Multimodal speech synthesis/ (N)GSLT - Speech Technology course

Björn Granström
CTT, KTH

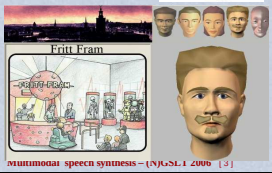
School for Computer Science and Communication

Department of Speech, Music and Hearing

CTT - Centre for speech technology

Research areas

- Speech production
- Speech perception
- Communication aids
- Multimodal speech synthesis
- Speech understanding
- Speaker characteristics
- Interactive dialog systems
- Affective computing
- Language learning



CTT - modes of operation

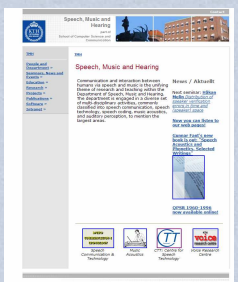
- long-term research projects,
- participation by CTT personnel in international projects,
- research exchange with leading foreign groups,
- graduate-level research training,
- dissemination of competence through the exchange of researchers within Sweden and collaboration with other Swedish research groups
- special information dissemination efforts

CTT partners phase 4 - 2004-2006

Babel-Infovox
EnglishTown
GN Resound
Hjälpmedelsinstitutet
HoneySoft
IcePeak
Dolphin Audio Publishing
LingTek
Phoneticom
Polycom Technologies
SaabTech
SpeechCraft
STTS
Sveriges Radio
Sveriges Television
TeliaSonera
TPB - Talboks- och punktskriftsbiblioteket
Vattenfall
VoiceProvider

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



Februari 1998

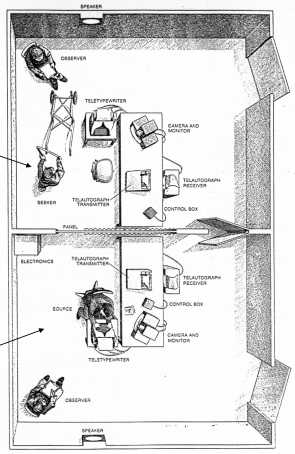
Why Speech Tech?

User

"Interactive Human Communication"

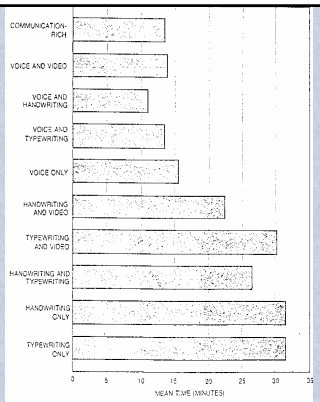
Chapanis
Scientific American 1975!

"Computer"



More efficient and convenient!

Average time for solving divers 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 10 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 voice is excluded in the five slower modes.

Speech is fast and verbose

	COMMUNICATION-RICH	VOICE	HANDWRITING	TYPEWRITING EXPERIENCED TYPISTS	TYPEWRITING INEXPERIENCED TYPISTS
SOLUTION TIME IN MINUTES	29	33	53.3	66.2	69
NUMBER OF MESSAGES	230.4	163.8	15.9	27.2	31.5
NUMBER OF SENTENCES	372.6	275.9	24.9	45.8	44.1
TOTAL NUMBER OF WORDS	1,563.8	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	10.2

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.

Speech technology is making money!

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

- 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

The first pen phone

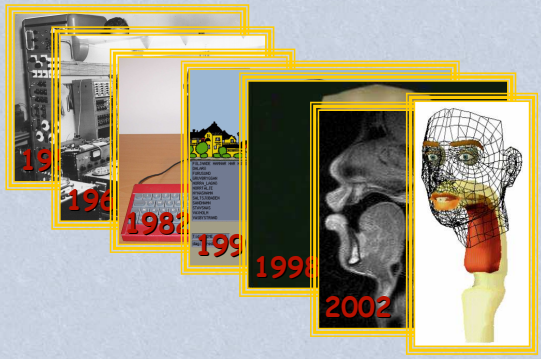
Why no commercial success?

Too small keys?

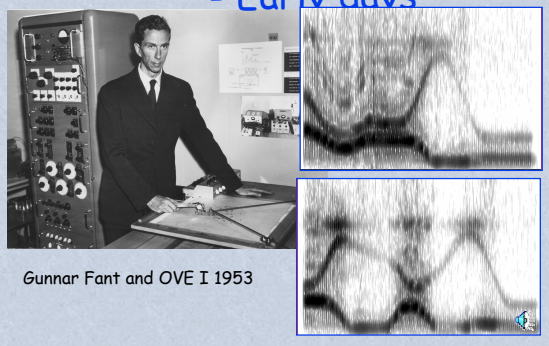
Too small display!?

1 april 1998

Speech synthesis developments at KTH



The KTH speech group - Early days



Gunnar Fant and OVE I 1953

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`

Ove II, 1958

1961

1962

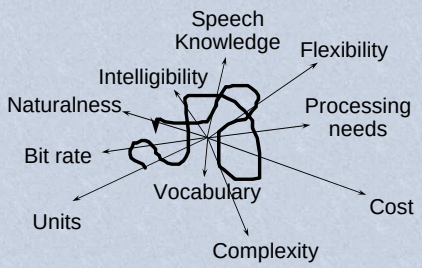
KTH/TTS history

- 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
  - MC 68000, NEC 7720
- 1983, Founding of Infovox Inc.

