TMH-QPSR ABSTRACTS

Speech Communication and Speech Technology

ENABL – Enabler for engineering software using language and speech

Corine Bickley and Sheri Hunnicutt*
**Concentra Ltd., Royal Leamington Spa, Concentra, UK*

The objective of the ENABL project has been to develop a Speech Understanding System, using speech recognition, for computer-based engineering design, analysis, and configuration tasks. In this project, tools have been developed that can be used by engineers who are motorically disabled to the extent that they cannot use their hands. The central technical problem has been the implementation of an accessible user interface for rule-based engineering design that is controlled by speaking. Issues of efficiency and ease of use as well as accessibility for persons with dysarthric speech have been evaluated. Voice care for users of speech recognition technology has been monitored, and guidelines for healthful use of voice with speech recognition developed.

TMH-QPSR 1/2000: 1-11

Perceptual analysis of dysarthric speech in the ENABL project

Elisabet Rosengren

This paper presents the perceptual analysis of dysarthric speech recorded for use in the ENABL project. Twelve dysarthric speakers were tested with a Swedish dysarthria test that evaluates several speech functions; attention was focused on articulation and intelligibility. Articulation was tested with sentences representing different articulation types, and articulation deviation rated for each type. Word intelligibility was measured both in isolated words and in sentence context. The severity of the dysarthria was found to vary from very mild to moderate /severe. The results of the perceptual analysis were later put to use in other project tasks for investigating relationships between these perceptual findings, acoustic-phonetic data and results on speech recognition.

TMH-QPSR 1/2000: 13-18

Acoustic analysis of dysarthric speech with some implications for automatic speech recognition

Tina Magnuson and Mats Blomberg

This paper describes one part of an EU project, ENABL, in which speech recognition systems are used to provide access by voice to computer software. An acoustic analysis of the speech of fourteen motorically disabled persons, twelve with dysarthria and two without, is presented. The analysis is based on a standard description of dysarthric speech and parameters such as fundamental frequency, syllable rate, speech/articulation rate, reading speed, dysfluencies and articulatory deviations are measured. Implications for speech recognition systems are discussed.

TMH-QPSR 1/2000: 19-29

Improving the speech recognition in the ENABL project

Nathalie Talbot

One main objective of the ENABL project was to provide access by voice, via speech recognition, to an engineering design system, ICAD. Since persons with manual disabilities also often have a vocal disability, a part of the ENABL project was dedicated to evaluating the performance of dysarthric speech on a speech recognition engine. After a brief overview of the ENABL system, this

paper is divided in two parts. The first part is dedicated to the speech recognition of a future user of the system. The results of different experiments made in order to achieve a better recognition rate for this user are presented. Among those experiments are the use of triphone models trained on telephone speech compared to the use of monophone models trained on microphone speech. In addition, an experiment in adaptation of the acoustic models in order to fit the characteristics of the user's speech better is presented. The second part of the paper concentrates on speech recognition of dysarthric speech. Perceptual analysis and recognition results are compared and experiments in improving the recognition rate are presented.

TMH-QPSR 1/2000: 31-38

Example based shallow semantic analysis in the August spoken dialogue system

Nikolaj Lindberg and Joakim Gustafson

In this paper, the semantic analyser component of the August spoken dialogue system is presented. The task of the semantic analyser is to transform the output of a speech recogniser into a flat semantic representation, used by the dialogue manager. The analyser produces a non-compositional shallow semantic representation for each of the hypotheses in an N-best list produced by the speech recogniser. The analyser uses a machine-learning system to build its analyses from an example database. An analysis is obtained by running three independent classifiers in parallel, and concatenating the results from the different classifiers. One of the classifiers recognises unlikely utterances, and is used as a negative filter to identify (semantically) implausible hypotheses. Another predicts the topic of an utterance, while the third one returns a flat feature-value representation.

