


Multimodal speech synthesis/ NGSLT - Speech Technology course

Björn Granström
CTT, KTH

Multimodal speech synthesis - NSGLT 2004 [1]

The work presented in this lecture is the result of many researchers' efforts at the Department of Speech, Music and Hearing

Reports and further information can be found on our home page www.speech.kth.se



Multimodal speech synthesis - NSGLT 2004 [2]

LET'S TALK!

SPEECH TECHNOLOGY IS THE NEXT BIG THING IN COMPUTING

Februari 1998

Multimodal speech synthesis - NSGLT 2004 [3]

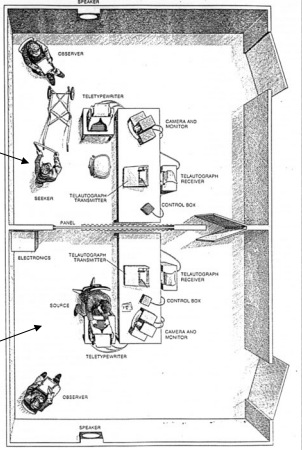
Why Speech Tech?

User

"Interactive Human Communication"

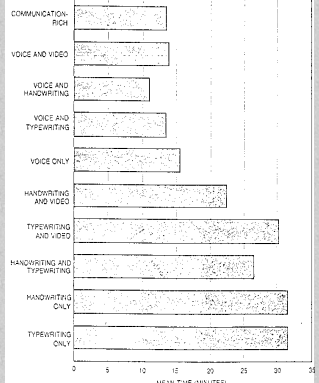
Chapanis
Scientific American 1975!

"Computer"



More efficient and convenient!

Average time for solving diverse problems using different combinations of means of communication



MODE OF COMMUNICATION influenced the time required to solve problems. Here the average time taken by teams to solve problems is charted for 18 different modes of communication. The data fall into two fairly distinct groups. The faster five all involve the use of the voice in communication, whereas the slower five exclude the voice.

Multimodal speech synthesis - NSGLT 2004 [4]

Speech is fast and verbose

	COMMUNICATION-RICH	VOICE	HANDWRITING	EXPERIENCED TYPISTS	INEXPERIENCED TYPISTS
SOLUTION TIME IN MINUTES	29	33	53.3	66.2	69
NUMBER OF MESSAGES	230.4	163.6	15.9	27.2	31.5
NUMBER OF SENTENCES	372.6	275.9	24.5	45.9	44.1
TOTAL NUMBER OF WORDS	1,563.9	1,374.8	224.8	322.9	257.4
TOTAL NUMBER OF DIFFERENT WORDS	397.5	305.9	118.5	150.5	133.4
TYPE-TOKEN RATIO	3	3	6	5	6
NUMBER OF WORDS PER MINUTE	190.3	171.2	17.3	18.1	19.2

EXPERIMENTAL RESULTS are enumerated for the solution of problems by various modes of communication. "Type-token ratio" is ratio of different words to total words. Problem solving by voice takes the least time but is wordier than the other modes are.

Multimodal speech synthesis - NSGLT 2004 [5]

Speech technology is making money!

Classic example : AT&T and Lucent Technologies VRCP, "Voice Recognition Call Processing" service

- Selection of payment method
- Vocabulary only five words : collect, calling card, third number, person, operator
- More than 5 000 000 calls per day
- Earnings already (1999?) more than AT&T/Bell labs' total investment in speech research

Multilingual speech synthesis - NG SLT 2004 [7]

Världens första penntelefon!



The first pen phone

Why no commercial success?

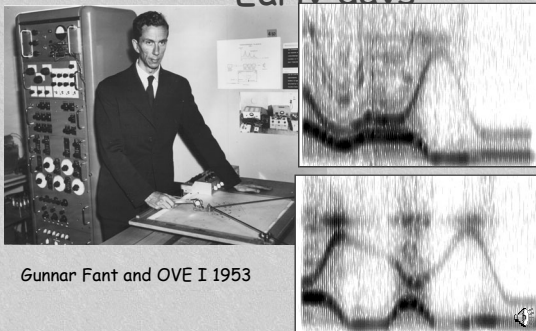
Too small keys?

Too small display!?

1 april 1998

Multilingual speech synthesis - NG SLT 2004 [8]

The KTH speech group - Early days



Gunnar Fant and OVE I 1953

Multilingual speech synthesis - NG SLT 2004 [9]

OVE I, in the WaveSurfer tool



- Interface is based around WaveSurfer, a general purpose tool for speech and audio viewing, editing and labelling
- TTS and Talking Head functionality is added as plug-ins
- WaveSurfer (presently without TTS & TH) works on all common platforms and is freely available as open source
- Modules from Waves available - formants and F0 in present release - thanks to Microsoft and AT&T
<http://www.speech.kth.se/wavesurfer>

Multilingual speech synthesis - NG SLT 2004 [10]

Ove II, 1958



1961

1962

Multilingual speech synthesis - NG SLT 2004 [11]

KTH/TTS history

- 1967, Digitally controlled OVE III
- 1974, Rule-based system RULSYS - transformation rules
- 1979, Mobile text-to-speech system - used by a non-vocal child
- 1982, Portable TTS (ICASSP, Paris) - Multilingual - M C 68000, NEC 7720
- 1983, Founding of Infovox Inc.

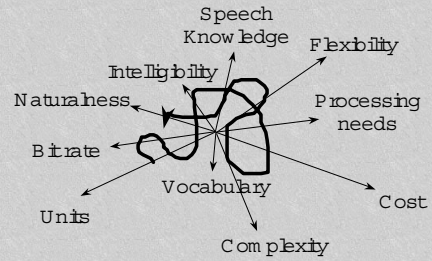
Multilingual speech synthesis - NG SLT 2004 [12]

What do we mean by Speech synthesis?

- Recorded speech
 - Words or phrases (telephone banking)
 - Fixed vocabulary - maintenance problems..
- Concatenative speech synthesis
 - Diphones or larger units (unit selection)
 - LPC: source filter model (too simple?)
 - PSOLA/MBROLA/HNM - and mixes for prosody
 - One speaker
- Parametric synthesis
 - Formant synthesis
 - Articulatory synthesis
 - flexible
 - But lower quality - today
- Multimedial synthesis

Multimedial speech synthesis - NGSLT 2004 [13]

The synthesis space



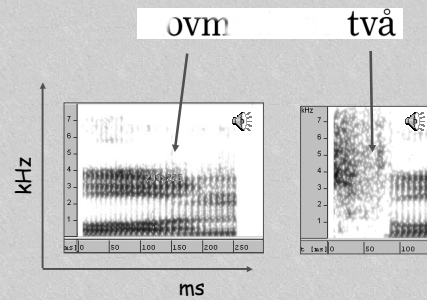
Multimedial speech synthesis - NGSLT 2004 [14]

TEXT vs SPEECH

- Parallel vs. sequential
- permanent vs disappearing
- Text as transcription of speech?
- Example - WAVE Surferdem o :
 - Palindrome
 - "V ar sak hart två sidor"

