

Interactive Virtual Agents

Catharine Oertel

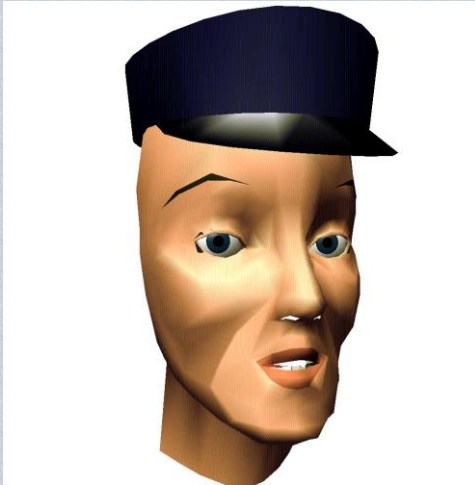
TMH, KTH

School for Computer Science and Communication



Introduction

Multimodal speech synthesis



Motivation: using talking heads to improve...

HUMAN-MACHINE
INTERACTION



HUMAN-HUMAN
INTERACTION



CONVENTIONS? - use same as
for person-to-person
communication



SONY SDR



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

Why Talking Heads?

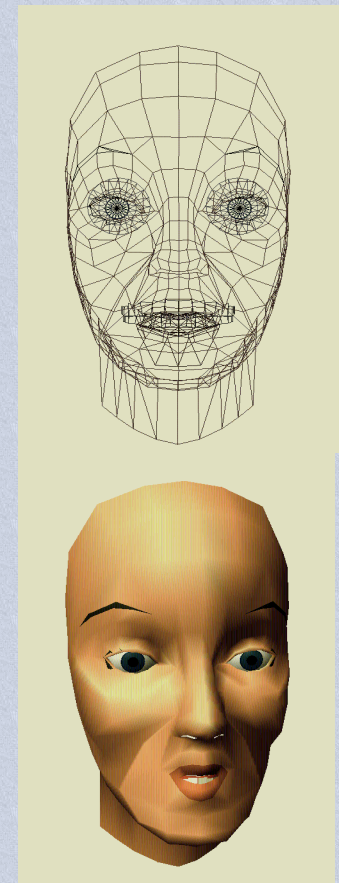
- The most natural form of communication that we know of is face-to-face interaction
- A virtual talking head can improve information transfer
 - Through expression, gaze, head movements
 - Through lip- cheek and tongue movements

Tasks of an Animated Agent

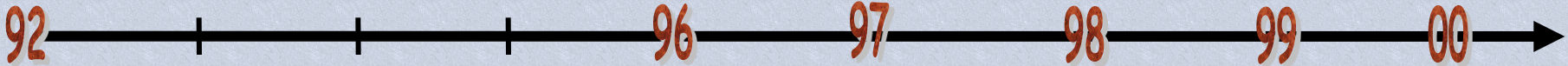
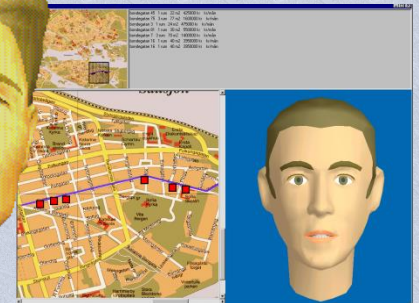
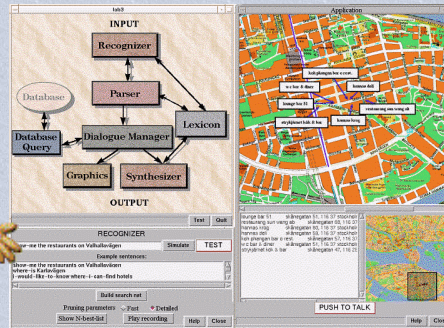
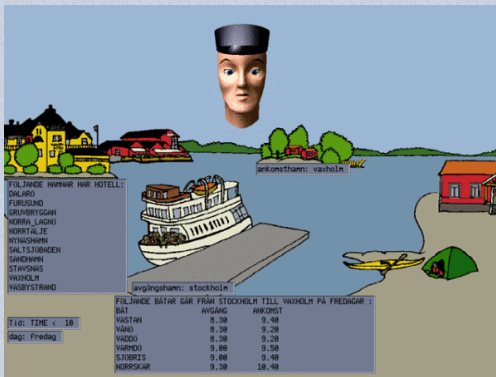
- Provide intelligible synthetic speech
- Indicate emphasis and focus in utterances
- Support turn-taking
- Give spatial references (gaze, pointing etc)
- Provide non-verbal back-channeling
- Indicate the system's internal state

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.



Dialog systems at KTH



Waxholm

Olga

GULAN

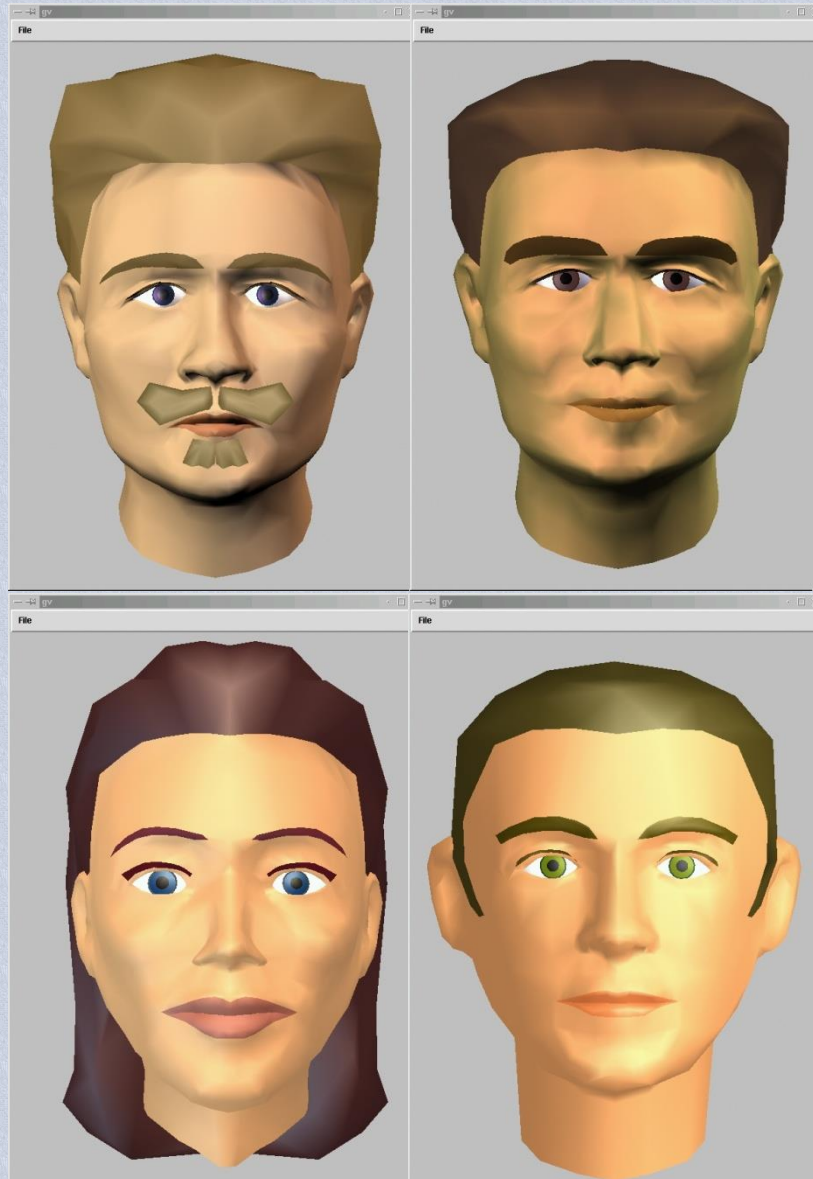
AUGUST

Adapt

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



Different characters



Facial animation techniques

Techniques for facial animation

- Talking head synthesis requires:
 - *A signal model*
 - *A control model*

The signal model

- Video-based (2D)
 - Enables realistic reproduction
 - Lacking flexibility
- Muscle based (3D)
 - Highly flexible
 - Difficult to gather anatomical data
 - Computationally intensive (although less of a problem today...)
- Direct parametrisation/morphing based (3D)
 - Good compromise between flexibility and realism
 - Simple data acquisition using optical methods