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 rules for prosody
  - One speaker
- **Parametric synthesis**
  - Formant synthesis
  - Articulatory synthesis
  - flexible
  - But lower quality - today
- **Mulitmodal synthesis**

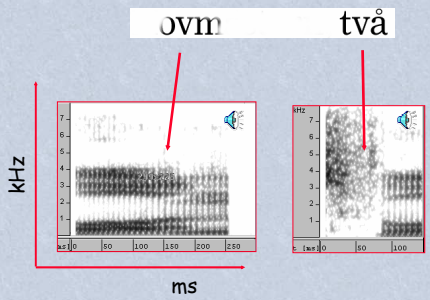
The synthesis space



TEXT vs SPEECH

- Parallel vs. sequential
- permanent vs disappearing
- Text as transcription of speech?
- Example - WaveSurfer demo :
  - Palindrome
  - ”Var sak har två sidor”

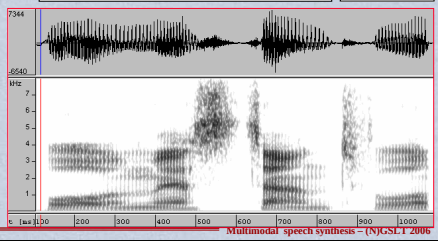
Phonetics



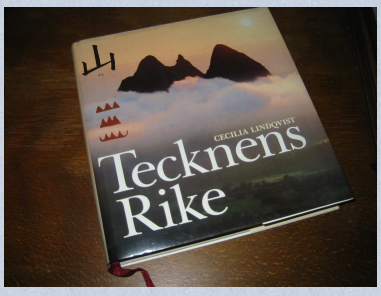
Phonology, Morphology

Hovmästarn, två

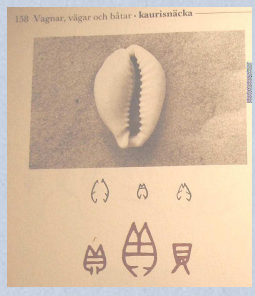
hov#mäst-ar-n | två

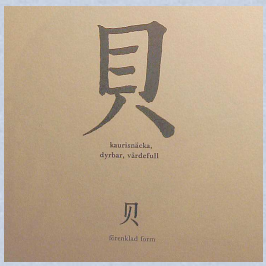


Non-phonetic writing

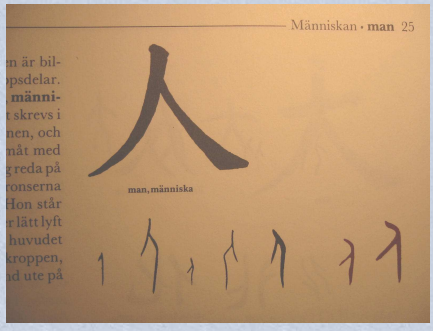


Signs vs. technology



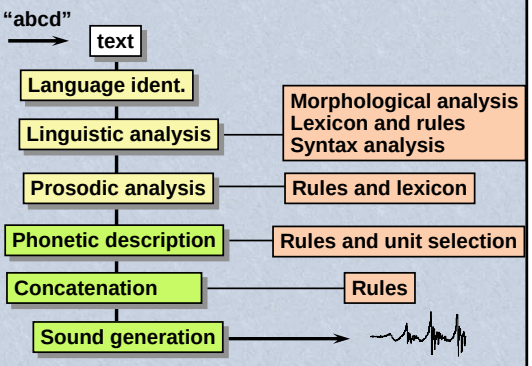


Is the sign what you think?

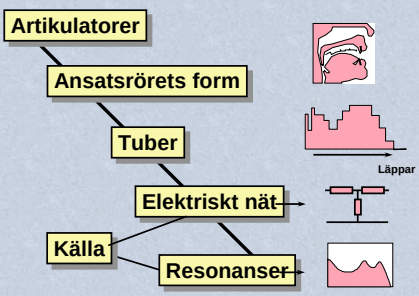


How is a Chinese lexicon organized?

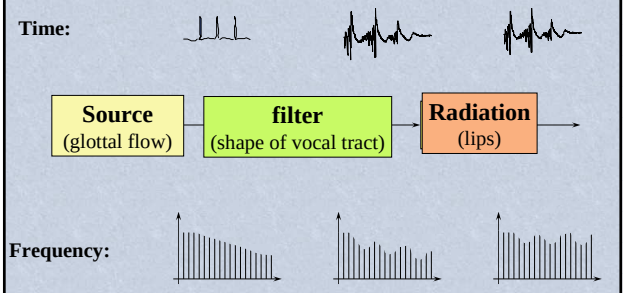
Text-to-speech (TTS)



Synthesis methods



Source filter theory



RULSYS Rules - Features

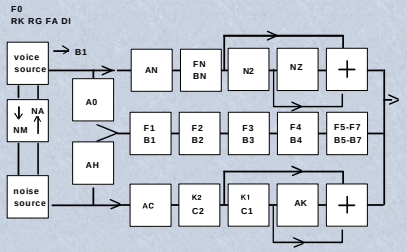
- Is IPA synthesiser possible?
- Based on generative phonology
- Language specific definitions
- Rules for contextual modifications
- Examples
- Interactive rule manipulations

(a) a → e / __ <cons> e #

(b)

morph		root	
graph	a	e	
phon	<cons>	→	e
	^ ^		^ ^

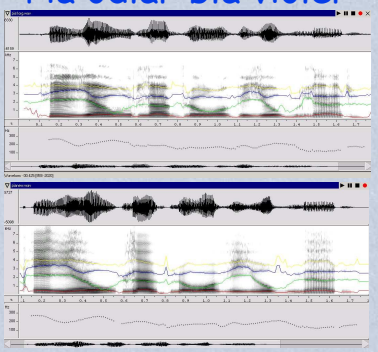
Glove



"Pia odlar blå violer"