Multimedial speech synthesis - NGSLT 2004 [15]

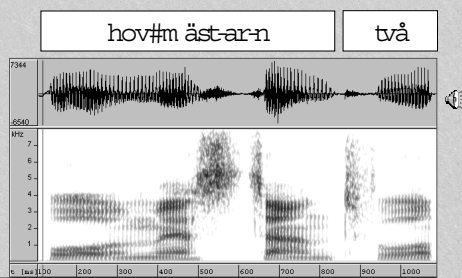
Phonetics



Multimedial speech synthesis - NGSLT 2004 [16]

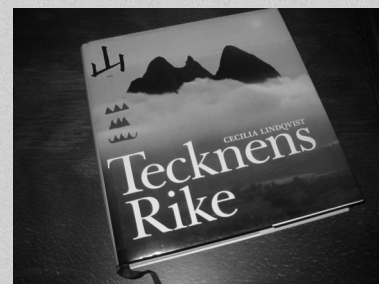
Phonology, Morphology

Hovmästarn, två



Multimedial speech synthesis - NGSLT 2004 [17]

Non-phonetic writing



Multimedial speech synthesis - NGSLT 2004 [18]

Signs vs. technology

Multimedial speech synthesis - NG SLT 2004 [19]

Is the sign what you think?

Multimedial speech synthesis - NG SLT 2004 [20]

How is a Chinese lexicon organized?

man, människa	人 / 人
stor	大 - 大
öga	目 月 月 月 目
ansikte, yta	面 - 一 一 一 一 而 而 而 而 面
öra	耳 - 一 一 一 一 耳
nått, själv	自 ' 一 一 一 一 自
mun	口 口 口
tand, tänder	齒 一 一 一 一 止 止 止 止 齒 齒 齒 齒

Multimedial speech synthesis - NG SLT 2004 [21]

Text-to-speech (TTS)

Multimedial speech synthesis - NG SLT 2004 [22]

Synthesis methods

Multimedial speech synthesis - NG SLT 2004 [23]

Source filter theory

Multimedial speech synthesis - NG SLT 2004 [24]

RULSYS Rules - Features

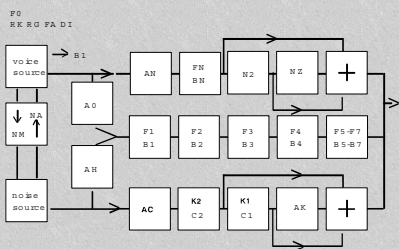
- Is IPA synthesis possible?
- Based on generative phonology
- Language specific definitions
- Rules for contextual modifications
- Examples
- Interactive rule manipulations

(a) $a \rightarrow e / _ \langle \text{cons} \rangle e \#$

(b)

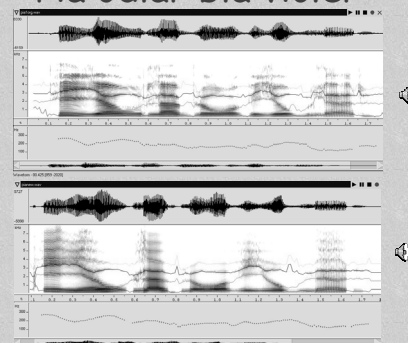
morph		root	
graph	a	e	
phon	^	^	→ e

Glove



"Pia odlar blå violer"

Original
Synthesis



Carlson, R., Granström, B., and Karlsson, I. (1990): "Experiments with voice modeling in speech synthesis."

Speaker characteristics

- Speaker
dialect, sex, social, education, age
- Situation
formality, style, interspeaker relation
- Complex description
many dependent variables

Speaker characteristics

- TIDE/Voices project (Voices, attitudes and emotions in synthetic speech)
- "Voice fitting"
- Software, Globe synthesis
- 10 user controlled parameters
- Rule specified connection to synthesis parameters

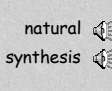



Emotional/ubiquitous computing - do we want it?



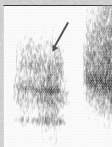


Early BBC vision - the conversational toaster
Thanks to Mark Huckvale, UCL, for the video clip

Multimedial speech synthesis - NG SLT 2004 [31]

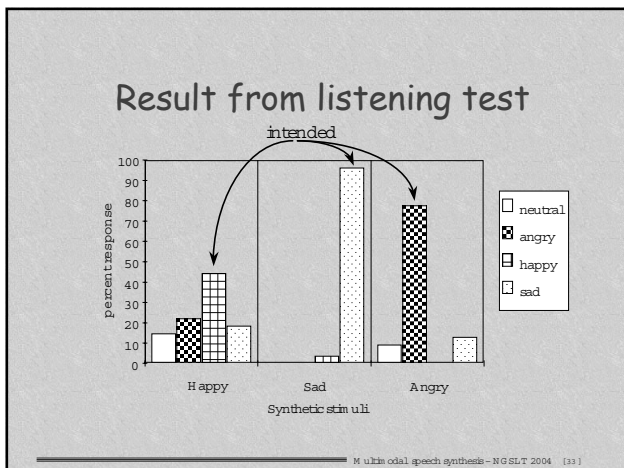
Emotions

natural synthesis   Neutral  natural synthesis 

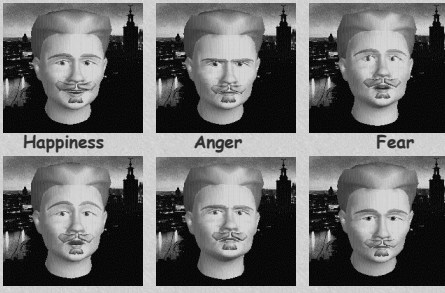
Happy

Angry   natural synthesis 

Multimedial speech synthesis - NG SLT 2004 [32]



Emotions

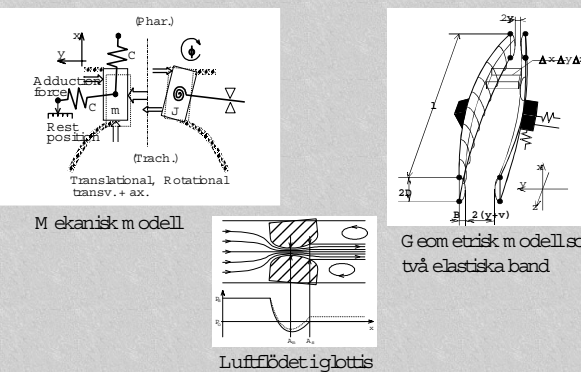


Happiness Anger Fear

Surprise Disgust Sadness

Multimedial speech synthesis - NG SLT 2004 [34]

Källasimulering



Mekanisk modell

Geometrisk modell som två elastiska band

Luffbådet i glottis

Multimedial speech synthesis - NG SLT 2004 [35]