Direct parameterisation

- 3D-modelling
- Deformation through high-level parameters
- Different possible parameterisations:
 - Articulatory oriented
 - MPEG-4 (low-level)

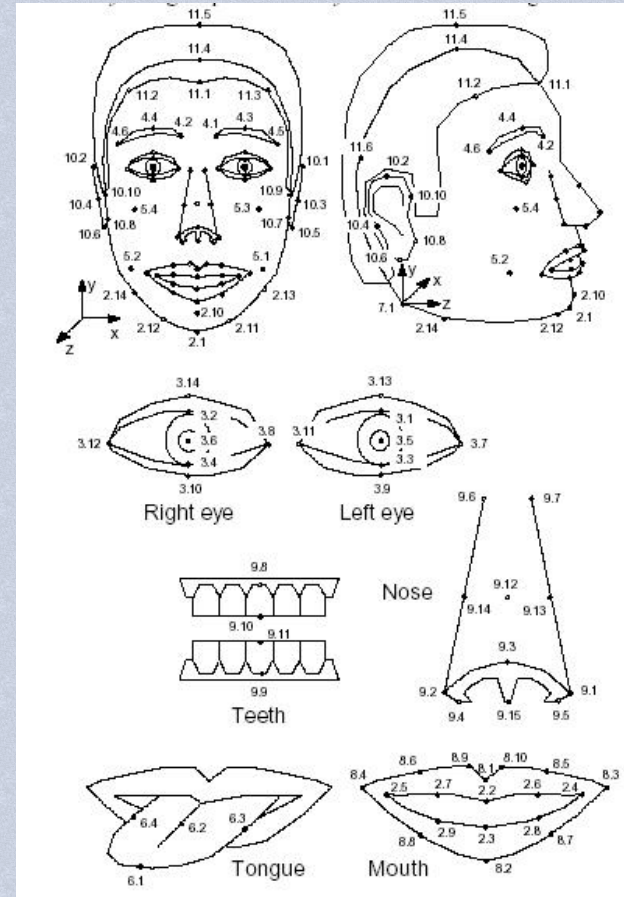


Articulatory oriented direct parameterisation

- High-level parameterisation tailored for visual speech animation
- Parameter set includes
 - Jaw opening
 - Lip rounding
 - Bilabial closure
 - Labiodental closure
- Parameters are normalized relative to spatial targets

MPEG-4 direct parameterisation

- Original purpose: model based video coding
- The standard defines a generic face object
- 84 feature points (FPs)
- 68 facial animation parameters (FAPs)
- FAPs are normalized relative to distances in the face
- Expressed in FAPU (FAP unit)

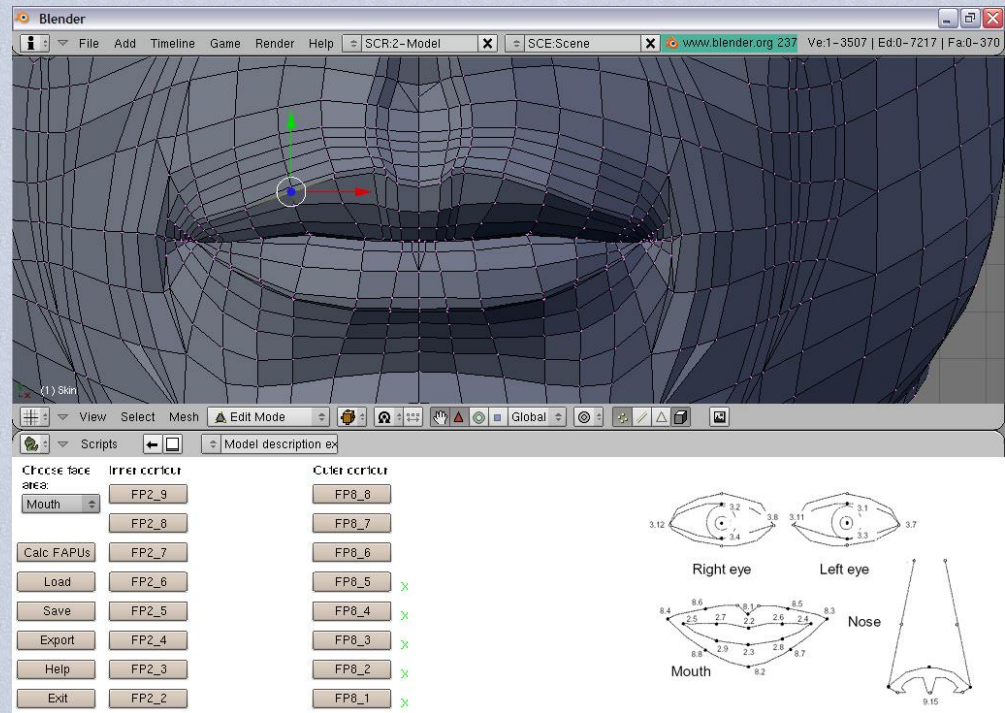


MPEG-4 FA

- Advantages:
 - A standard
- Disadvantages
 - Difficult to know the perceived result of a certain expression or articulation
 - Difficult to control manually (ex: lip rounding involves manipulation of 20 FAPs...)

Automatic creation of MPEG-4 models

- Input: static face model
- Annotation of FPs
- Calculation of deformation weights
- Output: animatable MPEG-4-modell

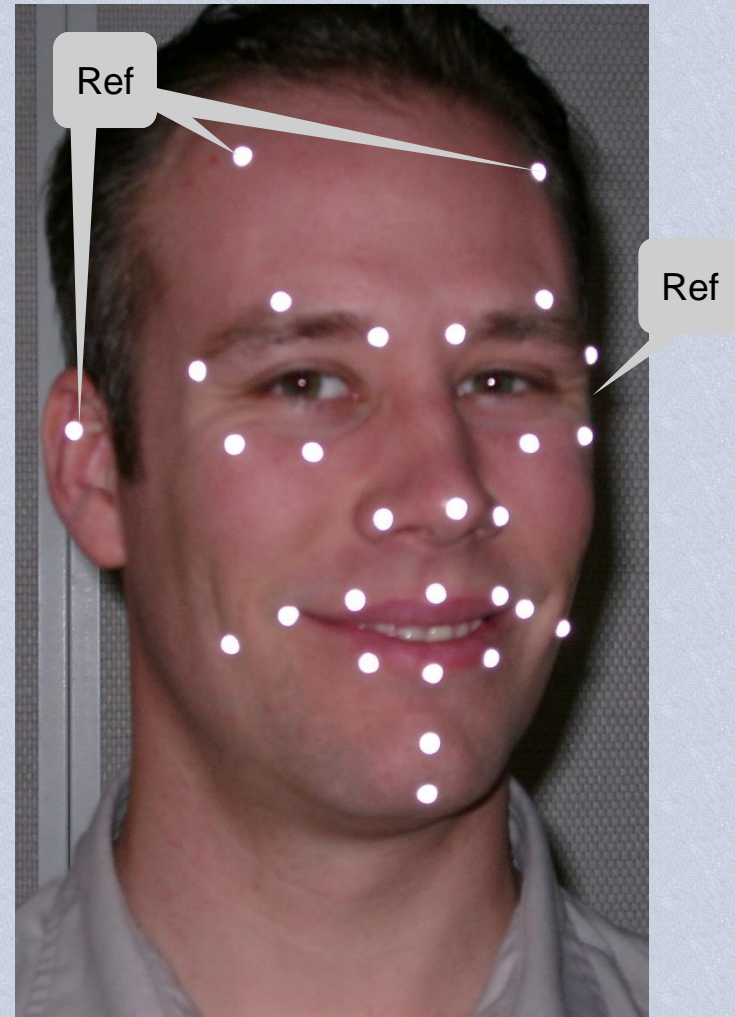


Measuring talking faces

- Video analysis
 - Can provide texture, contours and points
 - Inexpensive hardware
 - Typically less accurate than motion capture
- Cyberware scanning
 - High definition 3D shape and texture
 - Static
 - Expensive equipment
- Motion capture
 - Captures points in 3D
 - Dynamic
 - Starting to become affordable