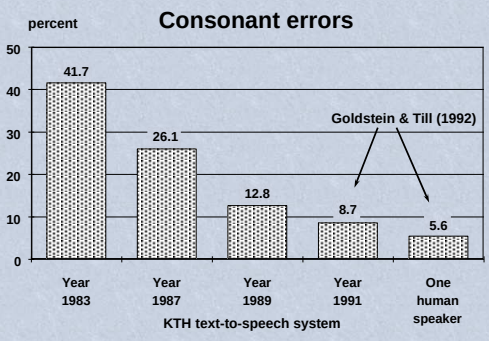
Original

Synthesis



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

Evaluation of synthesis- VCV test



Speaker characteristics

Speaker
dialect, sex, social, education, age

Situation
formality, style, interspeaker relation

Complex description
many dependent variables

Speaker characteristics

- TIDE/Vaess project (Voices, attitudes and emotions in synthetic speech
- "Voice fitting"
- Software, Glove 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

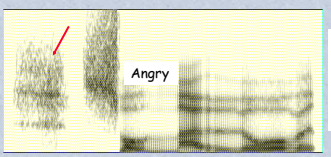
Emotions

natural synthesis

Neutral

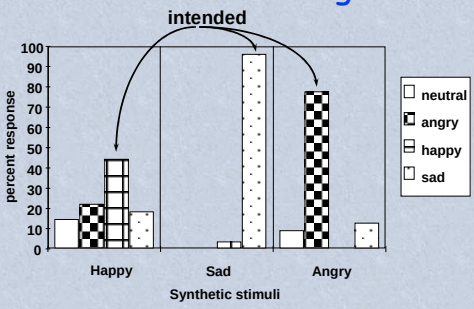
Happy

natural synthesis



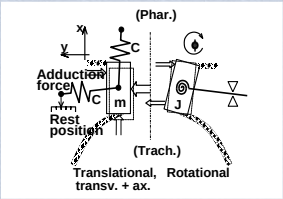
natural synthesis

Result from listening test

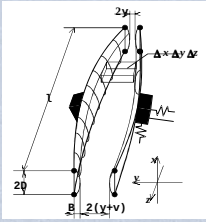


Emotions

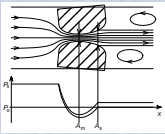
Voice source simulation



Mechanical model



Geometric model with two elastic rods



Air flow through glottis

Voice source simulation examples

Samplingsintervall:	1 parameter
Stämbandsanatomi:	18 parametar, t ex längd, vikt, dämpningsfaktor
Stämbandsartikulation:	7 parametrar, t ex restglapp, lungtryck
Högtrycksartikulation:	4 parametrar, t ex tonhöjd, ljudstyrka
Talrörsartikulation:	8 parametrar, t ex area, dämpningsfaktor

Exempel	*originalyttrande*
fördubblad stämbandsmassa	
förlängning av stämband 13->17 mm	
assymetriförändring 1.0->1.8	
assymetriförändring 1.0->2.0	
restgap 0.1->1.0 mm	

Predictable word accents

- Word accent and stress
  - 'anden (bird)
  - 'ànden (spirit)
  - 'ànk-pasta
  - 'ànk-pastej
  - 'ànk-pastejen
  - 'ànk-pastejs-mål-tid
- Name pronunciation
  - karl-'èrik (karl-'erik) + 'ànne-marie
  - Onomastica/EU
- Reduction
  - be'tòng-väg(g)s-konstruktion (C-C Elert)

Letter-to-sound vs. lexicon

Size vs. Precision

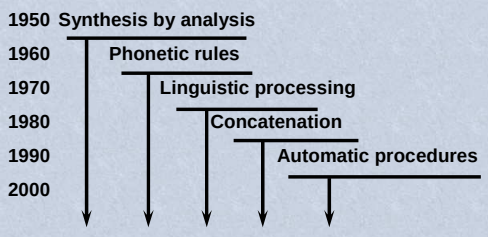
Maintenance - "half of the words are unique"

Name pronunciation (Onomastica Project)

Misspellings

Use rules, morphology, analogy…as fall back

Research trends in speech synthesis



Concatenative synthesis

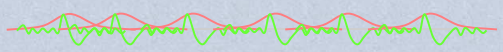
- Already Peterson et al. (1958)
- Dixon and Maxey (1968)
- "Diadic Units", (Olive, 1977)
- "PSOLA", Charpentier and Stella (1986)
- Review, Möbius (2000)
- Unit selection
  - Concatenation cost
  - Quality/size of database
  - Prosody/speaking styles

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

Speak&Spell - TexasInstrument Christmas 1978

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
- Annotation 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.ims.unistuttgart.de/~moehler/synthspeech/examples.html

also e.g. *Synthesis examples*:
http://www.naturalvoices.att.com/
http://www.acapela-group.com/demos/demos.asp
http://www.naturalvoices.att.com/
http://www.nextup.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)*

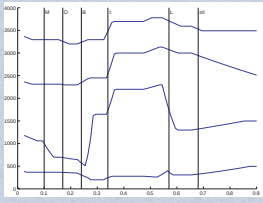
- *All available on* www.speech.kth.se

Aim

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

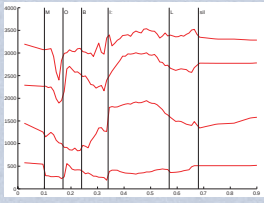
Rule-driven formant synthesis

- 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.)