Källasimuleringar med återsyntetisering

Samplingsintervall: 1 parametrer






Stämbandsanatom i: 18 parametrer, tex längd, vikt, dämpningsfaktor


Stämbandsartikulation: 7 parametrer, tex restgap, lungtryck

Högtrycksartikulation: 4 parametrer, tex tonhöjd, ljudstyrka

Talrörsartikulation: 8 parametrer, tex area, dämpningsfaktor

Exempel

- fördubblad stämbandsassa 
- förlängning av stämband 13 -> 17 mm 
- assymetriförändring 1.0 -> 1.8 
- assymetriförändring 1.0 -> 2.0 
- restgap 0.1 -> 1.0 mm 

originalyttrande 

Multimedial speech synthesis - NG SLT 2004 [36]

Predictable word accents

- Word accent and stress
 - bän-den (bird)
 - bän-den (spirit)
 - bän-k-pasta
 - bän-k-pastej
 - bän-k-pastejen
 - bän-k-pastejs-m ål-tid
- Name pronunciation
 - Karl-Erik (karl-erik) + ämne-m arie
 - Onomastica/EU
- Reduction
 - betöng-väg (g)s-konstruktion (C-C E,ert)

Multimedial speech synthesis - NG SLT 2004 [37]

Letter-to-sound vs. lexicon

Size vs. Precision

Maintenance - "half of the words are unique"

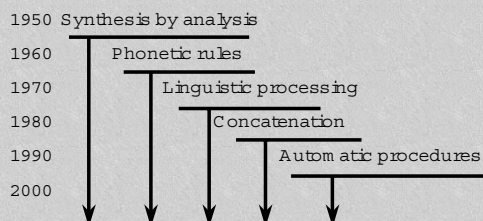
Name pronunciation (Onomastica Project)

Misspellings

Use rules, morphology, analogy... as fallback

Multimedial speech synthesis - NG SLT 2004 [38]

Research trends in speech synthesis



Multimedial speech synthesis - NG SLT 2004 [39]

Concatenative synthesis

- Already Peterson et al. (1958)
- Dixon and Maxey (1968)
- "Diacic Units", Olive, 1977)
- "PSOLA", Charpentier and Stella (1986)
- Review, Möbius (2000)
- ICSLP 2002
 - Unit selection
 - Concatenation cost
 - Prosody

Multimedial speech synthesis - NG SLT 2004 [40]

Concatenative synthesis Signal manipulations

- Prosodic modifications
 - Possibility to modify F0
 - Possibility to lengthen or shorten segments
- Spectral modifications
 - Interpolation of spectrum at joints
- Early technique - LPC

Multimedial speech synthesis - NG SLT 2004 [41]

Speak&Spell - TexasInstrument Christmas 1978



Multimedial speech synthesis - NG SLT 2004 [42]

PSOLA



- Pitch pulses moved in time to fit F0 contour
- Conceptually simple and computationally efficient
 - Need for precise pitch pulse marking
 - Could not handle spectral interpolation



Unit selection

- Large databases of recorded natural speech
- Minimal processing
- A notation of database – what information is needed?
- Synthesis defaults to transcription and search problem
- Few cuts > maximally long units selected (but context and prosody must fit well)
- Target and concatenation costs

Synthesis methods

- Unit selection – minimal processing
 - chatr(ATR) weather (CSTR) good, less good
- Diphone synthesis
 - Svensk (Mbrola), Fransk (ICP) (CNET)
- Formantsynthesis
 - Svensk (KTH), Fransk (KTH)

Unit selection - BrightSpeech

- Swedish 
- Norwegian 

Examples of Synthesized Speech
 Universität Stuttgart
 Institut für Maschinelle Sprachverarbeitung

[German] [English] [French] [Dutch] [Spanish] [Italian]
 [Portuguese] [Swedish] [Norwegian] [Finnish] [Estonian]
 [Icelandic] [Czech] [Russian] [Greek] [Croatian] [Romanian]
 [Japanese] [Chinese] [Korean] [Hebrew] [Arabic]

<http://www.institut-stuttgart.de/~moehler/synthespeech/examples.htm>

also e.g. <http://www.naturalvoices.att.com/>

To make it work today: Hybrid systems
 e.g. Who has number nn...?
 Key input - Synthesis mixed with
 recorded speech (Call +46 118 999)



Hybrid methods, cont.

- Rolf Carlson, Tor Sigvardson, Arvid Sjölander (2002). Data-driven formant synthesis. *Proc of Fonetik 2002, TMH-QPSR*
- David Öhlin and Rolf Carlson Data-driven Formant Synthesis, *Proc of Fonetik 2004*
- + Four MSc theses (Sigvardson, Sjölander, Vinet, Öhlin)
- All available on www.speech.kth.se

Multimedial speech synthesis - NG SLT 2004 [49]

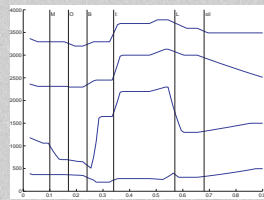
Aim

- Keeps the flexibility of the formant synthesis
- More natural sounding than rule-driven synthesis
- Speaker adaptation

Multimedial speech synthesis - NG SLT 2004 [50]

Rule-driven formant synthesis

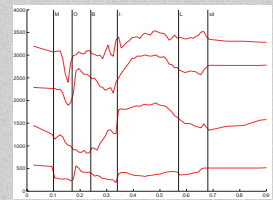
- Parameters are generated by rule (RULSYS, Carlson et al)
- Formant values are generated by interpolating between target frequencies
- Parameters are fed to a synthesizer (GLOVE, Carlson et al)



Multimedial speech synthesis - NG SLT 2004 [51]

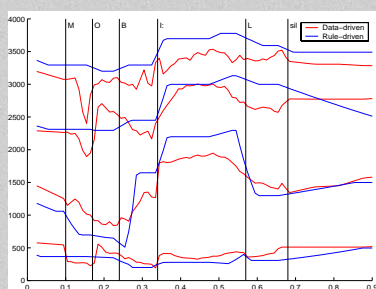
Data-driven formant synthesis

- Some parameters (namely, the first four formants) are replaced by data
- Same synthesizer



Multimedial speech synthesis - NG SLT 2004 [52]

Synthesis comparison



Multimedial speech synthesis - NG SLT 2004 [53]

Data-driven formant synthesis

- Formants are replaced through unit selection from a formant diphone library
- Formant trajectories are scaled and interpolated to fit the rule-generated durations

Multimedial speech synthesis - NG SLT 2004 [54]

Cost function

- Designed to promote probable formant candidates
- Penalizes:
 - Large bandwidths
 - Large frequency deviations (given the current phoneme)
 - Large frequency jumps (promotes smooth trajectories)

Multimedial speech synthesis - NG SLT 2004 [55]