Data recording for talking head animation

- 3D motion capture
- Marker placement corresponding to MPEG-4 FAPs
- 4 reference markers capture rigid head movement
- 25 markers capture facial deformation



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



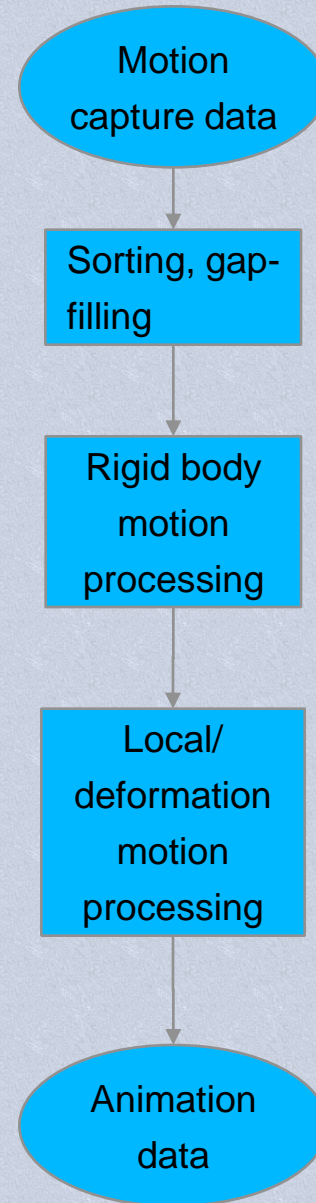
How it works

- 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



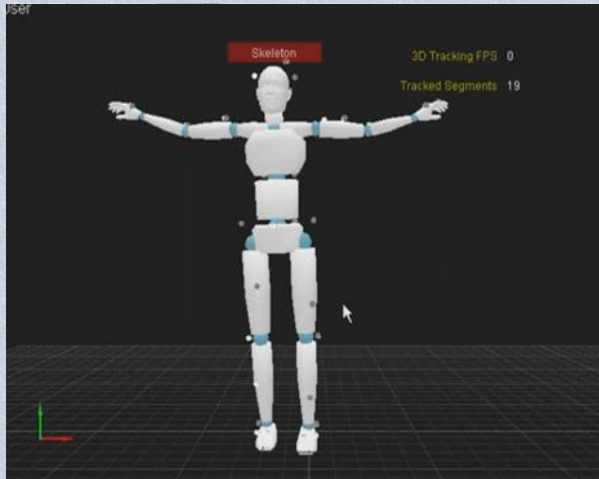
Data processing

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

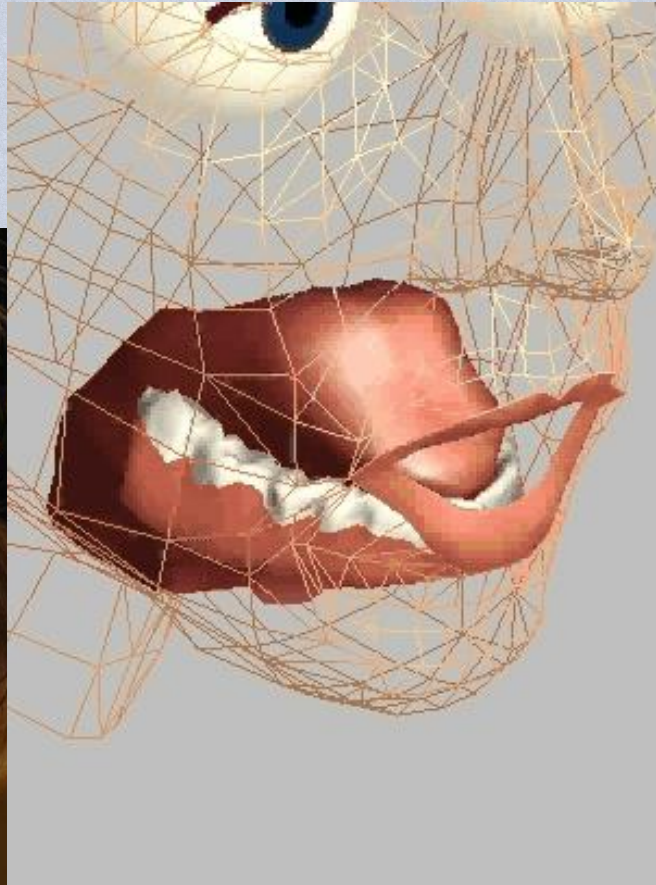
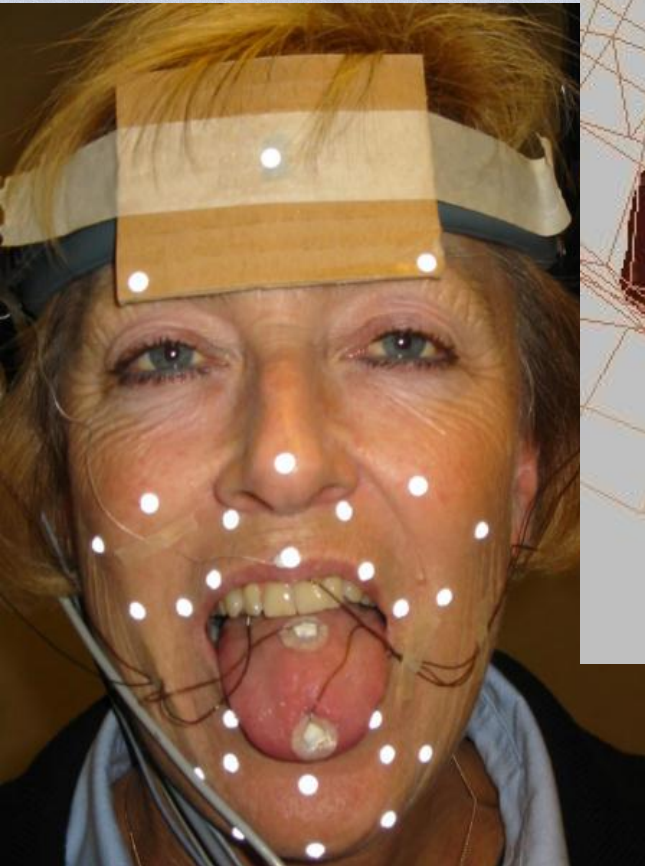


