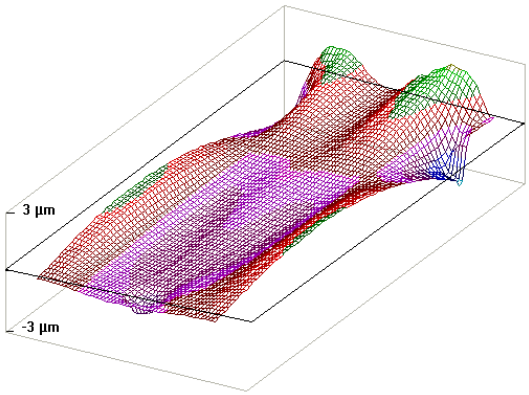


Figure 5:
Unrolled vibration pattern of a thin-walled organ pipe

The wall thickness of this pipe is 0,7 mm. As a rule, the thinner the wall, the more the inner air column stimulates the wall to vibrations. In the majority of cases the eigenfrequencies of column and wall are not the same and influence each other by feedback [1].

To investigate the generation of sound in organ pipes and it's propagation inside the pipe body a model pipe with dimensions analogous to a natural pipe was built. Figure 6 shows the active model pipe with the sound inside and outside the pipe body according to the set-up of Figure 3.

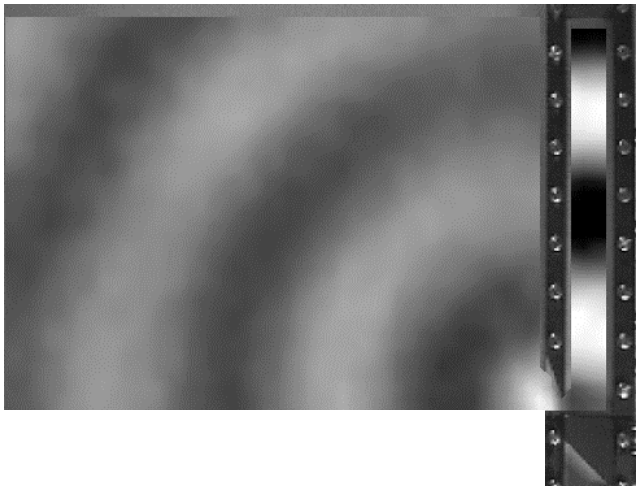


Figure 6:
Visualization of sound inside and outside the body of a blown organ pipe model

Positive and negative pressure amplitudes are presented as light and dark shades, respectively. Clearly recognizable by their contrast are the high sound amplitudes within the pipe on the right hand and the lower amplitudes outside, the latter decreasing proportionally to the distance from the pipe mouth. Obvious is also the phase difference of $\lambda/2$ between the wave periods inside and outside the pipe.

Increasing the supply pressure of the pipe, different wave modes inside the pipe can be excited and visualized.

Figure 7 shows a vortex distribution near the labium of the model pipe. On the right hand is the acoustic resonator of the pipe, left is the free space. The roughness at the left labium edge in Figure 7 results alone from the scanning pitch. The different gray tones represent the varying amplitudes of the generated sound. Along the outer edge of the labium a sequence of vortexes is fading away. As is expected, the appearance of every single low-speed vortex is connected with the generation of a new period of the acoustic wave, which propagates with sound velocity. In this way the transformation of the vortex energy into sound energy can be observed.

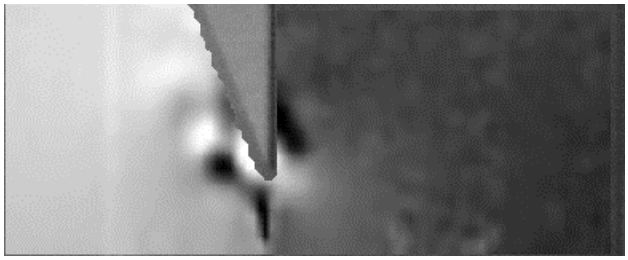


Figure 7:
Visualized vortexes near the labium of a model organ pipe

4. MEASUREMENTS ON BUGLES AND TRUMPETS

Another purpose of our investigations of wind instruments was to find the characteristics of sound propagation in the interior of wind instruments with dominating straight or curved shapes, respectively, e.g. trumpets and bugles. Models were manufactured with predominantly straight and curved elements, whose dimensions were taken from original musical instruments and reduced by means of the Helmholtz number. The models were acoustically excited by a small loudspeaker at the inner ends of the tubes.