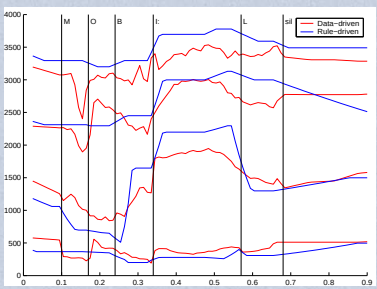


Data-driven formant synthesis

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



Synthesis comparison



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

Cost function

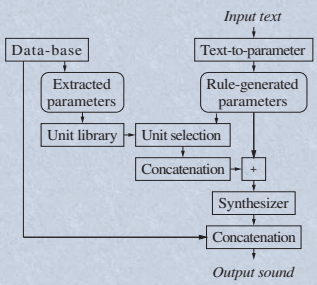
- Designed to promote probable formant candidates
- Penilizes:
  - Large bandwidths
  - Large frequency deviations (given the current phoneme)
  - Large frequency jumps (promotes smooth trajectories)

Voiceless consonants

- Replace the voiceless fricatives and plosives with recorded versions
- Voiceless fricatives in Standard Swedish: /f/, /s/, /sj/, /tj/, and /rs/
- Voiceless plosives: /k/, /p/, /t/, /rt/

(/h/ is excluded)

Text-to-speech synthesis



Listening test evaluation 1

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

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

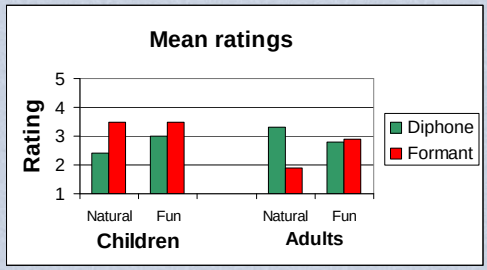
Sound samples

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

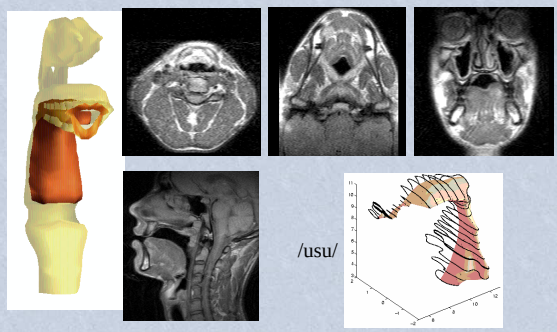
- ”Pytteliten”, rule-driven:
- ”Pytteliten”, data-driven:

Child-directed speech synthesis

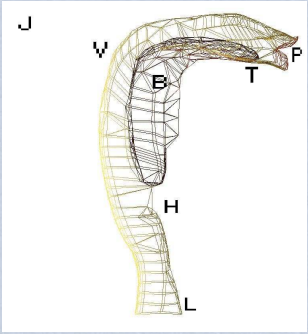
- Increase prosodic variation in synthesis
- How do children between the ages of 9 and 11 react to:
  - default, F0, duration?
  - diphone synthesis vs formant synthesis?



Articulatory synthesis

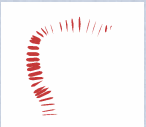


Articulatory parameters



- Jaw opening
- Lip rounding
- Lip Protrusion
- Tongue position
- Tongue height
- Tongue tip
- Velum
- Hyoid

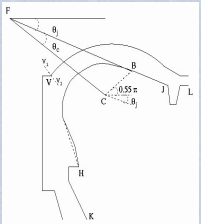
Articulatory synthesis potential use



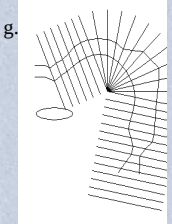
- Articulatory synthesis
  - Calculations directly from cross sectional areas
  - Fluid dynamics calculations

- Visual synthesis
  - Articulation training
- Demonstrations and research

Why a 3D model?

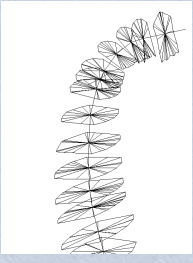


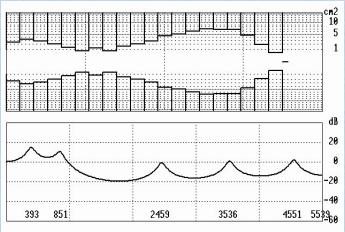
- 2D articulatory models, e. g.
  - Mermelstein
  - Maeda
- Imposing a third dimension: area=a•(width)^b
- A 3D model:
  - Direct calculation
  - laterals etc.

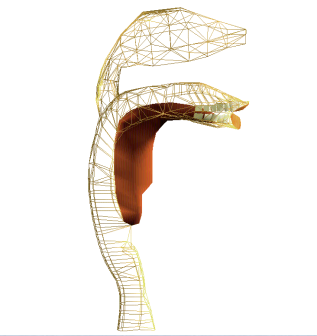


From areas to formants

A transfer function is determined from the cross-sectional areas



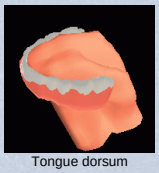




Ranges of parameter activations

Jaw height

Tongue body



Tongue dorsum

Multimodal speech synthesis

Talking heads - Applications

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

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

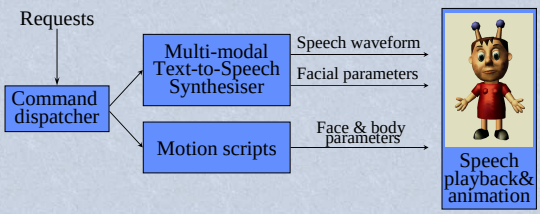
Conventions? – use same as for person-to-person communication