Combining model and data

Re-synthesis
using speech
movement
recorded
with Qualisys



Combining several motion capture techniques

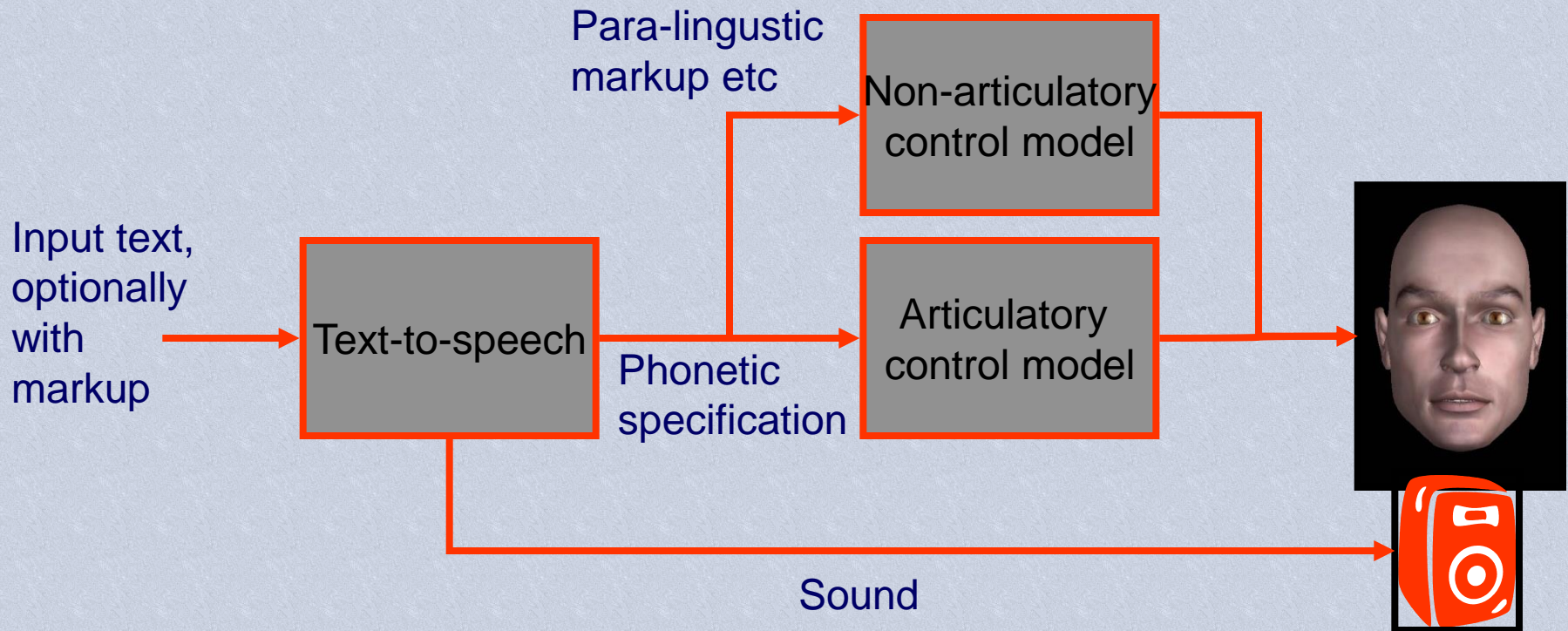


EMA & Qualisys

Control models for talking heads

- A control is what makes that talking head move
- Input is typically time stamped events, e.g. phonemes, gestures symbolic, or an audio signal
- Output is control parameters for the signal model
- May work in real-time or off-line

Text-to-animation



Articulatory control models for visual synthesis

- Concatenation of
 - Diphones
 - Non-uniform units
- Co-articulation model
 - Rule-based
 - Trainable models (Cohen/Massaro, Öhman)
- General machine learning methods
 - Hidden markov models
 - Artificial neural networks

Experiment: which control model should we use?

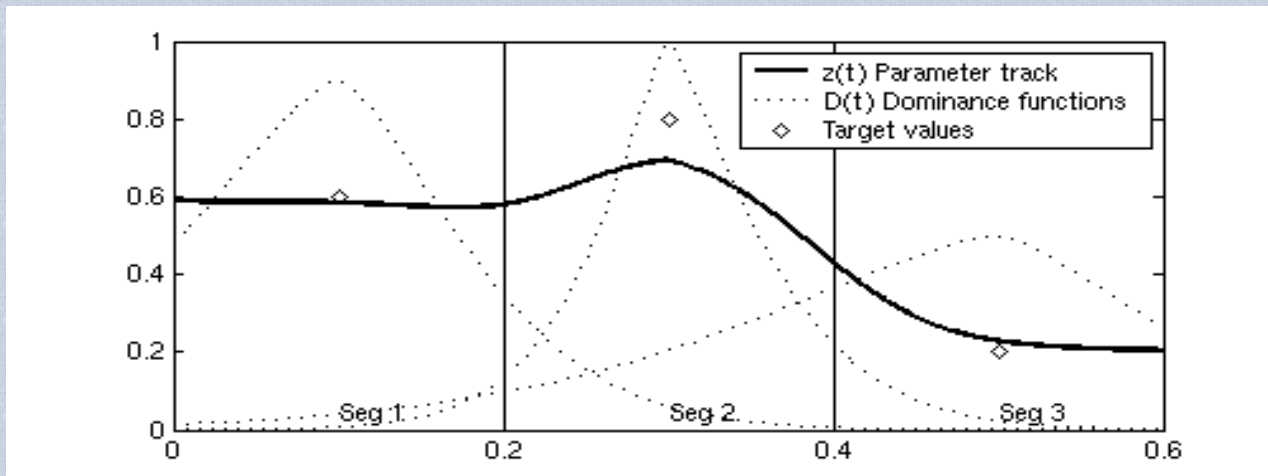
- Head-to-head comparison of 4 models:
 - KTH Rule-based (hand-crafted)
 - Cohen massaro (trained)
 - Öhman (trained)
 - ANN (trained)

KTH Rule-based model

- Each target is either assigned a value or is left undefined
- Undefined values are interpolated from context
- Example:
 - /r/ has undefined lip rounding, will be inferred from context

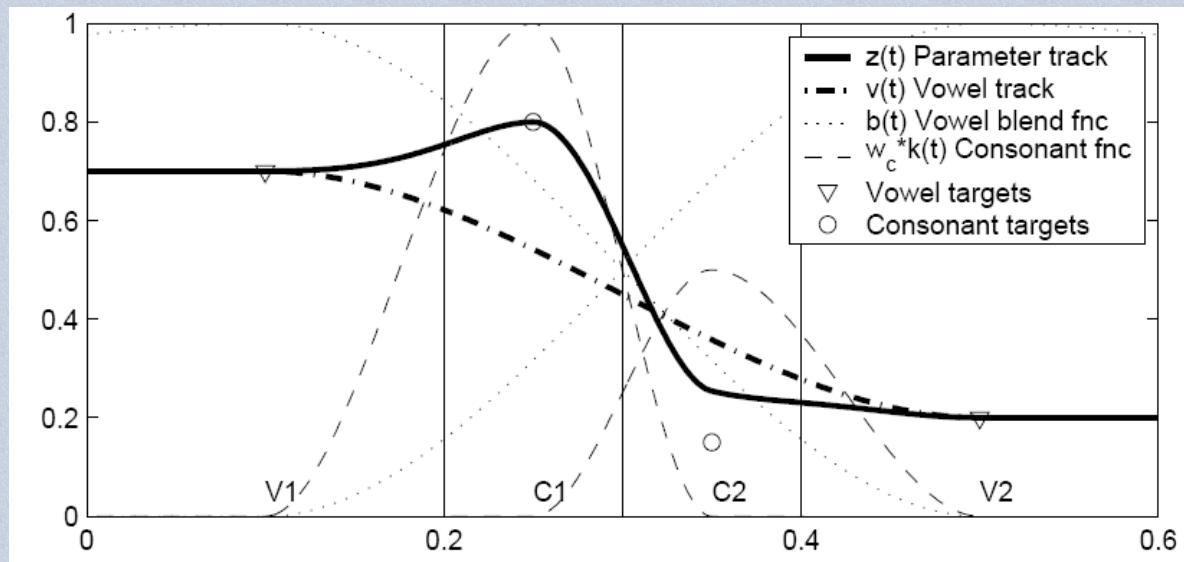
Cohen/Massaro model

- Each segment has
 - A target value
 - An exponentially decaying dominance function
- Parameter trajectory is given as a weighted sum of targets and dominance functions
- 3800 free parameters to be trained using error minimisation



Öhman's model

- A vowel track $v(t)$ is formed by interpolation between successive vowels
- Consonant movements are superimposed/blended into the vowel track using a co-articulation factor
- 1040 free parameters to be trained using error minimisation