TMH-QPSR 1/2000: 39-44

Acoustic-phonetic studies of prominence in Swedish

Gunnar Fant, Anita Kruckenberg, Johan Liljencrants and Stellan Hertegård

This is an integrated study of the role of subglottal pressure and other acoustic parameters as correlates of perceived syllabic prominence in Swedish text reading. The covariation and interdependencies of sub- and supraglottal pressure, articulation and source parameters, F0, and intensity have been studied in short lab speech sentences and in the reading of a prose text of about one minute length. Continuously scaled syllable prominence, Rs, has been determined from listening tests and is added to our standard measurement display in synchrony with oscillogram, spectrogram, F0, and two intensity parameters, the sound pressure level SPL, and a high frequency pre-emphasised measure, the SPLH. For documentary purposes, the complete material from the text reading has been included. Most of the principal results from the analysis have been published in our earlier publications but are reviewed here. Of special interest is the predictability of SPL from sub-glottal pressure and F0 and their temporal patterns, in particular in relation to stress and prominence. A novelty is the analysis of sub- and supraglottal pressures in vowels and consonants in stressed and unstressed context. Another object of analysis is the spectral dynamics associated with pre-occlusion aspiration of stressed vowels.

TMH-QPSR 2-3/2000: 1-51

Replicating three-dimensional tongue shapes synthetically

Olov Engwall

A new three-dimensional tongue model has been developed within the KTH 3D vocal tract project using MR Images of a reference subject producing 43 artificially sustained Swedish articulations. The tongue contour was manually extracted from the MR images and the three-dimensional tongue shape was reconstructed for each articulation. The tongue shape for a neutral tongue position with closed jaw served as reference in the modeling process and the deviation from this reference determined the values of the articulatory control parameters for each articulation. The six linear parameters jaw height, tongue body, tongue dorsum, tongue tip, tongue advance and tongue width were determined using an ordered linear factor analysis controlled by articulatory measures.

88% of the variance in the midsagittal plane and 78% of the overall sagittal variance was explained by the first five factors of the analysis. The six parameter model is able to reconstruct the modeled articulations in 3D with an overall RMS reconstruction error of 0.13 cm sagittally and 0.12 cm laterally, and it specifically handles lateral differences and the observed asymmetries in tongue shape.

TMH-QPSR 2-3/2000: 53-64

Voice changes after using a voice input system: an acoustic study

Whiteside SP, de Bruijn CG*, Rosen KM**, Hunnicutt S**, Nord L** & Syder D**

** University of Sheffield, UK*
*** KTH, Stockholm, Sweden*

Automatic speech recognition systems, and voice input systems in general, are becoming more widely used, and more people are opting for speech driven computer interface as an alternative input method to the keyboard, both in the home and office environment. There is some evidence to suggest that the use of speech recognition based human computer interfaces could potentially lead to vocal fatigue, or even to symptoms associated with dysphonia. It has therefore become necessary to qualify any potential risks of voice damage. This study reports on a case study that was carried out to investigate acoustic changes in the voice, after the use of a discrete speech recognition system. Acoustic analyses were carried out on two Swedish users of such a system. These results are presented and discussed. A set of voice care guidelines for regular users of voice-input systems is provided. In addition, recommendations for improving recognition and hence vocal care are provided as an adjunct to the voice care guidelines.

TMH-QPSR 4/2000: 33-48

Dynamical aspects of coarticulation in Swedish fricatives – a combined EMA & EPG study

Olov Engwall

An electromagnetic articulography (EMA) system and electropalatography (EPG) have been employed to study five Swedish fricatives in different vowel contexts. Articulatory measures at the onset of, the mean value during, and at the offset of the fricative were used to evidence the coarticulation throughout the fricative. The contextual influence on these three different measurements of the fricative are compared and contrasted to evidence how the coarticulation changes. Measures were made for the jaw motion, lip protrusion, tongue body with EMA and linguo-palatal contact with EPG. The data from the two sources were further combined and assessed for complementary and conflicting results.

TMH-QPSR 4/2000: 49-73

Music Acoustics

A method for describing different styles of singing.

