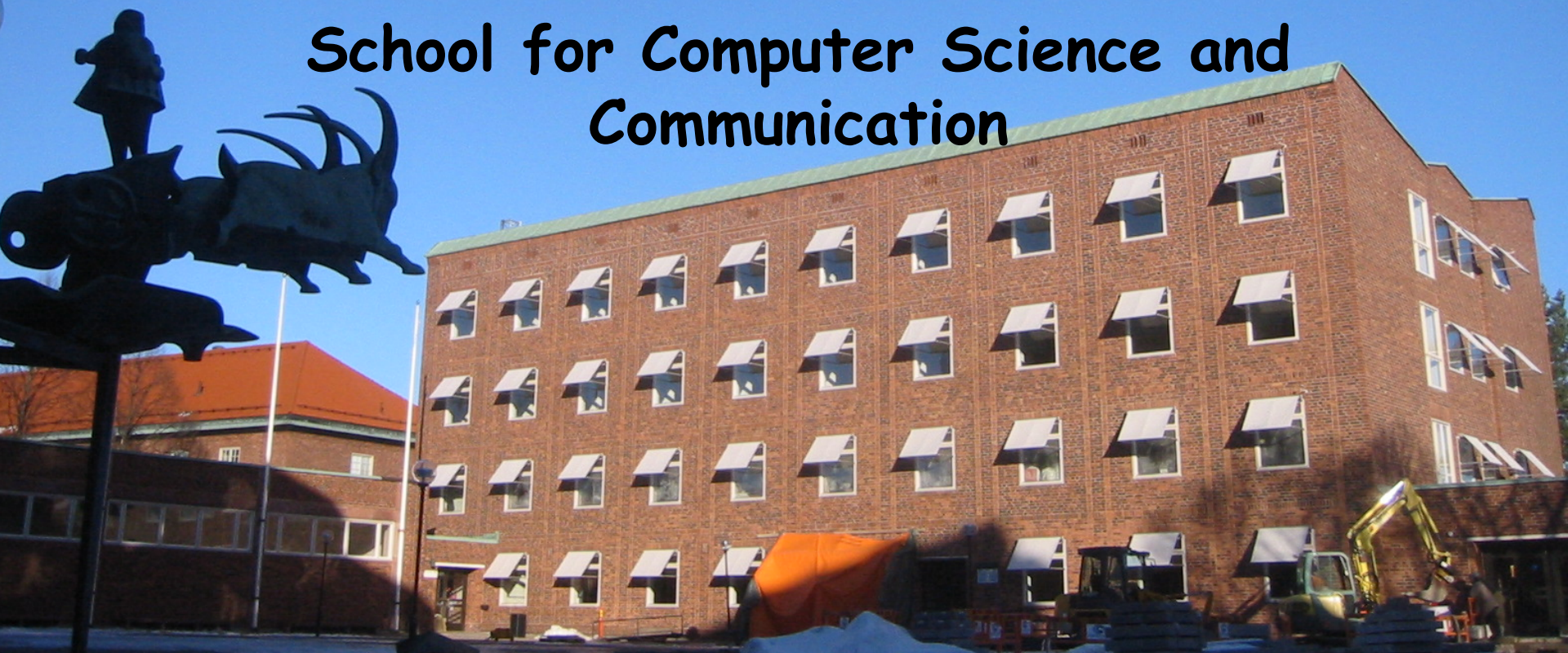
Multi-modal speech for talking interactive agents and their use in spoken dialogue systems

Björn Granström
CTT, KTH

School for Computer Science and Communication



KTH - Kungliga tekniska högskolan



Department of Speech, Music and Hearing

Kungliga tekniska högskolan, KTH

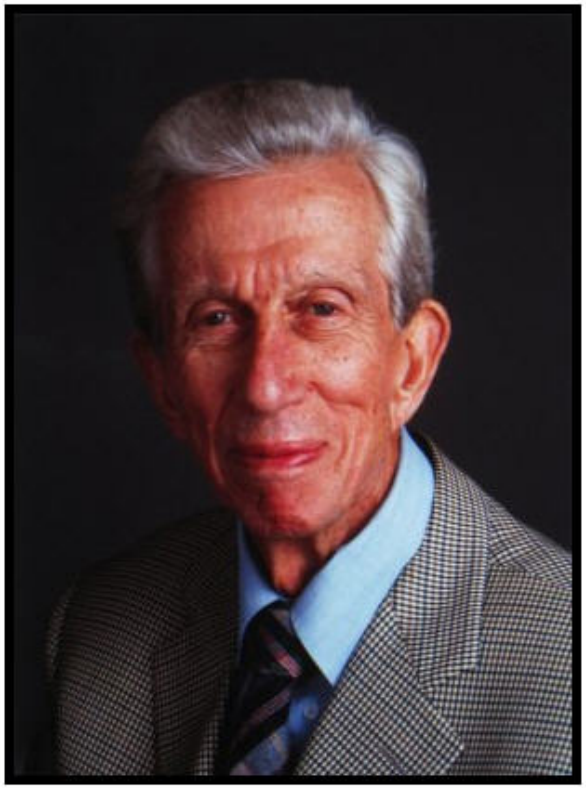


Speech, Music and Hearing

part of
School of Computer Science and Communication



TMH



Gunnar Fant

October 8, 1919 - June 6, 2009

http://www.speech.kth.se/~gunnar/

Speech, Music and Hearing
CTT, Centre for speech technology
The KTH speech group



http://www.speech.kth.se

- Headed by Björn Granström & Rolf Carlson
- 25 employees (some part time)
  - including 10 PhD students
  - Multidisciplinary staff
- Activities
  - Education – mostly post grad.
  - National and EU projects

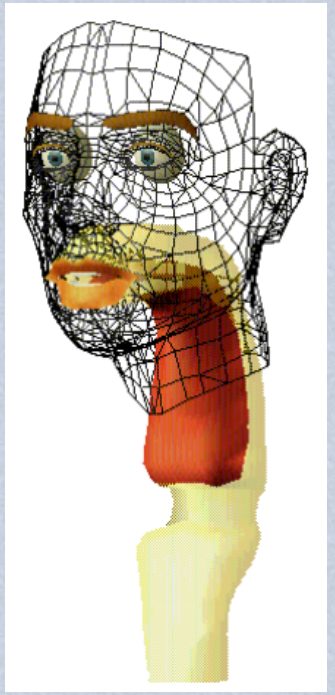
Centre for Speech Technology

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



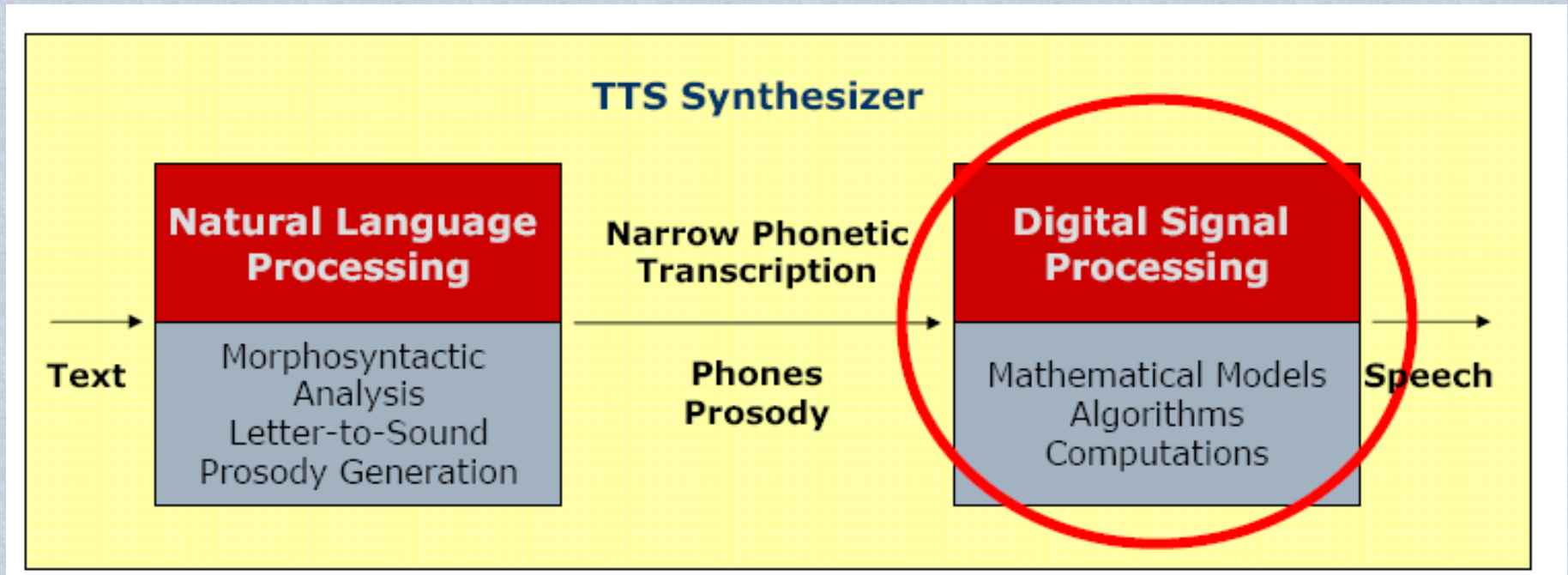




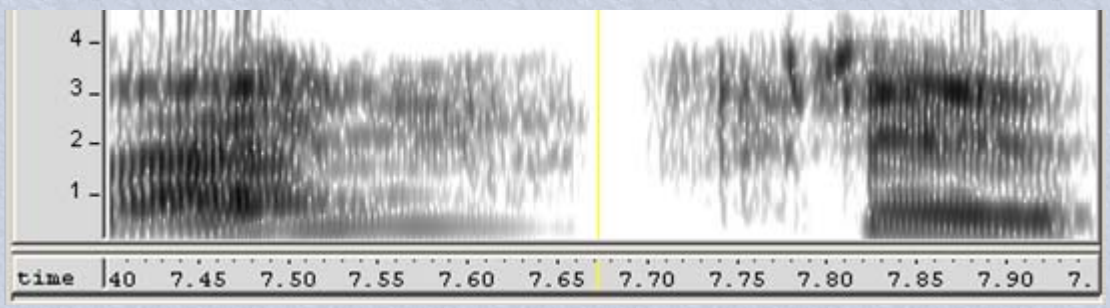
ACKNOWLEDGEMENTS

The work at KTH reported here was carried out by a large number of researchers at the Centre for Speech Technology which is gratefully acknowledged. The work has also been supported by the EU/IST projects SYNFACE, PF-Star, CHIL, MonAMI, MUSCLE and HaH.

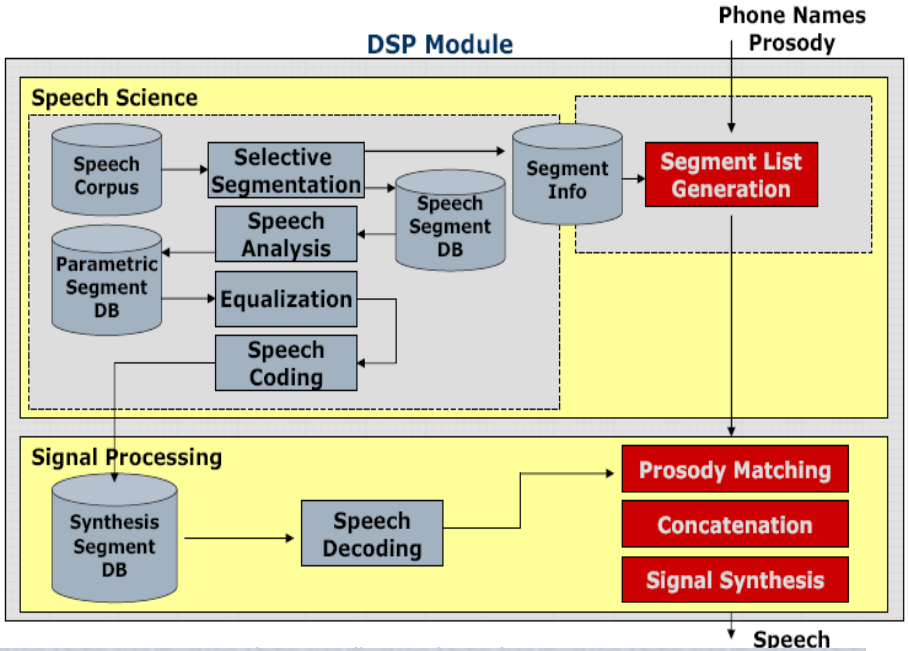
Definition & Main scope

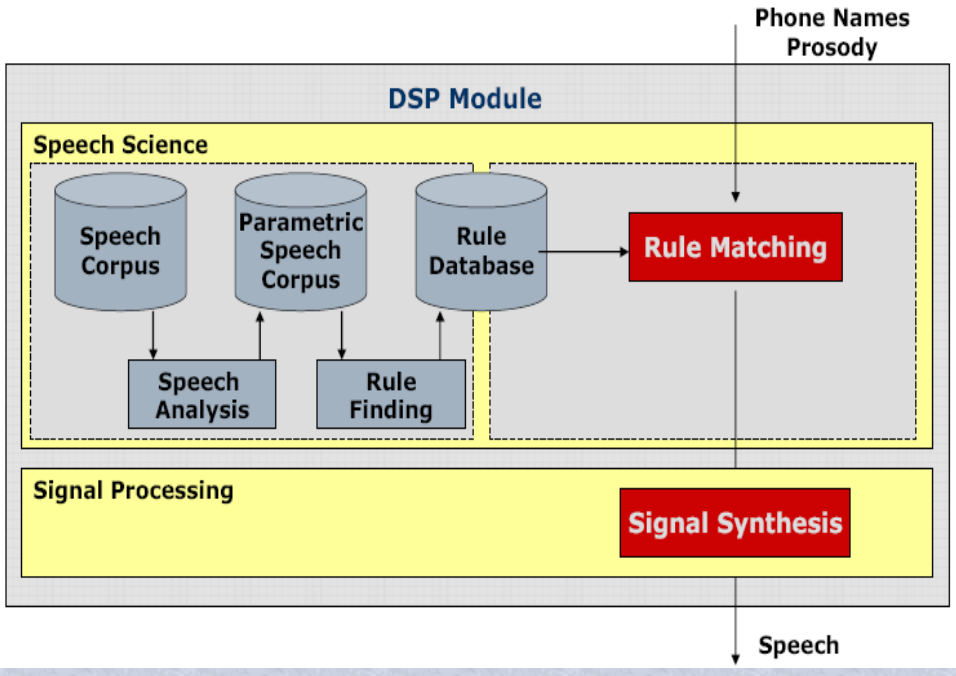


The *automatic* generation of synthesized sound or visual output from *any* phonetic string.



Synthesis approaches





By Concatenation

Elementary speech units are stored in a database and then concatenated and processed to produce the speech signal

By Rule

Speech is produced by mathematical rules that describe the influence of phonemes on one another

History: van Kempelen

Wolfgang von Kempelen's book

Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine (1791).

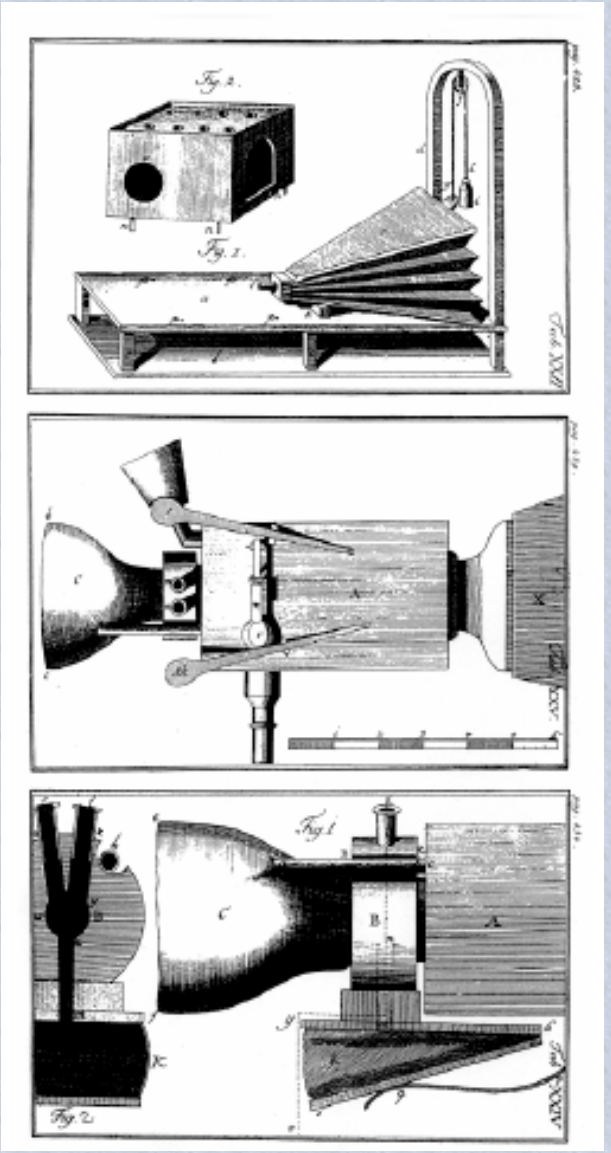
The essential parts

- pressure chamber = lungs,
- a vibrating reed = vocal cords,
- a leather tube = vocal tract.

The machine was

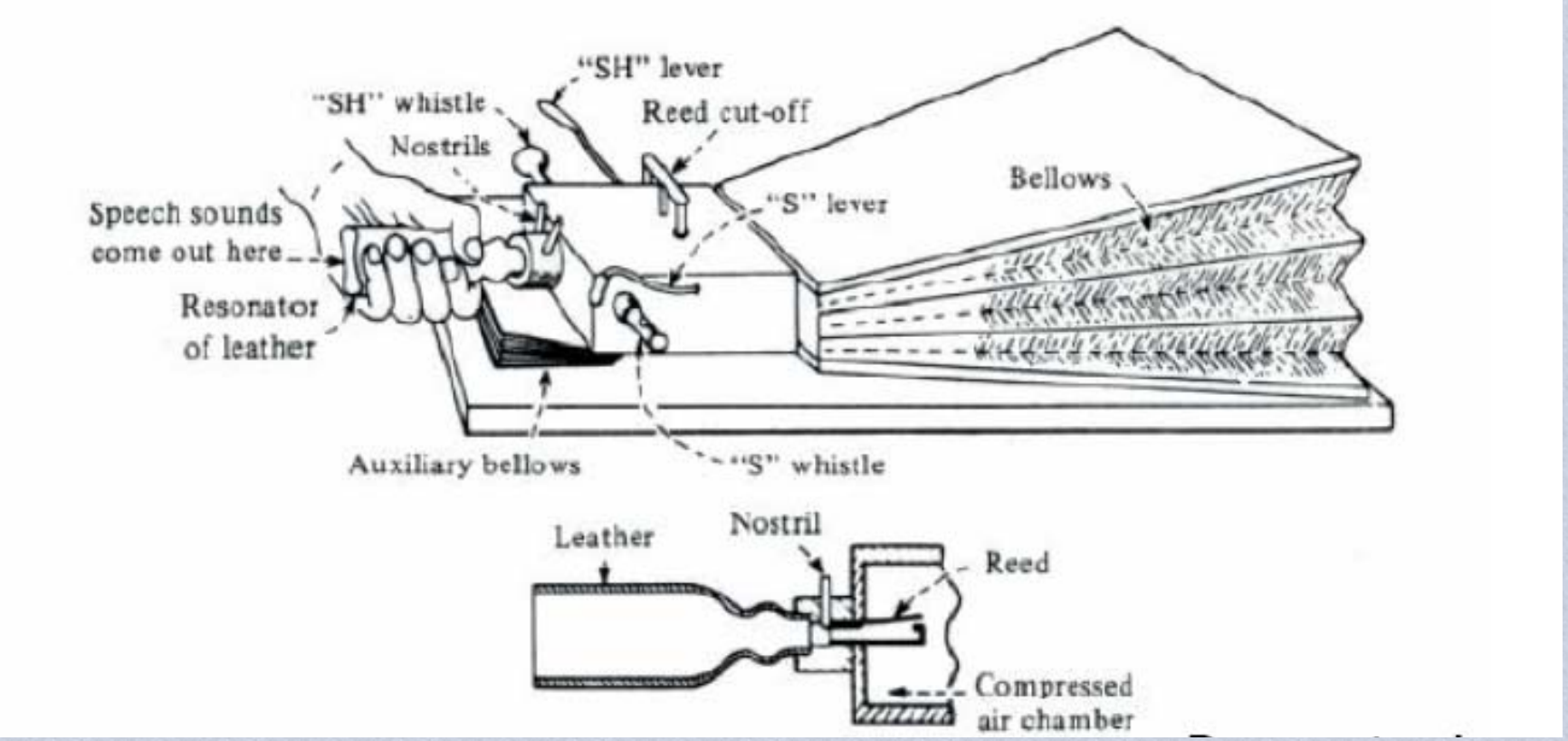
- hand operated
- could produce whole words and

short phrases.

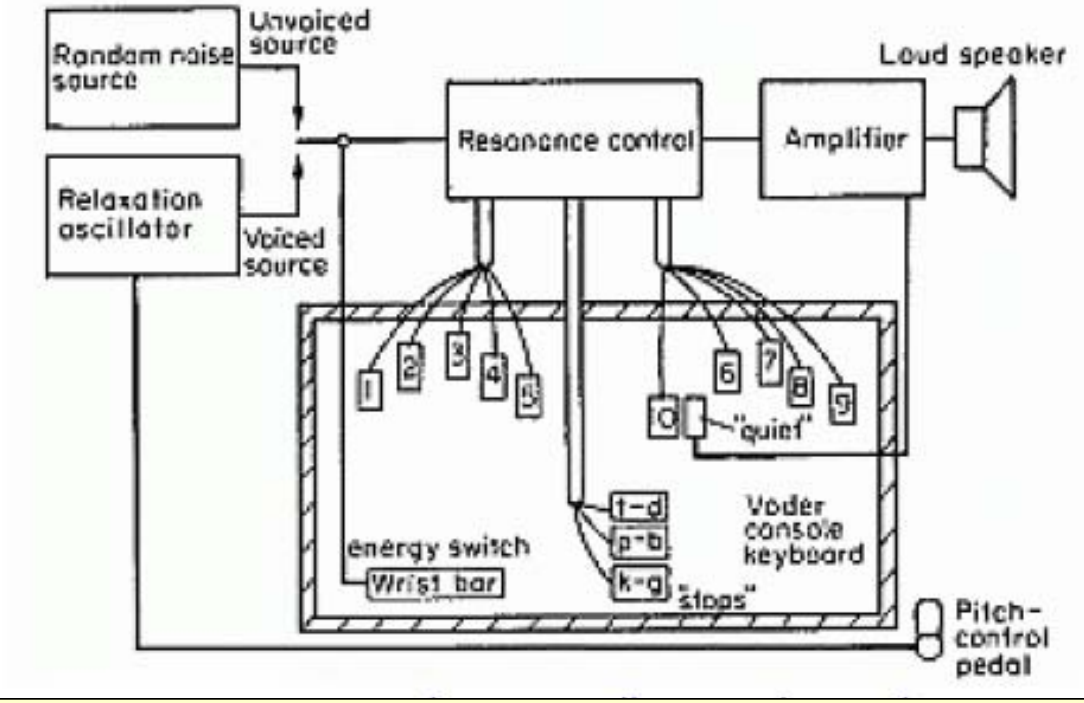


Wheatstone's version

Charles Wheatstone's version of von Kempelen's speaking machine



First electronic synthesis





- Homer Dudley presented VODER (Voice Operating Demonstrator) at the World Fair in New York in 1939
- The device was played like a musical instrument, with voicing/noise source on a foot pedal and signal routed through ten bandpass filters.

First formant synthesizers

- 1950's PAT (Parametric Artificial Talker), Walter Lawrence

  3 electronic formant resonators input signal (noise)

  6 functions to control 3 formant frequencies, voicing, amplitude, fundamental frequency, and noise amplitude.

- 1950's OVE (Orator Verbis Electris) by Gunnar Fant

- From 1950's: other synthesizers including the first articulatory synthesis DAVO (Dynamic Analog of the Vocal tract)

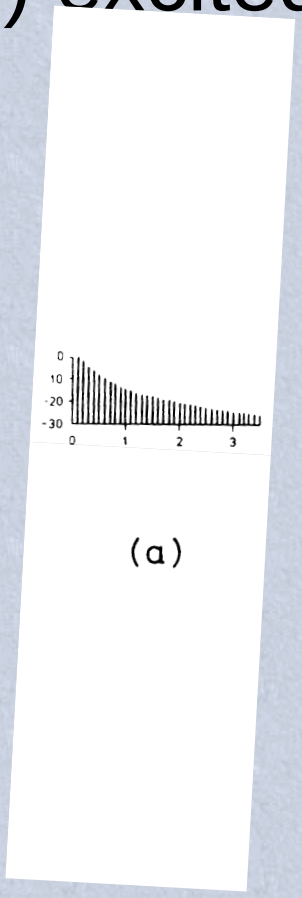
An excellent historical trip of speech synthesis:

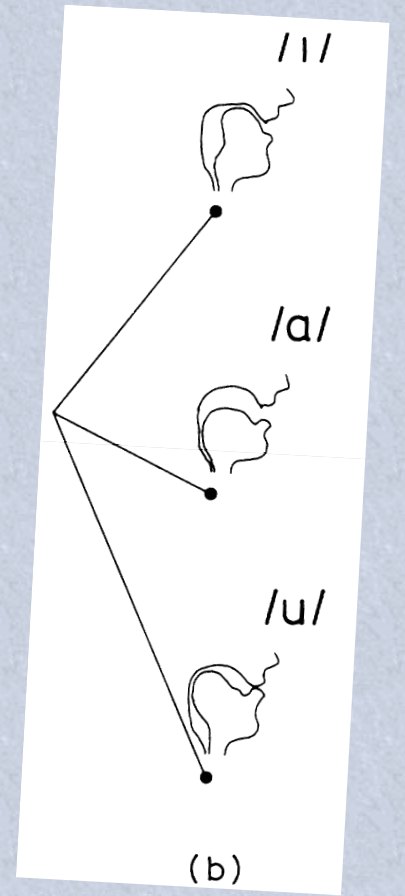
Dennis Klatt's History of Speech Synthesis at

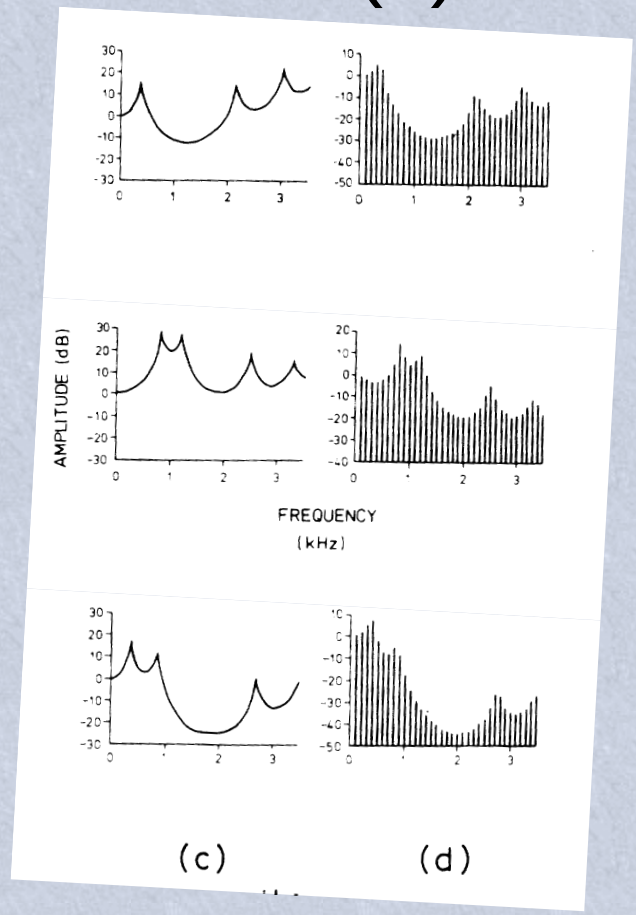
http://www.cs.indiana.edu/rhythmsp/ASA/Contents.html

The source-filter theory

The **signal** (c-d) is the result of a **linear filter** (b) excited by one or several sources (a).