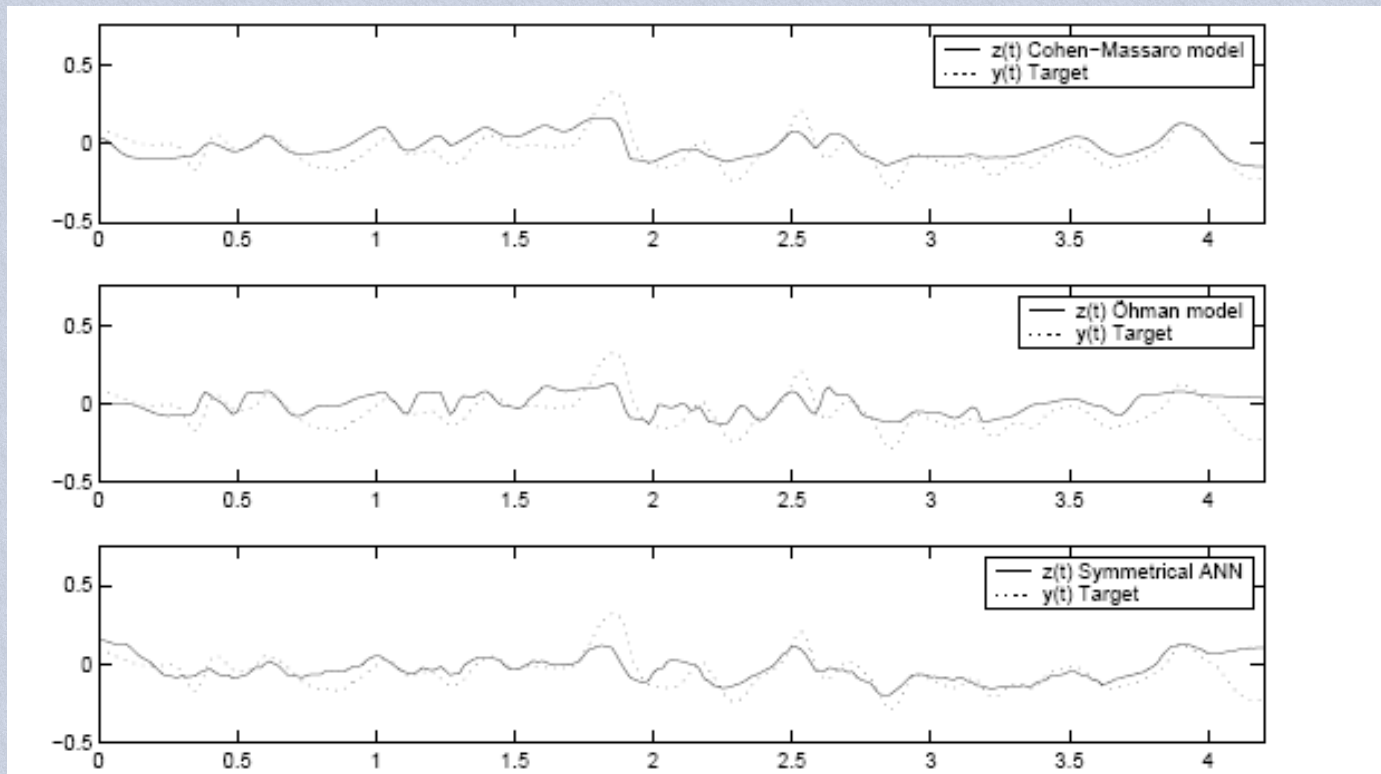


ANN-model

- An artificial neural network may be used to predict animation parameters
- Input vectors with 17 prototypical phonetic features specifying place and manner of articulation, phoneme class etc.
- 250 ms forward and backward context
- Recurrent connections in hidden layer
- 123870 free parameters trained using back propagation

Model training

Training set of 87 phonetically balanced sentences
Training aims to minimize error between measurement and prediction



Intelligibility comparison

- All 4 models were used to synthesize short sentences
- Intelligibility in noise with/without talking head was measured

	Audio-visual condition				Rule-based
	Audio only	Cohen-Massaro	Öhman	ANN 1	
Keywords correct (%)	62,7	74,8	75,3	72,1	81,1

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.

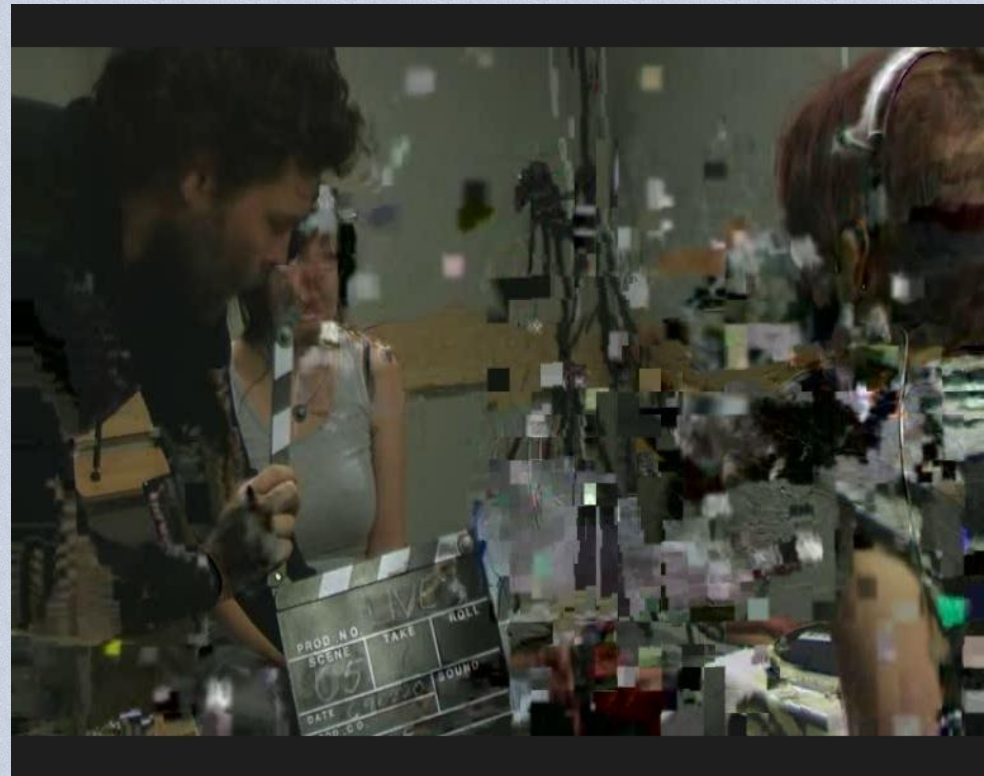


Transcription and annotation

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

The spontal recordings

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



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>

Conversation with agent



Face Robot - a commercial facial animation system



Emotional animation

Emotional/ubiquitous computing – do we want it?



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

Basic emotions



Happiness



Anger



Surprise



Disgust

Vision from audio

original



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

mimic
from audio

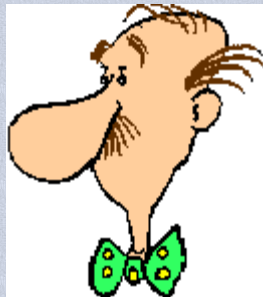


Abb. 2. a) Der ung. Satz *Tiz* in gehässigem Ton; mit zusammengebißnen Zähnen gesprochen.
b) Derselbe Satz, wiederholt von der zweiten Schauspielerin.