A comparison of a female singer's voice source in "classical", "pop", "jazz" and "blues"

Margareta Thalén and Johan Sundberg*
** SMI (University College of Music Education in Stockholm)*

The voice is apparently used in quite different manners in different repertoires of singing. Also, it differs between individuals. Some of these differences concern the voice source, which, however, varies considerably with loudness, pitch, and mode of phonation. Therefore, when comparing different types of voice use, it is necessary to analyze how the voice source varies with these parameters. This investigation attempts to describe voice source differences between classical, pop, jazz and blues styles of singing as produced by a professional female singer and voice pedagogue at the pitches A3, C#4, E4 and G4 in soft middle and loud phonation. The voice source was analyzed by inverse filtering the flow signal. Four parameters were considered: (1) subglottal pressure, captured as the oral pressure during p-occlusion; (2) closed quotient; (3) the level difference between the two lowest source spectrum partials; and (4) the glottal

compliance, defined as the ratio between the air volume contained in a voice pulse divided by the underlying subglottal pressure. The method was first to analyze how these voice source characteristics varied with pitch and loudness in the different modes and styles. Then averages across pitch and loudness for each mode and style were compared and related to their total range of variation in the subject. It was found that for most of the voice source parameters, classical was most similar to flow and leaky phonation, pop and jazz to neutral and flow phonation, and blues to pressed phonation.

TMH-QPSR 1/2000: 45-54

Fourier analysis applied to high-speed laryngoscopy

*Svante Granqvist and Per-Åke Lindestad **
** Dept of Logopedics and Phoniatrics, Karolinska Institute, Huddinge University Hospital*

A new method for analysis of high-speed recordings is presented. The method is based on extraction of light intensity time-sequences from consecutive images, which in turn are Fourier-transformed. The spectra thus acquired can be displayed in four different modes, each having its own benefits. When applied to the larynx the method visualises oscillations in the entire laryngeal area, not merely the glottal region. The method was applied to three laryngo-scopic image sequences. These examples revealed co-vibrations in the ventricular folds and in the mucosa covering the arytenoid cartilages. In some cases, the co-vibrations occurred at other frequencies than those of the glottis.

TMH-QPSR 1/2000: 55-60

Legato, staccato, and repeated tones in expressive piano performance

*Roberto Bresin and Giovanni Umberto Battel**
** Conservatorio di Musica "Benedetto Marcello", Venezia, Italy*

Articulation strategies applied by pianists in expressive piano performances of a piano piece are analysed. Measurements of key overlap time and its relation to the inter-onset-interval are collected for notes marked legato and staccato in the first sixteen bars of the Andante movement of W A Mozart's Piano Sonata in G major, K 545. Five pianists played the piece nine times. First, they played in a way that they considered "optimal". In the remaining eight performances they were asked to represent different expressive characters, as specified in terms of different adjectives. Legato, staccato, and repeated notes articulation applied by the right hand were examined by means of statistical analysis. Although the results varied considerably between pianists, some trends could be observed. The pianists generally used similar strategies in the renderings intended to represent different expressive characters. Legato was played with a key overlap ratio that depended on the inter-onset-interval (IOI). Staccato was realised by means of a key detached time, during which no key was depressed, that amounted to approximately 60% of the IOI. Repeated notes were played with a key detached time of about 40 % of the IOI. The results seem useful as a basis for articulation rules in grammars for automatic piano performance.

TMH-QPSR 1/2000: 61-71

Motion in music: Sound level envelopes of tones expressing human locomotion

Anders Friberg, Johan Sundberg and Lars Frydén

The common association of music with motion was investigated in a direct way. Could the original motion quality of different gaits be transferred to music and be perceived by a listener? Measurements of the ground reaction force by the foot during different gaits were transferred to sound by using the vertical force curve as sound level envelopes for tones played at different tempi. Three listening experiments assessed the motion quality of the resulting stimuli. In the first experiment, where the listeners were asked to freely describe the tones, 25% of answers were direct references to motion and with more motion references for faster tempi. In the second experiment, where the listeners were asked to describe the motion quality, about half of the answers directly related to motion could be classified as

belonging to one of the categories dancing, jumping, running, walking or stumbling. Most gait patterns were clearly classified as belonging to one of these categories independent of presentation tempo. In the third experiment, the listeners were asked to rate the stimuli on 24 adjective scales. A factor analysis yielded four factors that could be interpreted as Swift vs. Solemn (factor 1), Graceful vs. Stamping (factor 2), Limping vs. Forceful (factor 3) and Springy (factor 4, no negative adjective). The results from the three experiments were consistent and indicated that each tone (corresponding to a particular gait) could clearly be categorised in terms of motion.