SONY SDR

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

Animated Character - architecture



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

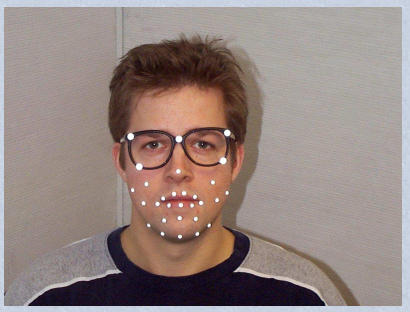
The WaveSurfer Tool

- 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

`http://www.speech.kth.se/wavesurfer`

WaveSurfer Tool Demo

How to obtain data?



Qualisys recordings in Linköping

Combining model and data

Re-synthesis using speech movement recorded with Qualisys

pf-star

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/

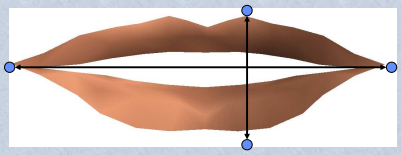
Vision from audio

original

mimic from audio

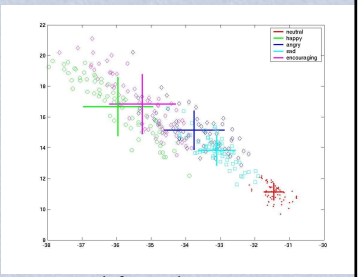
Fónagy, 1967 "Hörbare Mimik" , Phonetica

Measurement points for lip coarticulation analysis



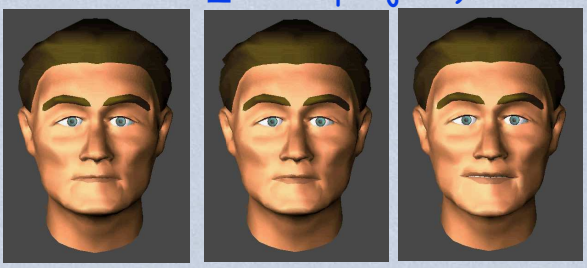
The expressive mouth

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

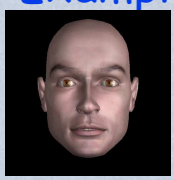


"left mouth corner"

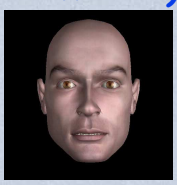
Interactions: emotion and articulation (from AV speech database – EU/PF_STAR project)



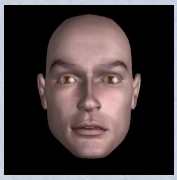
Examples (automatic)



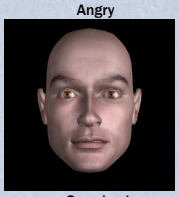
Happy



Angry

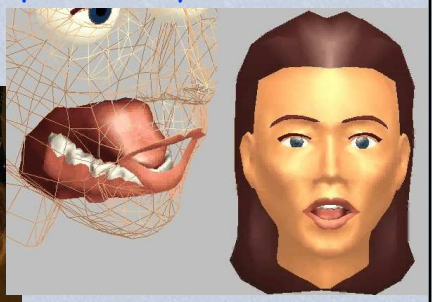


Sad



Surprised

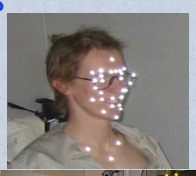
Combining motion capture techniques Example of resynthesis



EMA & Qualisys

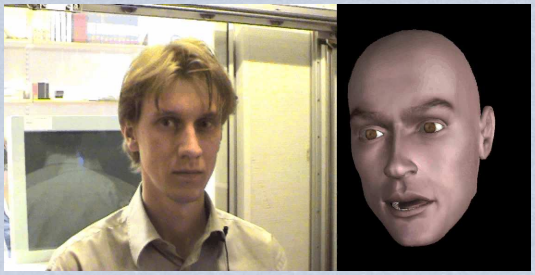
Collection of audio-visual databases: interactive spontaneous dialogues

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



Recording and model

Conversation with agent



Eyebrow vs intonation

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.

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

Eyebrow movement

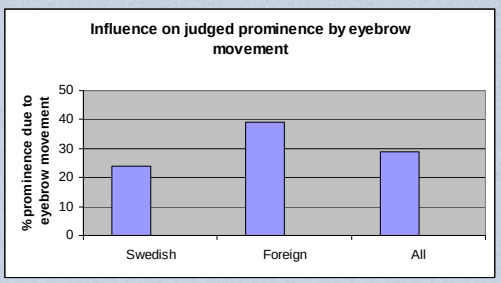
- Hand edited with a synthesis parameter editor
- 500 ms
  - 100 ms dynamic rise
  - 200 ms static raised
  - 200 ms dynamic lowering

TWO EXAMPLES

No eyebrow movement (neutral)

Eyebrow movement

Prominence increase due to eyebrow movement



Feedback experiment

- Mini dialogues (two turns)
- Travel agent application
- Both visual and acoustic feedback cues
- Affirmative cues – agent understands/accepts the request
- Negative cues – agent is unsure about the request (seeks confirmation)
- Six cues hypothesised

Granström, House & Swerts (2002)