Fónagy, 1967 "Hörbare Mimik", *Phonetica*



Emotions

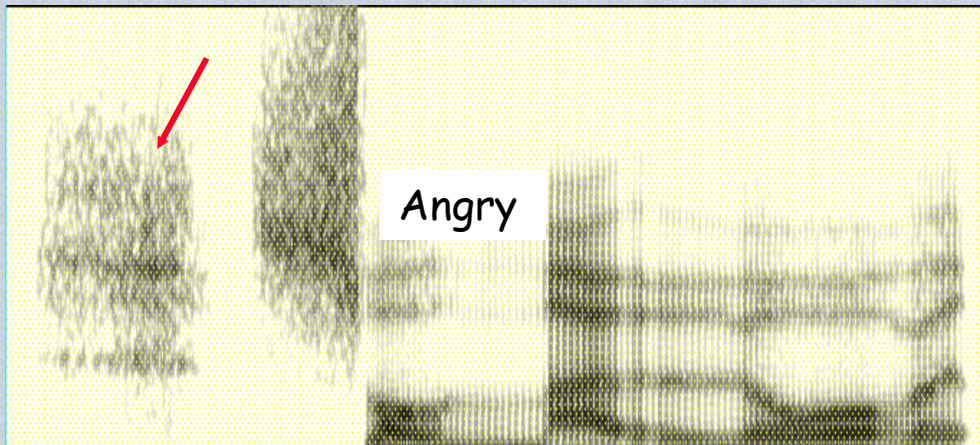
natural 
synthesis 






Neutral
Happy

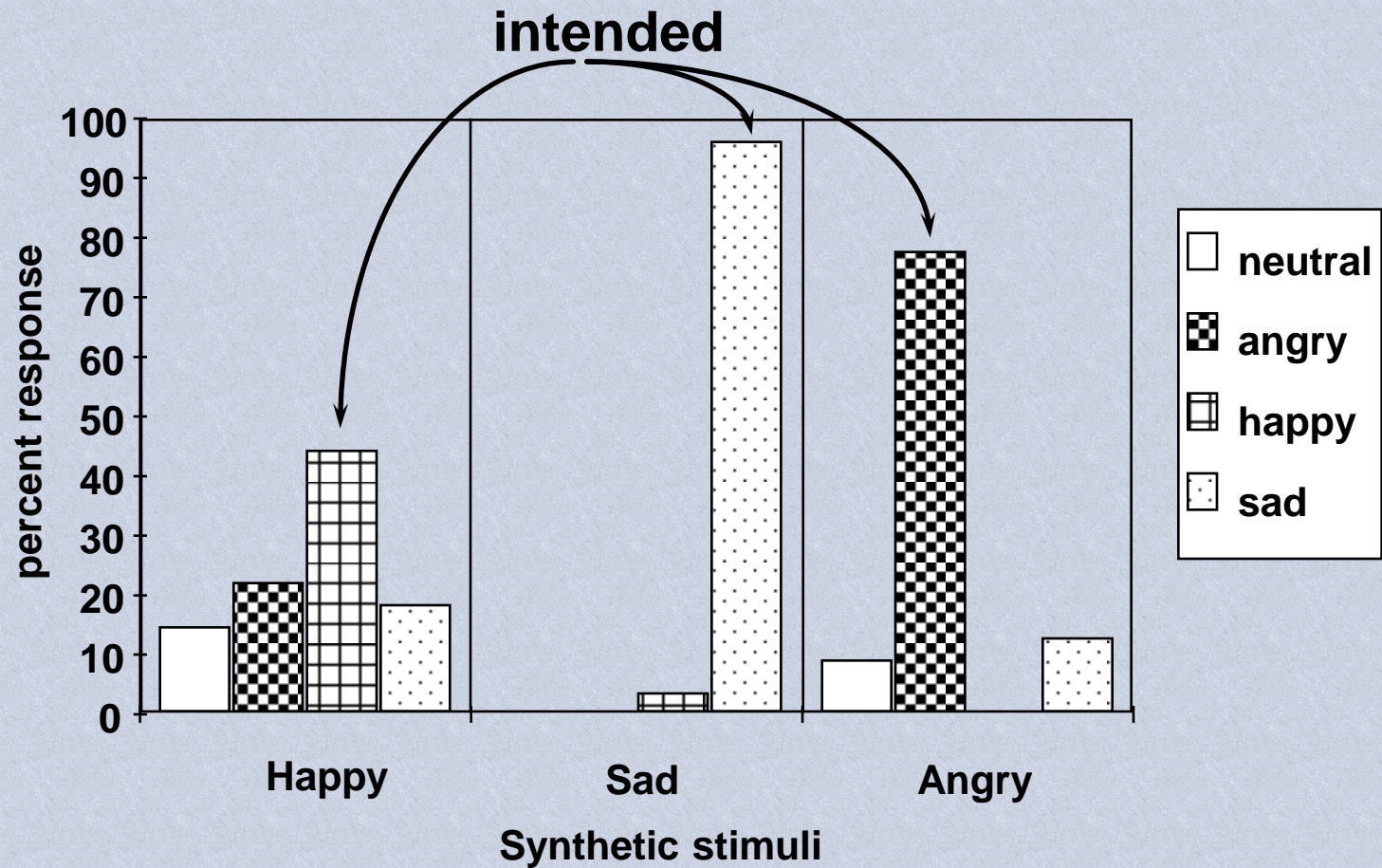


natural 
synthesis 



natural 
synthesis  

Result from listening test





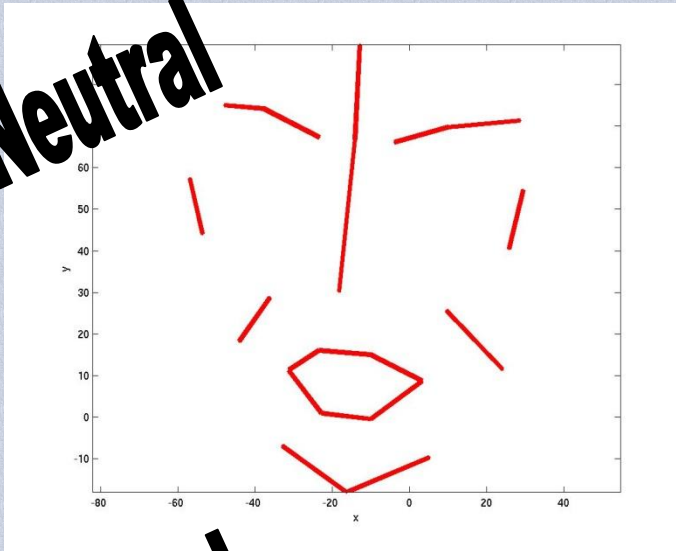
Preparing future multisensorial
interaction research

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

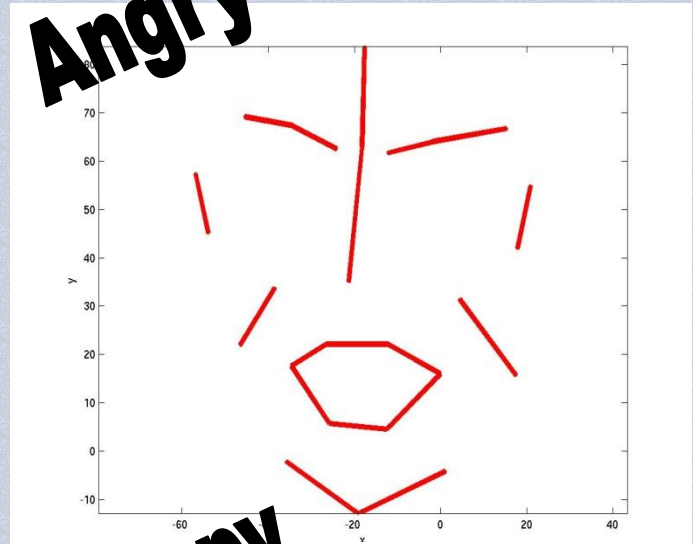
EU project: start October 2002, duration 2 YR
ITC-IRST (Trento) co-ordinates + 3*Germany
+ Italy + UK + Sweden