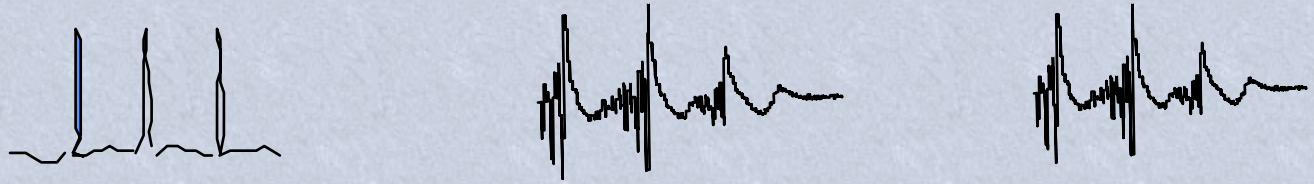


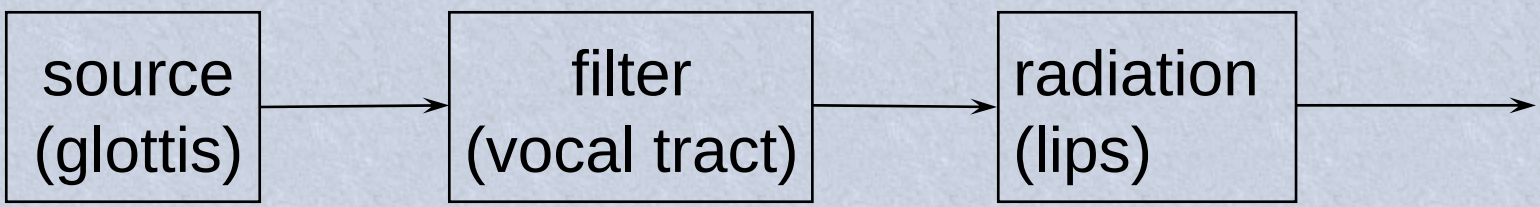




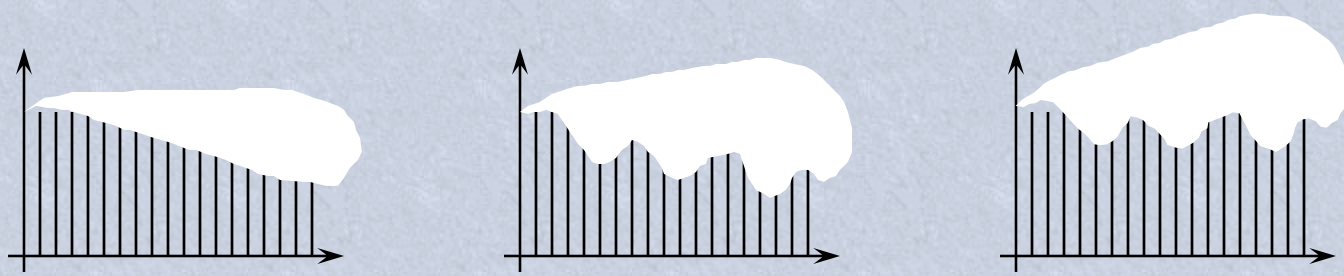
The source-filter theory

TIME:

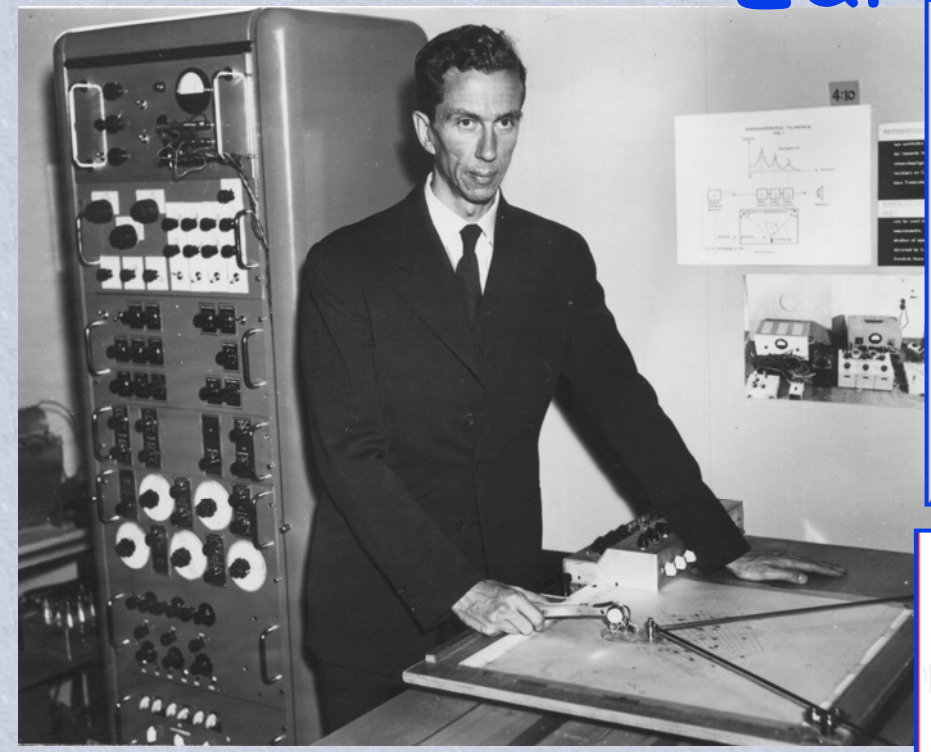


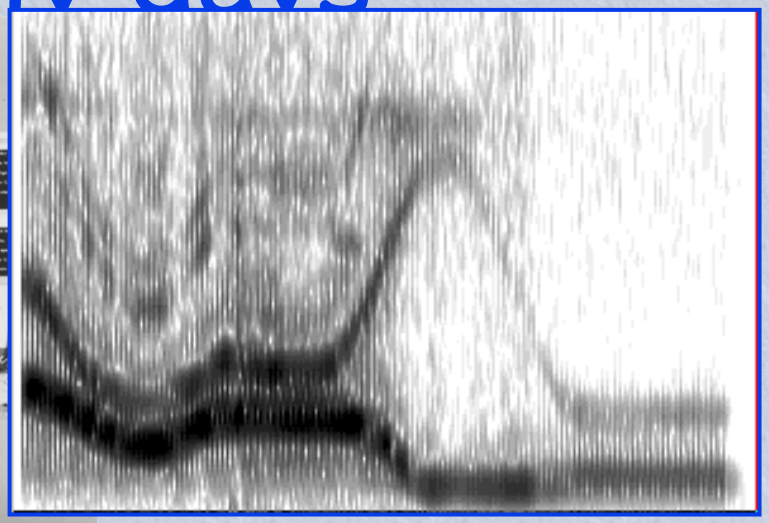


FREQUENCY:



The KTH speech group - Early days





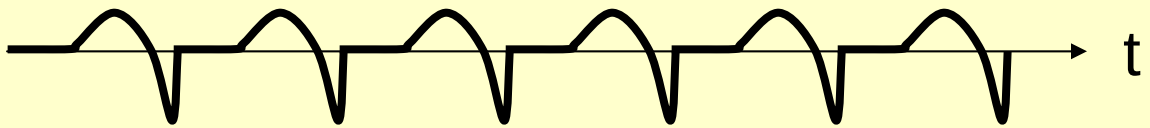


Gunnar Fant and OVE I 1953

OVE I in freeware Snack

Source functions

- The voiced quasi-periodic source (glottis pulses) – vowels



Parameters:
  - on/off,
  - Fundamental frequency F0,
  - intensity,
  - shape

- Frication source – fricatives
- Transient noise – plosives
- No source – voiceless occlusions

Ove II, 1958

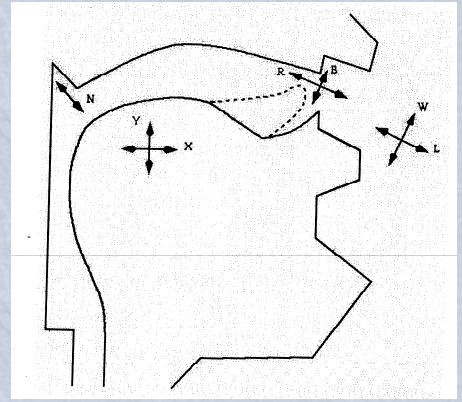


1961

1962

Articulatory synthesis

Filter I: Articulation model



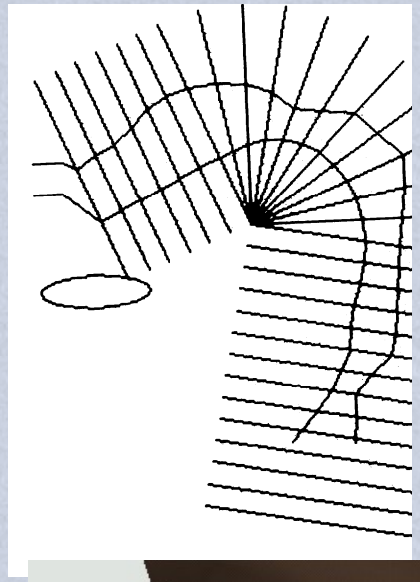
Benefits:

- Speech production in the same way as humans
- Can be made with few parameters
- The changes are intuitive
  (raise the tongue tip, round the lips)

Disadvantages:

- Computationally demanding
- Problems with consonants
- Articulatory measurements required
- State-of-the-art articulatory synthesis still sounds bad
- Hard to get data

3D articulatory syntesis



Why?

• Two-dimensional models simulate the third dimension as $area = a \cdot (distance)^d$.
a & *d* are decided empirically and vary through the tube.



A three-dimensional model gives

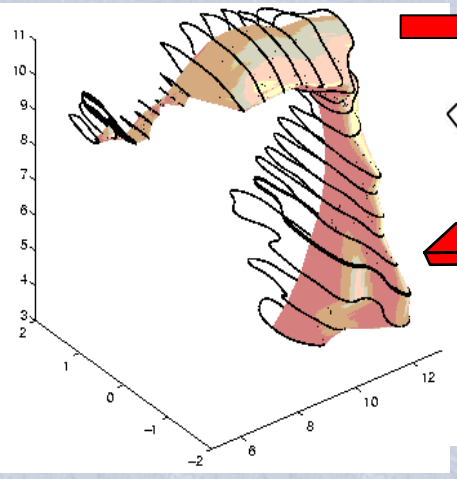
• the cross-sectional area directly

• lateral modeling (/l/)

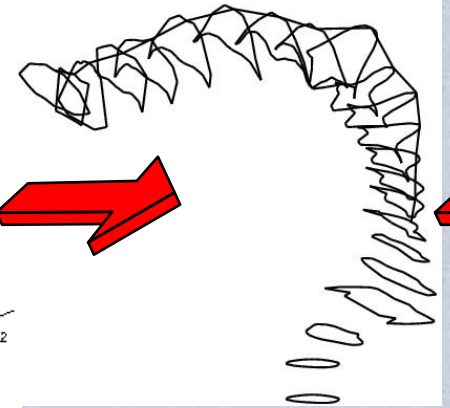
• visual synthesis (pronunciation training)



From articulation to acoustics

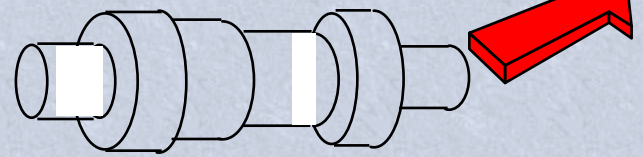


Vocal tract model

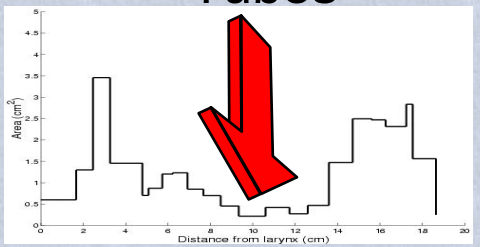


Cross-sections

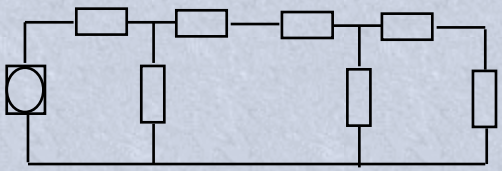
3D air flow calculations



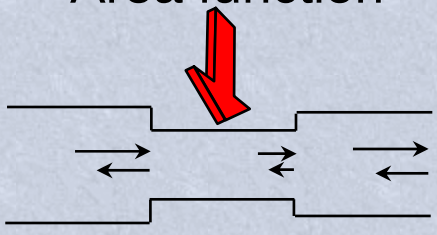
Tubes



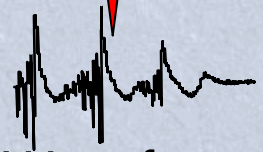
Area function



Electric circuit equivalent



2D airflow dynamics

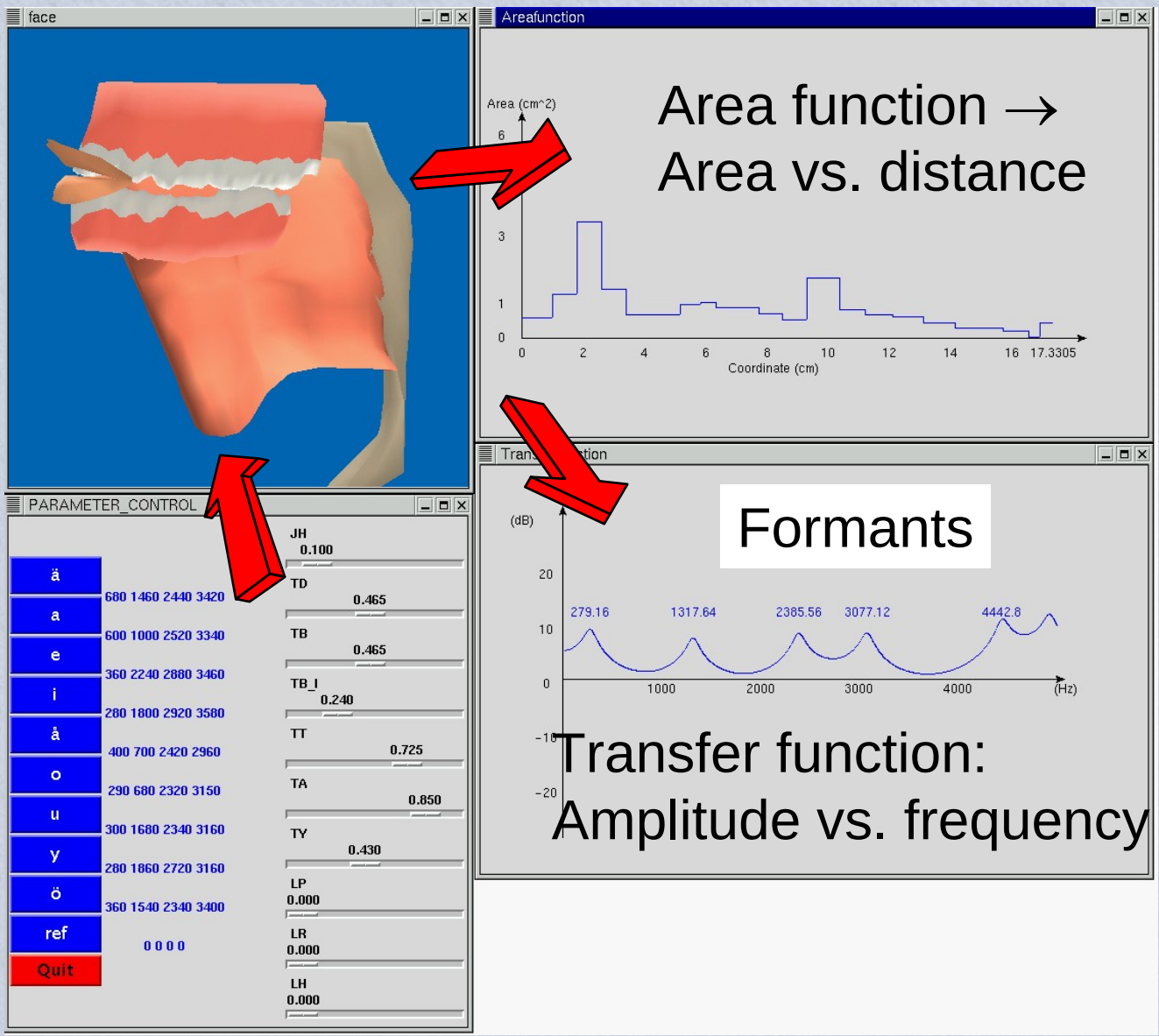


Waveform

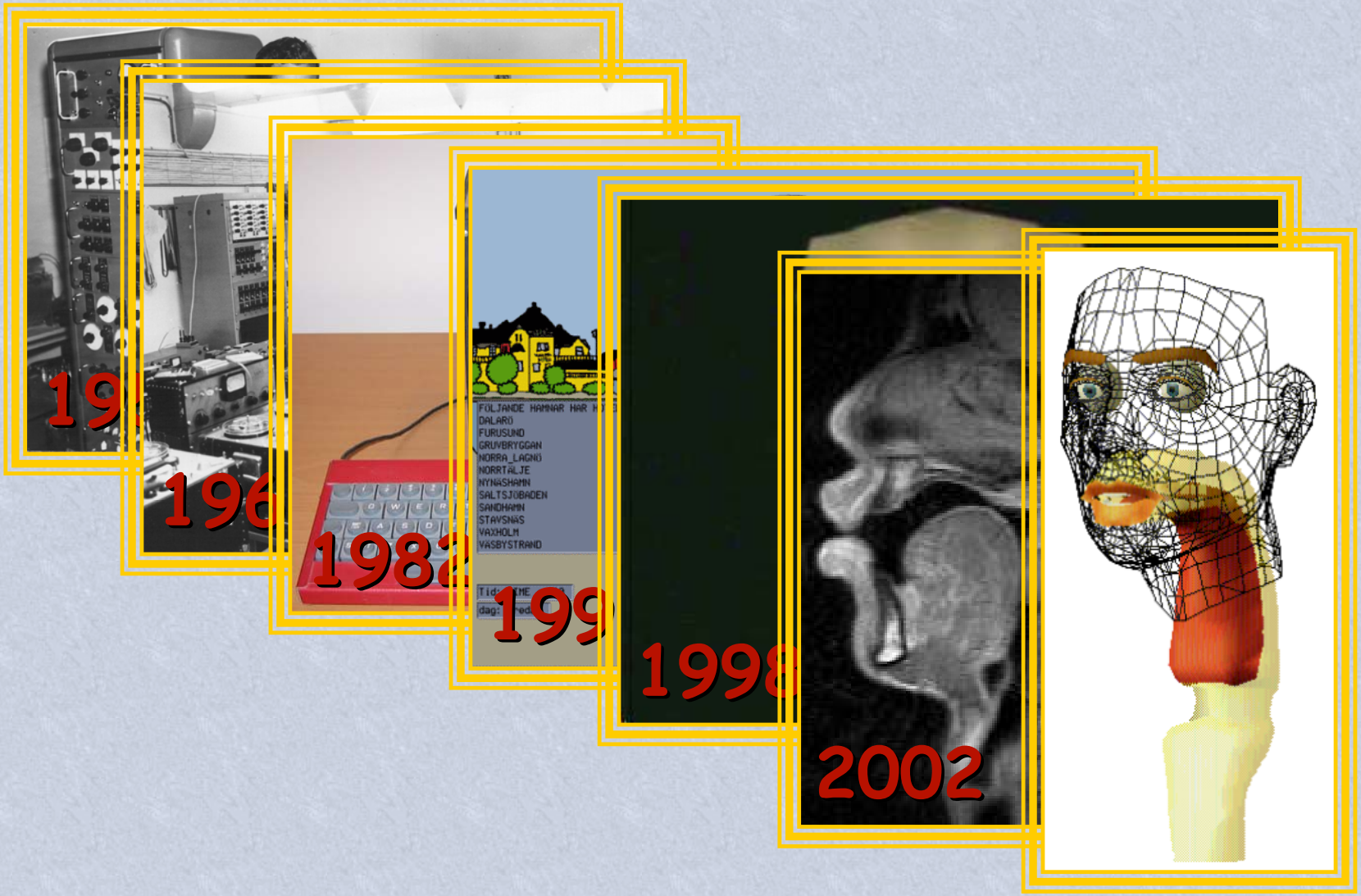
Area & transfer functions

Articulatory model

Parameter settings



Speech synthesis developments at KTH



The voice terminal?!

Motorola H5 MINIBLUE with Bluetooth®



Designed around the way your body works, the in-ear microphone and speaker of the Motorola H5 MINIBLUE pick up your voice through the ear canal so conversations come in loud and clear – even in a noisy environment. And measuring only 33 x 41 millimeters this barely-there headset fits perfectly inside your ear for pristine, unobtrusive communication. This petite powerhouse packs in battery power so powerful - you can chat wire free for up to approximately 7.5 hours using the included portable charging base.

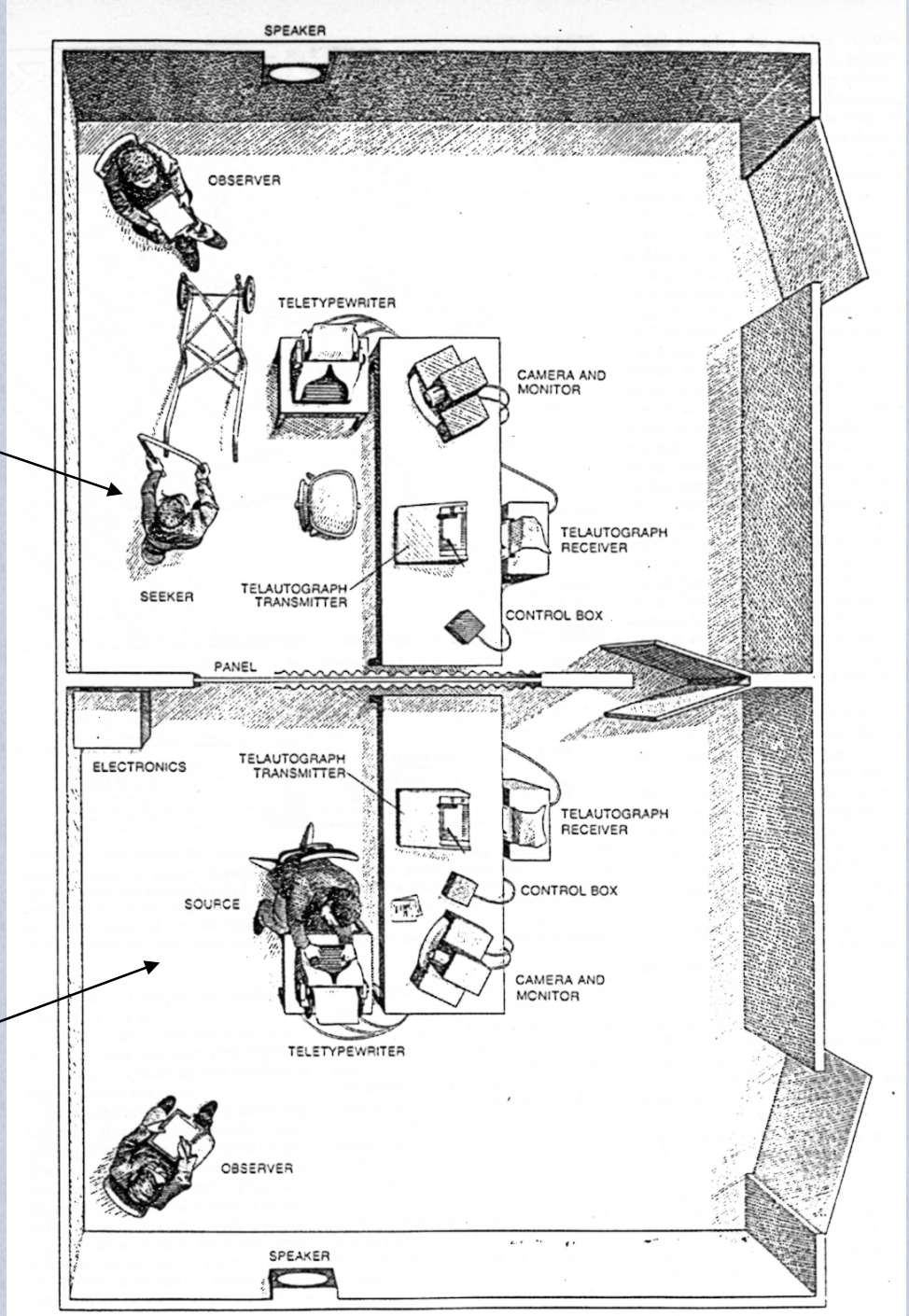
Why Speech Tech?

User

"Interactive Human Communication"

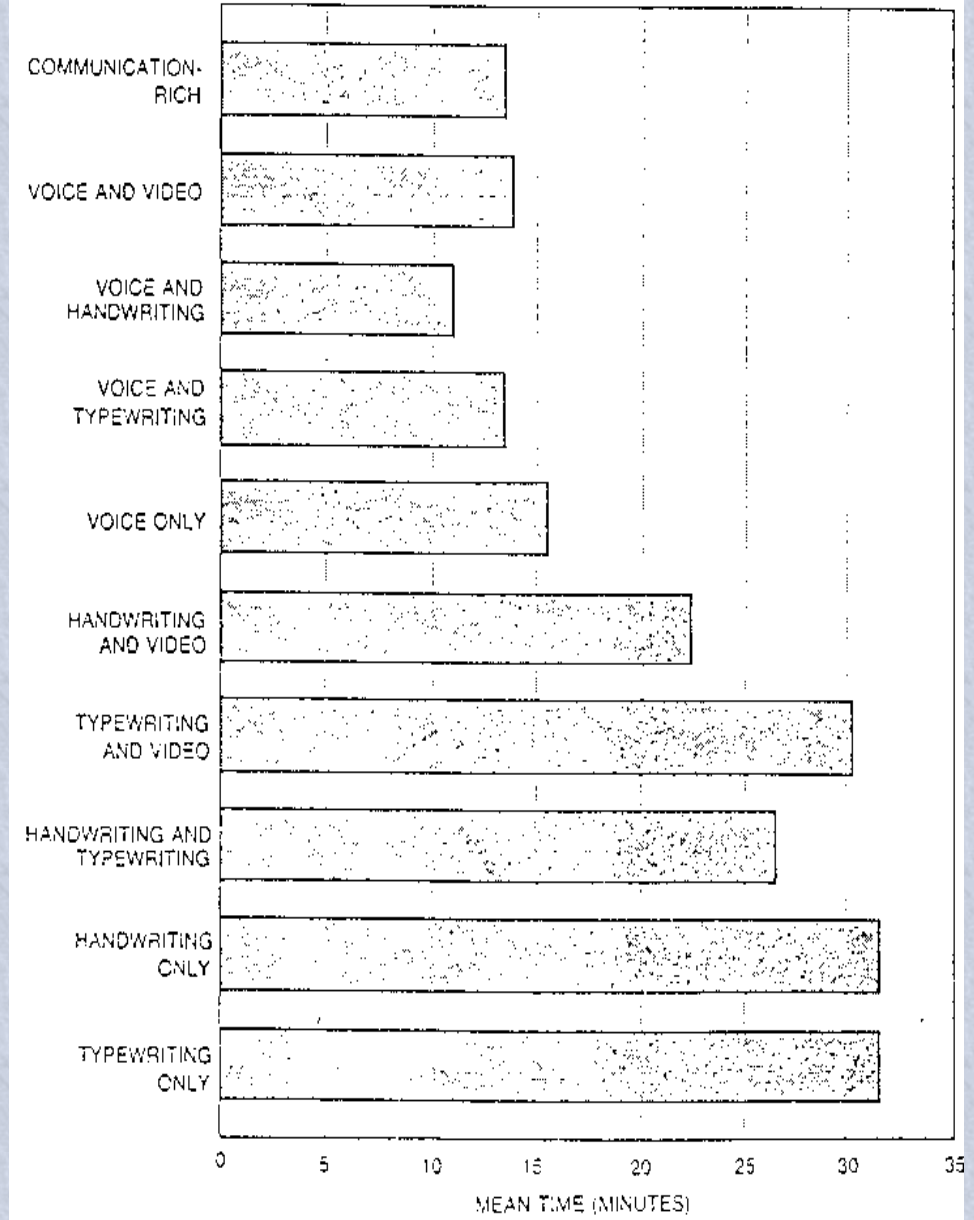
Chapanis
Scientific American 1975!

"Computer"



More efficient and convenient!

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



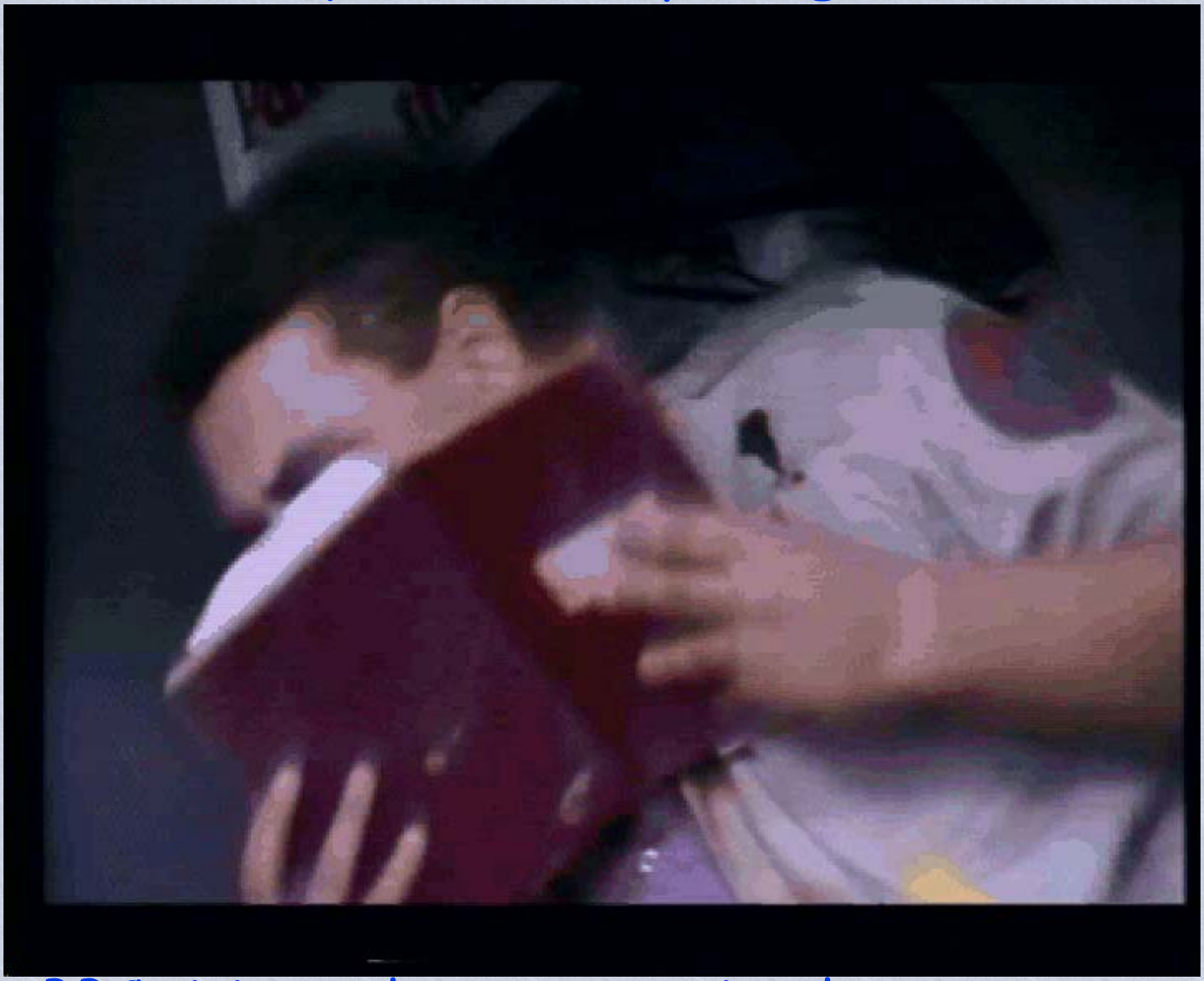
MODE OF COMMUNICATION influenced the time required to solve problems. Here the average time taken by teams to solve problems is charted for 10 different modes of communication. The data fall into two fairly distinct groups. The faster five all involve the use of the voice in communication, whereas the voice is excluded in the five slower modes.

Speech is fast and verbose

	COMMUNICATION-RICH	VOICE	HANDWRITING	TYPEWRITING	
				EXPERIENCED TYPISTS	INEXPERIENCED TYPISTS
SOLUTION TIME IN MINUTES	29	33	58.3	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.

Emotional/ubiquitous computing – do we want it?