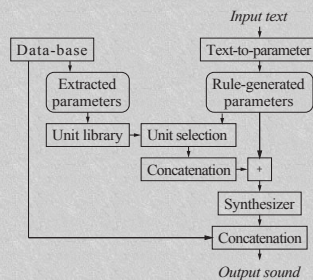
Voiceless consonants

- Replace the voiceless fricatives and plosives with recorded versions
- Voiceless fricatives in Standard Swedish: /f/, /s/, /sʃ/, /tʃ/, and /ts/
- Voiceless plosives: /k/, /p/, /t/, /tʰ/

(/h/ is excluded)

Multimedial speech synthesis - NG SLT 2004 [56]

Text-to-speech synthesis



Multimedial speech synthesis - NG SLT 2004 [57]

Listening test evaluation 1

- 15 subjects, 20 sentences, continuous scale
- Data-driven synthesis with non-connected formant data was judged more natural sounding than rule-driven synthesis

Multimedial speech synthesis - NG SLT 2004 [58]

Listening test evaluation 2

- 12 subjects, 10 sentences, binary scale
- Data-driven synthesis with manually corrected formant data was preferred in 73 % of the cases over rule-driven synthesis

Multimedial speech synthesis - NG SLT 2004 [59]

Sound samples

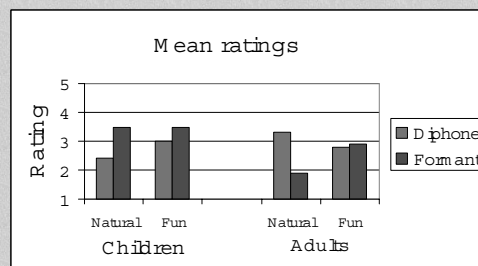
- Strindberg, rule-driven:
- Strindberg, data-driven:
- Strindberg, MBROLA:
- "Pytteliten", rule-driven:
- "Pytteliten", data-driven:

Multimedial speech synthesis - NG SLT 2004 [60]

Child-directed speech synthesis

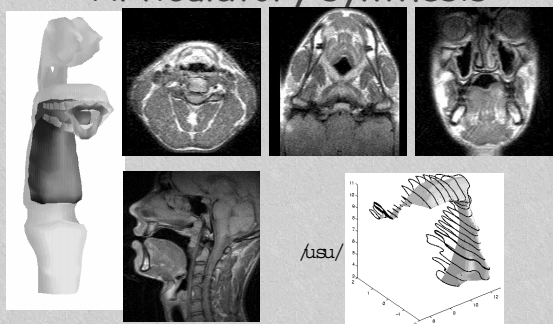
- Increase prosodic variation in synthesis
- How do children between the ages of 9 and 11 react to:
 - default, F0, duration?
 - diphone synthesis vs formant synthesis?

Multimedial speech synthesis - NG SLT 2004 [61]



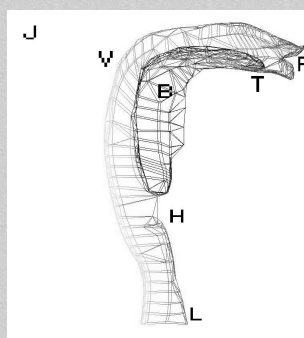
Multimedial speech synthesis - NG SLT 2004 [62]

Articulatory synthesis



Multimedial speech synthesis - NG SLT 2004 [63]

Artikulatoriska parametrar



- Käköppning
- Läpprundning
- Protusion
- Tungplacering
- Tunghöjd
- Tungspets
- Velum
- Hyoid

Multimedial speech synthesis - NG SLT 2004 [64]

Potentiell användning



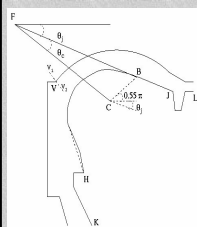
- Artikulatorisk syntes
 - beräkningar direkt från tvärsnittareorna
 - flödesmekaniska beräkningar



- Visuellt syntes
 - artikulationstråning
 - demonstrationer

Multimedial speech synthesis - NG SLT 2004 [65]

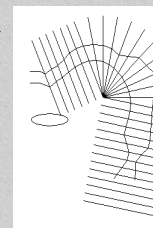
Why a 3D model?



- 2D artikulatoriska modeller, eg. Mermelstein
 - Mæda

- In posing a third dimension: $area = a \cdot (width)^b$

- A 3D model:
 - Direct calculation
 - lateral etc.

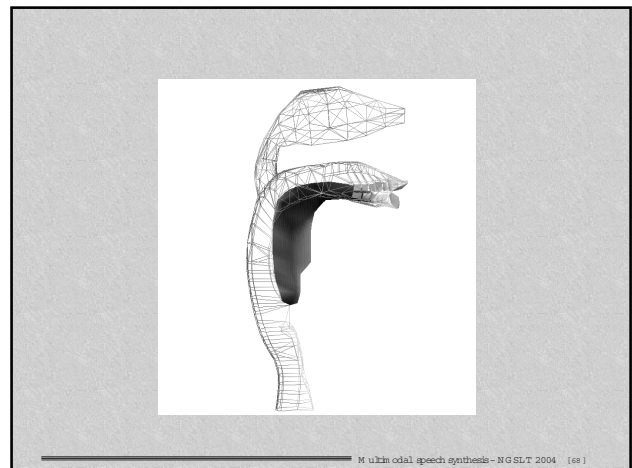


Multimedial speech synthesis - NG SLT 2004 [66]

From areas to formants

A transfer function is determined from the cross-sectional areas

Multimedial speech synthesis - NG SLT 2004 [67]



Ranges of parameter activations

Jaw height Tongue body Tongue dorsum

Multimedial speech synthesis - NG SLT 2004 [69]

Multimodal synthesis

Multimedial speech synthesis - NG SLT 2004 [70]

Talking heads - Applications

- Improved speech synthesis
- Human-Computer Interface in spoken dialogue systems
- Aid for hearing in paired
- Educational software
- Stimuli for perceptual experiments
- Entertainment: games, virtual reality, movies etc.

Multimedial speech synthesis - NG SLT 2004 [71]

A new paradigm for human-computer interaction

- Shift from desktop-metaphor to person-metaphor
- Spoken dialogue as well as non-verbal communication
- Take advantage of the user's social skills
- Strive for believability, but not necessarily realism

Listening & Thinking

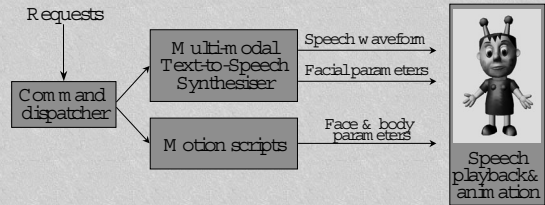
Multimedial speech synthesis - NG SLT 2004 [72]

Tasks of an Animated Agent

- Provide intelligible synthetic speech
- Indicate emphasis and focus in utterances
- Support turn-taking
- Give spatial references (gaze, pointing etc)
- Provide non-verbal back-channeling
- Indicate the system 's internal state

M ultim odal speech synthesis - NG SLT 2004 [73]

Animated Character - architecture



M ultim odal speech synthesis - NG SLT 2004 [74]

Parameters used for articulatory control of the face.