TMH-QPSR 1/2000: 73-82

Long-term average spectrum (LTAS) analysis of developmental changes in children's voices

Peta White

Long-term average spectrum (LTAS) analysis has been found to offer representative information on voice timbre. It provides spectral information averaged over a period of time and is particularly useful when persistent spectral features are under investigation. The aim of this study was to compare perceived and actual sex of the recorded voices of children to the LTAS characteristics. A total of 320 children, 20 boys and 20 girls in each of eight age groups (range 3 to 12 years), were recorded singing a nursery rhyme. In an earlier analysis, the recorded voices were evaluated with respect to perceived sex by expert listeners. Mean LTAS analysis for boys and girls groups revealed a peak at 5 kHz for children consistently perceived as boys (whether male or female in actuality), and a flat spectrum at 5 kHz for children consistently perceived as girls.

TMH-QPSR 2-3/2000: 85-87

Long-term-average spectrum characteristics of country singers during speaking and singing

Cleveland TF, Sundberg J & Stone RE**
**Vanderbilt Voice Center, Vanderbilt School of Medicine, Nashville, TN*

Five premiere male country singers involved in our previous studies spoke and sang the words of both the national anthem and a country song of their choice. Long-term-average spectra were made of the spoken and sung material of each singer. The spectral characteristics of country singers' speech and singing were similar. A prominent peak in the upper part of the spectrum, that has been previously described as the "speaker's formant," was found in the country singers' speech and singing. The singer's formant, a strong spectral peak near 2.8 kHz and an important part of the spectrum of classically trained singers, was not found in the spectra of the country singers. The results support the conclusion that the resonance characteristics in speech and singing are similar and that country singing is not characterized by a singer's formant

TMH-QPSR 2-3/2000: 89-94

Effects of inhalatory abdominal wall movement on vertical laryngeal position during phonation

Jenny Iwarsson

Department of Speech, Music, Hearing, Royal Institute of Technology, KTH, SE-100 44 Stockholm, Sweden, Phn +468 790 7876,Fax +468 790 7854, email: jenny@speech.kth.se, &

Department of Logopedics and Phoniatrics, Karolinska Institute, Huddinge University Hospital, SE-141 86 Huddinge, Sweden

The configuration of the body resulting from inhalatory behaviour is sometimes considered a factor of relevance to voice production in singing and speaking pedagogy and in clinical voice therapy. The present investigation compares two different inhalatory behaviours; 1) with a "paradoxical" inward movement of the abdominal wall, and 2) with an expansion of the abdominal wall, both with regard to the effect on vertical laryngeal position during the subsequent phonation. Seventeen male and 17 female healthy, vocally untrained subjects participated. No instructions were given regarding movements of the rib cage. Inhaled air volume as measured by respiratory inductive plethysmography, was controlled to reach 70% inspiratory capacity.

Vertical laryngeal position was recorded by two-channel electroglottography during the subsequent vowel production. A significant effect was found; the abdomen-out condition was associated with a higher laryngeal position than the abdomen-in condition. This result apparently contradicted a hypothesis that an expansion of the abdominal wall would allow the diaphragm to descend deeper in the torso, thereby increasing the tracheal pull, which would result in a lower laryngeal position. In a post-hoc experiment including six of the subjects, the body posture was studied by digital video recordings, revealing that the two inhalatory modes were clearly associated with postural changes affecting laryngeal position. The "paradoxical" inward movement of the abdominal wall was associated with a recession of the chin towards the neck, such that the larynx appeared in a lower position in the neck, for reasons of a postural change. The results suggest that the laryngeal position can be affected by the inhalatory behaviour if no attention is paid to posture, implying that instructions from clinicians and pedagogues regarding breathing behaviour must be carefully formulated and adjusted in order to ensure that the intended goals are reached.