Parameter settings to create different stimuli

	Affirmative setting	Negative setting
Smile	Head smiles	Head has neutral expression
Head movement	Head nods	Head leans back
Eyebrows	Eyebrows rise	Eyebrows frown
Eye closure	Eyes close a bit	Eyes open widely
F0 contour	Declarative intonation	Interrogative intonation
Delay	Immediate reply	Slow reply

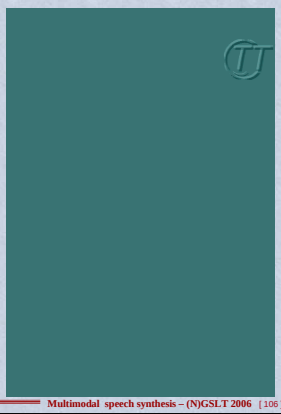
Cue strength - demonstration

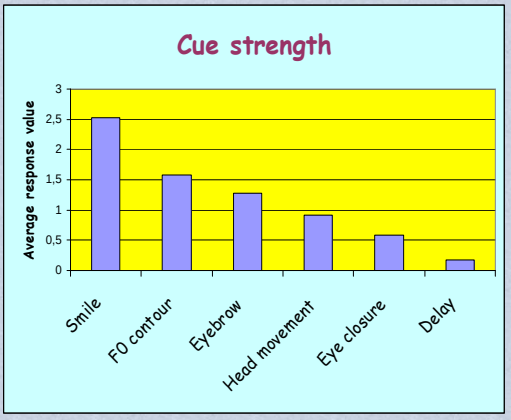
Human: "Jag vill åka från Stockholm till Linköping"

Agent: "Linköping"

Affirmative cues:

- None
- All
- Smile (only)
- F0 contour (only)
- No brow frown (only)
- Head movement (nod) (only)
- Eye closure (only)
- Delay (only)
- All





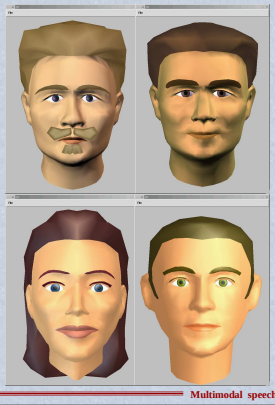
Conclusions

- 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)

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)

Different characters

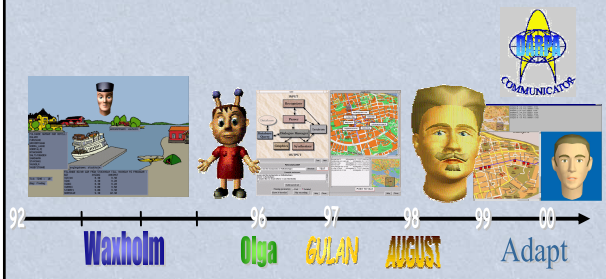


Talking heads on the Science Museum, Stockholm(
(Exhibition: FrittFram)

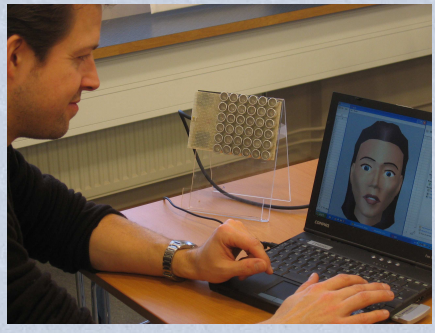
Speech synthesis applications

- Dialogue systems (at next NGSLT at KTH)
- Translators
- Aids for disabled
- Mediated communication
- New HMI
- Language learning
- ....

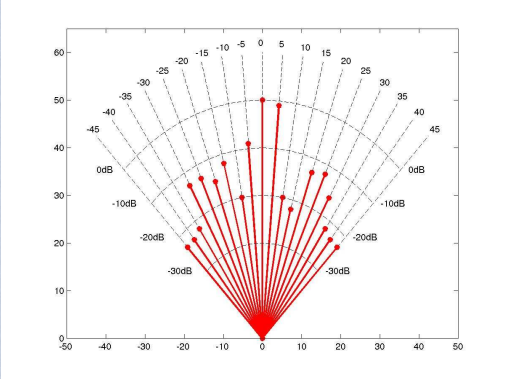
Dialog systems at KTH



Targeted audio & talking head for personal announcements (EU/Chil project)



Directionality



Speech technology for visually impaired persons

- First synthesis application
- Intelligibility vs. naturalness
- Screen reader vs. GUI
- Talking books and newspapers
- "Design for all" or special demands
  - E.g. Rapid speech - 500 wpm

Aid for speech impaired The EU/Vaess project

Vaess - Voices, attitudes and emotions in speech synthesis
The Vaess Text-to-Speech converter enables speech impaired persons to communicate by individualised speech

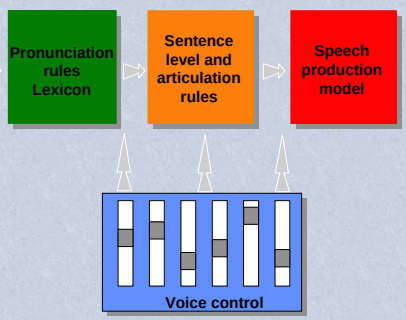
Synthesis Software Development

- Improved speech synthesis for disabled and elderly people
- Danish, British English, Spanish and Swedish speech synthesis
- Development of new voices
- Development of different speaking styles
- Experiments with new synthesis strategies for emotive speech
- Uses an extension of the KTH/Infovox Speech Synthesizer
- Based on analysis of human speech databases

User controlled "voice fitting"