- Jaw rotation
- Lip rounding
- Lip protrusion
- Mouth width
- Bilabial closure
- Labiodental closure
- Upper lip raise
- Lower lip depression
- Apex
- Tongue length
- + more for prosody, attitude, emotions, turn-taking, back-channeling, pointing

M ultim odal speech synthesis - NG SLT 2004 [75]

The WaveSurfer Tool



- Interface is based around WaveSurfer, a general purpose tool for speech and audio viewing, editing and labelling
- TTS and Talking Head functionality is added as plug-ins
- WaveSurfer (presently without TTS & TH) works on all com m on platform s and is freely available as open source

<http://www.speech.kth.se/wavesurfer>

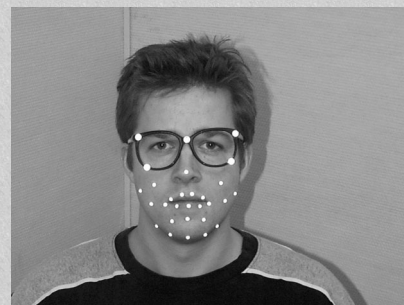
M ultim odal speech synthesis - NG SLT 2004 [76]

WaveSurfer Tool Demo



M ultim odal speech synthesis - NG SLT 2004 [77]

How to obtain data?




Qualisys recordings in Linköping


M ultim odal speech synthesis - NG SLT 2004 [78]

Combining model and data

Re-synthesis using speech movement recorded with Qualisys



Multimedial speech synthesis - NG SLT 2004 [79]



Preparing future multisensorial interaction research

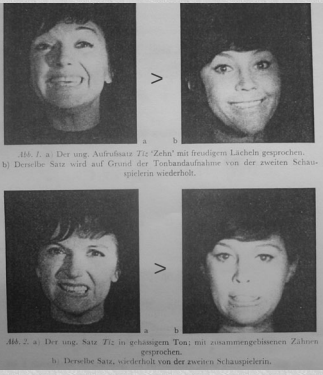
1. technologies for speech-to-speech translation
2. detection and expressions of emotional states
3. core speech technologies for children

EU project: start October 2002, duration 2 YR
 ITC-IRST (Trento) co-ordinates + 3*Germany + Italy + UK + Sweden
<http://pfstar.itc.it/>

Multimedial speech synthesis - NG SLT 2004 [80]

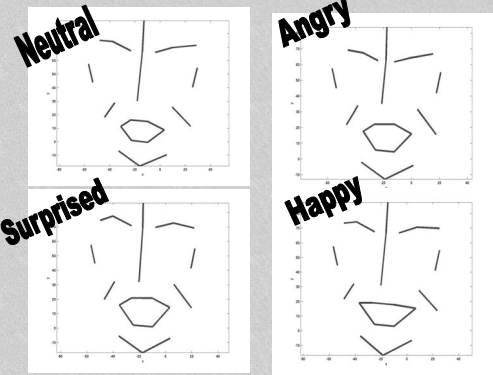
Vision from audio

original > m i n i c from audio



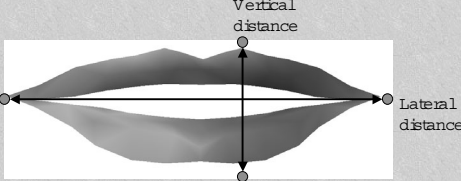
Fónagy, 1967 "Hörbare Mimik", Phonetica

Multimedial speech synthesis - NG SLT 2004 [81]



Multimedial speech synthesis - NG SLT 2004 [82]

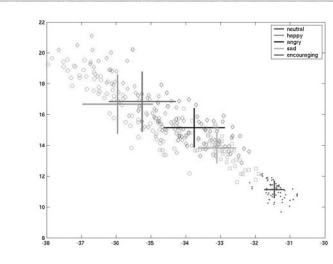
Measurement points for lip coarticulation analysis



Multimedial speech synthesis - NG SLT 2004 [83]

The expressive mouth


- All vowels (sentences)
 - Encouraging
 - Happy
 - Angry
 - Sad
 - Neutral



Multimedial speech synthesis - NG SLT 2004 [84]

Interactions: emotion and articulation

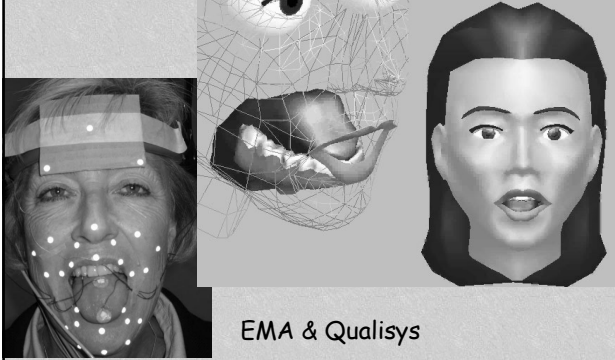
(from AV speech database - EU/PF_STAR project)



M ultim odal speech synthesis - NG SLT 2004 [85]

Combining motion capture techniques

Example of resynthesis




EMA & Qualisys

M ultim odal speech synthesis - NG SLT 2004 [86]


Collection of audio-visual databases: interactive spontaneous dialogues

- * Eliciting technique: information seeking scenario
- * Focus on the speaker who has the role of information giver
- * The speaker sits facing 4 infrared cameras, a digital video-camera, a microphone. The other person is only video recorded.



M ultim odal speech synthesis - NG SLT 2004 [87]

Recording and model



M ultim odal speech synthesis - NG SLT 2004 [88]

Conversation with agent



M ultim odal speech synthesis - NG SLT 2004 [89]

Eyebrow vs intonation

Eyebrow Motions in
Multimodal Speech Synthesis

- 1 No eyebrow motion
- 2 Eyebrow motion controlled by the fundamental frequency of the voice
- 3 Eyebrow motion at focal accents +
- 4 Eyebrow motion at the first focal accent +

"Jag heter Axel, inte Axell" (translation: "My name is Axel, not Axell"). In Sweden Axel is a first name as opposed to Axell, which is a family name.