In Figure 8 the sound distributions for a low and a higher frequency within the instruments are shown. The line diagrams present the sound pressure intensities along the central axes. In the first measurements, no standing waves were found. On the contrary, the wave periods are wandering from the input towards the outlet, for low frequencies as well as for higher ones. This can be seen impressively in the animated frames. In the outlet region of the bugle model appeared, beginning from medium frequencies, a vibration mode orthogonal to the propagation direction of the emitted sound. This effect is the object of further investigations.

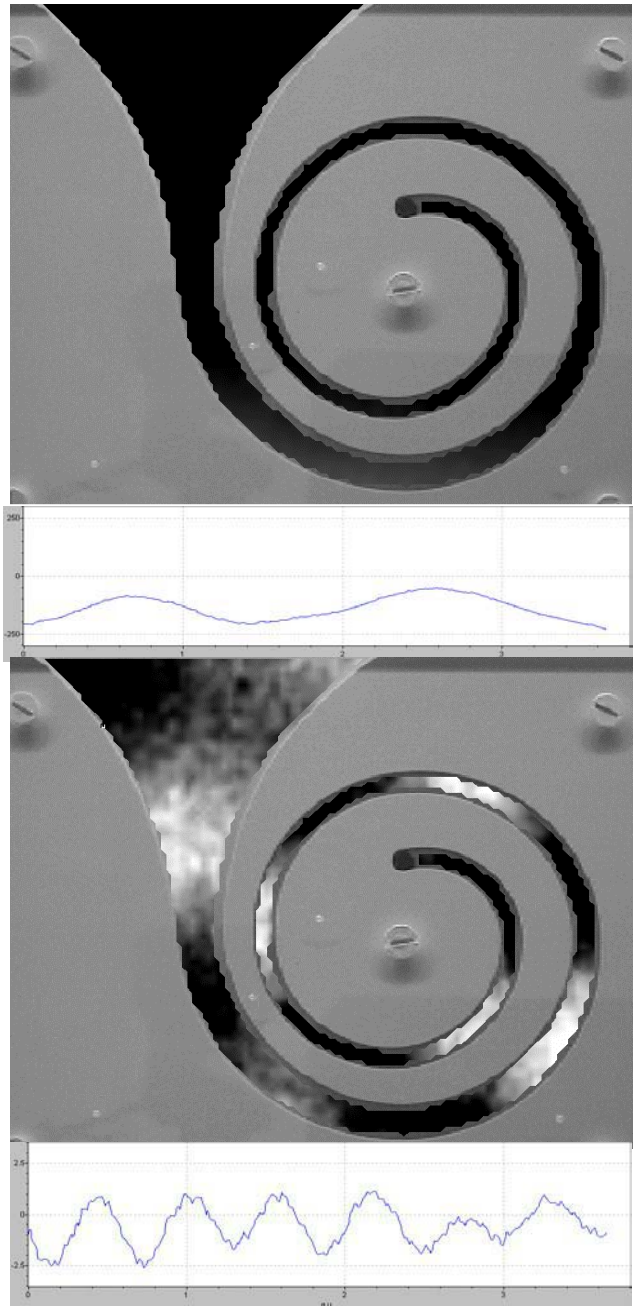


Figure 8:
Visualized sound waves in a bugle model

Measuring results for the trumpet model are shown in Figure 9. Different from the bugle, at low frequencies standing waves appear. Curves with radii significantly smaller than λ/4 seem to act as reflecting elements. The behavior at higher frequencies is similar to the bugle.