<http://pfstar.itc.it/>

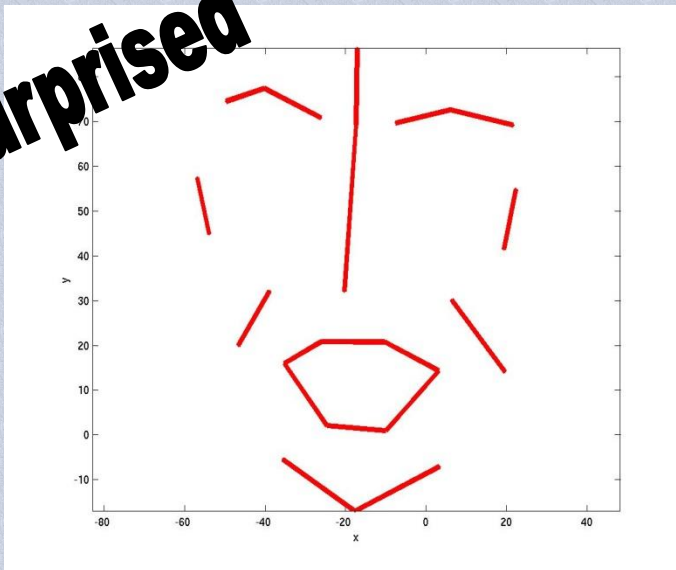
Neutral



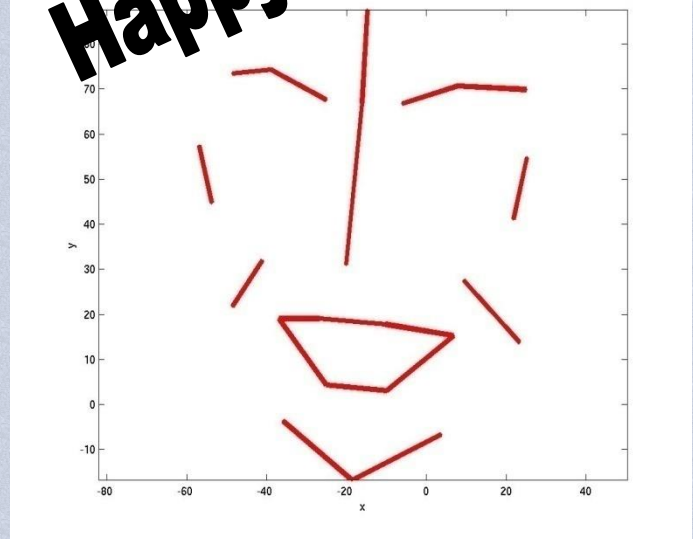
Angry



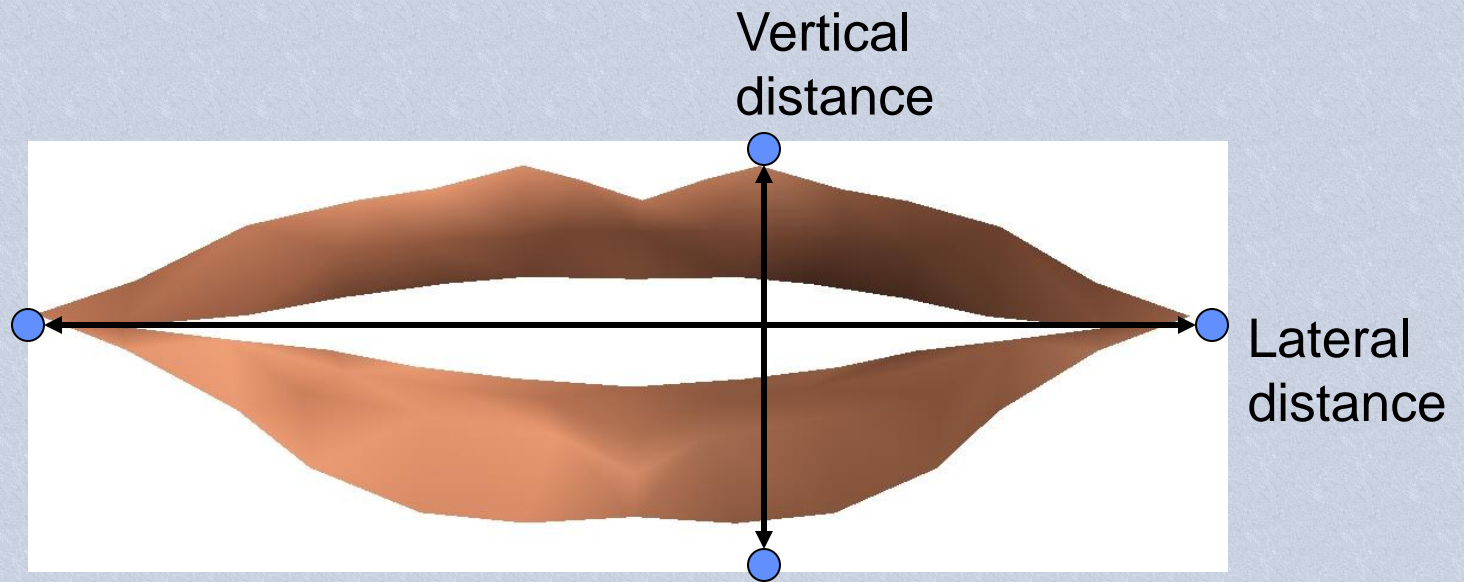
Surprised



Happy

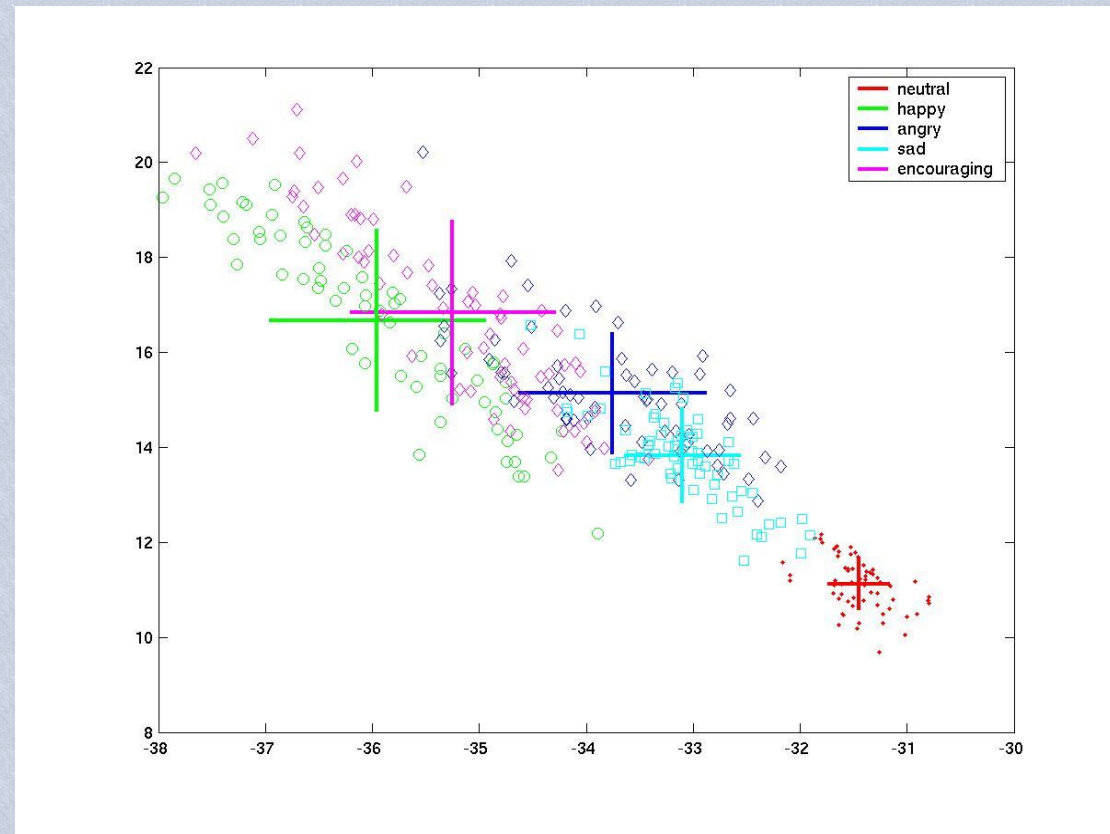


Measurement points for lip coarticulation analysis



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)



Datadriven facial synthesis with MPEG4 model



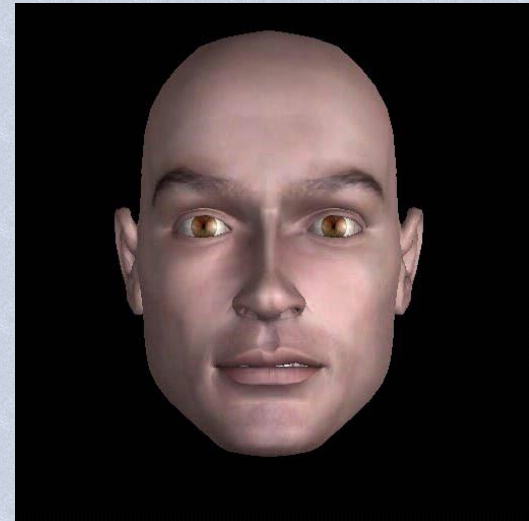
Happy



Angry



Sad



Surprised

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)