M ultim odal speech synthesis - NG SLT 2004 [90]

Experiment

- Speech material
 - När pappa fiskar stör, piper Putte
When dad is fishing sturgeon, Putte is whimpering
 - När pappa fiskar, stör Piper Putte
When dad is fishing, Piper disturbs Putte
- 6 versions
 - 1 static, 5 eyebrow raising on successive content words
- 20 stimuli (6 x 3) plus first and last
- Subjects: 21 students at KTH
 - 14 native Swedish, 7 non-Swedish

Multimedial speech synthesis - NG SLT 2004 [91]

Eyebrow movement

- Hand edited with a synthesis parameter editor
- 500 ms
 - 100 ms dynamic rise
 - 200 ms static raised
 - 200 ms dynamic lowering

Multimedial speech synthesis - NG SLT 2004 [92]



Multimedial speech synthesis - NG SLT 2004 [93]

TWO EXAMPLES

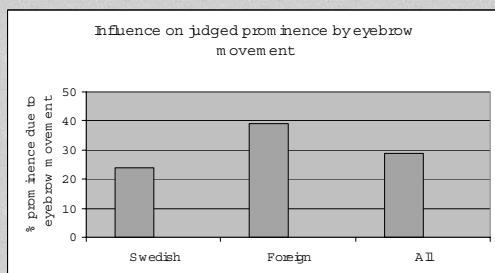


No eyebrow movement (neutral)

Eyebrow movement

Multimedial speech synthesis - NG SLT 2004 [94]

Prominence increase due to eyebrow movement




Multimedial speech synthesis - NG SLT 2004 [95]

Conclusions

- Eyebrow movement can be an independent cue to prominence
- Non-native Swedish listeners rely more on the visual cues
- Interaction
 - visual and acoustic cues
 - visual cues and prominence expectation
- Further work on interaction
 - prominence, mood and attitude (demo)

Multimedial speech synthesis - NG SLT 2004 [96]


Examples on the use of eyebrow and head motion (from the August dialogue system)



Translation: "Symmetrical works of art easily become dull just like symmetrical beauties; impeccable or flawless people are often unbearable." (Strindberg 1907)


Multimedial speech synthesis - NG SLT 2004 [97]

Different characters



Multimedial speech synthesis - NG SLT 2004 [98]

Talande ansikten på Tekniska Museet - Utställningen Fritt Fram



Multimedial speech synthesis - NG SLT 2004 [99]

Talteknologitillämpningar - exempel

- Talarverifiering
 - säkerhet
- Översättning eller översättningshjälp
 - språkidentifiering, ämnesbestämning, temabögi
- Studiehjälp
 - talmodig lärare
 - språkinläring, uttalundervisning
- Interaktiva informationssystem /diagnosystem
 - även textgenerering
- Indexering och sökning
 - radio och TV program

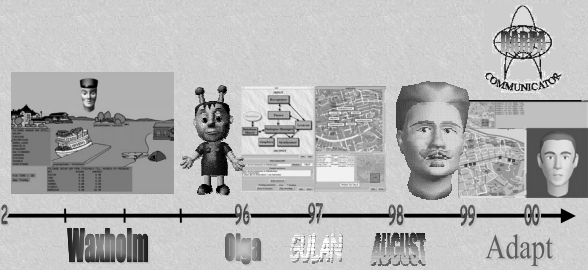
Multimedial speech synthesis - NG SLT 2004 [100]

Talteknologitillämpningar - exempel

- Handikapphjälpmedel
 - synskadade, tal-skadade, omgivningskontroll
- Telefonsjänster
 - kundhjälp
 - intelligenta "telefonsvarare", vem skickar "e-mail"
 - informationssökning, telefonhandling
- "Fria händer"
 - diktering
 - mobiltelefon, trafikinformation
 - sortering, kvalitetskontroll

Multimedial speech synthesis - NG SLT 2004 [101]

Dialog systems at KTH



Multimedial speech synthesis - NG SLT 2004 [102]

Talteknologi för synskadade

- "Design for all" eller speciella behov
- Första talsyntestillämpningen
- Naturlighet vs. uppfattbarhet
- Skärm läsare vs. GUI
- Talböcker/talrättningar
- Speech browsing, ASR, dialogsystem
- Snabbsyntes - 500 wpm

Multimedial speech synthesis - NG SLT 2004 [103]

Talsyntes för talskadade Cameleon CV - talprotesen från Vaessprojektet



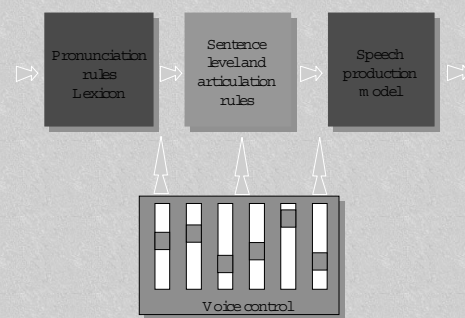
Multimedial speech synthesis - NG SLT 2004 [104]

User controlled "voice fitting"

- Direct access to selected voices
- Individual settings easy to use
- Phonetic rules use slide buttons as inputs
- Synthesizer implemented with great flexibility
- Examples of possible adjustments
 - Vocal tract size
 - Voice source characteristics
 - Pitch dynamics
 - Degree of clear or reduced speech

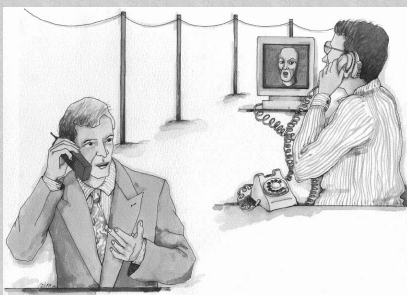
Multimedial speech synthesis - NG SLT 2004 [105]

System architecture



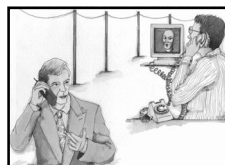
Multimedial speech synthesis - NG SLT 2004 [106]

Talsyntes för hörselskadade The Teleface application



Multimedial speech synthesis - NG SLT 2004 [107]

"The TELEFACE project" (simulated)



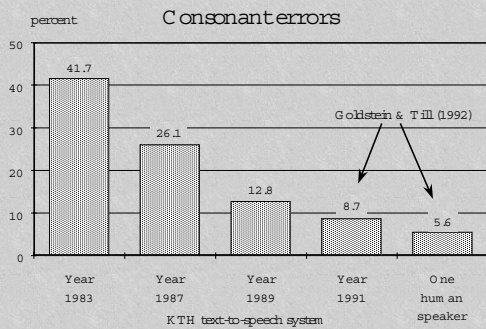
Multimedial Speech communication for the hearing impaired



Continues in EU project SYNFACE, aiming at a real-time demonstrator

Multimedial speech synthesis - NG SLT 2004 [108]

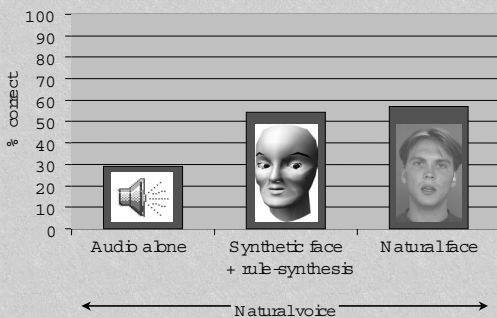
Evaluation of synthesis- VCV test



Formal intelligibility test

- Material: VCV (symmetric vowel context)
 - 2 vowels: /ʊ, a/
 - 17 consonants: /p, b, m, f, v, t, d, n, s, l, r, k, g, ɔ, ŋ, ç, j/
- Task: consonant identification
- Synthetic face with human speech
- hard of hearing subjects (or KTH students)
- Additive white noise, -3 dB SNR (if normal hearing)

Results for VCV-words (hearing in paired subjects)



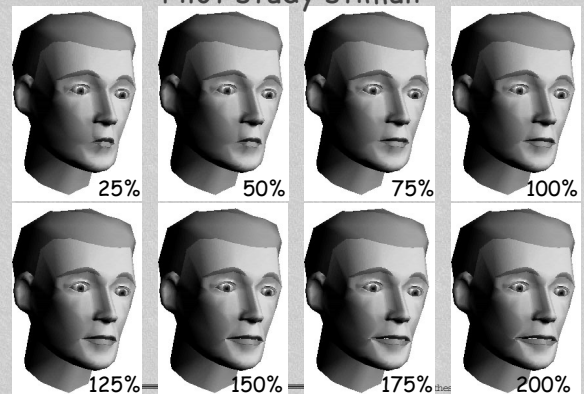
Better than humans?