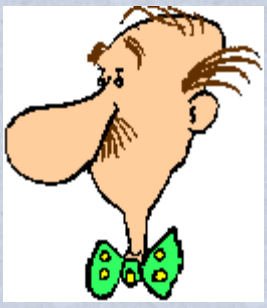
Early BBC vision – the conversational toaster

Thanks to Mark Huckvale, UCL, for the video clip

Emotions

natural

synthesis



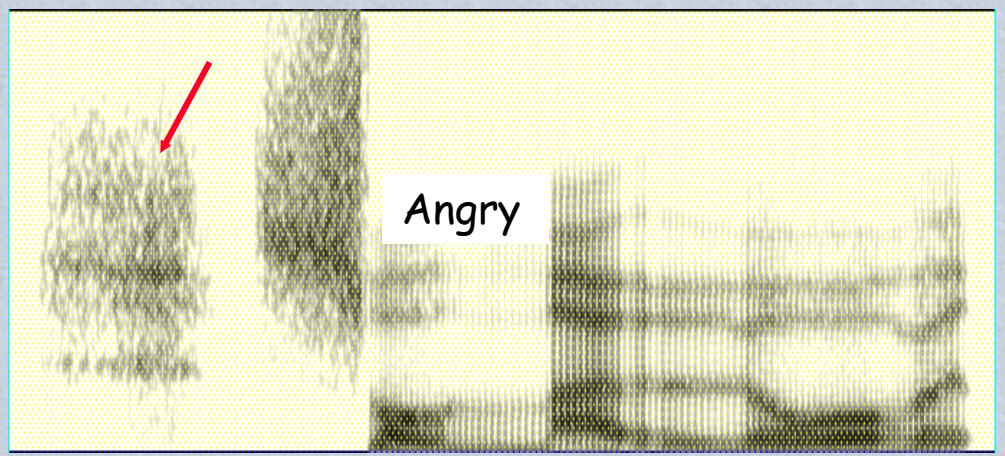
Neutral



natural

synthesis

Happy

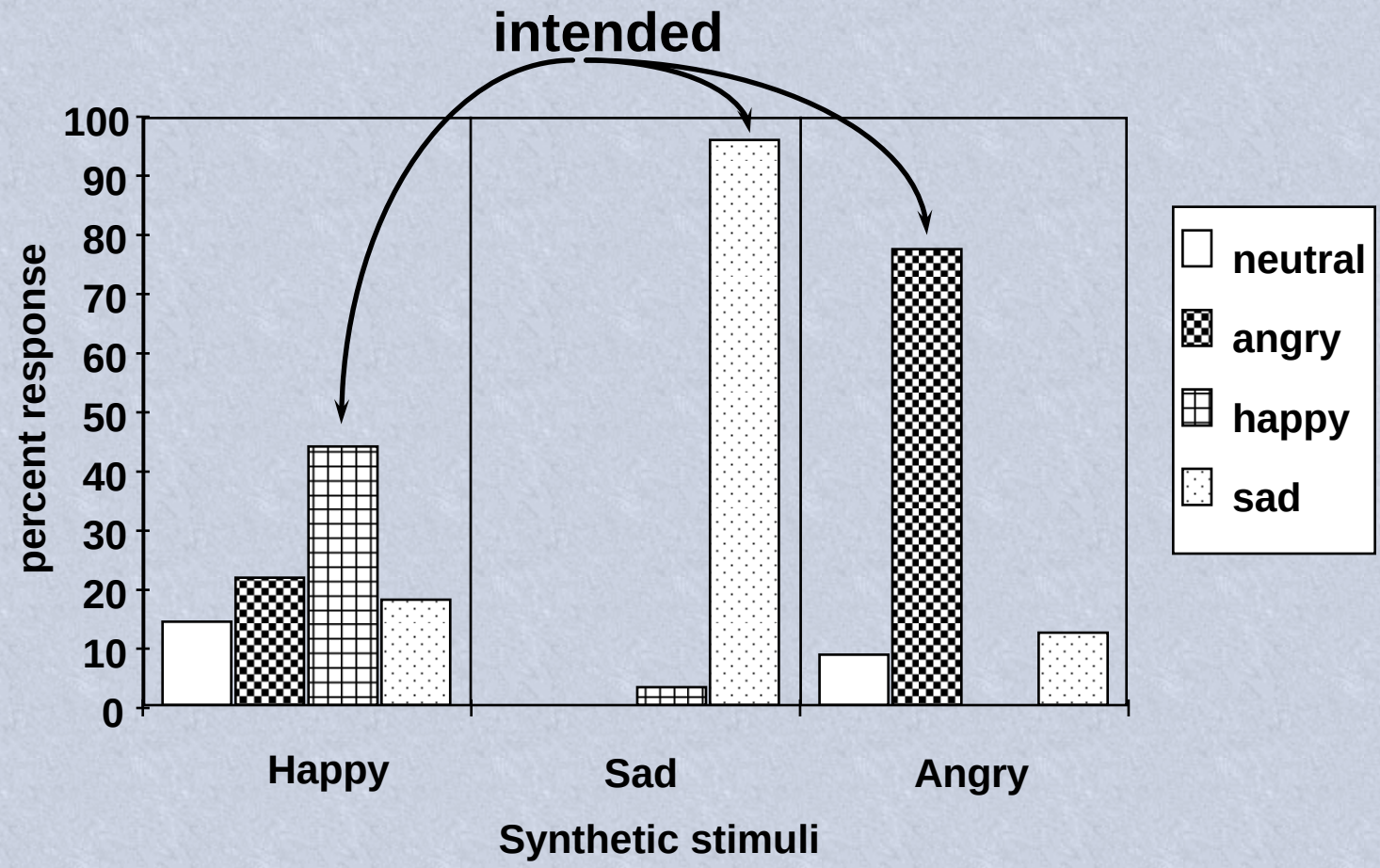




natural

synthesis

Result from listening test



Emotions



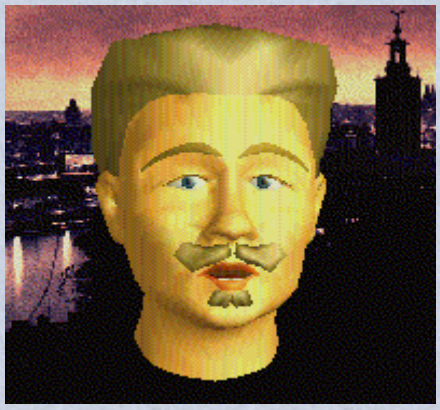
Happiness



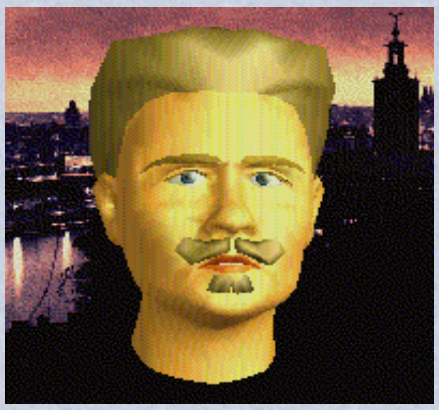
Anger



Fear



Surprise



Disgust



Sadness

Vision from audio

original

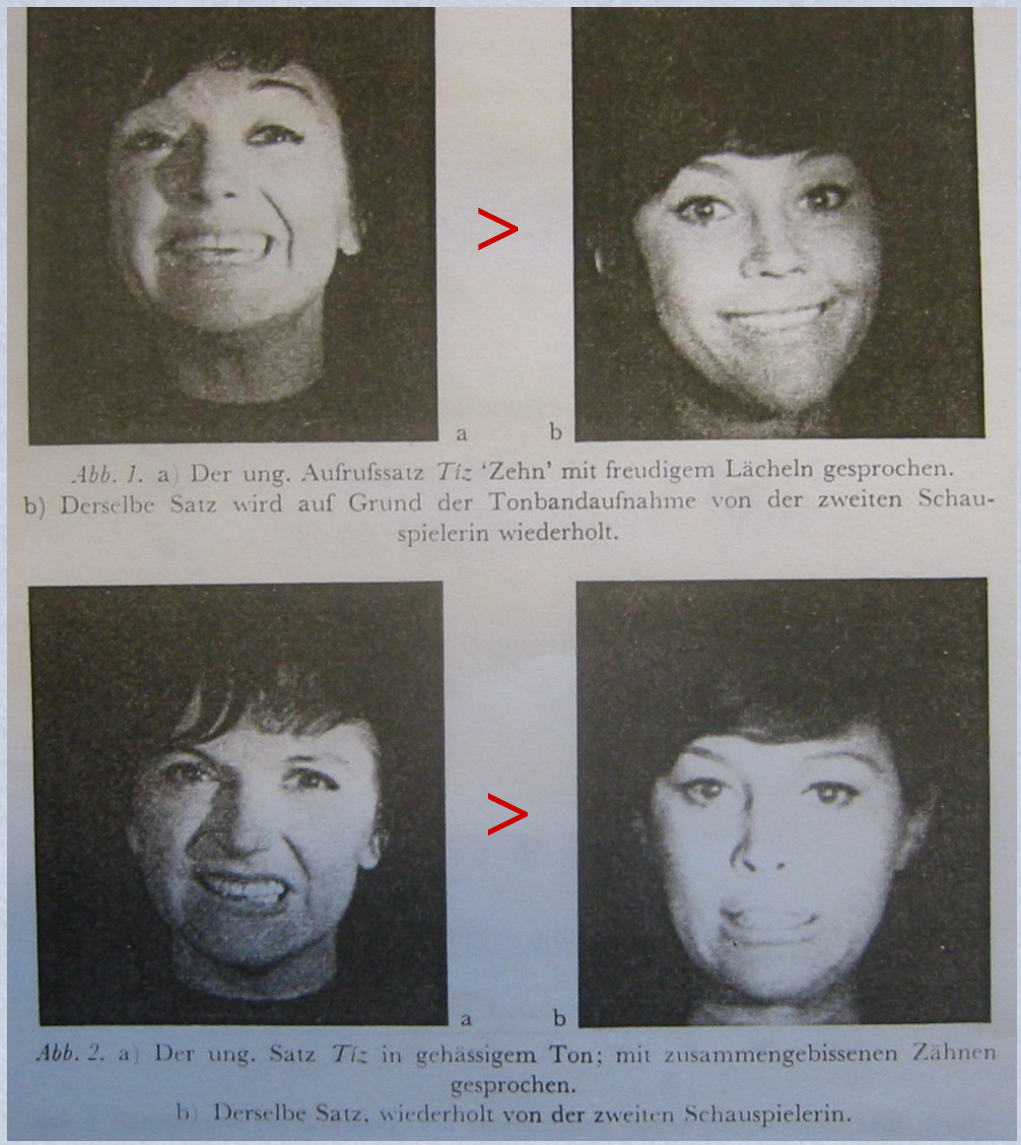


Abb. 1. a) Der ung. Aufrufssatz *Tíz* 'Zehn' mit freudigem Lächeln gesprochen.
b) Derselbe Satz wird auf Grund der Tonbandaufnahme von der zweiten Schauspielerin wiederholt.

Abb. 2. a) Der ung. Satz *Tíz* in gehässigem Ton; mit zusammengebissenen Zähnen gesprochen.
b) Derselbe Satz, wiederholt von der zweiten Schauspielerin.

mimic
from audio

Fónagy, 1967 “Hörbare Mimik” , Phonetica

Multimodal speech synthesis

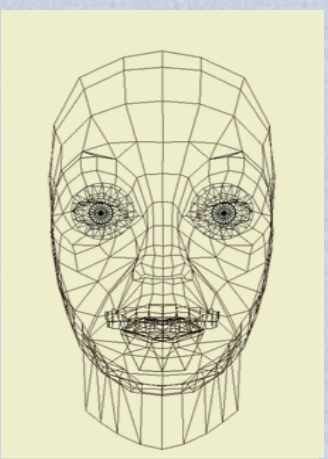






Talking heads - Applications

- 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.





Beyond 2001
Ready for the Odyssey?



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

The Next Generation Computer Interfaces
“A keyboard – how quaint!”



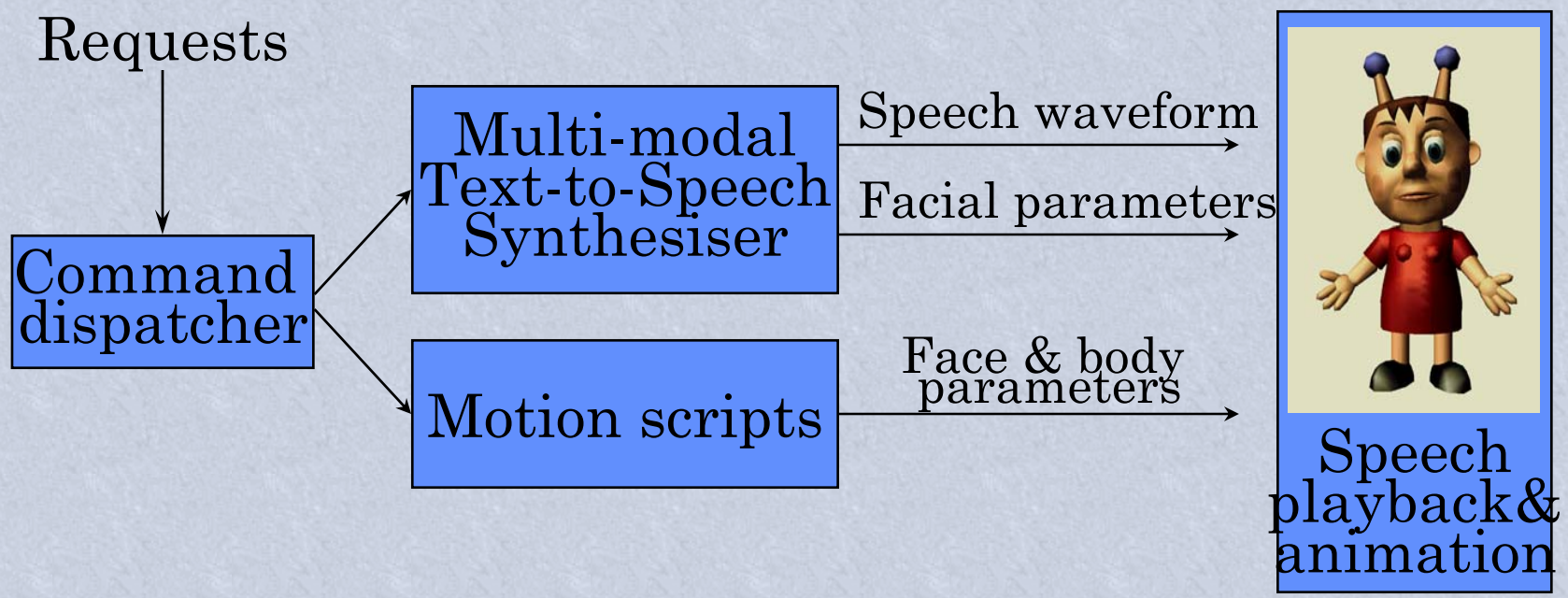
Star Trek IV: The Voyage Home (1986)

Applications of machines with human metaphor interaction

- Research tool
  - Cassel: “a machine that acts human enough that we respond to it as we respond to another human”
- Second Language Learning
- Social Robots
- Computer Games



Animated Character - architecture

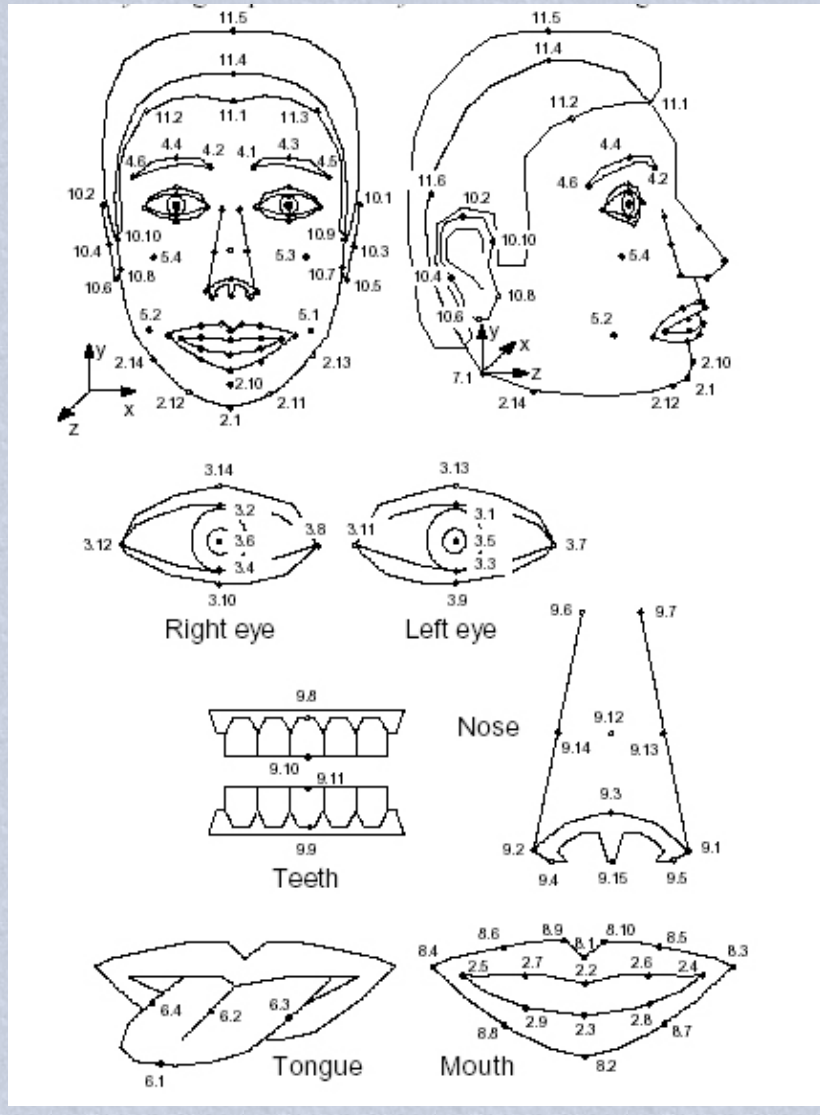


Parameters used for articulatory control of the face.

- 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

Face animation in MPEG-4

- 84 Feature points
- 68 FAP - Facial animation parameter



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

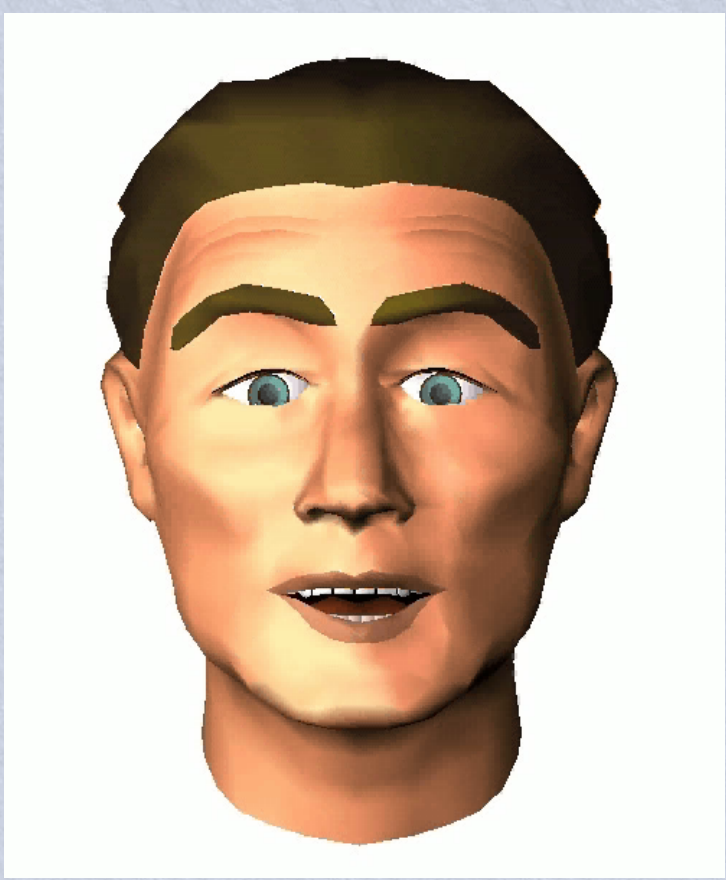
http://www.speech.kth.se/wavesurfer

Audio-visual interactions
The McGurk effect

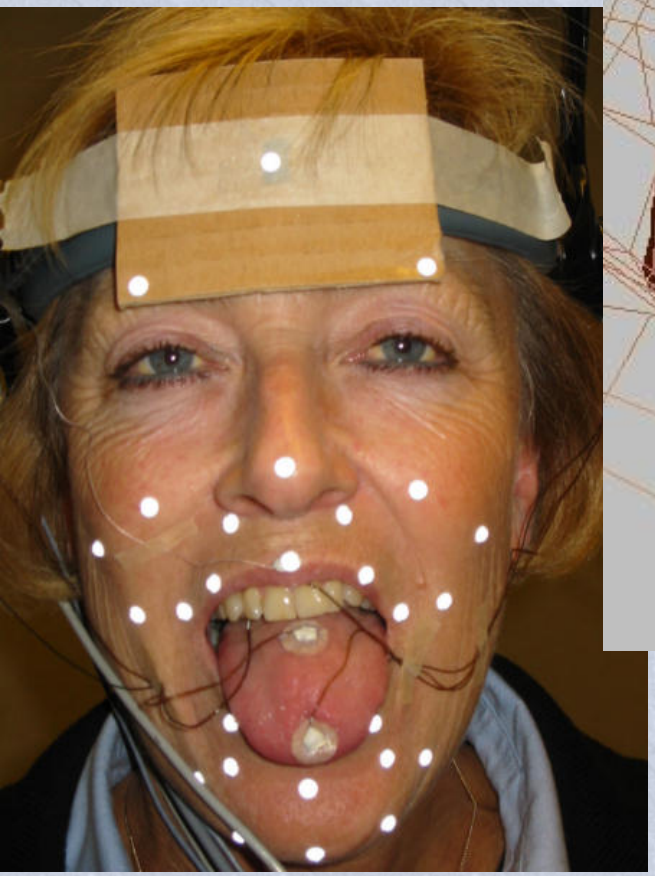


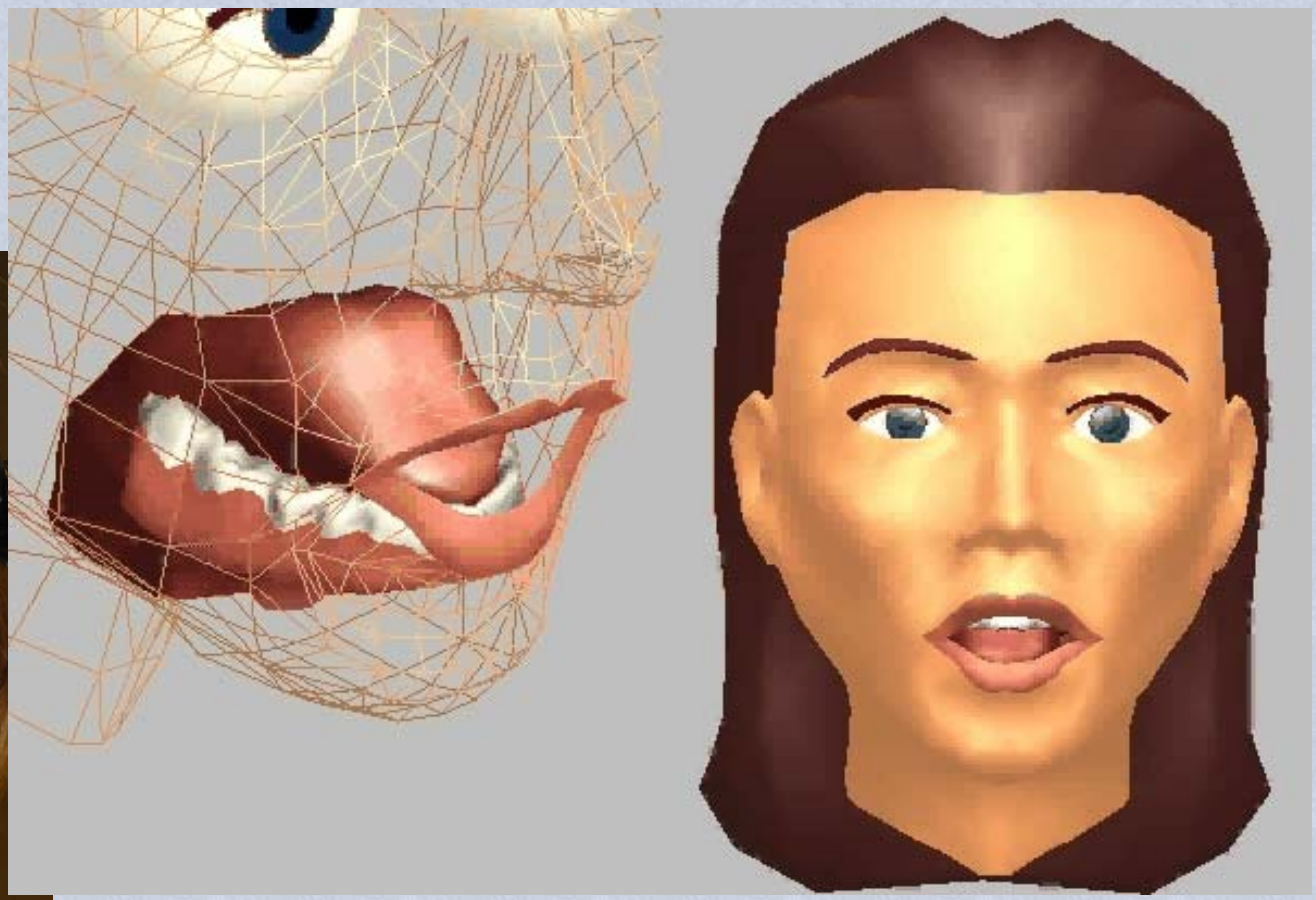
Visual appearance
Modeling wrinkles using bump-mapping

- Wrinkles have communicative relevance
- Traditional method ("the right way"): Physical simulation (FEM) – complicated and computationally expensive
- Alternative method: create a "wrinkle-map" weighted by the compression of the surface.
- Real-time rendering on consumer hardware using bump-mapping



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.



Combining model and data

Re-synthesis using speech movement recorded with Qualisys



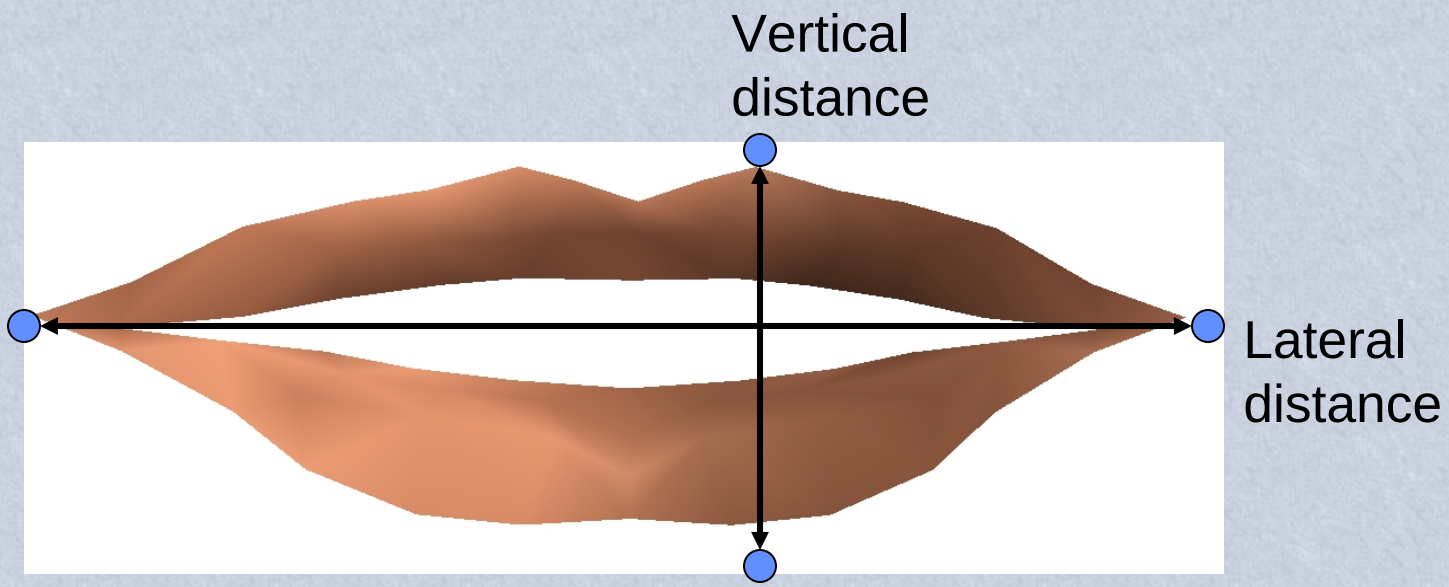
Recording and model



Conversation with agent



Measurement points for lip coarticulation analysis

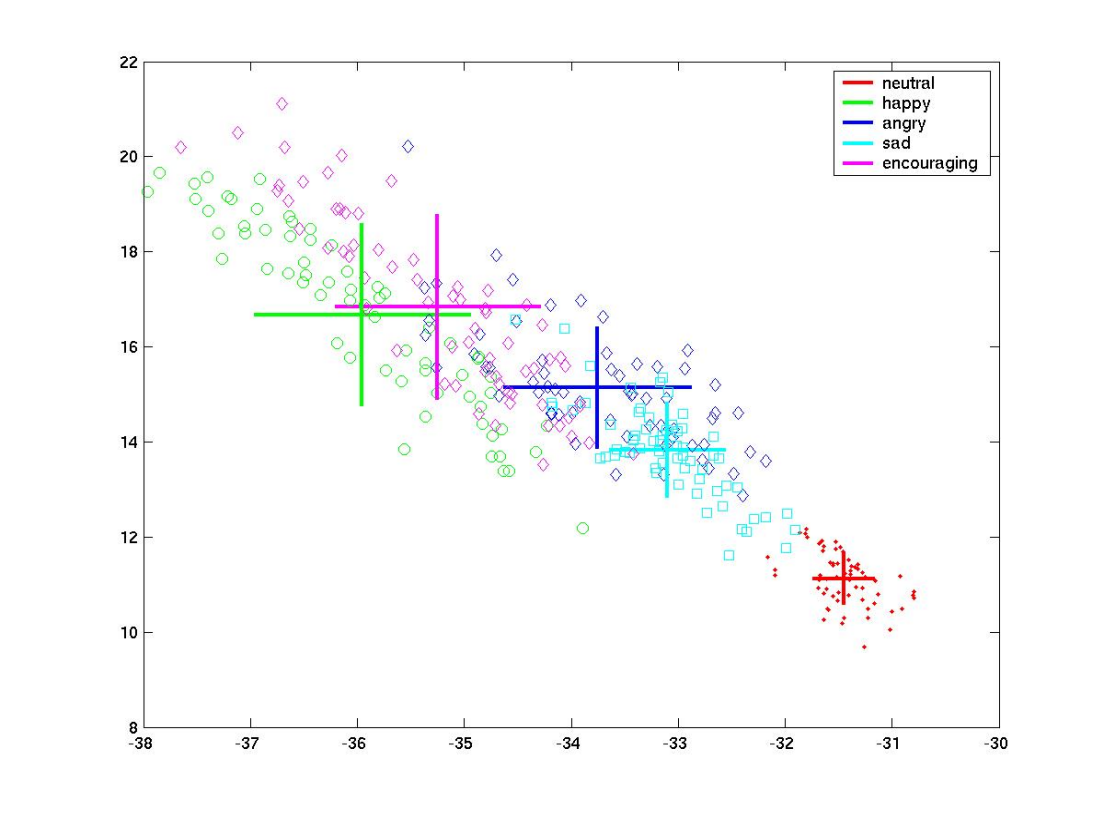


The expressive mouth

- All vowels

  (sentences)

  - Encouraging
  - Happy
  - Angry
  - Sad
  - Neutral



"left mouth corner"

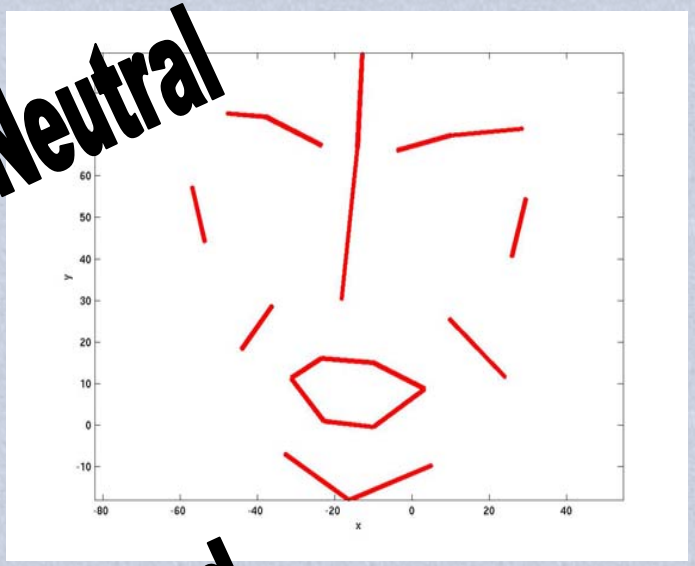
Interactions: emotion and articulation (from AV speech database - EU/PF_STAR project)