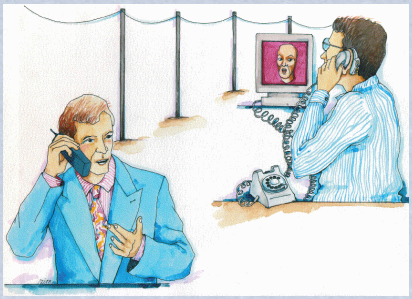
- Direct access to selected voices
- Individual settings easy to use
- Phonetic rules use slide buttons as inputs
- Synthesizer implemented with great flexibility
- Examples of possible adjustments
  - Vocal tract size
  - Voice source characteristics
  - Pitch dynamics
  - Degree of clear or reduced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

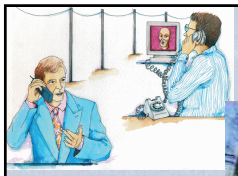
System architecture



Speech synthesis for non-vocal persons
Cameleon CV – a speech prosthesis (EU/Vaess project)

Speech synthesis for hard of hearing persons
The Teleface application





"The TELEFACE project" (simulated)

Multi-modal
Speech
communication
for
the hearing
impaired

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

EU-project Synface – Coordinated by KTH

n | Change edition

BBC NEWS WORLD EDITION

Last Updated: Saturday, 31 July, 2004, 23:58 GMT 00:58 UK

E-mail this to a friend Printable version

Phone success for hard of hearing

A computer that generates pictures of moving faces from speech is helping hard of hearing users.

The technology, known as Synface, was hailed a success by the 40 people with hearing problems who trialled a prototype in the UK.

Phone conversations are animated

The software can be installed on a regular computer and is used with a standard telephone.

European scientists will use the trial results to tweak the device before it is made available in coming years.

Synface - synthetic face - was developed by researchers at the Royal Institute of Technology in Stockholm, Sweden, and University College London.

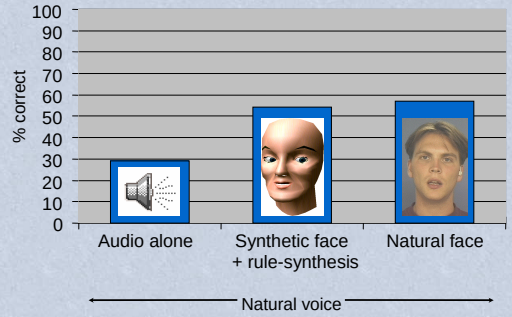
http://www.speech.kth.se/synface/

Instruction video

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, ŋ, ɧ, 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 impaired subjects)



Better than humans?

aCa

	bil	labd	den	pal	vel
bilabial	100				
labiodental		96,3	3,7		
dental		3,0	78,0	5,5	13,4
palatal			9,9	70,4	19,8
velar			4,9	16,0	79,0

Synthetic face

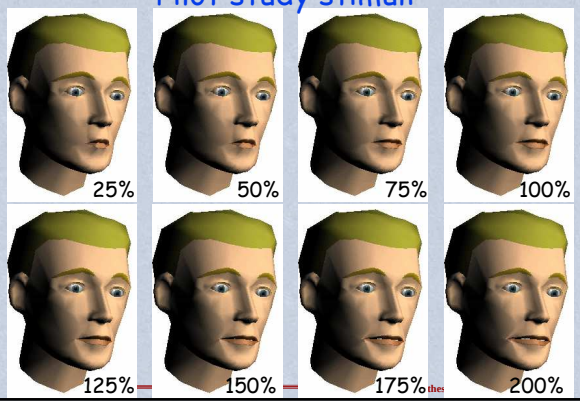
	bil	labd	den	pal	vel
bilabial	96,3		2,5	1,3	
labiodental		92,6	5,6	1,9	
dental			85,8	7,4	6,8
palatal		1,2	17,3	71,6	9,9
velar			2,5	25,0	72,5

Natural face

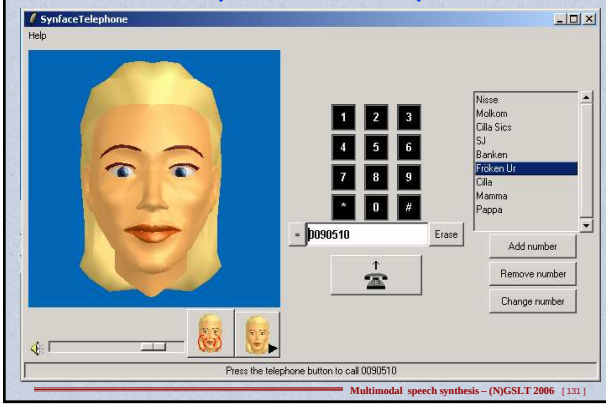
Possible improvements to "lip readability"

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

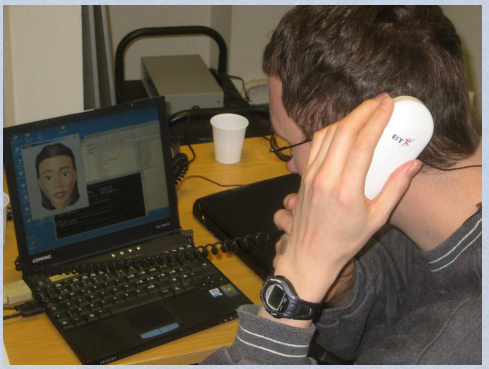
Hypo to hyper articulation
Pilot study stimuli