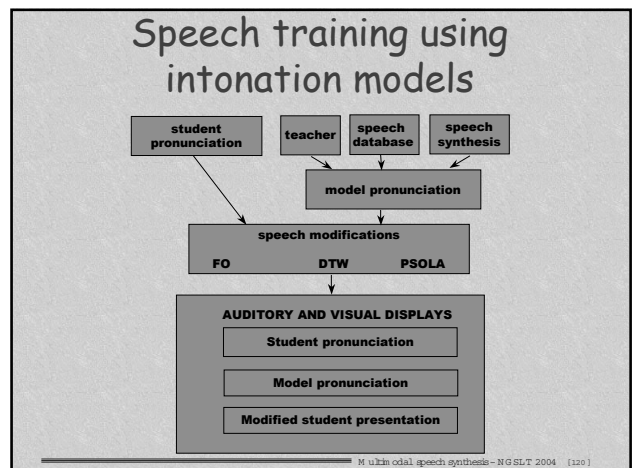
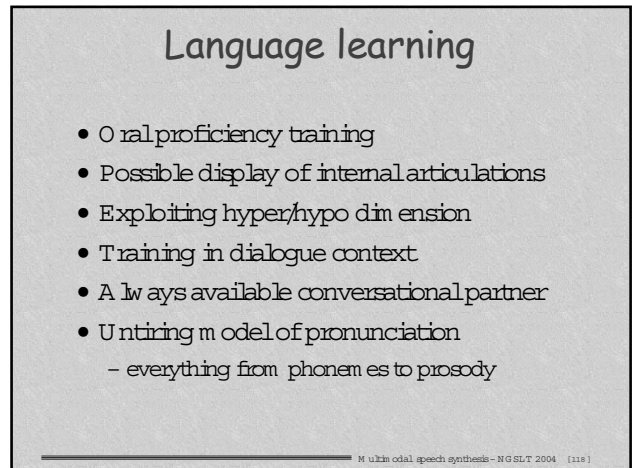
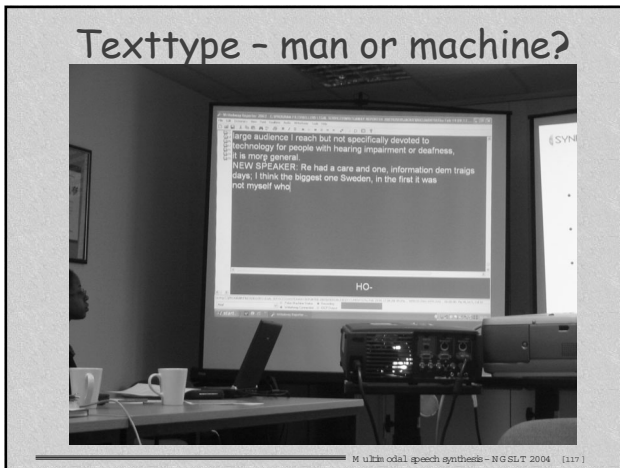
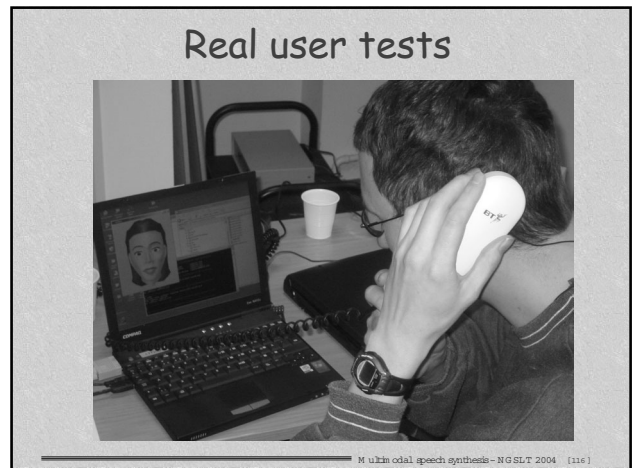
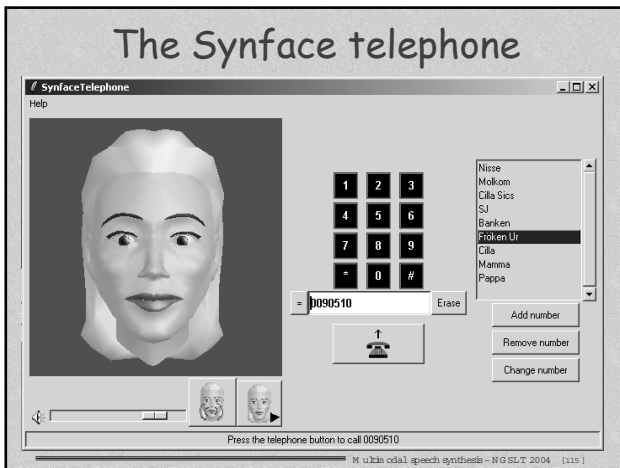
	bil	bbd	den	pal	vel		bil	bbd	den	pal	vel
aCa	bilabial	3,7						2,5	1,3		
	labiodental		3,7					5,6	1,9		
	dental	3,0	78,0	5,5	13,4			85,8	7,4	6,8	
	palatal		9,9	70,4	19,8			1,2	17,3	71,6	9,9
	velar		4,9	16,0	79,0			2,5	25,0	72,5	

Possible improvements to "lip readability"

- Great variation in human speakers due to for example
 - Speaking rate
 - Extent of articulatory movements (the hypo - hyperdimension)
 - Anatomy, facial hair...
 - Light, distance, viewing angle..