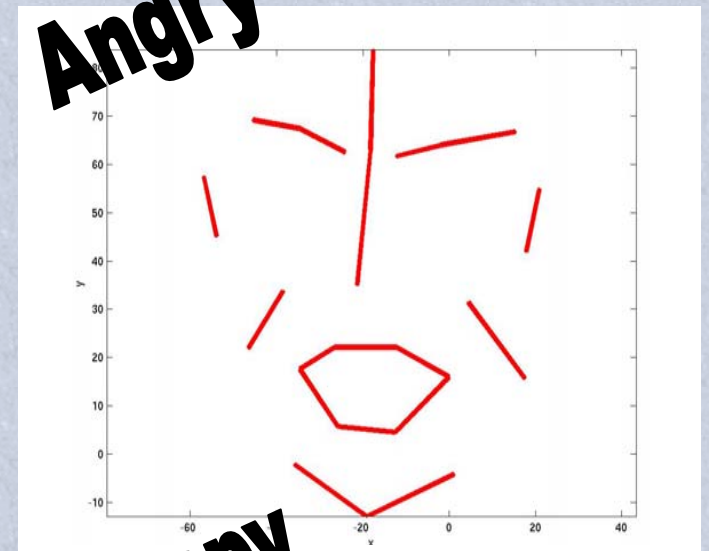




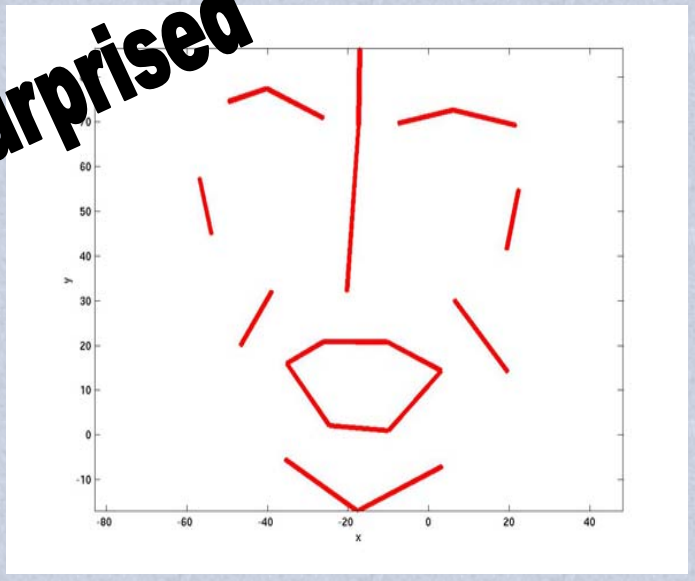
Neutral



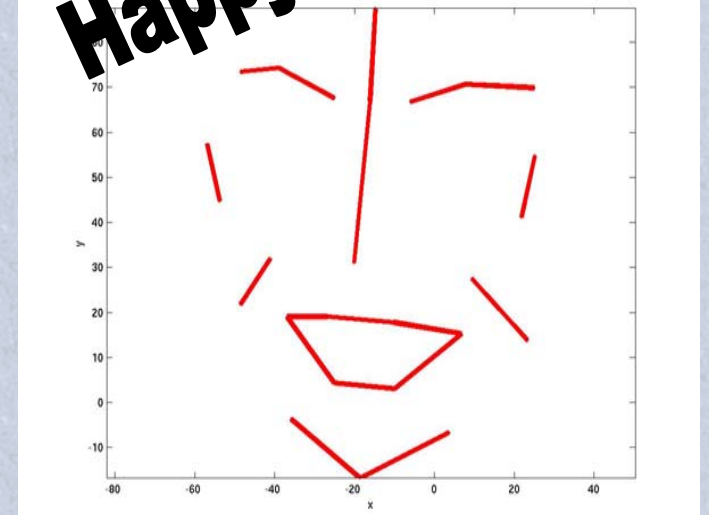
Angry



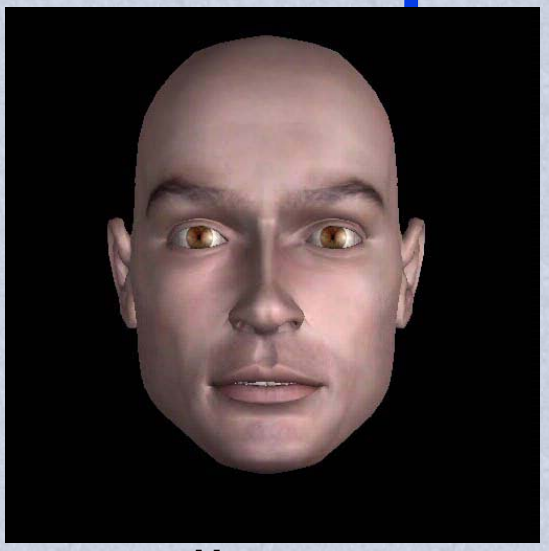
Surprised



Happy



Examples (automatic)



Happy



Angry



Sad



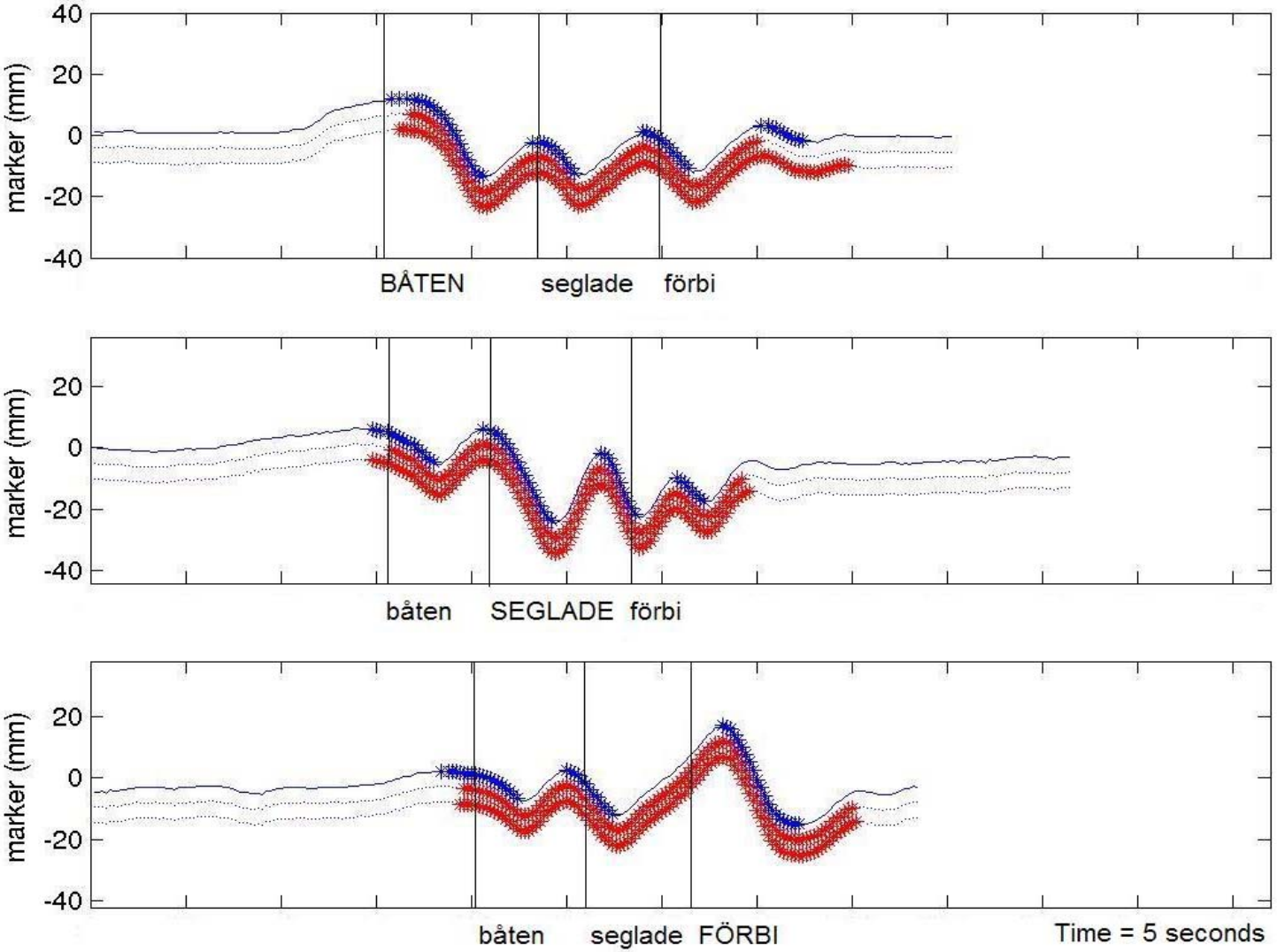
Surprised

Focal accent examples from the database

Focal accent on:

	Båten	seglade	förbi
Happy			
Confirming			

Nose marker traces with automatic (blue) and two human (red) annotated head nods (adapted from Cerrato & Svanfeldt 2006)

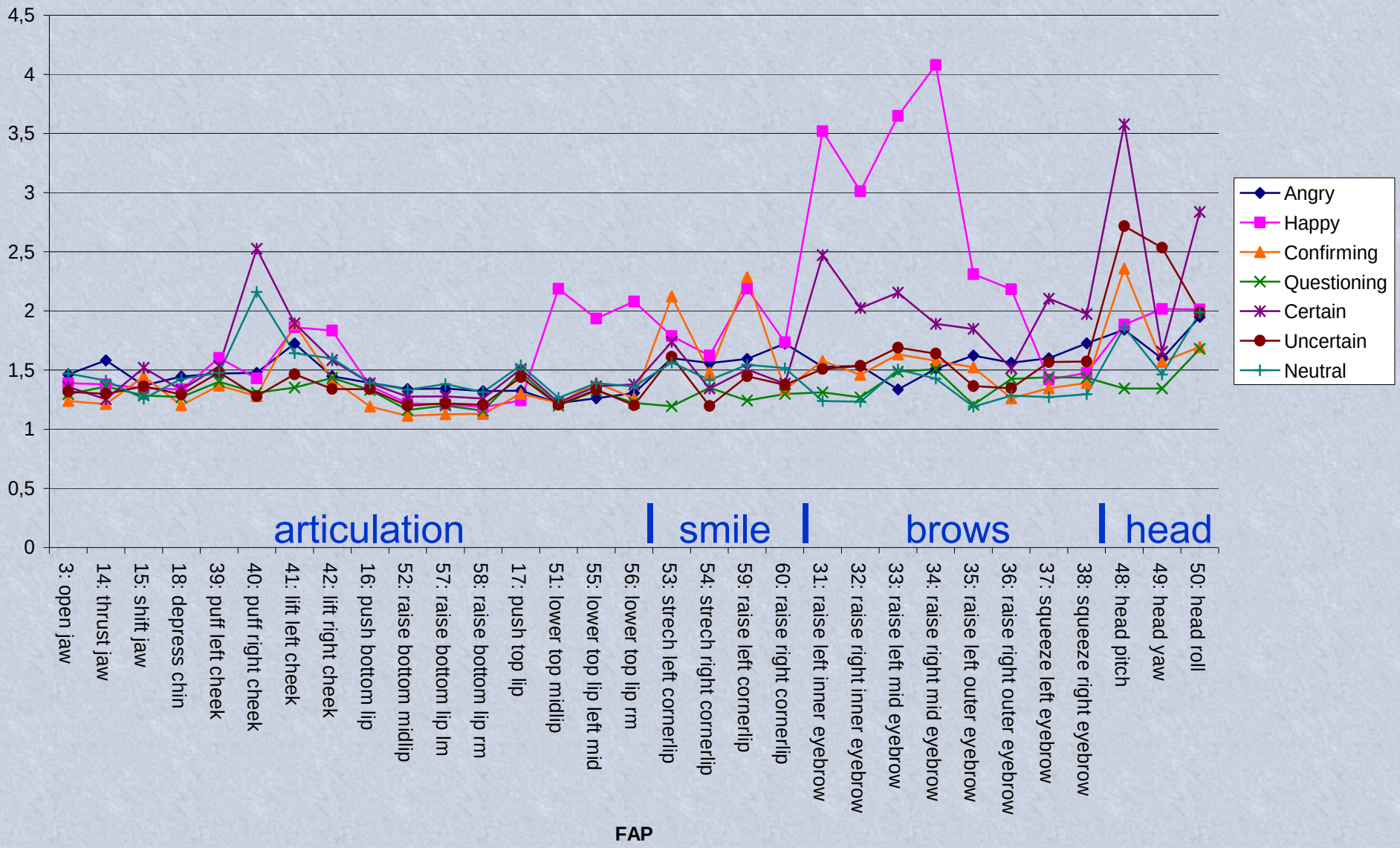


Analysis in terms of FAP and FMQ

MPEG-4 Facial Animation Parameter (FAP)
A subset of 31 FAPs out of the 68 FAPs defined in the MPEG-4 standard, including only the ones that we were able to calculate directly from our measured point data

Focal Motion Quotient, FMQ, defined as the standard deviation of a FAP parameter taken over a word in focal position, divided by the average standard deviation of the same FAP in the same word in non-focal position.

The focal motion quotient, FMQ, averaged across all sentences, for all measured MPEG-4 FAPs for several expressive modes



e-rektor – using an MPG-4 model

- Makes use of state-of-the-art 3D modelling software (XSI) for improved realism
- Java/OpenGL/XML based animation platform
- EU/PFStar project



First presented at Anders Flodström´s 60 year birthday

Interactive conversation

- When to take the turn?
- When to give feedback?

Situation dependent

The Hummer?



or



The Hummer

- Jens Edlund, Mattias Heldner and Rolf Carlson (not published)
- Demo based on pauses only
- Part of a larger project: Cues for possible location for feedback or turntaking, including duration and F0 analysis.

Task at VISPP 2008

- Create an Audio-Visual Hummer!
- With more refined triggers
- Based on realistic feedback in recorded interactions

Modelling the listener

- Portable 3D lab
- Used in VISPP summerschool in Estonia, August 2008
- Free conversation with focus on listener feedback
- Base for audio-visual hummer

VISPP summerschool, Aug 2008





Motion capture in movies

Facial motion capture was used extensively in Polar Express by Image Metrics where hundreds of motion points were captured.



Tom Hanks

Computer game animators



http://www.studiopendulum.com/alterego/index.html

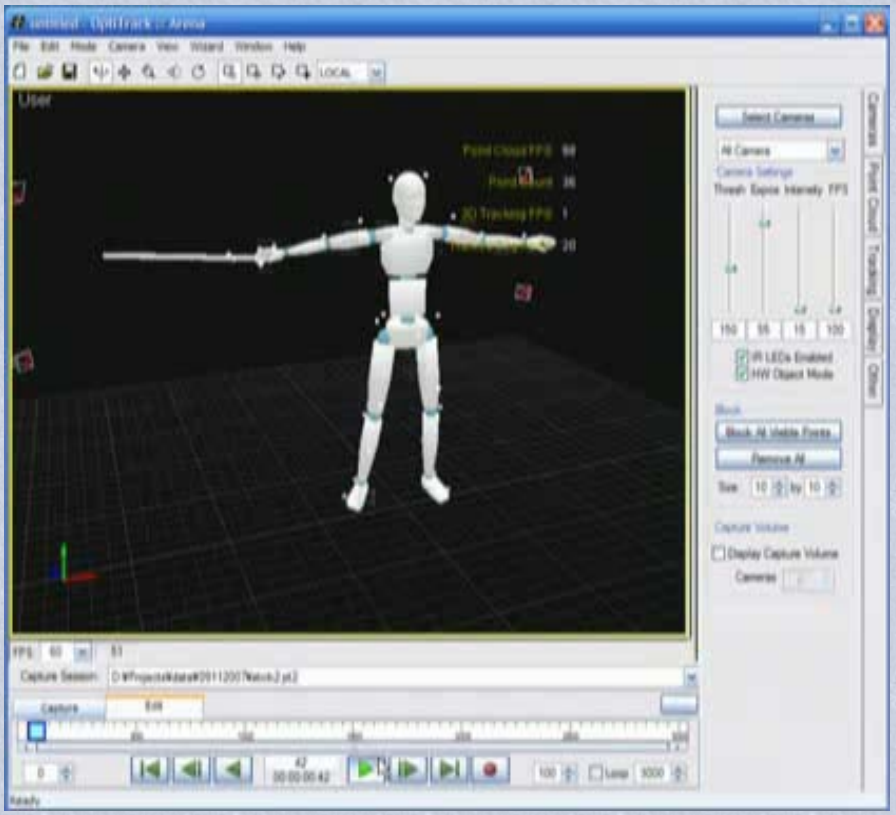
How it works

- IR-sensitive cameras
- Infrared light is emitted from each camera
- Reflective markers are attached to objects to be tracked
- Each camera outputs 2D coordinates of markers
- Data from multiple cameras are combined to provide 3D coordinates



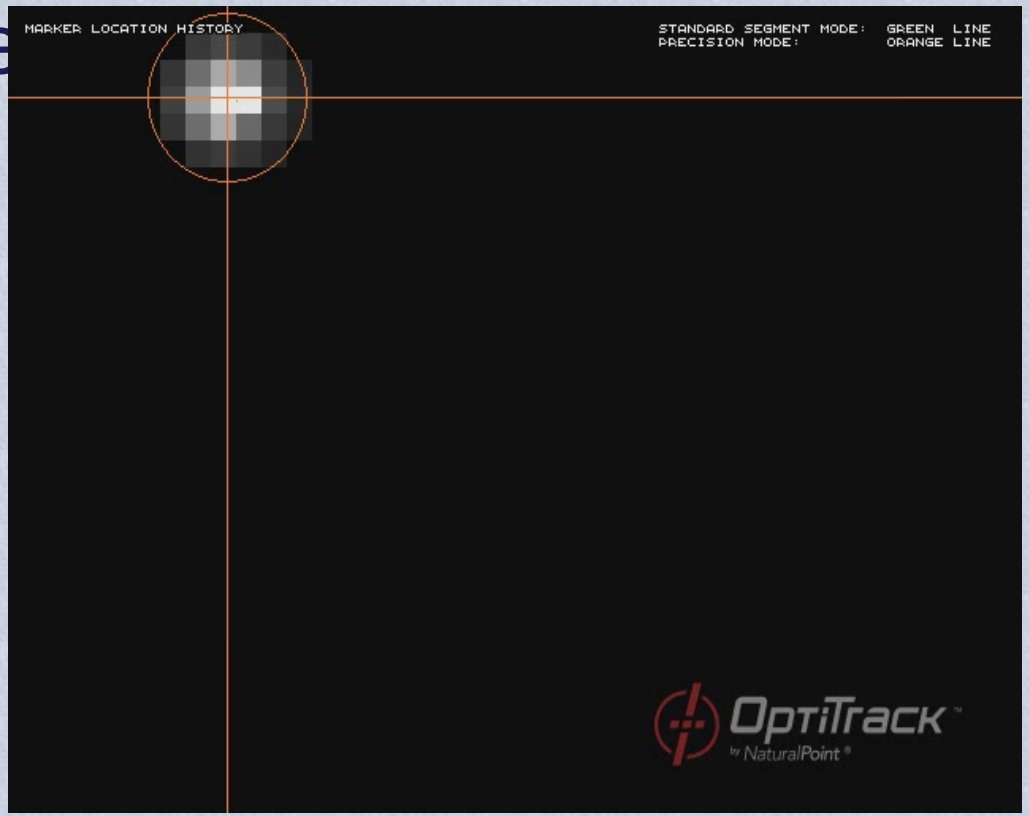
Frame Rate

- Frame rate is typically at least 100 Hz (for some systems >1000 Hz)
- Data can be output in real time on newer systems, allowing for interactive applications



Spatial accuracy

- High spatial accuracy thanks to the use of gray-levels in image proce



THE VISPP COOKING SHOW

The steps of the VISPP hands-on

- Data collection
  - Motion capture and voice recordings of an attentive listener
- Analysis
  - Labelling of data
  - Trigger pattern design
- Resynthesis

Mocap hands-on

- Setup
- Calibration
- Validation
- Recording

Mocap demo


OptiTrack FLEX:V100 Cameras
Sync Splitter
16' USB Uplink
16' USB Cable

What could possibly go wrong?

- **Cameras were**
  - underpowered;
  - physically disconnected;
  - programatically disconnected;
  - broken;
  - unlicensed.
- **Cables were**
  - broken;
  - lost or borrowed (standard cables is a blessing and a curse).
- **USB Hubs were**
  - broken;
  - powered through USB;
  - lost or borrowed.
- **Calibration caused the computer to hang**
  - halfway through;
  - on introduction of plane;
  - on start-up;
  - at some other point.
- **The calibration was off due to**
  - any of the above;
  - reflections from objects in the room;
  - reflections from objects on the person calibrating;
  - poor distribution of samples;
  - bad configuration settings;
  - lack of reflectors;
  - lack of time to validate the calibration;
  - forgetfulness – the results were never saved;
  - the cameras were moved at some point before or during recording.
- **The recording was useless due to**
  - bad calibration;
  - reflexions from objects in the room or on the subject(s);
  - light;
  - missing reflectors;
  - dirty reflectors;
  - concealed reflectors;
  - poorly placed reflectors;

...add it goes on...

Transcription and annotation

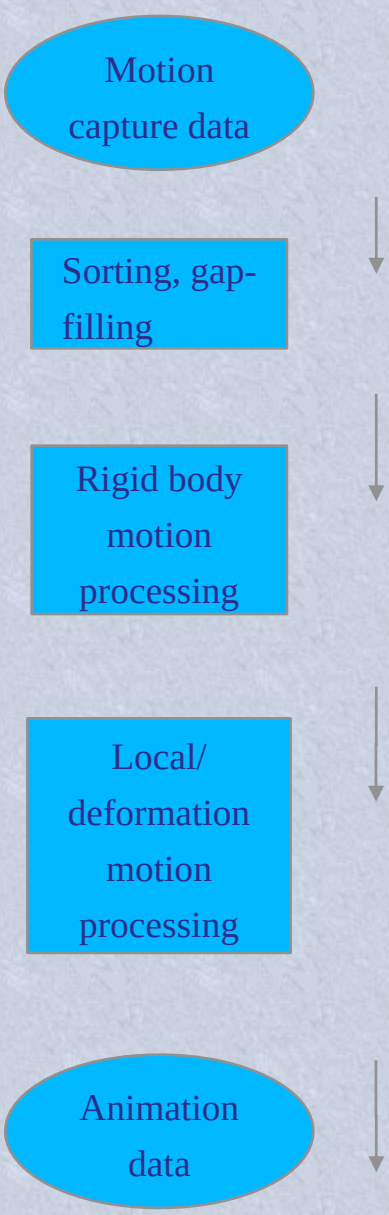
- Time consuming
- Maximize automation
- Iterative annotation
  - Initial annotations quickly
  - Detailed annotations when needed

The VISPP recordings

- Three speakers and three listeners where selected
- Listener wore markers on head, eyebrows and lips
- Speaker told listener something, for example a detailed description of route from home to work
- Listener payed attention, provided feedback
- Duration of recording ~5 min.

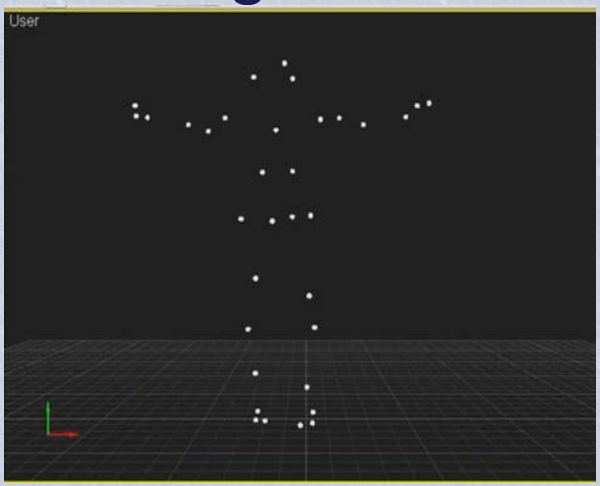
Data processing

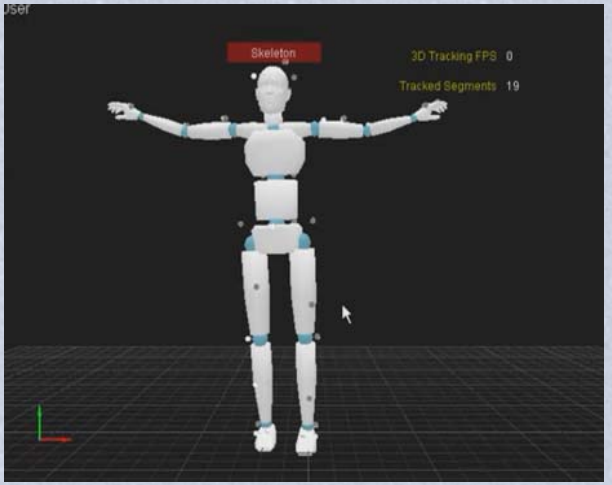
from measurements to animation parameters



Sorting

- Output is a point cloud – how can we tell which point is which?
- Two approaches:
  - Sorting based on motion
  - Sorting based on a model





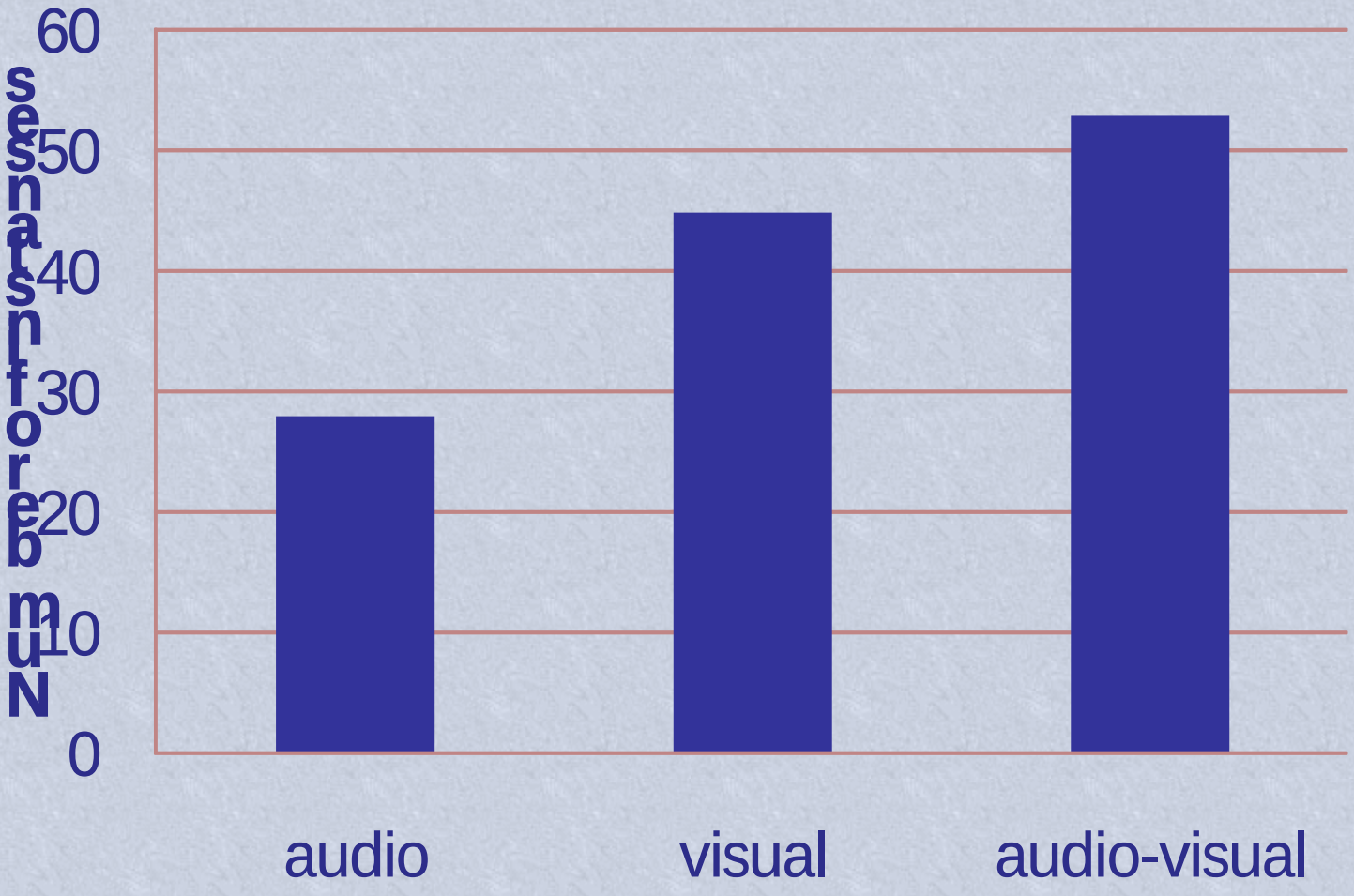
Synchronization of sound and points

Demo

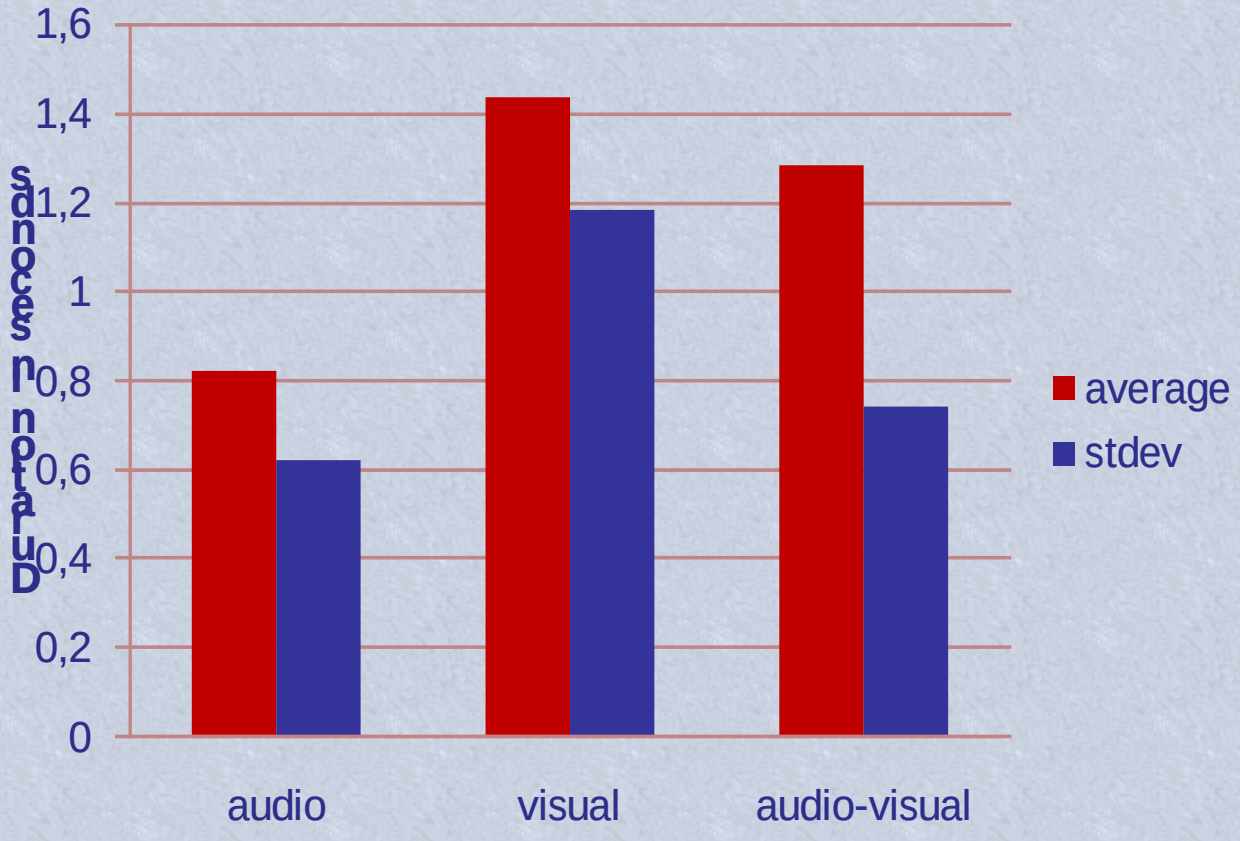
Annotation in VISPP

- 11 groups analysed 2 of the recordings
- Task: label feedback
- Used three categories/labels:
  - Acoustic feedback/af
  - Visual feedback/vf
  - Bi-modal feedback/bf
- Tool: wavesurfer with talking head plug-in