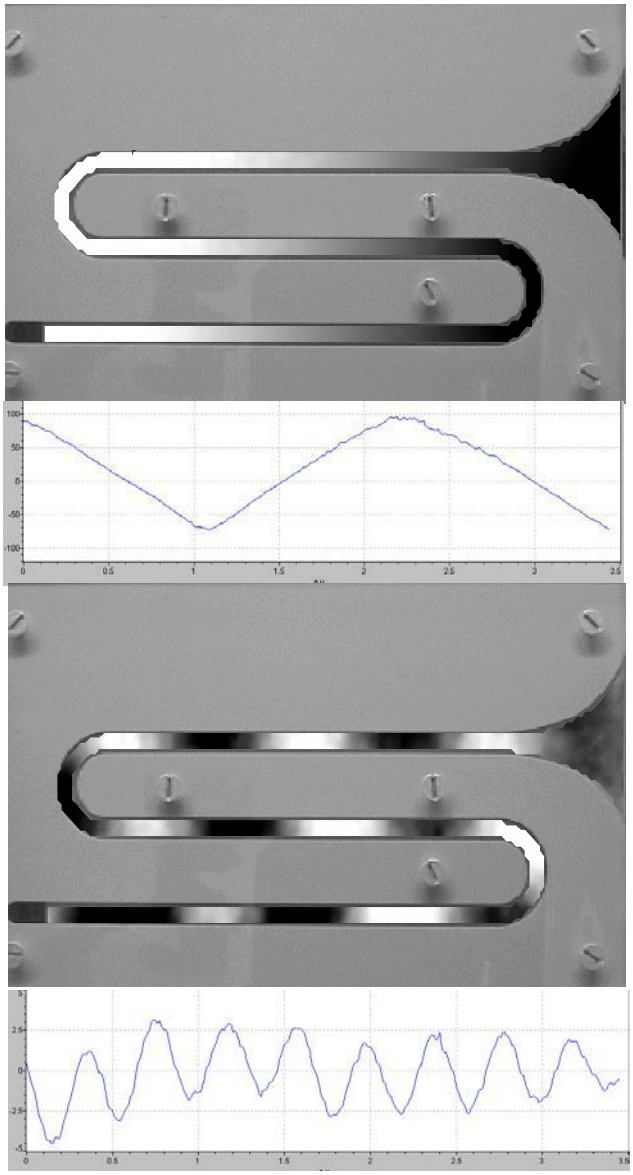


Figure 9:
Visualized sound waves in a trumpet model

5. REFERENCES

[1] Zipser, L. ; Franke, H. Vibration behaviour of organ flue pipes; Proc. of the Institute of Acoustics, Vol. 19 Part 5 (1997) Book 2 pp 359 – 366,ISMA'97 Conference ; University of Edinburgh,UK.

[2] Zipser, L. ; Franke, H. ; Visualisation and measurement of acoustic and fluidic phenomena. 5th Int. Conf. Vibr. Measurements, Ancona 2002, SPIE Vol 4827 pp 192-198

[3] Kreis, Th., Holographic Interferometry Akademie Verlag GmbH, Berlin 1996, p. 231

AUTHOR INDEX

Abel, J S 413
Adachi, S 313
Alipour, F 499
Almeida, A 243
Alonso-Moral, J 549
Amatriain, X 533
Amir, N 189
Ando, Y 475
Arcos, J L 533
Askenfelt, A 67, 91, 151
Astesano, C 671
Atig, M 267
Avanzini, F 95
Badeau, R 715
Bader, R 121
Bank, B 143
Barbancho, A M 517
Barbancho, I 517
Benckert, L 321
Bendikesen, B 561
Bennett, G 13
Bensa, J 365
Bensoam, J 369
Bergner, A 317, 707
Bergweiler, S 317, 707
Bertsch, M 193
Besson, M 671
Beurmann, A 167
Bilbao, S D 365
Bissinger, G 35
Bjerkvik, A 299
Bonada, J 439
Bonsi, D 751
Botros, A 247
Bresin, R 155, 159, 685, 689, 693
Bromage, S R 197
Brown, A W 39
Buen, A 43
Burns, C 373
Burtner, M 373
Campbell, D M 185, 197, 233, 251, 451, 747, 759
Camurri, A 561, 569, 631
Canazza, S 521
Carlsson, P 47, 103
Carral, S 251