TMH-QPSR 2-3/2000: 95-104

Production of staccato articulation in Mozart sonatas played on a grand piano. Preliminary results

*Roberto Bresin and Gerhard Widmer**

** Department of Medical Cybernetics and Artificial Intelligence, University of Vienna and Austrian Research Institute for Artificial Intelligence (ÖFAI)*
E-mail: gerhard@ai.univie.ac.at,
http://www.ai.univie.ac.at/~gerhard

Staccato articulation in piano performance is an important expressive means available to the interpreter and still almost unexplored in music performance research. In the present study, the performance of notes that are marked staccato in 13 Mozart´s piano sonatas is analysed. A professional pianist played the sonatas on a computer-monitored Bösendorfer grand piano. Results confirmed previous findings indicating that the relative amount of staccato for one tone is independent from the inter-onset-interval (IOI). The amount of staccato was found to vary with context, tempo indications and melody contour; isolated staccato tones and repeated staccato tones are played more staccato than other tones marked staccato in the score.

TMH-QPSR 4/2000: 1-6

Perceptual detection of inhalations in reading

Jenny Iwarsson

Dept of Speech, Music and Hearing, Royal Institute of Technology, Stockholm, and Dept of Logopedics and Phoniatrics, Karolinska Institute, Huddinge, University Hospital, Huddinge

Patients with vocal nodules seem to produce speech at lower lung volumes and inhale at syntactically deviant positions, as compared to normal subjects. Information about respiratory patterns during speech is thus of great value for the voice clinician. Such information can both add to the understanding of the patient's voice problems and help the clinician to design an adequate treatment. The purpose of the present study was to investigate the accuracy with which inhalations can be detected from audio recordings of reading. Six females with bilateral vocal nodules and six healthy females were recorded while reading a standard text. Inhalations were identified from the relative lung volume signal, recorded by respiratory inductive plethysmography. A listening panel of 31 students in speech pathology were asked to mark in a written text where they thought the readers inhaled; not only audible inhalations but also pauses implying an inhalation.

The syntactical distribution of inhalations showed that inhalations most often occurred at the initiation of a clause, for both subject groups. Only few inhalations were regarded as deviant from a linguistic point of view. These inhalations were missed by a high percentage of the listening panel. The overall ability to perceptually detect inhalations was fairly good but varied considerably between the listeners. On average 91.3% of the inhalations were correctly marked. The ability to perceptually detect inhalations was shown to be correlated to various factors, e.g. the syntactic location, the inhalation

frequency, the inhalatory volume change, and to some extent the sound levels of the inhalations.

It was concluded that perceptual detection of inhalations seems to offer a simple method to obtain information about inhalatory behaviour in voice patients. In order to include this method in clinical routinely measurements of voice disorders, further research seems necessary.

TMH-QPSR 4/2000: 7-18

Detection of drift in tempo

Sofia Dahl, Svante Granqvist, and Monica Thomasson

The detection threshold for a continuous drift in tempo in a click sequence was investigated. Eight stimuli with varying drift magnitudes were presented to musically trained and untrained listeners. The listening test used a modification of the method for Parameter Estimation by Sequential Testing (PEST) to present longer or shorter portions of stimuli to the listener, depending on the responses. The results showed that longer presentation times were needed for stimuli with lower drift magnitude. When presented with longer sequences listeners were able to detect drifts less than a tenth of the previously reported detection threshold for tempo changes.