What the students tagged



Length distributions



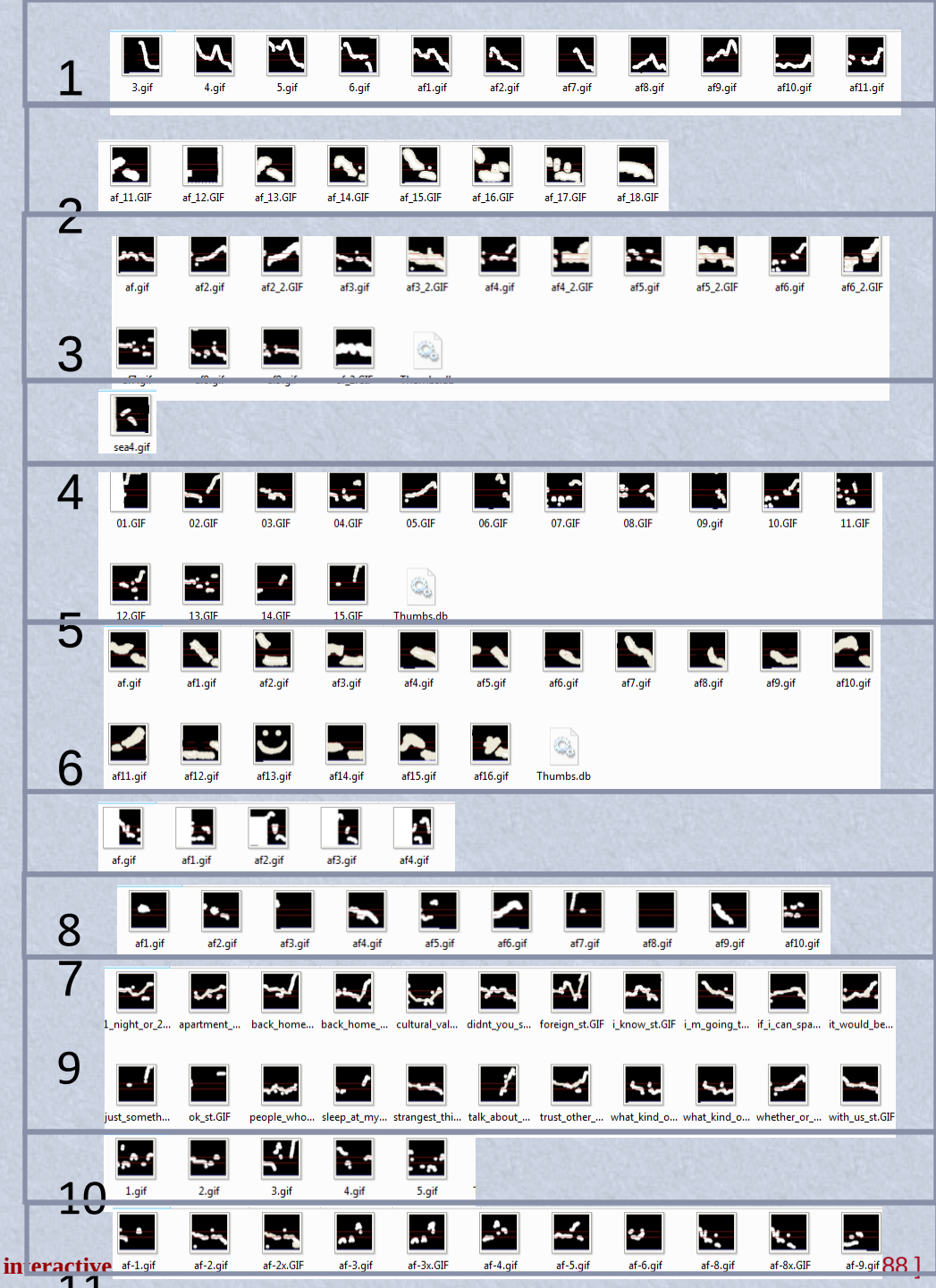
VISPP triggers

- Analysis of F0 preceding feedback
- Simple pattern recognition
- Draw in Paint; Photoshop, etc.

VISSP

trigger

patterns



SPONTAL conversational database

- Research Project: “Multimodal database of spontaneous speech in dialog” 2007-2010, funded by the Swedish Research Council, KFI - Grant for large databases

Research Program

- Both vocal signals and facial and body gestures are important for communicative interaction
- Signals for turn-taking, feedback giving or seeking, and emotions and attitudes can be both vocal and visual
- Our understanding of vocal and visual cues and interactions in spontaneous speech is growing, but there is a great need for data with which we can make more precise measurements
- A large Swedish multimodal database will enable researchers to test hypotheses covering a variety of functions of visual and verbal behavior in dialog
- Freely available for research

Research Goals

- Swedish multimodal spontaneous speech database
- Rich enough to capture speaker and speaking style variation
- High-quality audio and video recordings (HD)
- 5% motion capture for facial and head gestures
- Possibly 80-90% motion capture for body and head movements

Swedish Multimodal Database

15 dialogues* { female-female, female-male, male-female, male-male } { friends, strangers } motion capture

At least eight dialogues with motion capture for gesture and facial movements (one per configuration)

Motion capture for body gestures for the entire database

20 minutes free dialog, 10 minutes discussion of an artifact

Total database = 120 dialogues, 30 minutes for each dialog = 60 hours

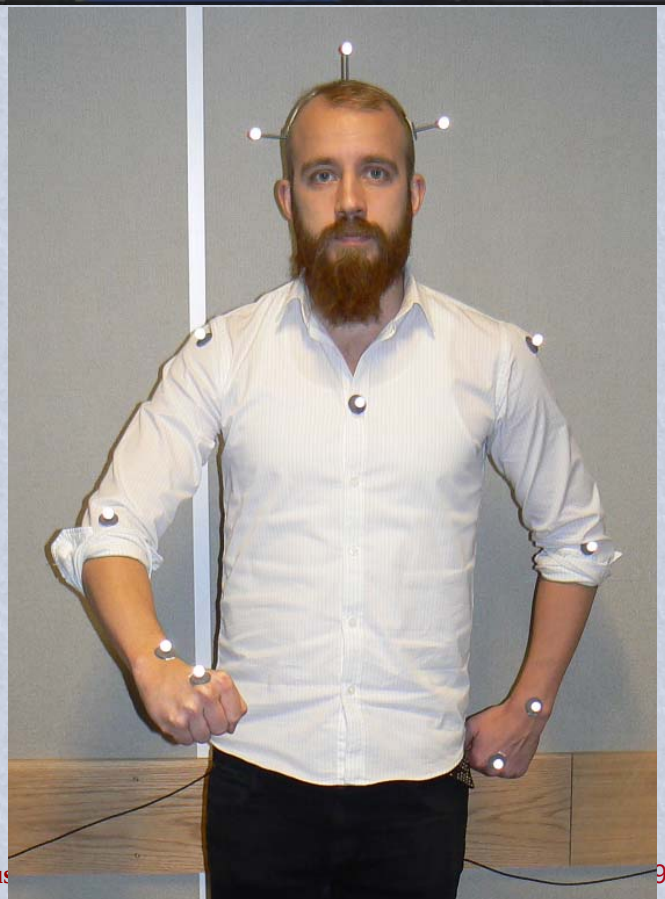






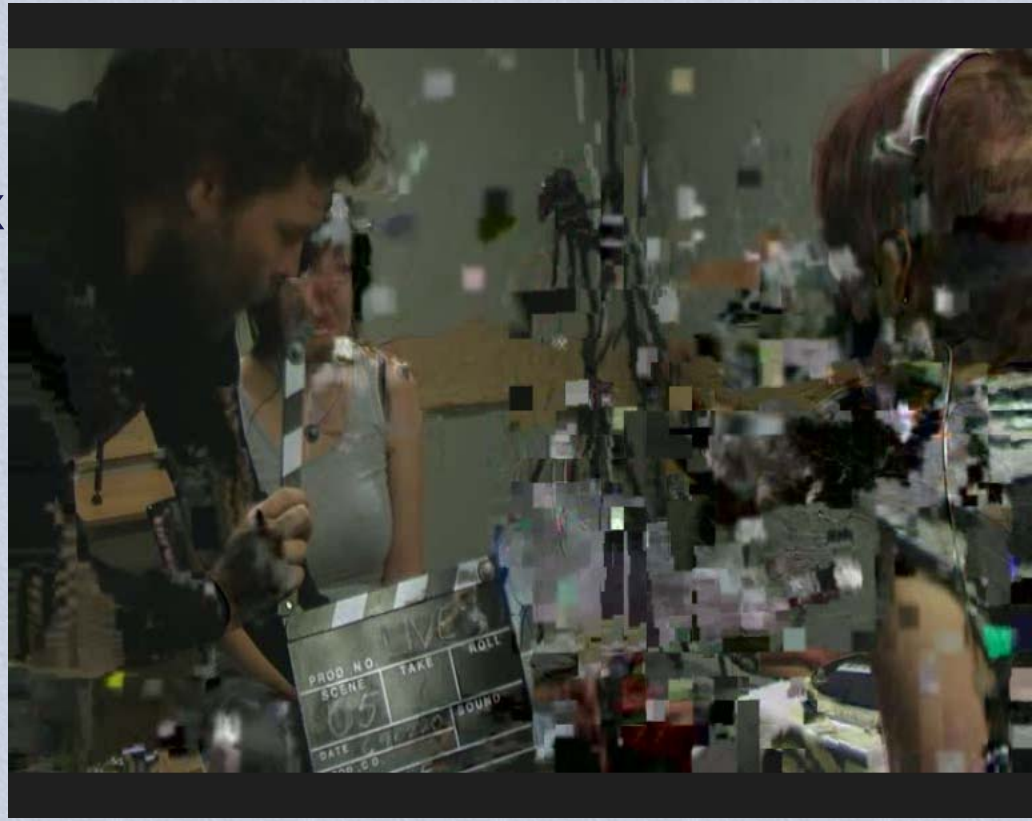






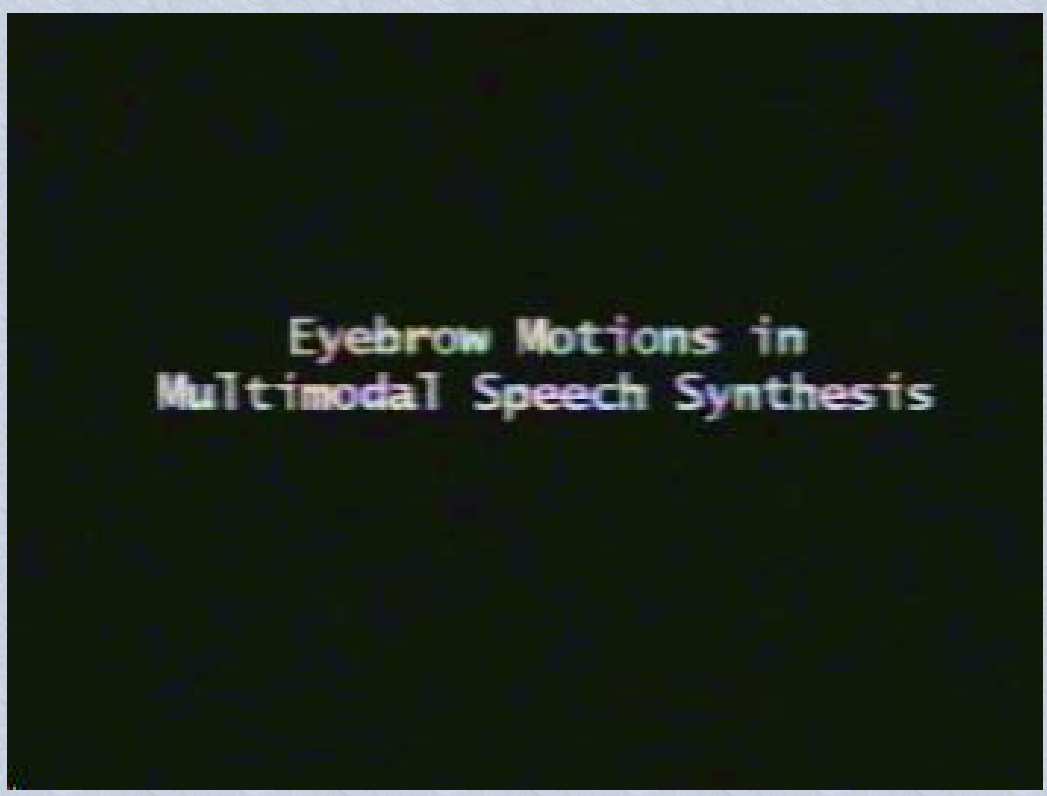
The real world...

- Fortunately, spoken interactions are complex
- Real example from the Spontal recordings: Evelina and Evelyn
  - Incrementality a must…
  - Tools need continuous adjustments, as do models
  - Subjects crippled by the laboratory setting?



Movie time

Some perceptual experiments with AV speech
Is glottal frequency and eyebrows related?



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.

Three more AV perception experiments

- Cues for prominence – conventional
- Cues for feedback – the eavesdropper
- Cues for feedback – the participant

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

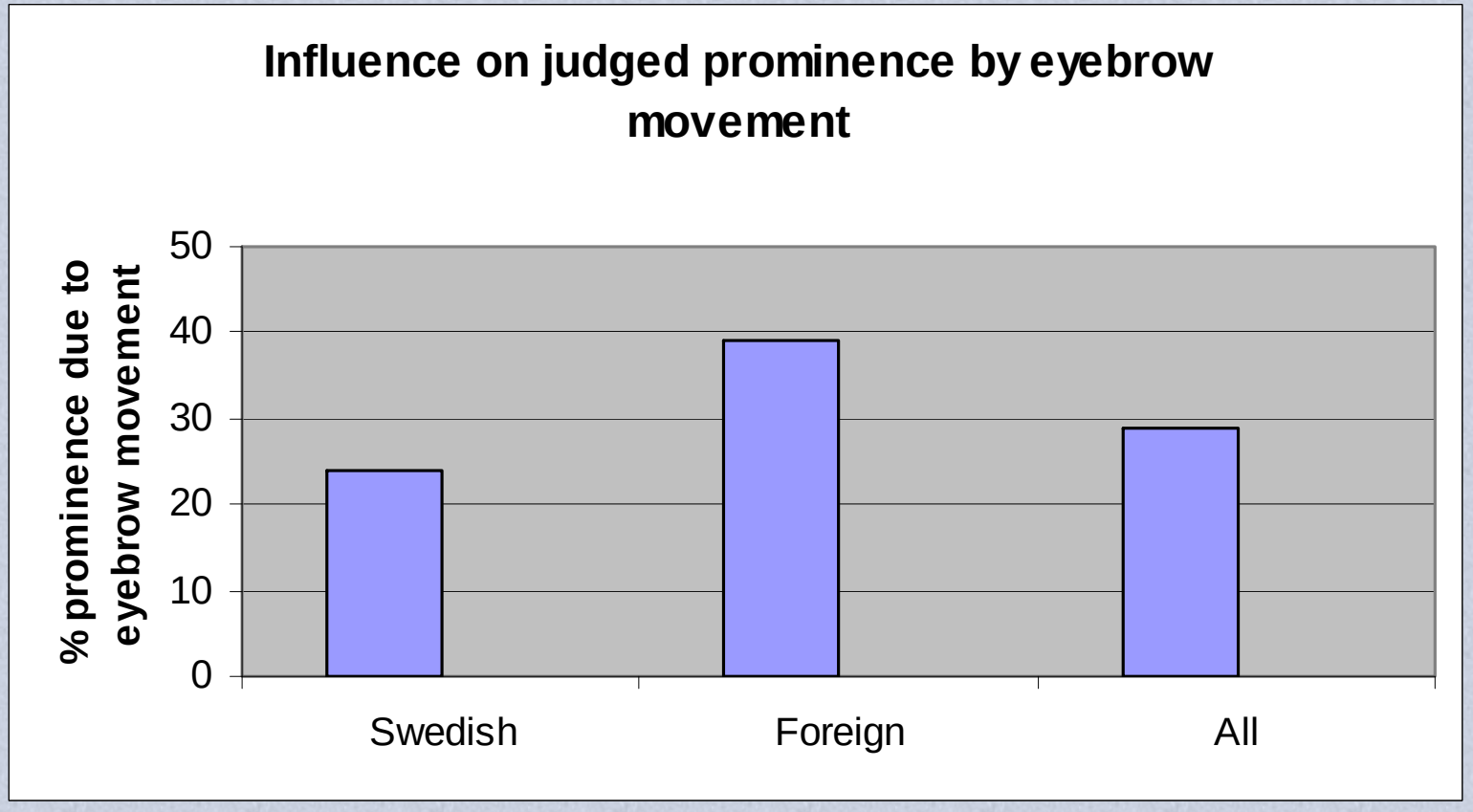
TWO EXAMPLES



No eyebrow movement (neutral)

Eyebrow movement

Prominence increase due to eyebrow movement



Cues for feedback – the eavesdropper

- Mini dialogues (two turns) Travel agent application
  - *Human: ”Jag vill åka från Stockholm till Linköping”*
  - *Agent: ”Linköping”*

  (subject sees and hears travel agent but only hears customer)
- 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

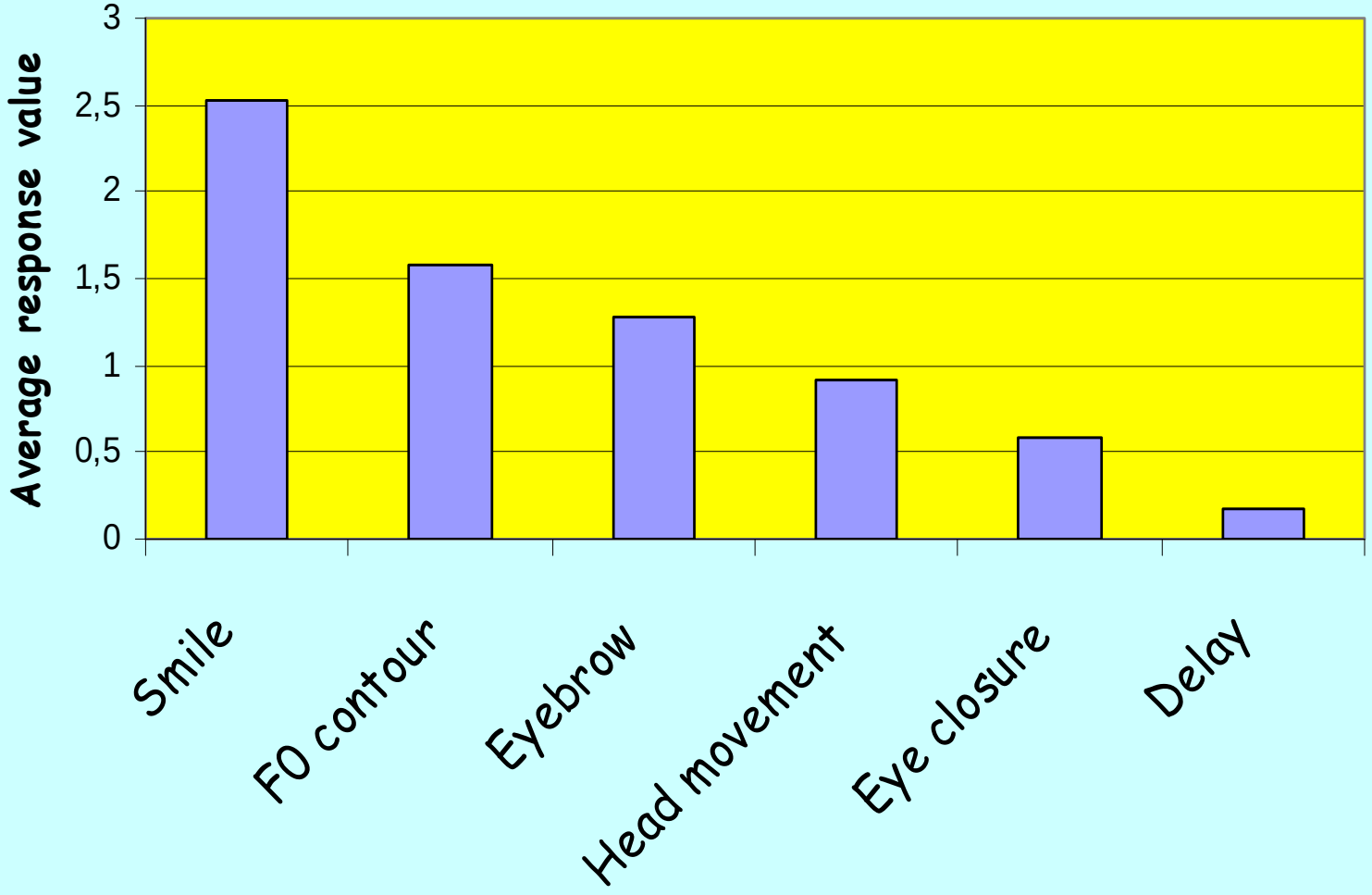
Granström, House & Swerts (2002)

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

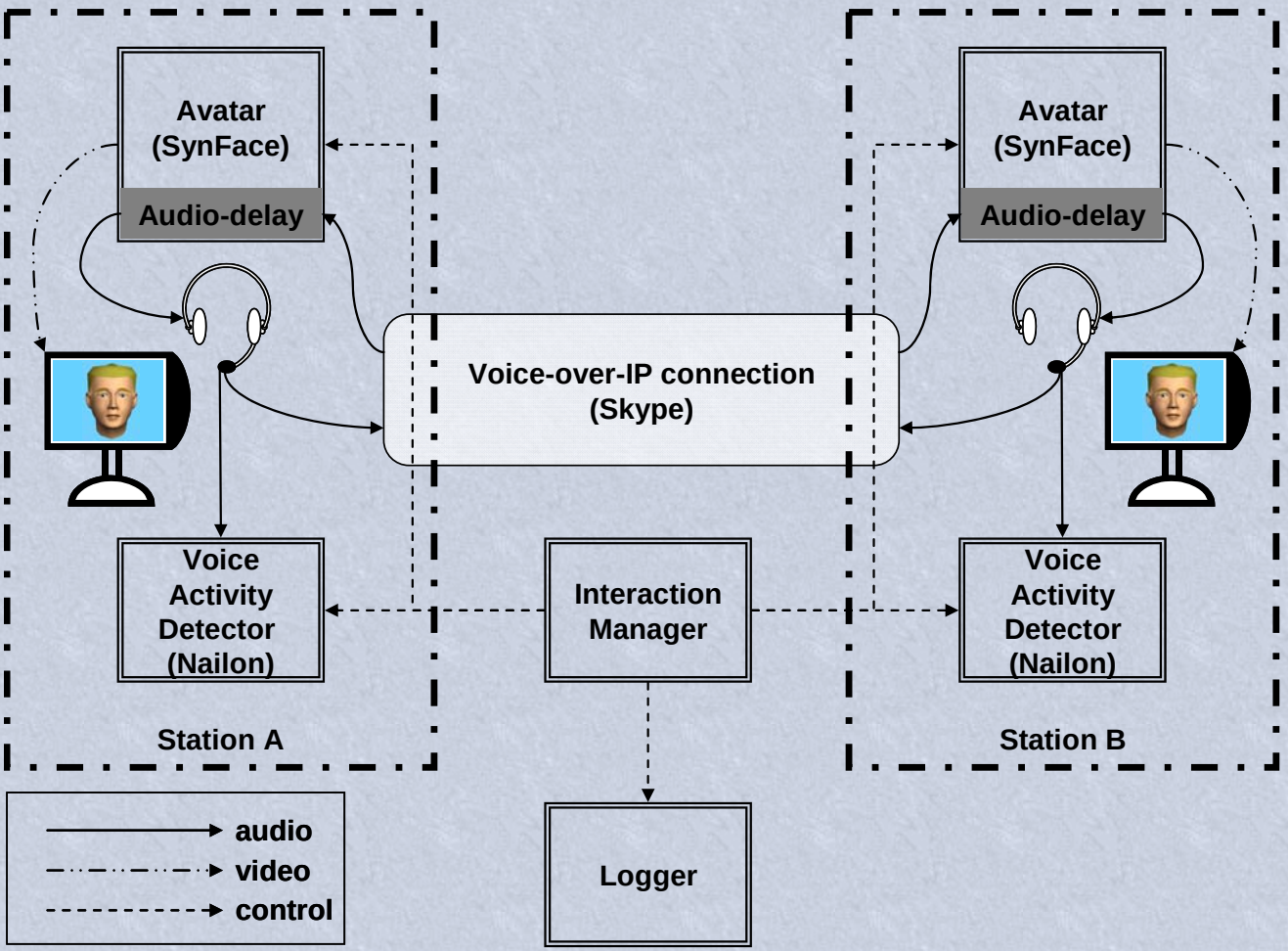


Cues for feedback – the participant

- Based on “SynFace” technology originaly used as lip reading support for hard-of-hearing making telephone calls
- Natural audio combined with synthetic (avatar) faces
- Can manipulated visual feedback affect the turntaking behaviour of subjects?

Edlund & Beskow, Interspeech 2007

The experimental system

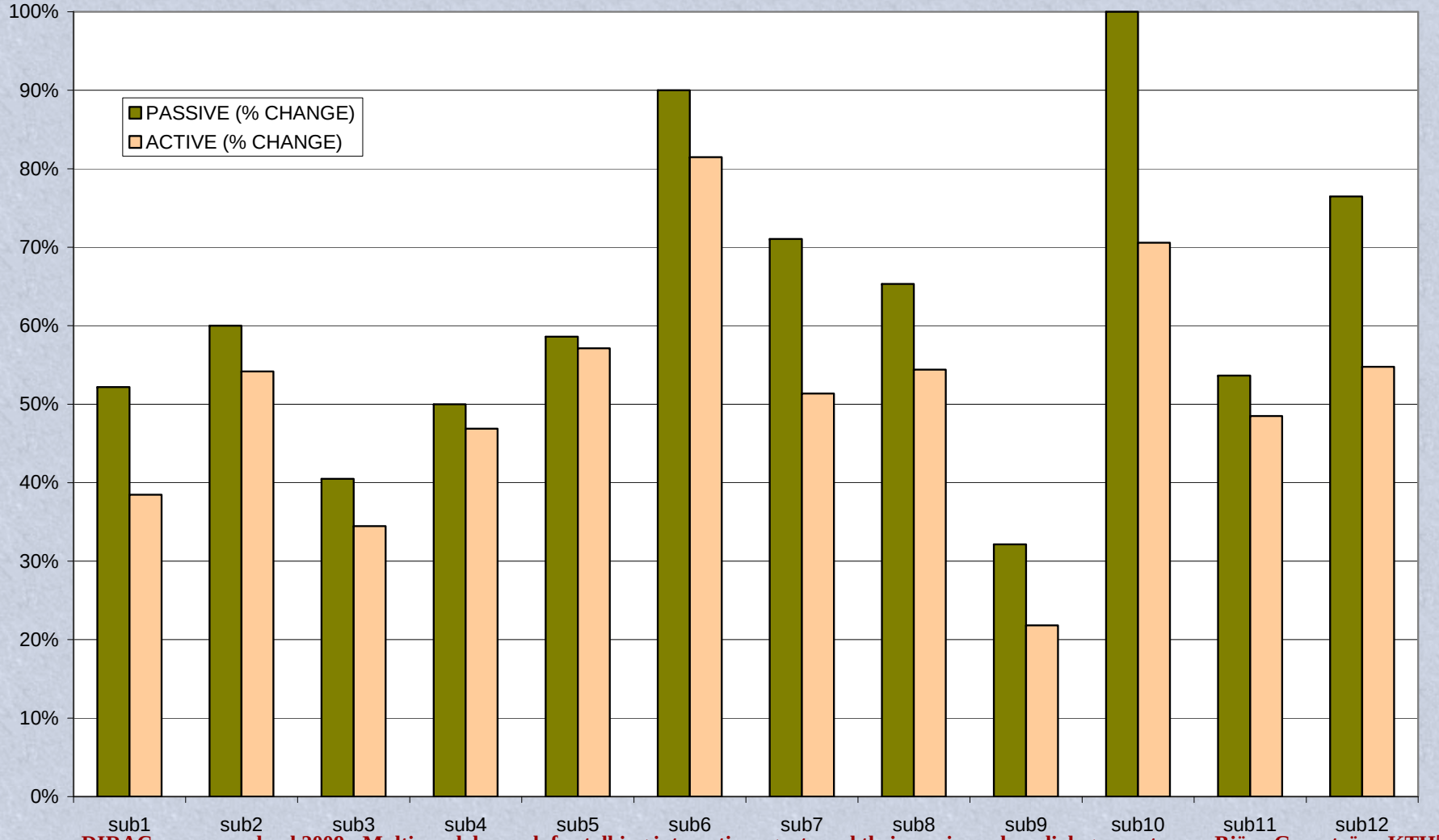


Subject A **Subject B**

The interaction manager

- Controls the visual turntaking behaviour of the avatar by controlling facial gestures:
  - Turntaking/keeping gesture (head turn and looking away (active)
  - Turn yielding/listening gesture (looks at the subject with slightly raised eyebrows (passive)
- Switching after ten detected silences between avatar A/B being neutral/neutral, active/passive, passive/active

Percentage of contributions followed by a change of turn for twelve subjects represented by passive vs. active avatars



Eye gaze experiment
"is this man looking at you?"