Hypo to hyper articulation Pilot study stimuli





Demo of prototype

Pohlmän - weatherman from the south

"Sen drar hela det här moln- och regnområdet i alla fall vidare österut" (~then, this whole cloud and rain system moves eastward)

1 Original recording

"Teacher" (sound only)-original-modified

2 Stockholm

3 South Swedish

4 Synthesis

Multimedial speech synthesis - NG SLT 2004 [121]

1 Original recording

2 Stockholm "Teacher" (sound only)-original-modified

3 South Swedish

4 Synthesis



Multimedial speech synthesis - NG SLT 2004 [122]

Articulatory training

- Stylized
- Program Fonem - Johan Liljencrants

Multimedial speech synthesis - NG SLT 2004 [123]

Reiko Yamada ATR, 1999



Multimedial speech synthesis - NG SLT 2004 [124]

new national project ARTUR

What?

Automatic articulatory feedback display using face and vocal tract models.

For whom?

Hearing impaired children, second-language learners, speech therapy patients.

How?

Contrasting the user's articulation with a correct one.



Multimedial speech synthesis - NG SLT 2004 [125]

CTT Virtual Language Tutor

- Practice dialogues
- Correct your pronunciation
- Keep track of your improvements
- Tailor lessons based on your interaction




Multimedial speech synthesis - NG SLT 2004 [126]

CTT Virtual Language Tutor

Different Types of Users:

- Swedish children learning English
- Adult immigrants learning Swedish
- Adult Swedes wanting to improve aspects of English (e.g. corporate English, technical English)
- Native Swedes with language disabilities wanting to improve their Swedish




Multimodal speech synthesis - NG SLT 2004 [127]

CTT Virtual Language Tutor

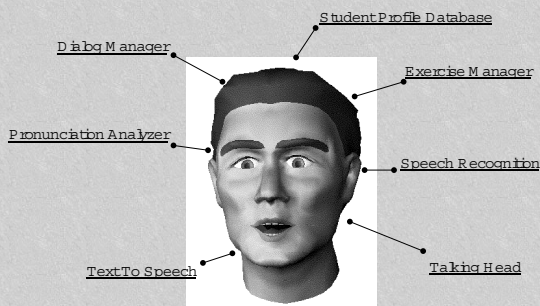
Separate:

- General tools from user specific tools
- Linguistically universal tools from language specific



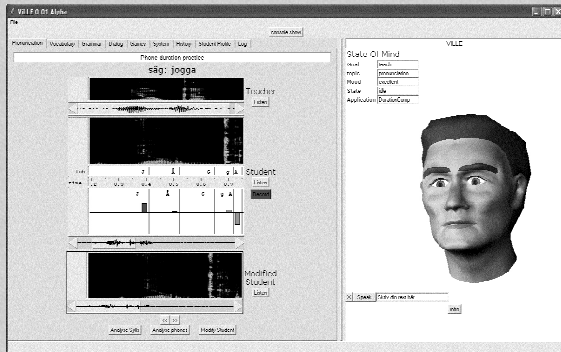
Multimodal speech synthesis - NG SLT 2004 [128]

CTT Virtual Language Tutor components



Multimodal speech synthesis - NG SLT 2004 [129]

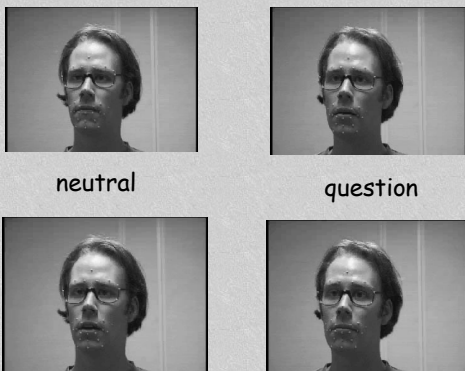
CTT Virtual Language Tutor



demonstration

Multimodal speech synthesis - NG SLT 2004 [130]

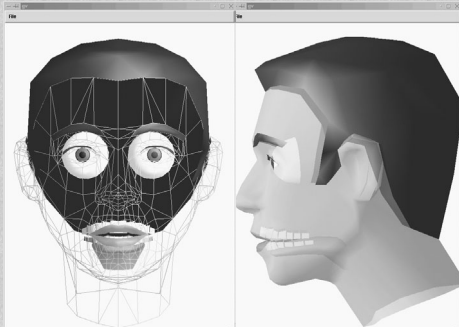
Scripted utterances



neutral question


Multimodal speech synthesis - NG SLT 2004 [131]

Different representations



Multimodal speech synthesis - NG SLT 2004 [132]

Automatic tutor simulation



no gestures some gestures

Multimedial speech synthesis - NG SLT 2004 [133]

Syntes: slutkommentar

- **Artikulatorisk syntes:** svårt med data, optimal kontroll med gränsvillkor givna!
- **Formantsyntes:** produktionsrelaterad, flexibel, svårt uppnå naturlighet
- **Konkateneringssyntes -PSOLA/MBROLA:** ej flexibel, hög naturlighet
- Dagens trend: **"Unit selection"** - stora databaser, minimal signalbehandling
- Ökad satsning på högre språkliga nivåer
- "Concept-to-speech" (t ex i dialogsystem)

Multimedial speech synthesis - NG SLT 2004 [134]

Bullet course at KTH

Nick Campbell, ATR 20-22 sept 9:30-11:30

- 1 Language, Speech, and Meaning

In this talk, I shall attempt to describe some of the roles played by prosody in speech communication, and will relate them to the requirements of computer speech processing. The talk covers phonetic, linguistic and paralinguistic aspects of speech.
- 2 Working with a Corpus of Expressive Speech

This talk describes the JST/CREST Expressive Speech Processing project, introduces a very large corpus of conversational speech and describes some of the main findings of our related research. The talk explores the roles of non-verbal and paralinguistic information in speech communication.
- 3 Synthesising Conversational Speech

This talk addresses the issues of synthesising non-verbal speech and describes a prototype interface for the synthesis of conversational speech. The synthesised samples are in Japanese, but I believe that they are sufficiently interesting that any inherent language difficulties might be overcome by higher-level speech-related interests.

Multimedial speech synthesis - NG SLT 2004 [135]

Summer school in Estonia, August 10-15, 2005

- Organized by the Nordic network VISPP
- VISPP - Variation in speech production and perception
- Focus on how to handle normal and unwanted variation - ASR, pathologies, second language
- Palmse conference centre
- <http://www.hfuiub.no/ilf/forskning/horfa/>

Multimedial speech synthesis - NG SLT 2004 [136]

Homework

- Experiment with concatenative speech synthesis
- Domain - three digit pronunciation
- Experiment with different unit sizes
- ...different speaking styles - emphatic, emotive, questioning etc.
- To be presented in Stockholm, Jan 2005
- Use WaveSurfer and/or own solutions

Multimedial speech synthesis - NG SLT 2004 [137]



HAL'S LEGACY: 2001'S COMPUTER AS DREAM AND REALITY

Chapter 6
"The Talking Computer": Text to Speech Synthesis
Joseph P. Olive

Chapter 7
When Will HAL Understand What We Are Saying?
Computer Speech Recognition and Understanding
Raymond Kurzweil

<http://mitpress.mit.edu/e-books/hal/>

Multimedial speech synthesis - NG SLT 2004 [138]