TMH-QPSR 4/2000: 19-28

Spectrum effects of subglottal pressure variation in professional baritone singers

Peta White & Johan Sundberg

The audio signal from five professional operatic baritone singers was analysed by means of spectrum analysis. Each subject sang a sustained diminuendo, from loudest to softest phonation, three times on the vowels [a:] and [æ:] at fundamental frequencies representing 25%, 50% and 75% of his total pitch range as measured in semitones. During the diminuendi the subjects repeatedly inserted the consonant [p] so that associated subglottal pressures could be estimated from the oral pressure during [p]-occlusions. Pooling the three takes of each condition, ten subglottal pressures (P_S), equidistantly spaced between highest and lowest, were selected for analysis along with the corresponding production of [a:] and [æ:] vowels. The levels of the first formant and the singer's formant, L_1 and L_{SF}, were measured as a function of increasing subglottal pressure. Averaged across subjects, an increase in P_S resulted in (a) an increase in L_1 and (b) a decrease in L_1-L_{SF}. This implies that a 10 dB increase at or near 600 Hz was, on average, accompanied by an increase of 17 dB of the level near 3 kHz.

TMH-QPSR 4/2000: 29-32

Hearing Technology

The behaviour of non-linear (WDRC) hearing instruments under realistic simulated listening conditions

Peter Nordqvist

This work attempts to illustrate some important practical consequences of the characteristics of non-linear wide dynamic range compression (WDRC) hearing instruments in common conversational listening situations. Corresponding input and output signal are recorded simultaneously, using test signals consisting of conversation between a hearing aid wearer and a non-hearing aid wearer in two different listening situations, quiet and outdoors in fluctuating traffic noise.

The effective insertion gain frequency response is displayed for each of the two voice sources in each of the simulated listening situations. The effective compression is also illustrated showing the gain adaptation between two alternating voice sources and the slow adaptation to changing overall acoustic conditions.

These non-linear effects are exemplified using four commercially available hearing instruments. Three of the hearing aids are digital and one is analogue.

TMH-QPSR 2-3/2000: 65-68

Comparison between the Multipeak (MPEAK) and Spectral Peak (SPEAK) speech coding strategies on objective and subjective speech tests by some cochlear implants

Eva Agelfors

Two groups of hearing-impaired adults with a severe to profound hearing loss participated in this comparison study and were tested at two sessions over a longer period with assistive devices, either cochlear implants (CI) or hearing aids (HA). One group consisted of four cochlear implanted subjects that for some years used the multipeak (MPEAK) speech coding strategy of the Nucleus 22-channel cochlear implant system and later was upgraded to the spectral peak (SPEAK) speech coding strategy of that system. Four experienced hearing-aid users, all with a profound hearing loss, represented the other group, and all of them had between the test periods exchanged their older hearing aids to newer ones. The aim of the study was to evaluate the change in performance when the SPEAK processing strategy replaced the MPEAK strategy, and the older hearing aid was replaced with new models. The test battery consisted of two parts: speech perception tests and self-rating performance inventory (PIPSL). The speech tests consisted of segmental test, test of prosodic contrasts and Connected Discourse Tracking (CDT) test, presented in two test situations, audiovisually and auditory only. All test were presented in quiet.

The obtained results showed small changes between MPEAK and SPEAK coding strategy on the speech recognition test, but there was a significant improvement at least on one test for the CI-group when using the SPEAK speech coding strategy. Average percentages of transmitted information of the vCv-syllables, presented audiovisually, was significantly greater for three of four subjects with the SPEAK speech coding strategy compared to MPEAK. The results on the CDT-test showed for the same presentation mode an improvement around 20 words/min for two of the poorer subjects with the SPEAK strategy, but no improvement was obtained by the two better subjects. In the test situation auditory alone, however, there was an improvement around 20% for the consonants by the two better subjects. Average and individual scores on the speech recognition test obtained by the HA-group showed no significant differences between the two hearing aids. One of the subjects scored a poorer result on the CDT-test auditory alone with HA:2 compared to HA:1.

The subjects' responses to the questionnaire concerning listening situations in everyday life showed for the CI-group significant changes between MPEAK and SPEAK for the category "Personal". The mean results rated by the HA-users showed significant changes between the two hearing aids for questions concerning speech. Poorer results were rated by the subjects fitted with HA:2 compared to HA:1 for the categories about speech recognition with or without visual cues.

TMH-QPSR 2-3/2000: 69-83