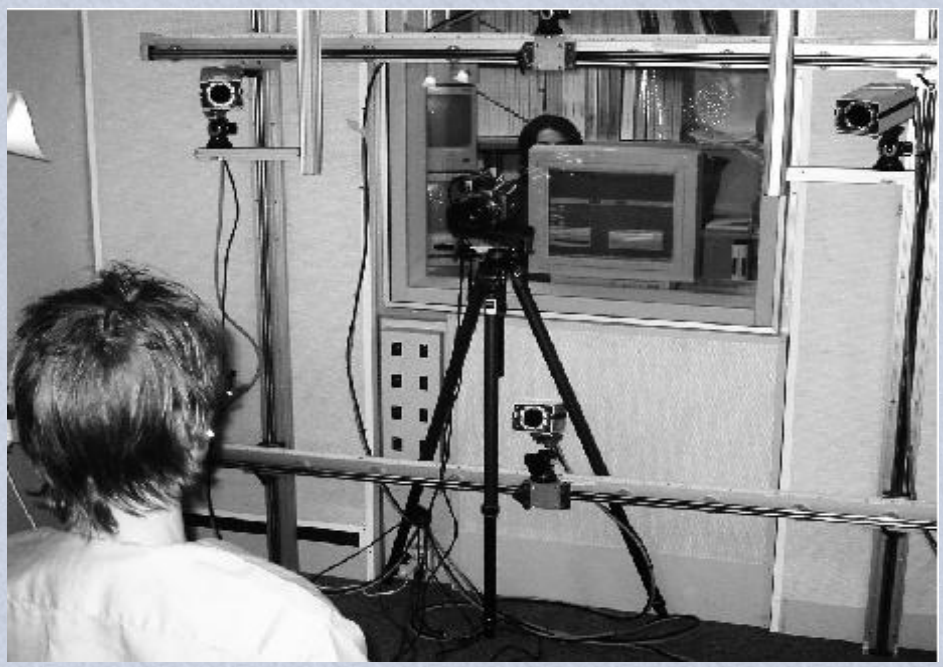
Mikael Nordenberg, Gunilla Svanfeldt & Preben Wik

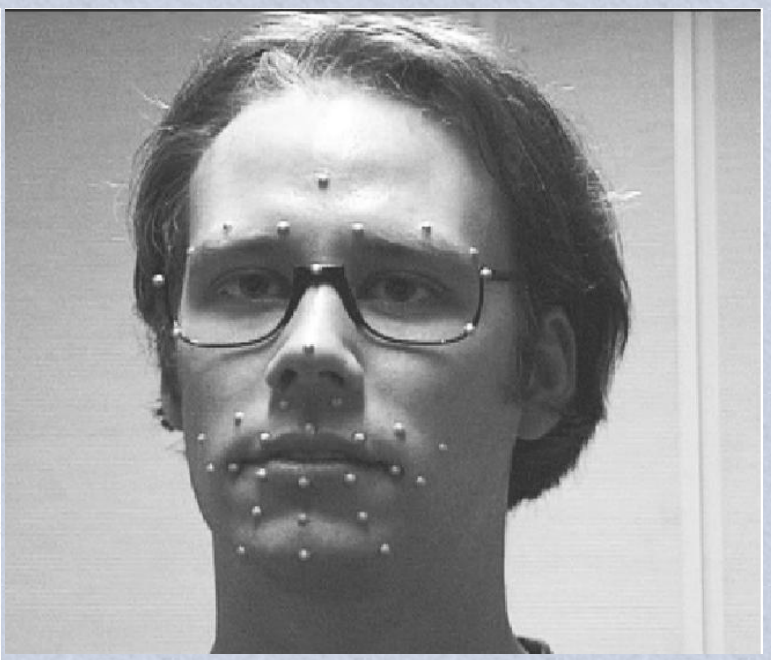
Improving believability in the talking heads

- Better brains, ASR, TTS etc.
- Improving facial animation
  - Articulation
  - Communicative gestures
  - Eye-movements
- Learn from human-human interaction
  - Eye-gaze for sending social signals
  - Synchronize speech



Data-driven articulation with 3D-mocap device for capturing articulation , but we can´t capture eye-movements with this technique





Investigating eye-movements

- Existing statistical models for eye-movement
- Requires the possibility of eye-contact
- How can we achieve eye-contact?
- How sensitive are we?

The experiment



How sensitive are we at detecting eye contact with the agent?

The experiment



Subjects were asked to look at pictures of the agent, and asked to answer the question:

"Does this man look you in the eyes?"

The experiment: Setup

- 2 head positions: front, side



Extremes: X-dimension



-20



+20

Extremes: Y-dimension



-20



+20

Extremes: Z-dimension



-95

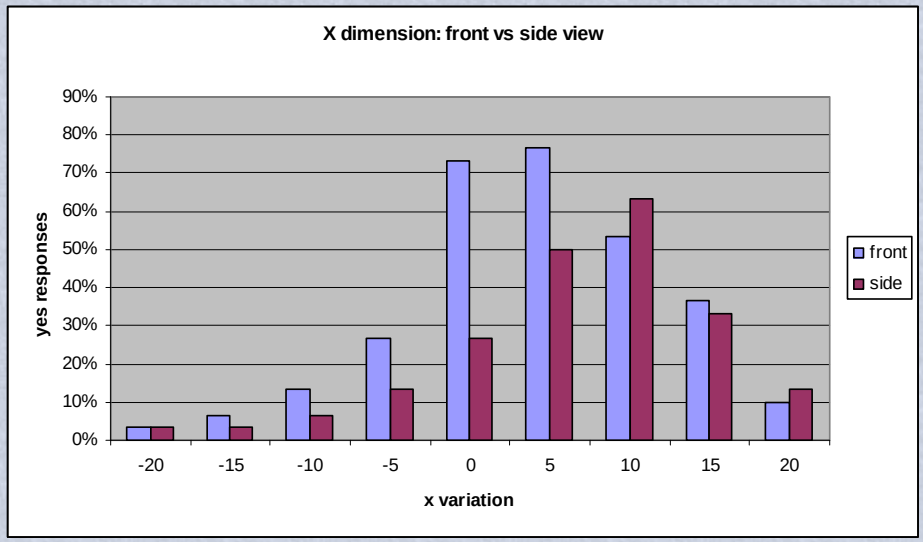


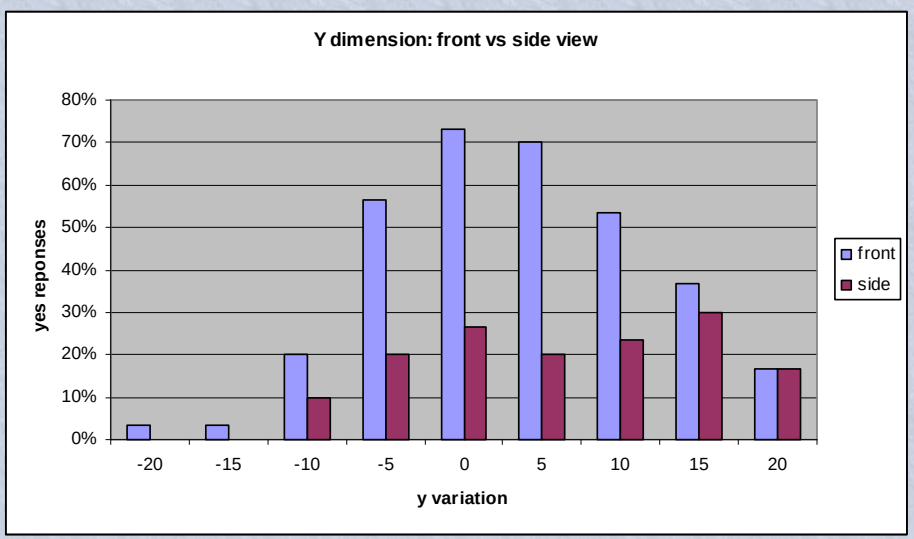
+175

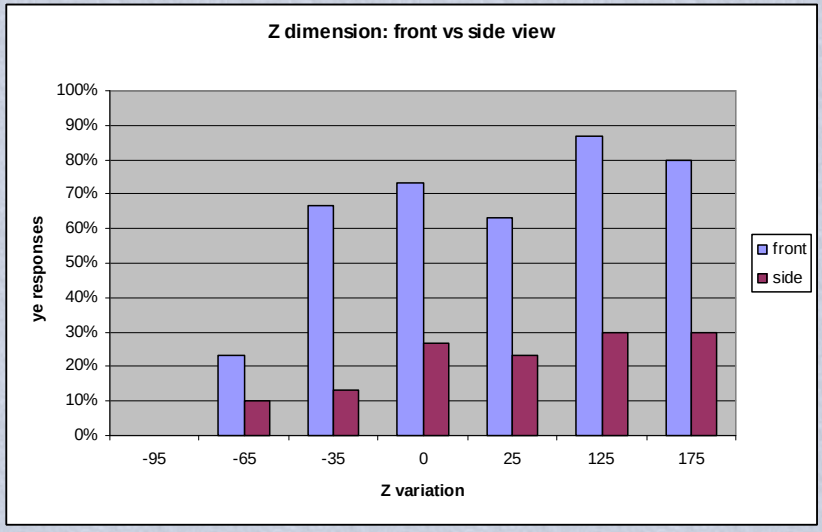
The experiment: Details

- Task: yes/no to "Does this man look you in the eyes?"
- Number of still pictures: 25
- 2 x each picture, random order
- Six pictures in the beginning as "dummies"
- 4 additional pictures – qualitative questions/answers
- The most prominent defect of the talking face?

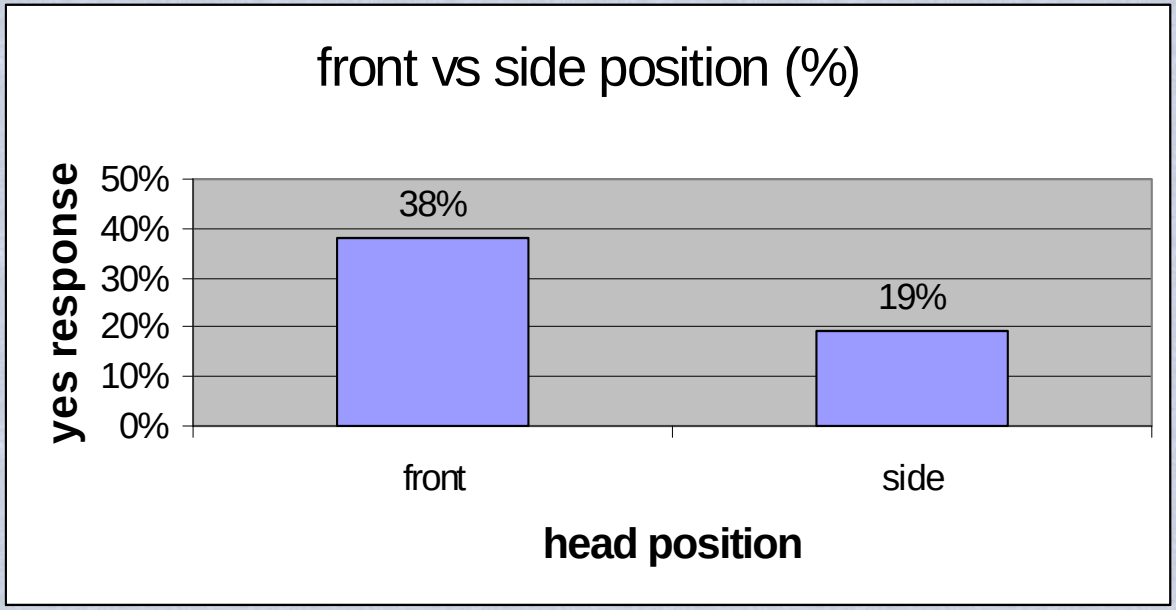
Distribution of yes responses





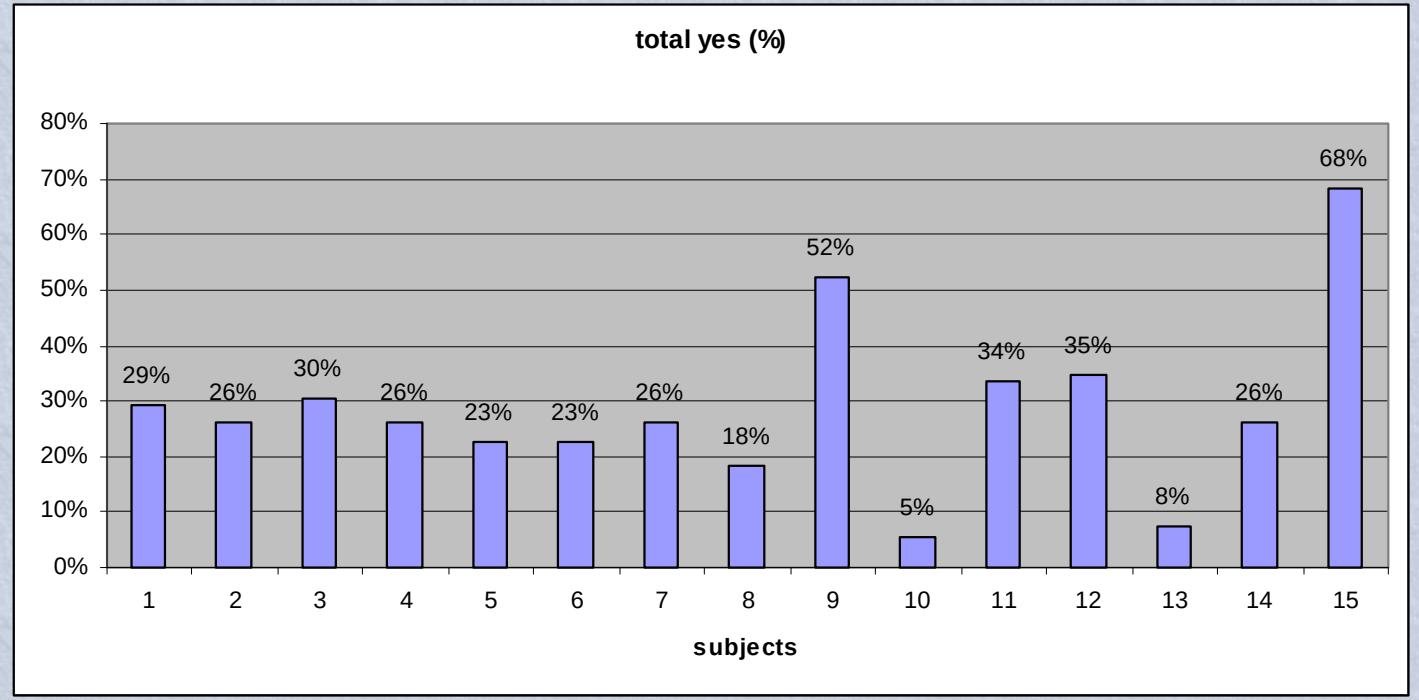


Head-angle influence





Test subject variation



Conclusions

- It is possible to produce eye gaze in the synthetic face that the observers perceives meets their gaze
- Could look subject in the eyes, but not always when expected
  - More studies needed to predictably achieve eye-contact.

Ideas for future work

- Try different combinations of viewpoints (x, y, z)
- Try different combinations of lighting
- Same test using pictures with real persons
- Focus point test instead of eye-contact test
- Static vs animated images
- Data driven methods for eye movements
- Use of Tobii eyetracker (in EU-HaH project – later)

Speech synthesis applications

- Dialogue systems
- Translators
- Aids for disabled persons
- Mediated communication
- New HMI
- Language learning
- ….

Spoken dialogue systems at CTT





















Multi-user In Museum

Computer Game

Pixie

92 Waxholm

96 Olga

97 GULAN

98 AUGUST

00 ADAPT

02

05

08 MonAMI

Embodied Pioneering Agent

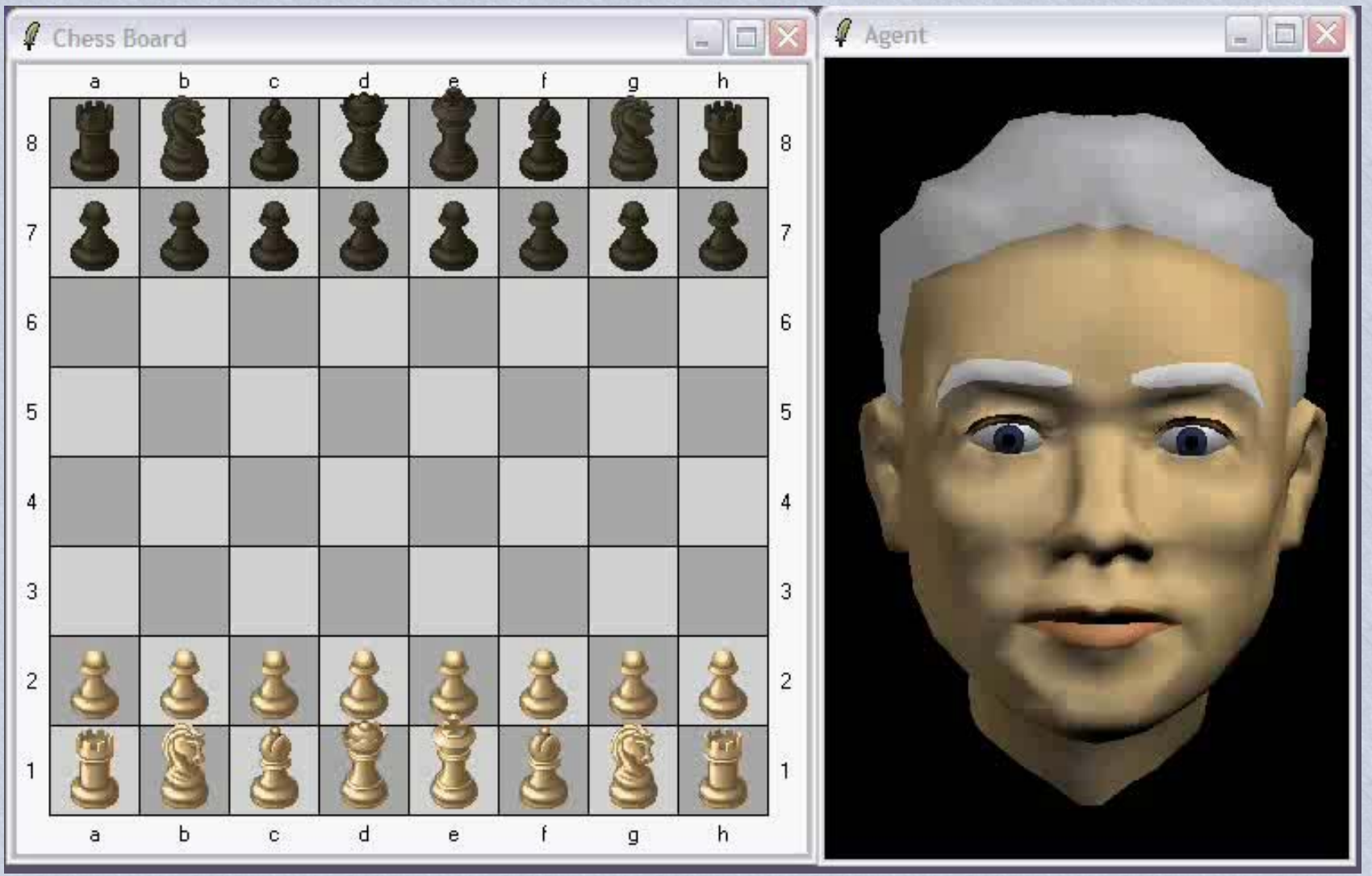
Education

Public Agent

Multimodal

Face-to-face

The CTT chess player



The virtual teacher vision already at STiLL, Marholmen 1998!



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

The cost of pronunciation errors

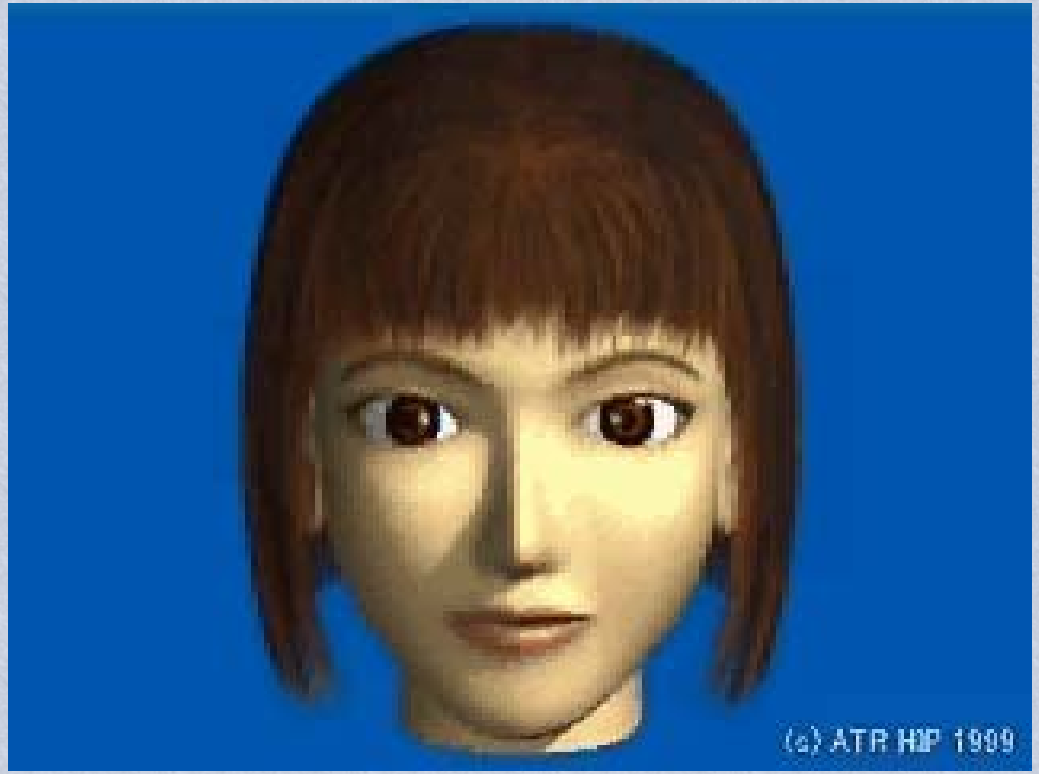
- Video

Automatic tutor simulation

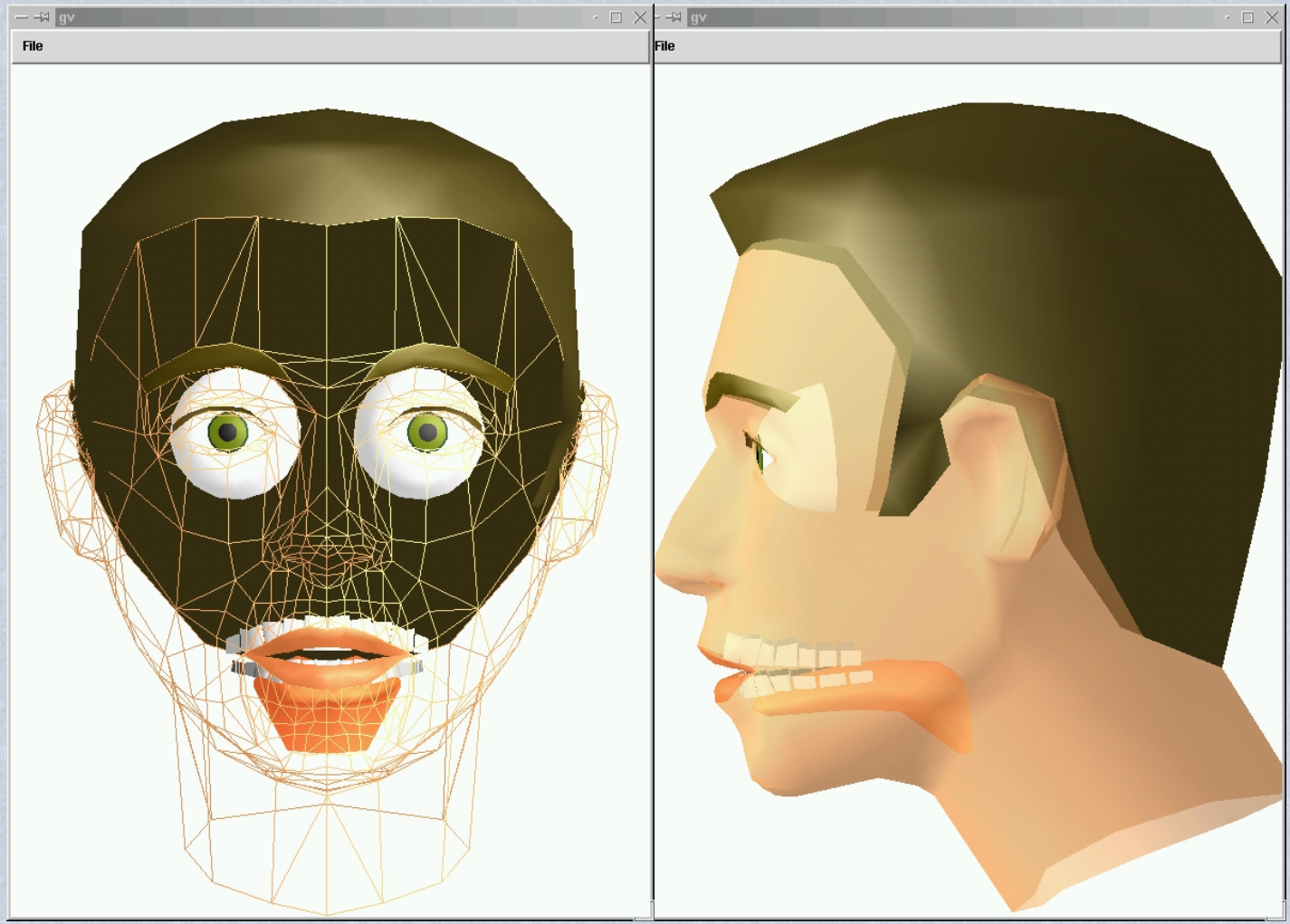




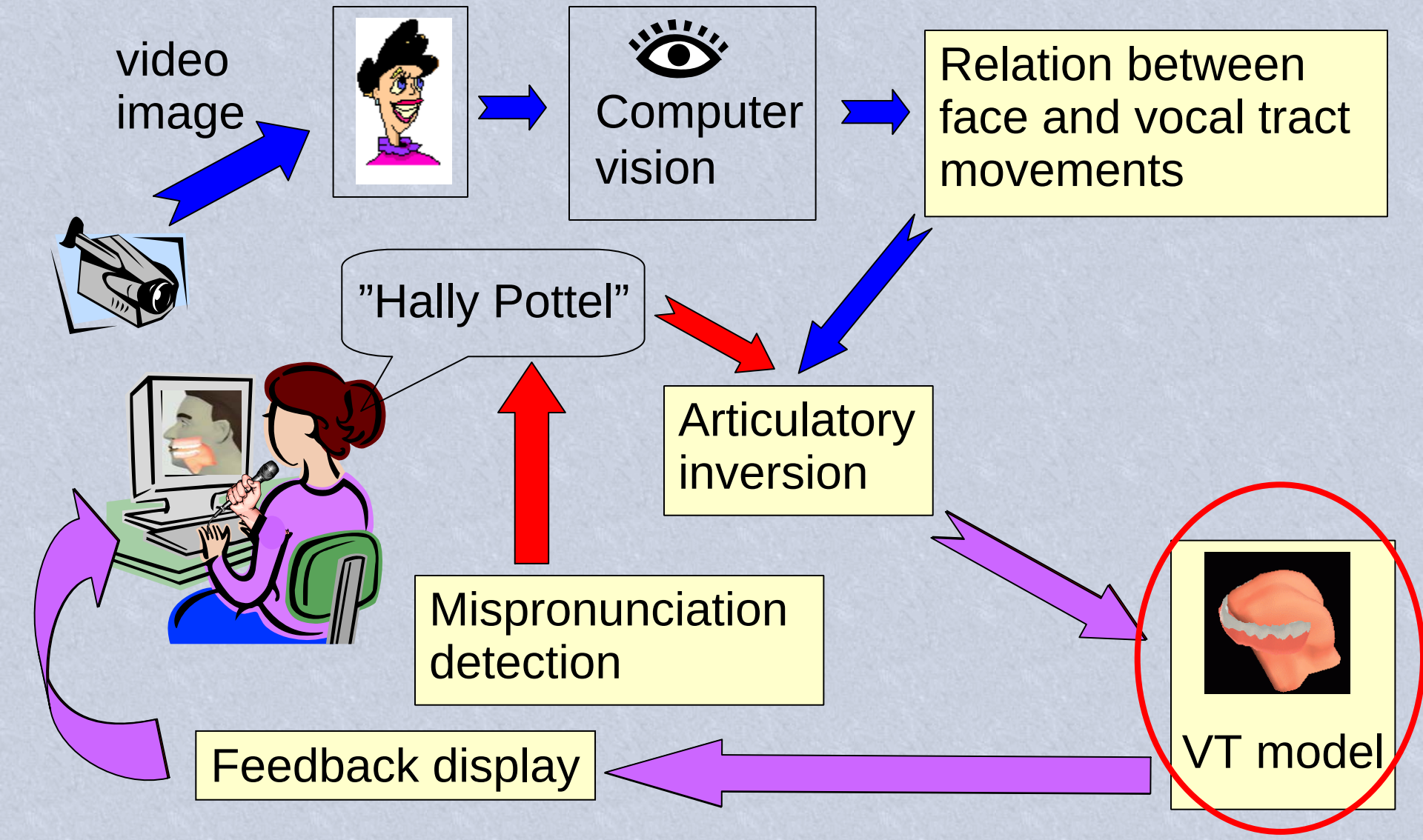
Reiko Yamada ATR, 1999



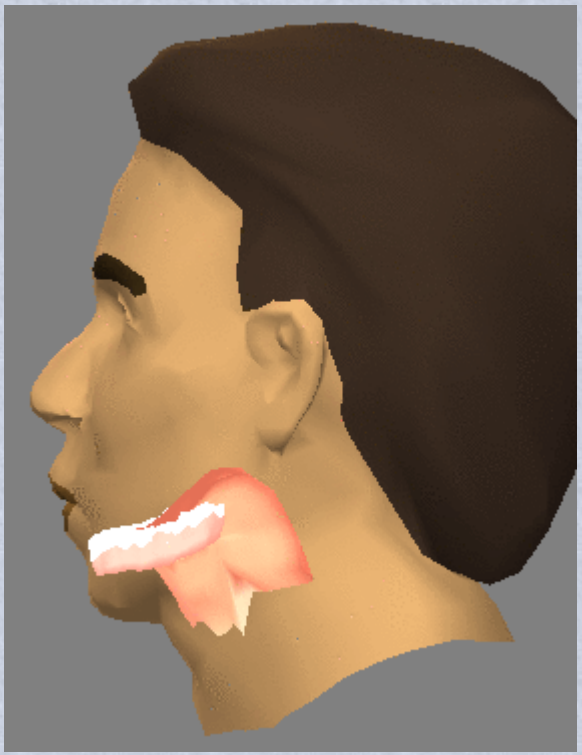
Different representations



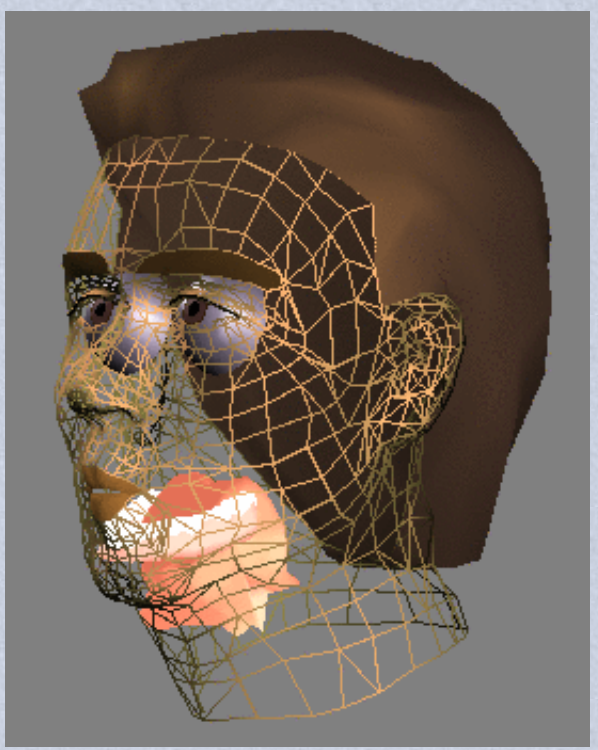
ARTUR: the articulation tutor



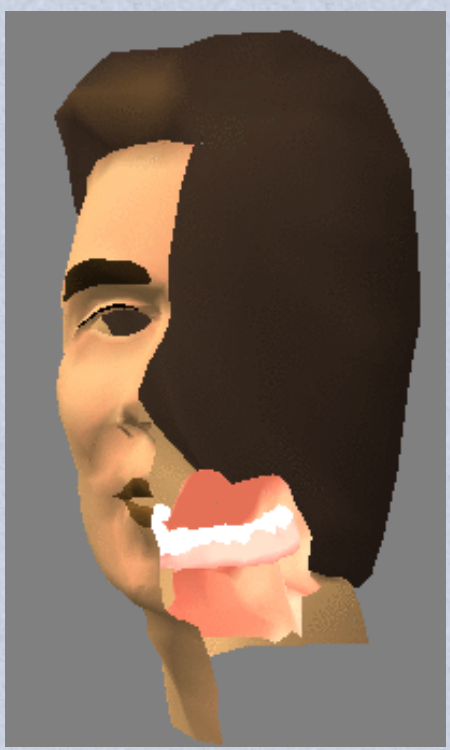
3D tongue movements



[s]