The Synface telephone



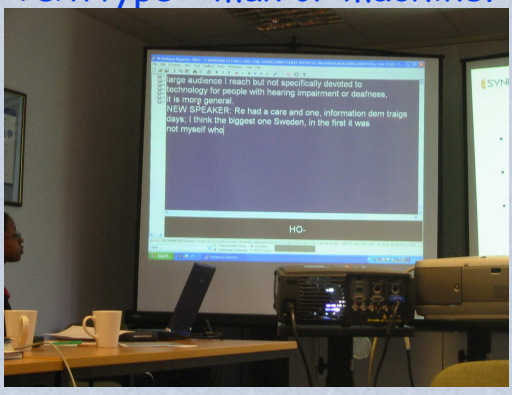
Real user tests



Jonas Beskow - mottagare av Chester Carlsons forskningspris 2 februari, 2006

Talking faces for speech reading support in TV

Texttype – man or machine?

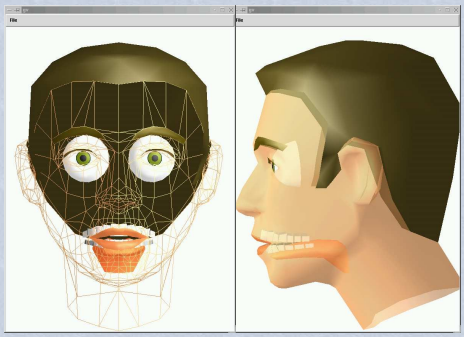


Language learning

- Oral proficiency training
- Possible display of internal articulations
- Exploiting hyper/hypo dimension
- Training in dialogue context
- Always available conversational partner
- Untiring model of pronunciation
  - everything from phonemes to prosody

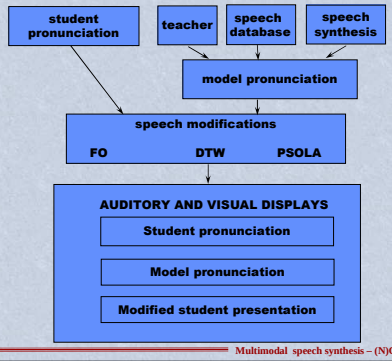
Automatic tutor simulation

Different representations



Speak&Spell - TexasInstrument Christmas 1978

Speech training using intonation models



Demo of prototype

Pohlman - 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

1 Original recording

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

3 South Swedish

4 Synthesis

Articulatory training

- Stylized
- Program Fonem –Johan Liljencrants

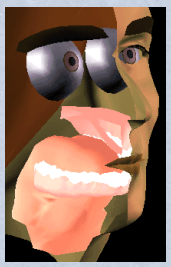
Reiko Yamada ATR, 1999

new national project ARTUR

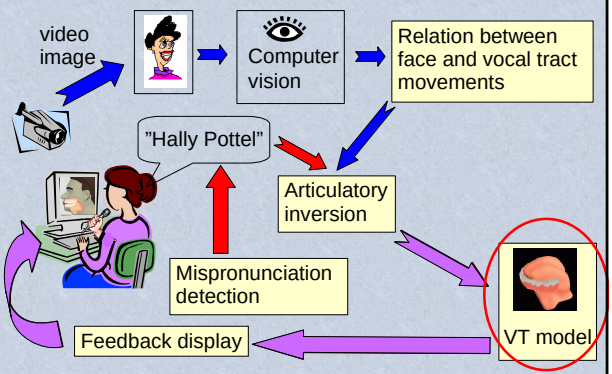
What?
Automatic articulatory feedback display using face and vocal tract models.

For whom?
Hearing impaired children, second-language learners, speech theraphy patients.

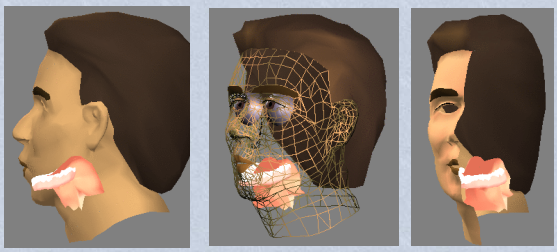
How?
Contrasting the user's articulation with a correct one.



ARTUR: the articulation tutor



3D tongue movements



[s] [k] vowels

"OK" 3D tongue movements (difficult to define an objective evaluation criteria - *suggestions?*)
The tongue movements can be presented from different views

CTT Virtual Language Tutor

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

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

Discriminate acceptable from unwanted variation

- How to do it (automatically)
- What are the aims of L2 learning
  - Less accentedness
  - Comprehensibility
  - Intelligibility
  - More? – Acceptability in context
- Economy of language learning
- Could a virtual tutor help?

The cost of pronunciation errors

- Video

The virtual teacher vision already at STiLL, Marholmen 1998!

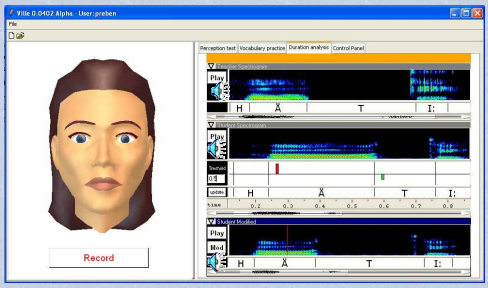
Unacceptable pronunciation needs to be identified

from a Vinnova video (on CTTs webpages)

First tried at the VISPP summerschool in Palmse, August 2005 (for Estonian)

Nordic project –NordPlus Sprog

Using VILLE – CTT Virtual Language Tutor



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/


Beyond 2001
Ready for the Odyssey?
[157]