Castellengo, M 455
Castro-Sierra, E 591
Caussé, R 243, 263, 369
Chiang, Y C 577
Chilekwa, V 711
Cirotteau, D 561
Cook, P R 385, 677
Cottingham, J P 255
Courtney, P E 107
Čulík, M 87
Czyzewski, A 259
Dahl, S 561, 595, 599, 685, 689
Dall'Osto, D 603
Dalmont, J-P 267, 283
Daniels, A E 213
David, B 715
De Baets, B 635
de Götzen, A 525
De Meyer, H 635
De Poli, G 17
De Voogdt, L 631
Deutsch, B 343
Dillon, R 529
Disley, A C 607
Divoir, S 397
Drioli, C 377
Drumm, I A 381
Dziubinski, M 623
Elejalde-Garcia, M J 549, 719
Ellis, N 369
Erkut, C 133
Esposito, E 125, 723, 755
Esquef, P A A 339
Essl, G 385
Fabre, B 239
Fagnan, L 443
Faraone, G 447
Farner, S 397, 671
Felce, R 343
Fletcher, N H 201, 307
Fontana, F 51
Franco-García, A 719
Franke, H 763
Friberg, A 13, 521, 561, 599
Fritz, C 263
Galembo, A 147, 151, 159
Galluzzo, P M 55

Gäumann, T 163
Geissler, P 59
Gerhard-Multhaupt, R 317, 707
Gilbert, J 229, 233, 267, 271
Giordano, B L 611
Goebl, W 155, 159
Gómez, E 533
Görne, T 317
Gough, C 63
Grachten, M 533
Granqvist, S 459, 595
Gray, C D 451
Greated, C A 451, 747
Gren, P 321
Green, M 295
Griffith, N J L 553
Guettler, K 67, 91
Guillemain, P 389
Hall, D E 727, 729
Hansen, K F 689, 693
Hansen, U J 351
Hanson, R J 731
Henrich, N 455
Hertegård, S 459
Hill, T J W 129, 275, 711
Hirschberg, A 239
Hollenberg, L C L 307
Howard, D M 463, 467, 607, 697
Huitric, H 619
Huovinen, E 615
Imagawa, H 471, 495
Janariz-Larumbe, J 719
Jansson, E V 71
Jelonek, J 75
Jerkert, J 537
Johansson, E-L 321
Johansson, Ö 647
Johansson, S 447
Jordà, S 681
Juslin, P N 513
Kamitani, T 541
Kanno, J 179
Karjalainen, M 133, 339, 393, 409, 739
Kato, K 475
Kausel, W 205
Kenmochi, H 439
Kergomard, J 263, 389, 397

Kim, Y E 479
Kirkwood, B 655
Kiss, J 619
Kjellmo, I 561
Klapuri, A P 587
Kob, M 431
Kostek, B 259, 623
Kouzuki, K 541
Krishnaswamy, A 627
Kristiansen, U 299
Kronland-Martinet, R 365, 671
Kühnelt, H 401
Kuuskankare, M 545
Larkin, B 355
Larsson, H 459
Laurson, M 545
Le Carrou, J-L 229
Lee, M E 483
Leman, M 631, 635
Lesaffre, M 635
Li, A 735
Liljencrants, J 325
Liu, R 553
Llanos-Vazquez, R 549
Løkberg, O J 43
Loscos, A 439
Ludwigsen, D O 209
Łukasik, E 75, 79, 83
Lyu, R Y 577
Macho- Stadler, E 549, 719
Macomber, H K 731
Madison, G 639
Magne, C 671
Marassi, M 723
Marshall, M 685
Martens, J P 635
Martner, O 59
Mathews, M V 13, 405
Matsuda, M 541
Mayer, A 279
Mazzarino, B 561, 631
McCarthy, H 561
Merker, B 639
Mikimoto, S 179
Millot, L 329
Ming, J 667
Mion, L 557