[k]



vowels

"OK" 3D tongue movements (difficult to define an objective evaluation criteria - *suggestions?*)

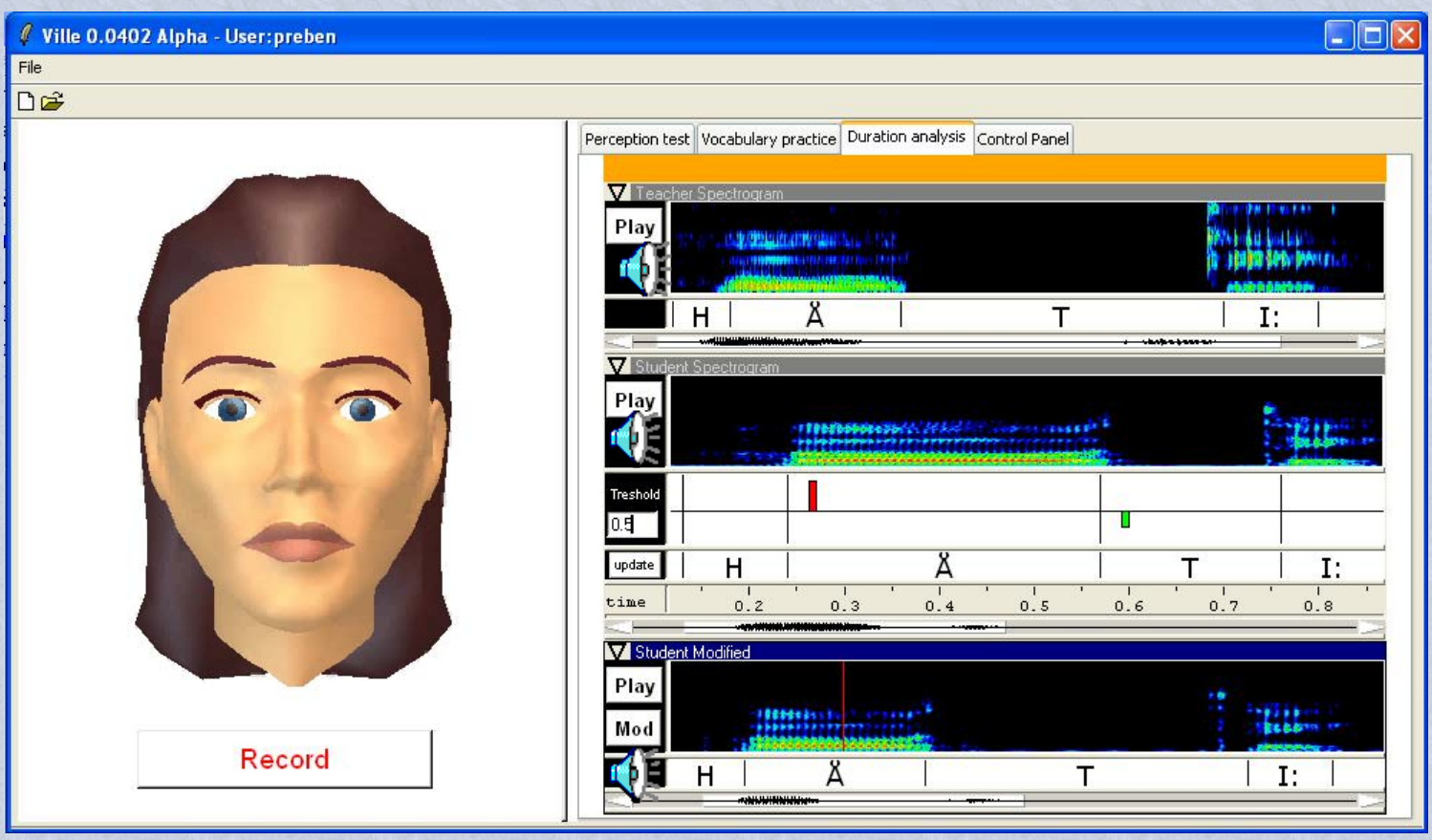
The tongue movements can be presented from different views

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

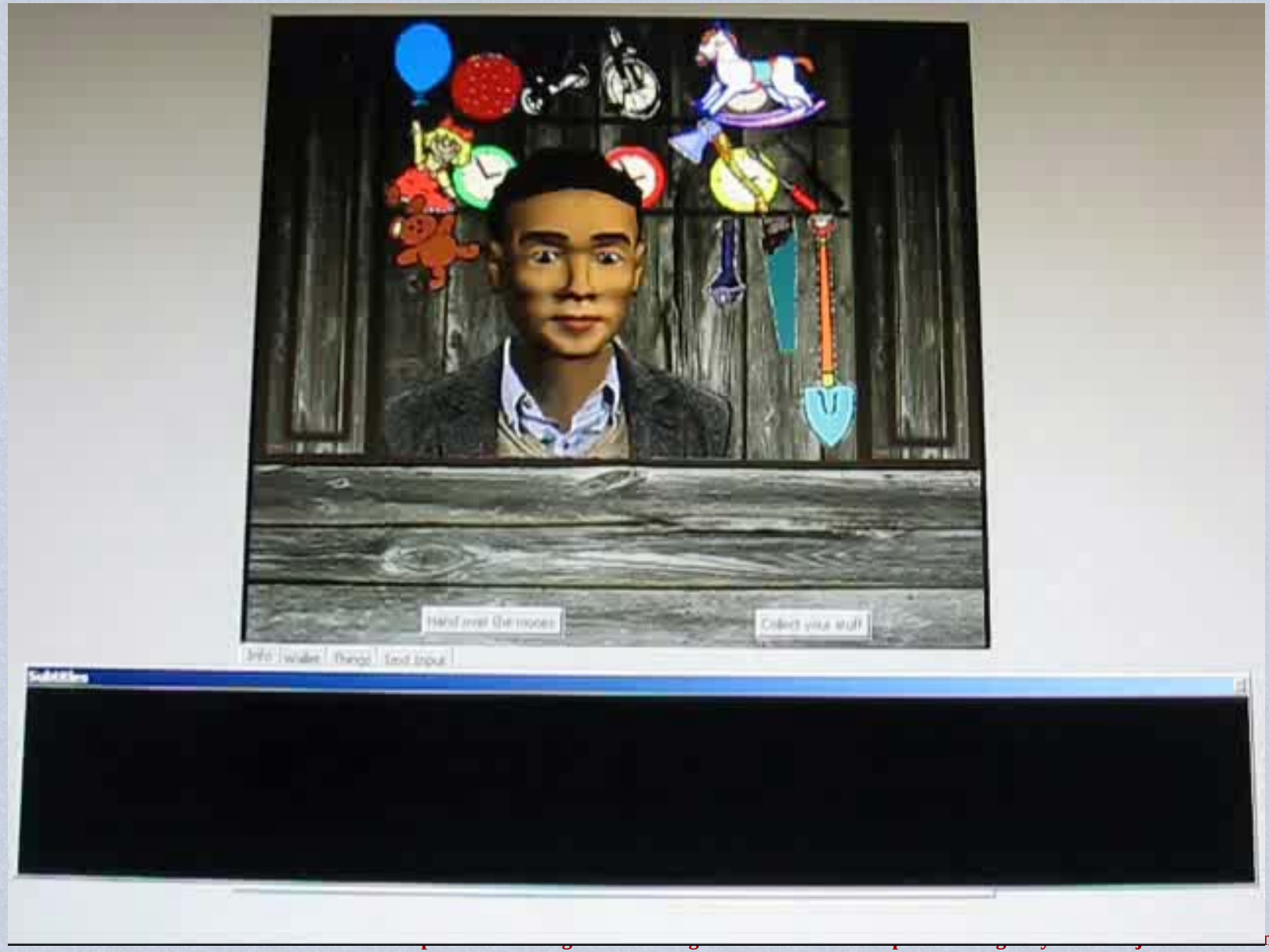


Nordic project –NordPlus Sprog

Using VILLE – CTT Virtual Language Tutor



Also for conversational training…



Speech tech for language learning





Disabled and elderly – a new market?

According to the Swedish Institute for Assistive Technology 1 200 000 persons have a disability (of a total population of 9 000 000!).

http://www.hi.se/templates/Page____2474.aspx

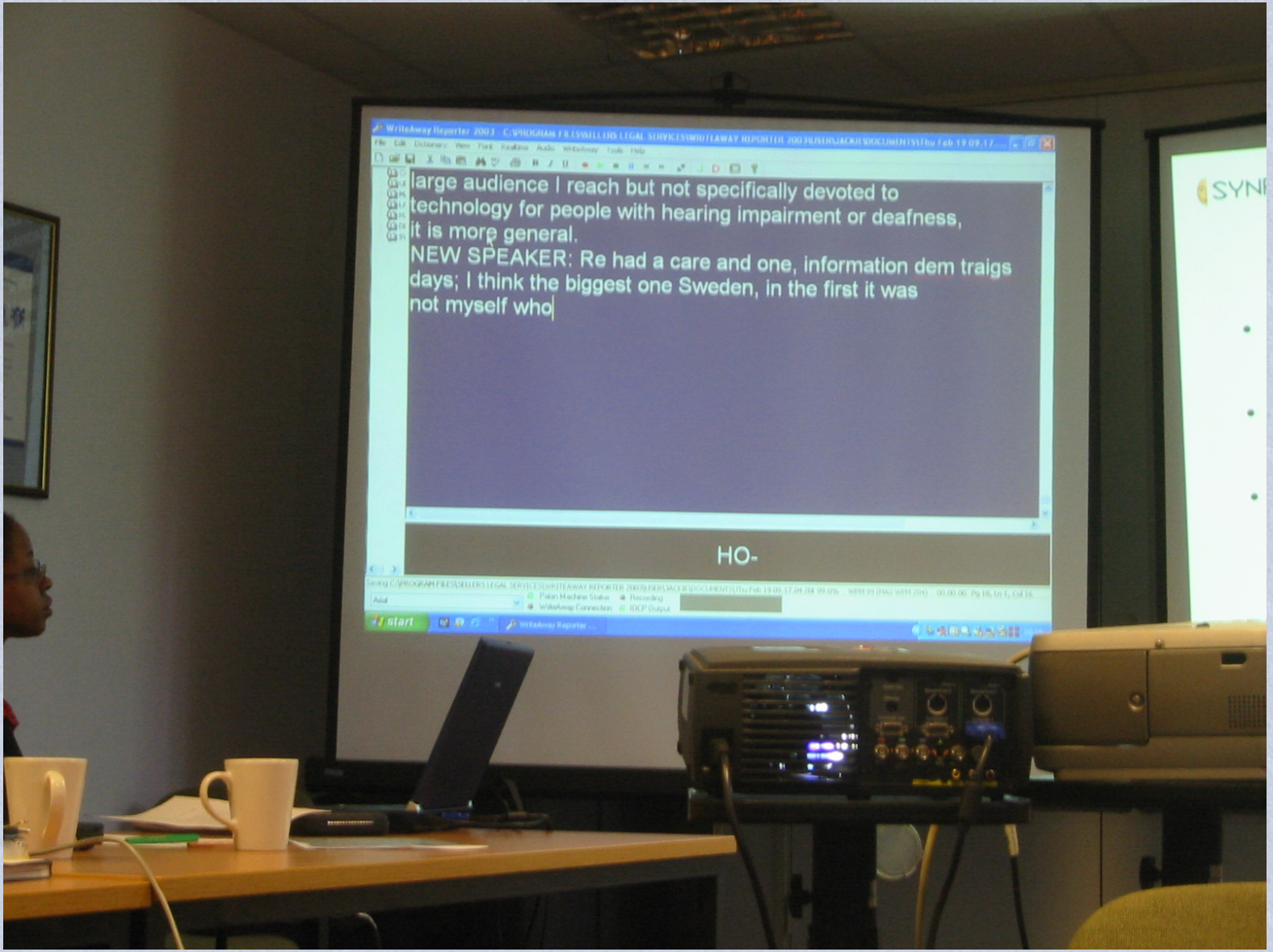
Funktionshinder	Disability	Antal
Hörselskadade	Hearing impaired	780 000
Döva	Deaf	14 000
- varav födda döva	Born deaf	10 000
Synskadade	Visually impaired	175 000
- varav gravt synskadade och blinda	Severe visual impairment and blindness	14 000
Dövblinda	Deaf blind	1 200
Rörelsehindrade	Mobility challenged	600 000
Reumatiker		800 000
Svåra tal- och språkskador	Severe speech and language disorder	40 000
Begåvningsnedsättning	Intellectually challenged	400 000
- varav utvecklingsstörda		40 000
- varav gravt utvecklingsstörda		16 000
Dyslexi	Dyslectic	600 000
Stamning	Stuttering	80 000
Neuropsykiatriska/neuropsykologiska problem hos barn (DAMP/ MBD, MPD, ADHD, autism)		120 000

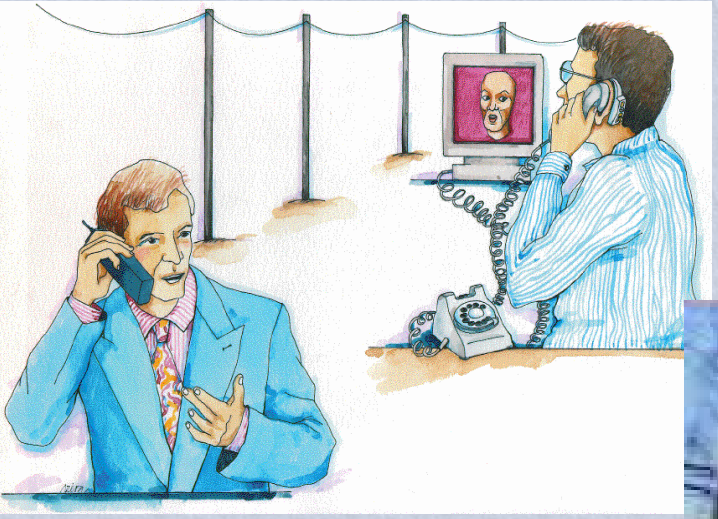
Källa: Hjälpmedelsinstitutet

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

Speech-to-text for hard-of-hearing persons – man or machine?





"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.



Phone conversations are animated

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

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.

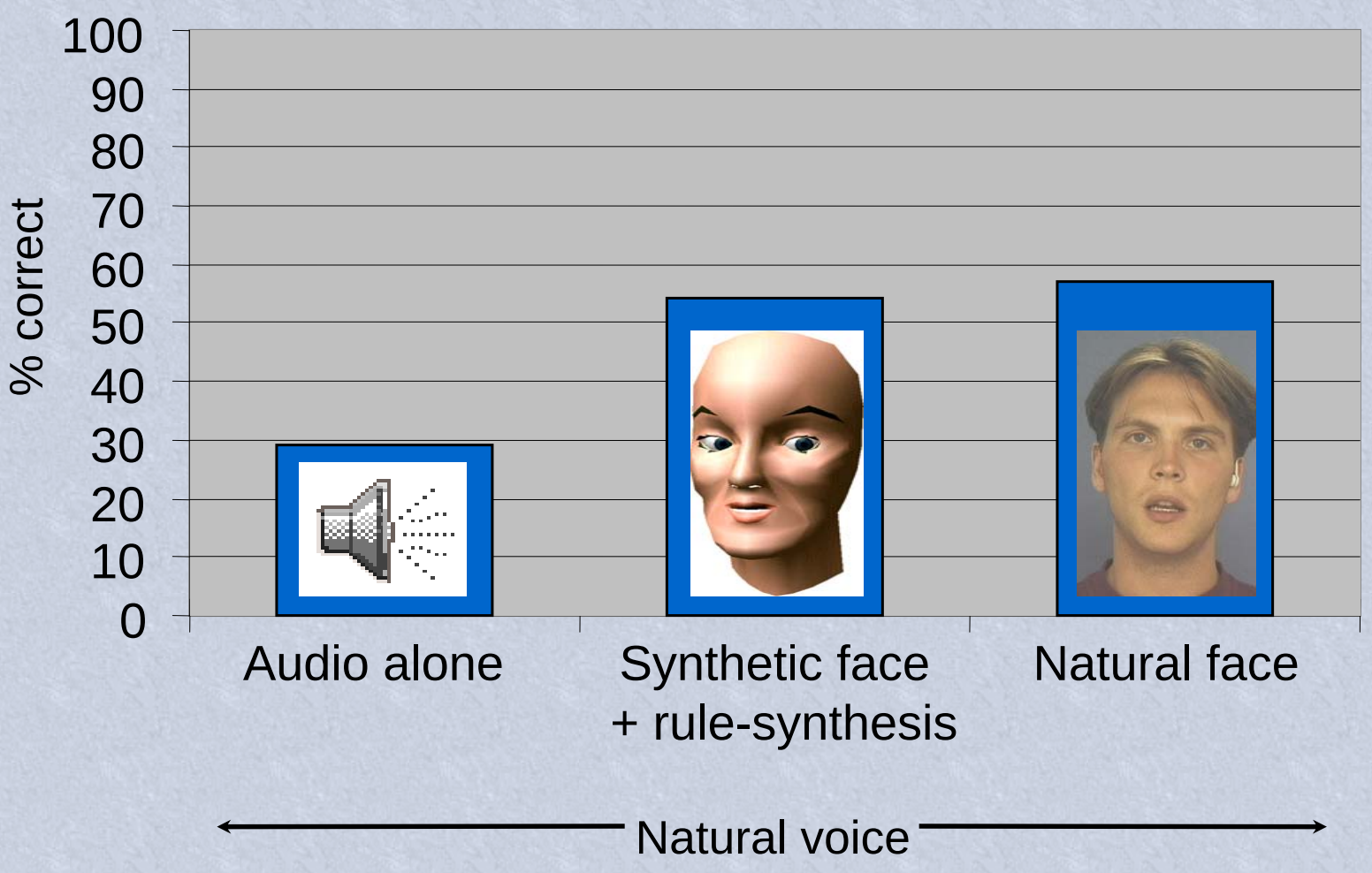
Synface - synthetic face - was developed by researchers at the Royal Institute of Technology in Stockholm, Sweden, and University College London.

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

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

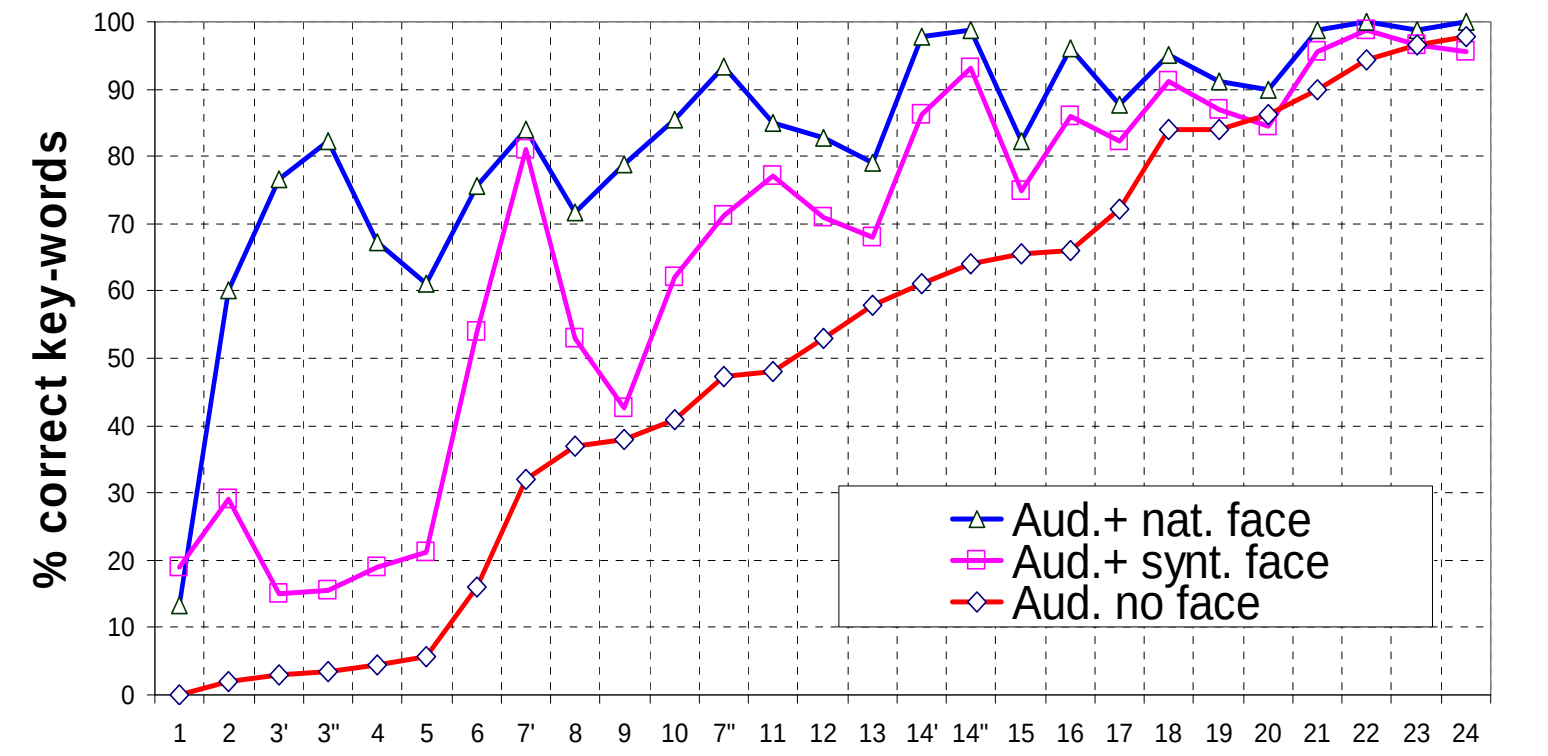
	bil	labd	den	pal	vel
bilabial	100				
labiodental		96,3	3,7		
dental		3,0	78,0	5,5	13,4
palatal			9,9	70,4	19,8
velar			4,9	16,0	79,0

Synthetic face

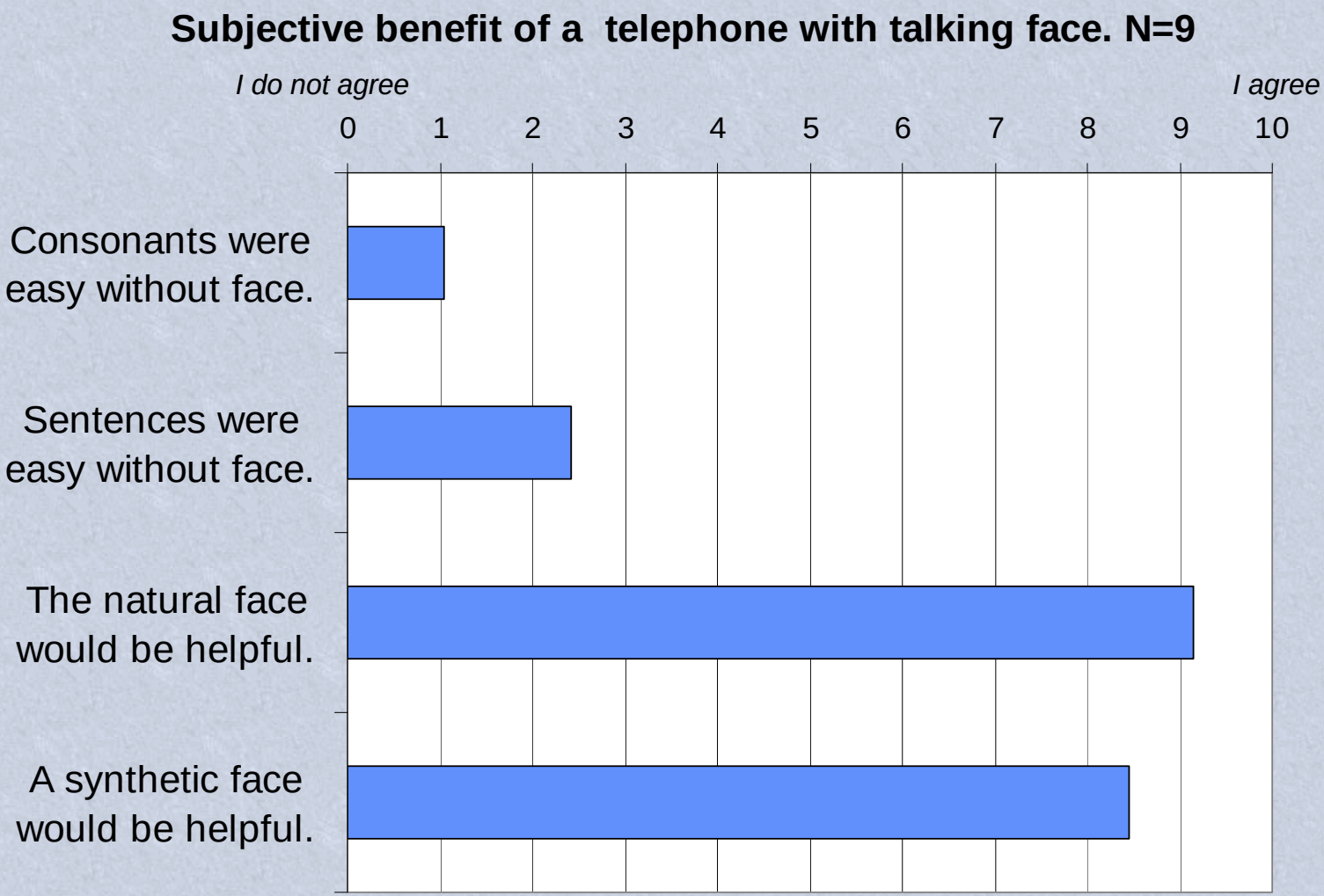
bil	labd	den	pal	vel
96,3		2,5	1,3	
	92,6	5,6	1,9	
		85,8	7,4	6,8
	1,2	17,3	71,6	9,9
		2,5	25,0	72,5

Natural face

Intelligibility tests (sentences).



Subjective benefit



Examples from other languages - trained for Swedish



English



Italian



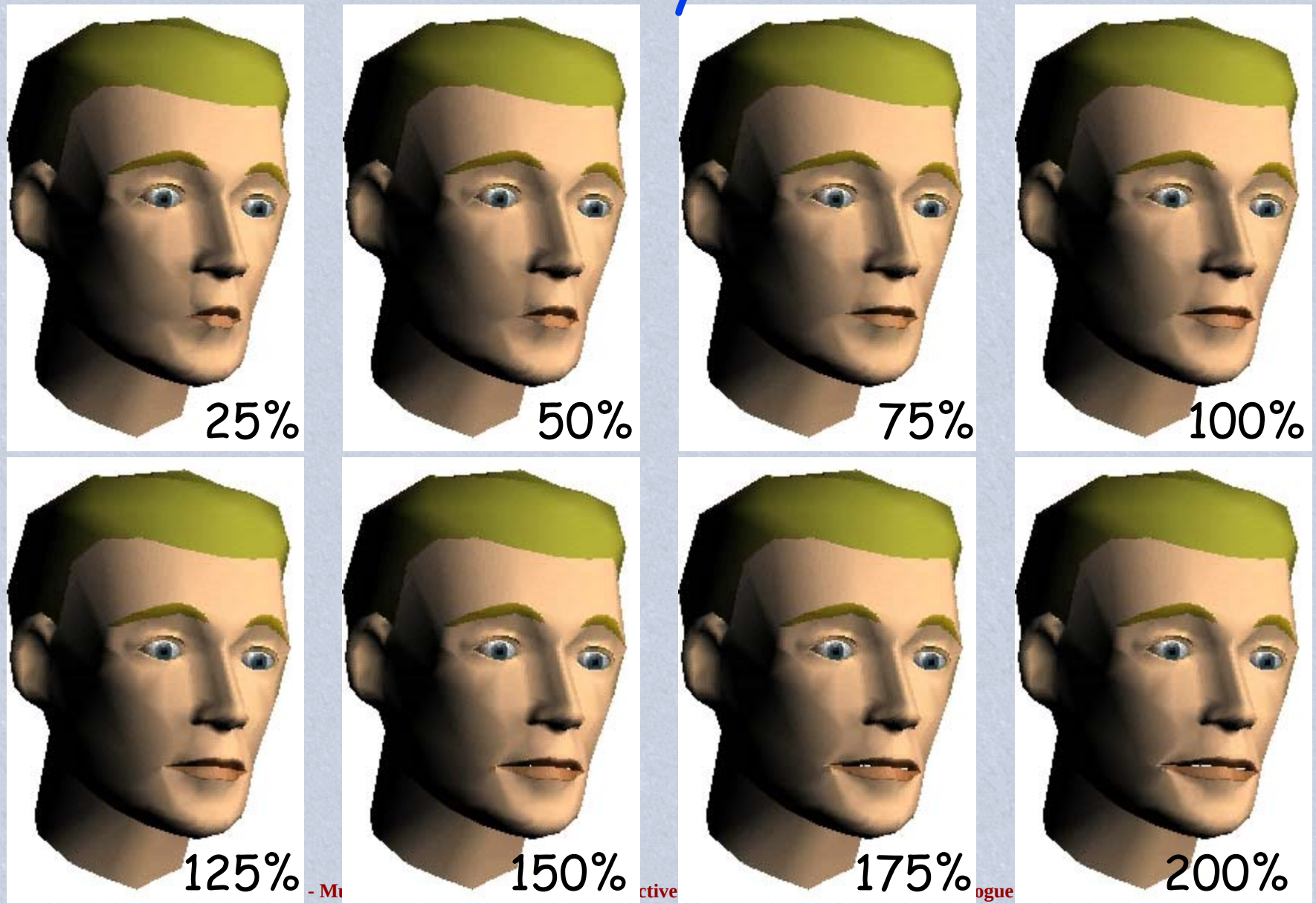
German

Shown at i3 Research Village, Comdex in Basel, 2001

Possible improvements to "lip readability"

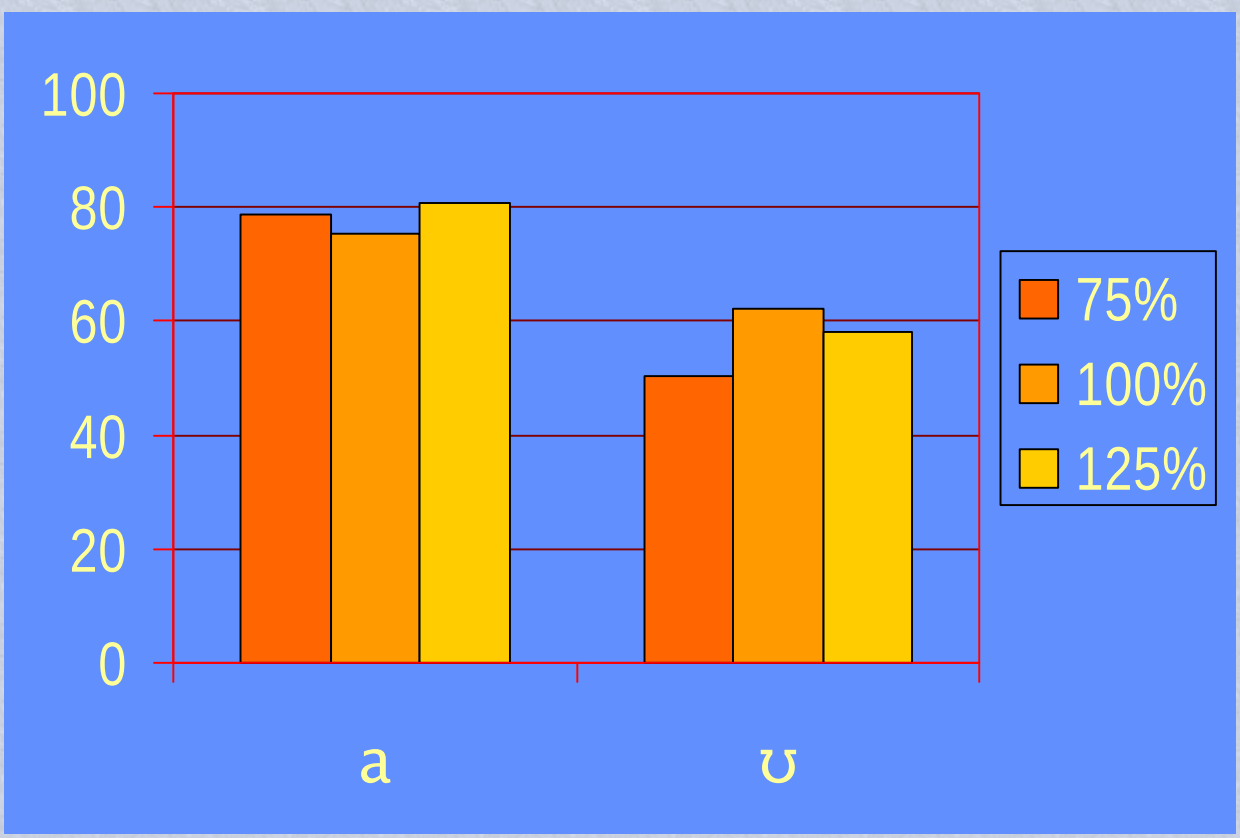
- Great variation in human speakers due to for example
  - Speaking rate
  - Extent of articulatory movements (the hypo – hyper dimension)
  - Anatomy, facial hair ….
  - Light, distance, viewing angle…

Hypo to hyper articulation
Pilot study stimuli



VCV - results

Percent correct place of articulation



Real user tests

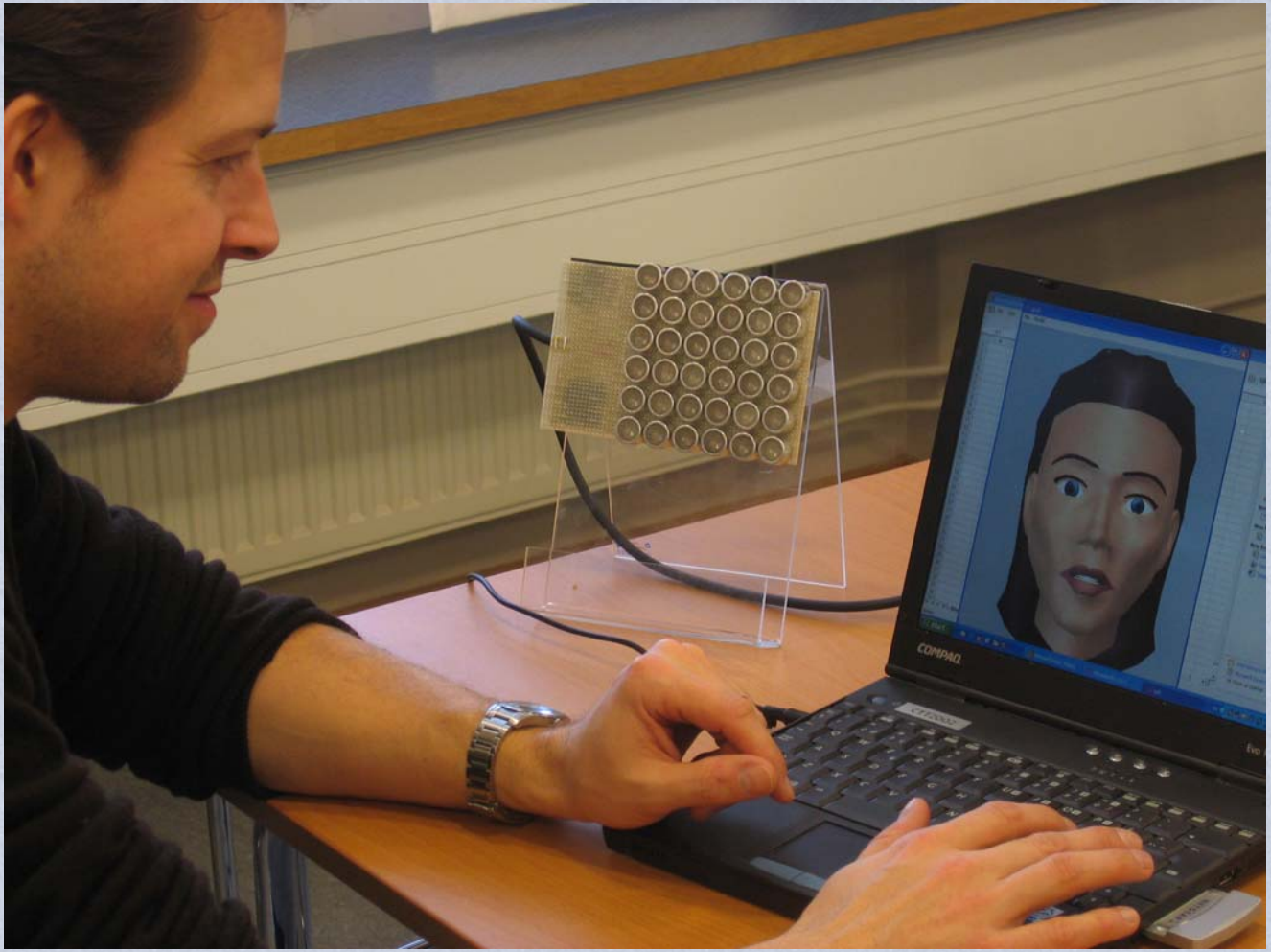


Jonas Beskow - Chester Carlson´s research award, 2 februari, 2006

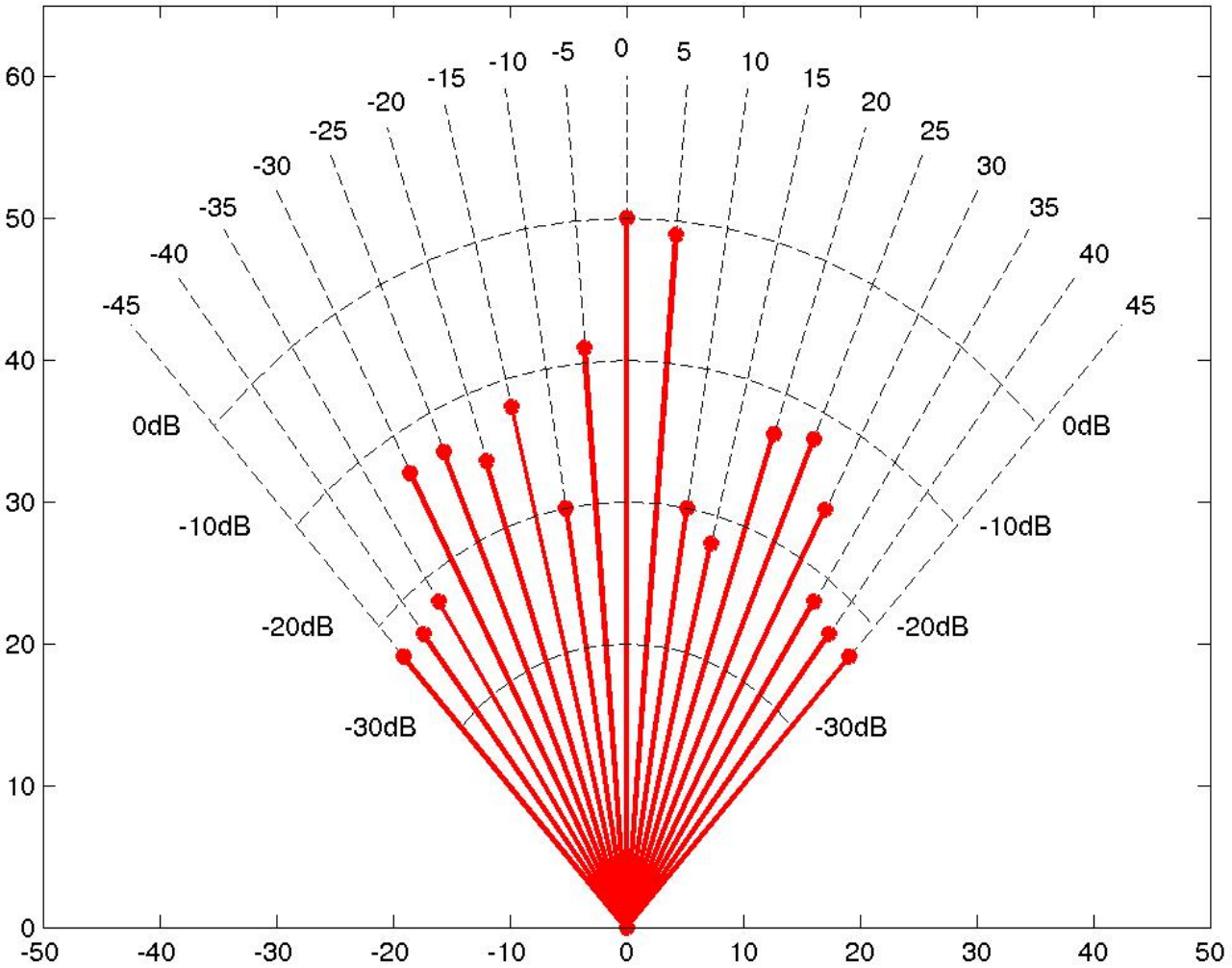


http://www.synface.com/

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



Directionality





Hearing at Home

http://www.hearing-at-home.eu/

OFFIS

HörTech
Kompetenzzentrum für Hörgeräte-Systemtechnik

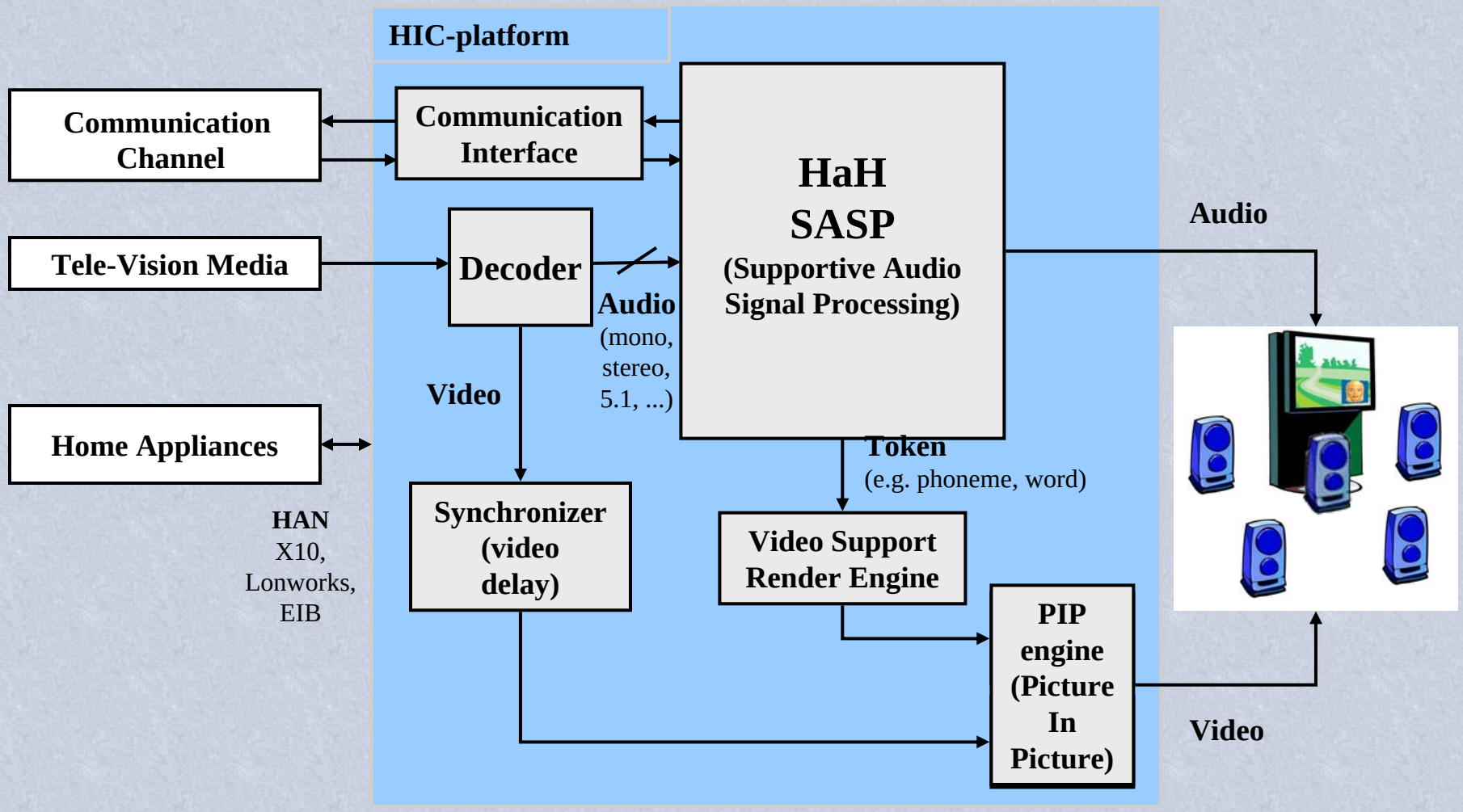
KTH VETENSKAP OCH KONST
ROYAL INSTITUTE OF TECHNOLOGY

Telefonica
TELEFÓNICA INVESTIGACIÓN Y DESARROLLO

ProSyst®

viataal

HaH System Architecture



Animated speaker for TV



Do non-verbal gestures help?

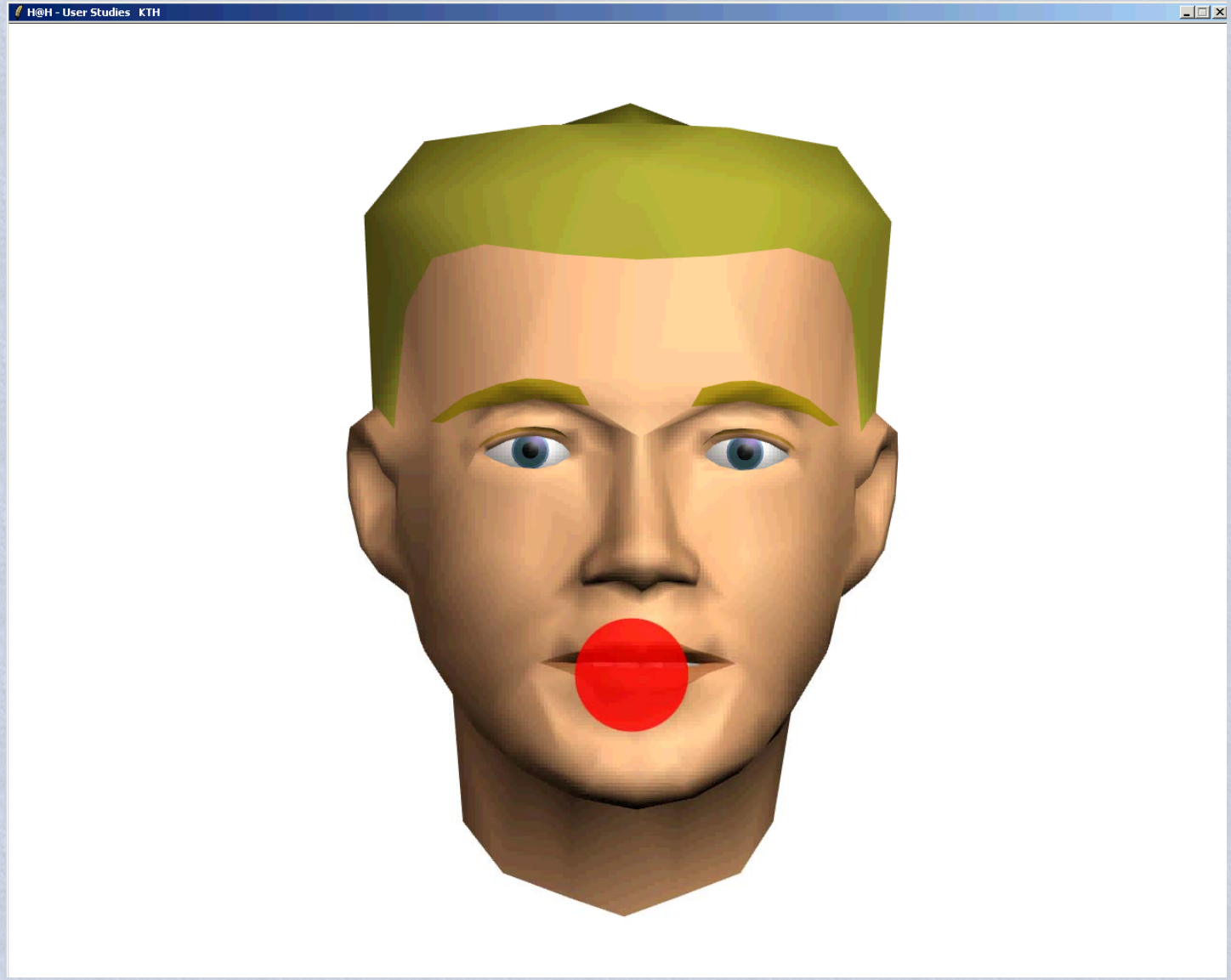
New evaluation setup: Using Eye-Tracking to study the subject gaze behavior.



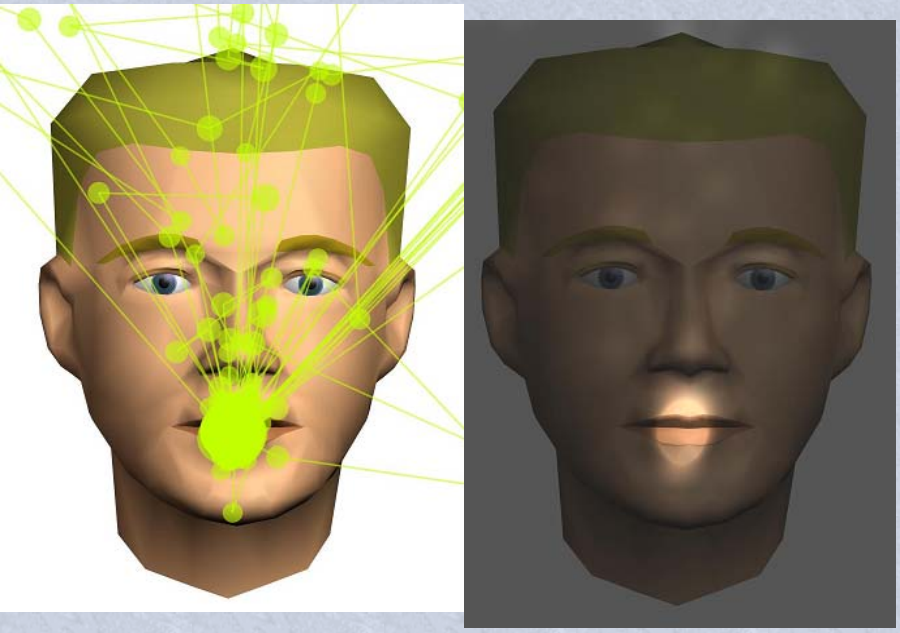
Setup

- Listening to an audio book articulated by SynFace by 10 moderately hearing impaired subjects.
- Two versions of SynFace:
  - Articulatory only
  - With randomized gestures over automatically detected prominence (head-nods, brows raise, eye-blinks.

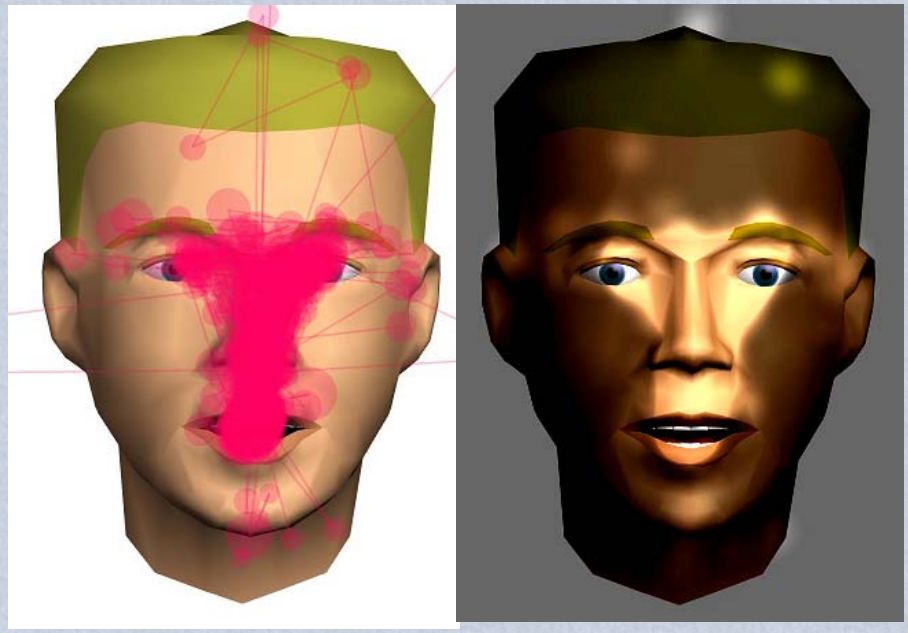
SynFace Eye-tracking sample video



Eye-tracks representation using Gaze Plots and Heat Maps



SynFace without gestures



SynFace with gestures

SynFace Eye-tracking

Results:

- There is a big variability among people in gaze behavior at a talking head.
- With articulatory only, people mainly focus only on the mouth when looking at the face.
- When adding gestures, gaze behavior at the synthetic face is closer to the gaze at a human face (alternating between eyes and mouth).
- Questionnaire show that subject believe that SynFace with gestures is significantly more natural and easier to read than SynFace without gestures.

MonAMI

Mainstreaming on Ambient Intelligence

LARGE IP – 4 years

KTH Work Package (WP8)

Innovative Interfaces

Björn Granström
Samer Al Moubayed, Jonas Beskow,
Mats Blomberg, Jens Edlund,
Joakim Gustafsson, Daniel Neiberg,
Alec Seward, Gabriel Skantze



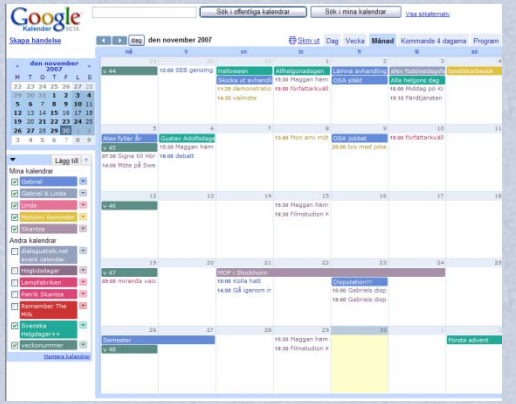
Innovative interfaces - vision

- Develop interface technology based on embodied conversational agents (ECA)
  - multimodal input: speech and other modalities
  - dialogue handling
  - multimodal output: agent animation and synthesis
  - Implement prototype system
  - Evaluation based on end user involvement

Google Calendar interfaces



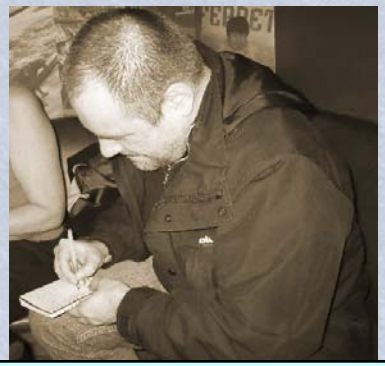




SMS notifications

Web GUI

Google Calendar



Handwriting



ECA

An example of tutoring: the MonAmi Reminder



Web GUI



SMS notifications

MonAMI Reminder

Handwriting





Talking Head



The visual attentional feedback in the MonAmi Reminder

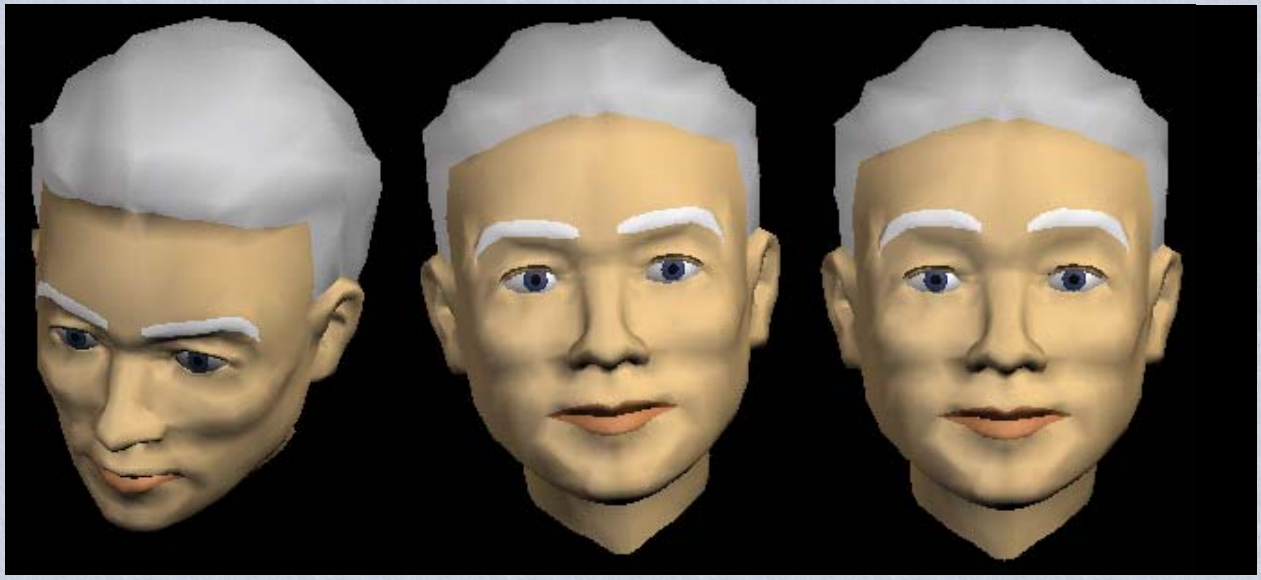


Figure 3 . Examples of facial animations triggered by the different states and events shown in Figure 2.

The system detects the user's head pose direction



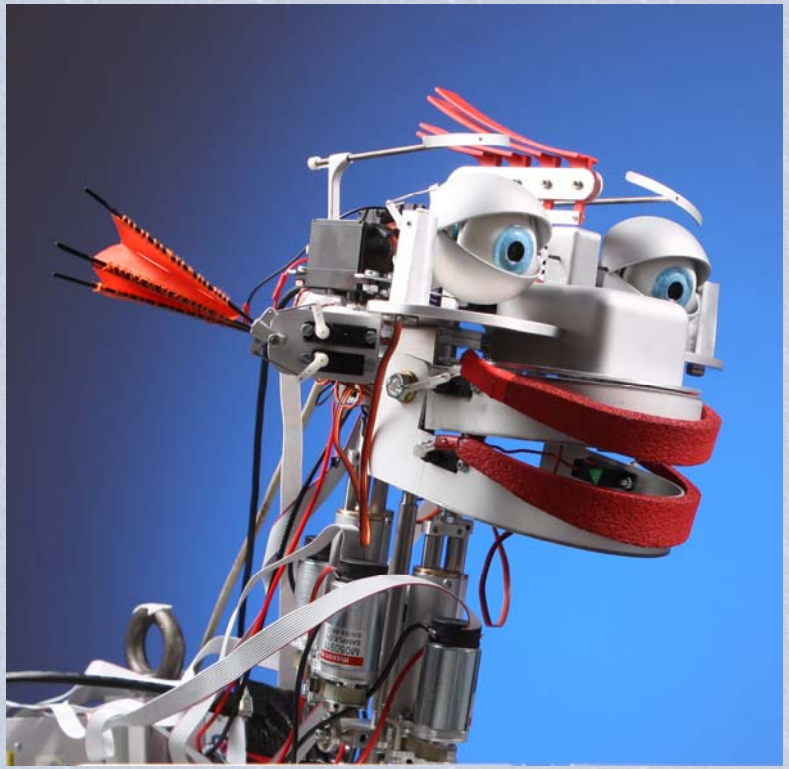
Tutoring session in



Next - the IURO project

- Not yet started
- Interactive Urban Robot
- EU/Strep – five partners: TUM, ETHZ, PLUS, KTH, ACCREA
- Robot carrying out tasks/moving in populated areas
- Information from conversations with people (in crowds)
- Using person-to-person interaction
- Natural interactions, including non-verbal etc…

Giving the robot a face



or