Misdariis, N 369
Molin, N E 703
Moore, T R 213, 343
Moravec, O 643
Morrison, A C 731
Moynihan, B 685
Murphy, P 553
Murtagh, F 667
Nahas, M 619
Naganovski, A 75
Nederveen, C J 217
Niimi, S 471, 495
Noreland, J O D 221
Noson, D 475
Nykänen, A J 647
Ollivier, S 283
Osaka, N 495
Otčenášek, Z 333, 651, 663
Otondo, F 487, 655
Pakarinen, J 409,739
Pallone, G 671
Parati, L 487
Pasdeloup, V 671
Penttinen, H 133
Perrin, R 343, 347
Peterson, D R 287
Petiot, J F 225, 229
Petrini, K 611
Pettersen, V 491
Poblano, H A 591
Poirson, E 229
Pölkki, J 133
Polotti, P 447
Poulimenos, A 499
Radhakrishnan, N 499
Rajčan, E 87
Rath, M 685
Richard, G 715
Richards, O F 197, 233
Richardson, B E 117, 129
Richardson, S J 129
Rimell, S 697
Rinman, M L 561
Robinson, A 343
Rocchesso, D 7, 95
Rodà, A 521
Rodet, X 243, 443

Rohner, F 351
Rossing, T D 251, 351, 355
Roubeau, B 455
Sakakibara, K-I 471, 495
Sapp, C 13
Scavone, G P 291
Schärer, S 351
Scherer, K R 25
Scherer, R C 499
Schleske, M 59
Schneider, A 167
Schoonderwaldt, E 67, 91
Serafin, S 51, 95, 99, 373
Sharp, D B 275, 711, 735
Shirado, T 659
Shirley, E T 213
Simon, L 271
Sjölander, K 743
Sjölander, P J 503
Skala, J S 171
Skulina, D J 747
Slowinski, R 75
Smith, J 247, 295, 307
Smith III, J O 3, 365, 405, 413
Smith, M J T 483
Smyth, T 413
Stanzial, D 751
Štěpánek, J 333, 643, 651, 663
Story, B H 435
Stulov, A 175
Sujbert, L 143
Sundberg, J 13
Susmaga, R 83
Suzuki, H 179
Swallowe, G M 343, 347
Syrový, V 333, 651
Taguti, T 565
Tanghe, K 635
Tarabusi, V 755
Tardón, L J 517
Tarnopolsky, A Z 307
Tayama, N 471
Terroir, J 271
Timmers, R 569
Tinnsten, M 47, 103
Titze, I R 21
Tro, J 299

Tronchin, L	755
Tsai, C G	303
Urgela, S	87
Välimäki, V	133, 339, 361, 409
van Walstijn, M O	759
Ventura-Miravet, R	667
Vergez, C	243, 369, 397
Vermeulen, V	631
Voinier, T	389
Volpe, G	569, 631
Waadeland, C H	573
Walker, J	553
Wang, C K	577
Wegener, M	317, 707
Westgaard, R H	491
Widmer, G	509, 581
Wirges, W	707
Wolfe, J	247, 263, 295, 307
Woodhouse, J	31, 55, 107, 137
Yanagida, M	659
Yoo, J	355
Young, D	99, 111
Ystad, S	671
Zanon, P	521, 581
Zerbs, C	59
Zipser, L	763
Zwan, P	623

GNU FREE DOCUMENTATION LICENSE

Version 1.2, November 2002

Copyright (C) 2000,2001,2002 Free Software Foundation, Inc.
59 Temple Place, Suite 330, Boston, MA 02111-1307 USA

Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties--for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements."

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

Copyright (c) YEAR YOUR NAME.
Permission is granted to copy, distribute and/or modify this document
under the terms of the GNU Free Documentation License, Version 1.2
or any later version published by the Free Software Foundation;
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
A copy of the license is included in the section entitled "GNU
Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

with the Invariant Sections being LIST THEIR TITLES, with the
Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.