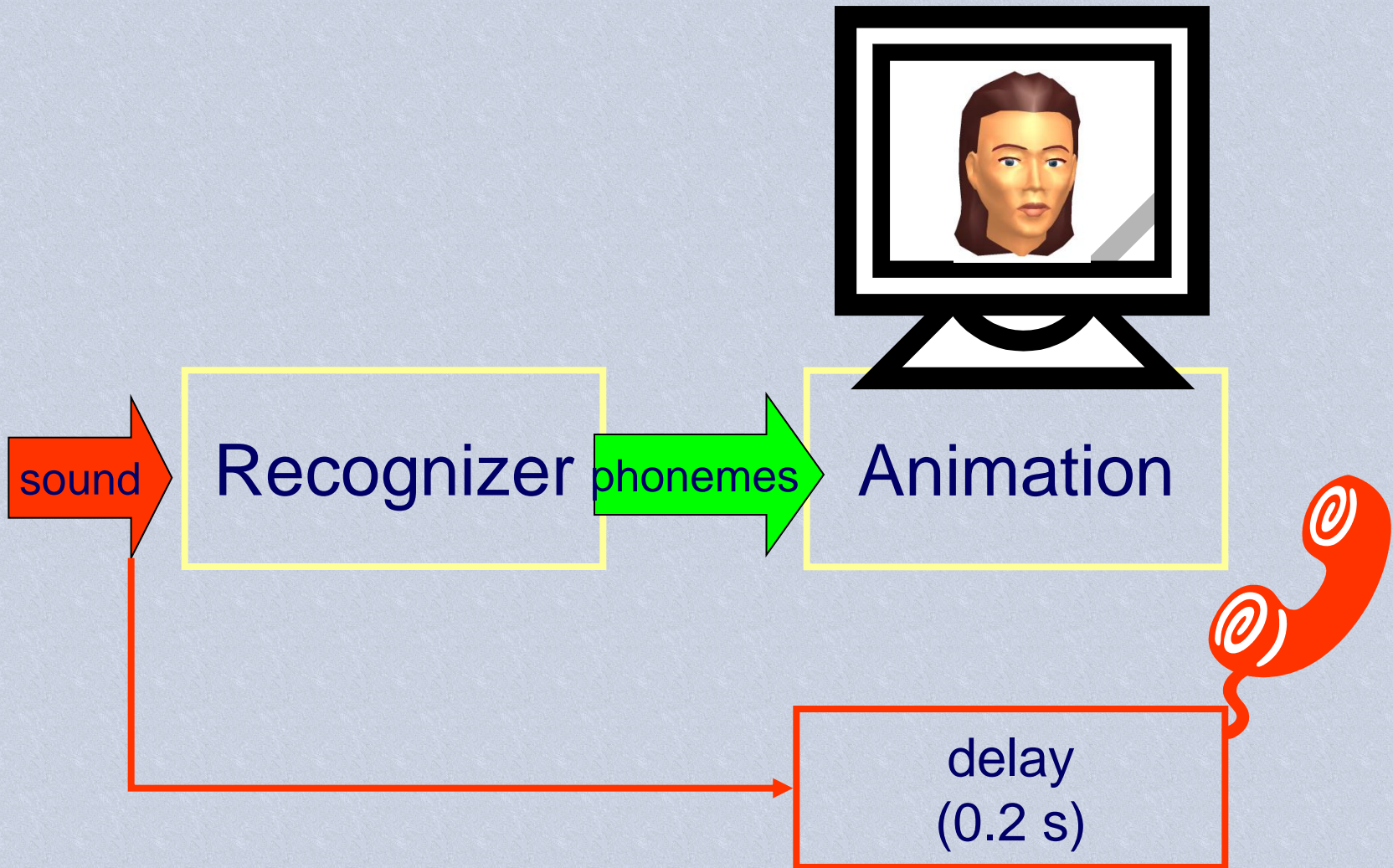
Talking heads as assistive technology

Talking heads as a visual hearing aid

- 2% of the population are severely hearing impaired
- They rely on lip reading
- Telephone conversations are difficult



SYNFACE



SYNFACE

- No need for special equipment at other end
 - Compare to video telephony, text telephony
- Only interprets sounds, not words
 - Single mis-interpretations often OK.
 - Graceful degradation

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.

A photograph showing a woman in profile on the left, holding a white mobile phone to her ear. In front of her is a desk with a laptop computer. The laptop screen displays a synthetic human face with a neutral expression. To the right of the laptop is a white corded telephone and some papers. The background is a simple office setting with a lamp and a wall.

Phone conversations are animated

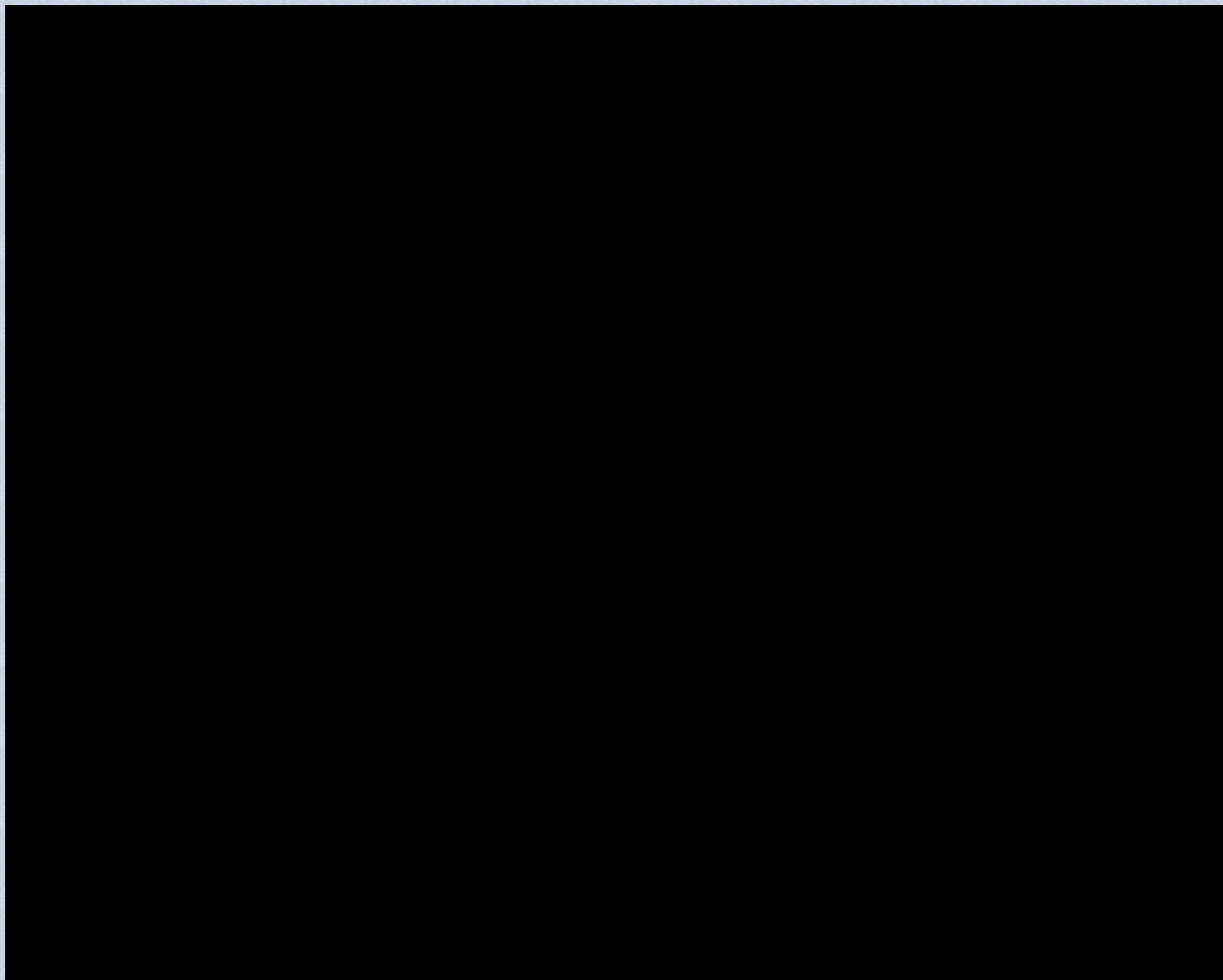
The software can be installed on a regular computer and is used with a standard telephone.

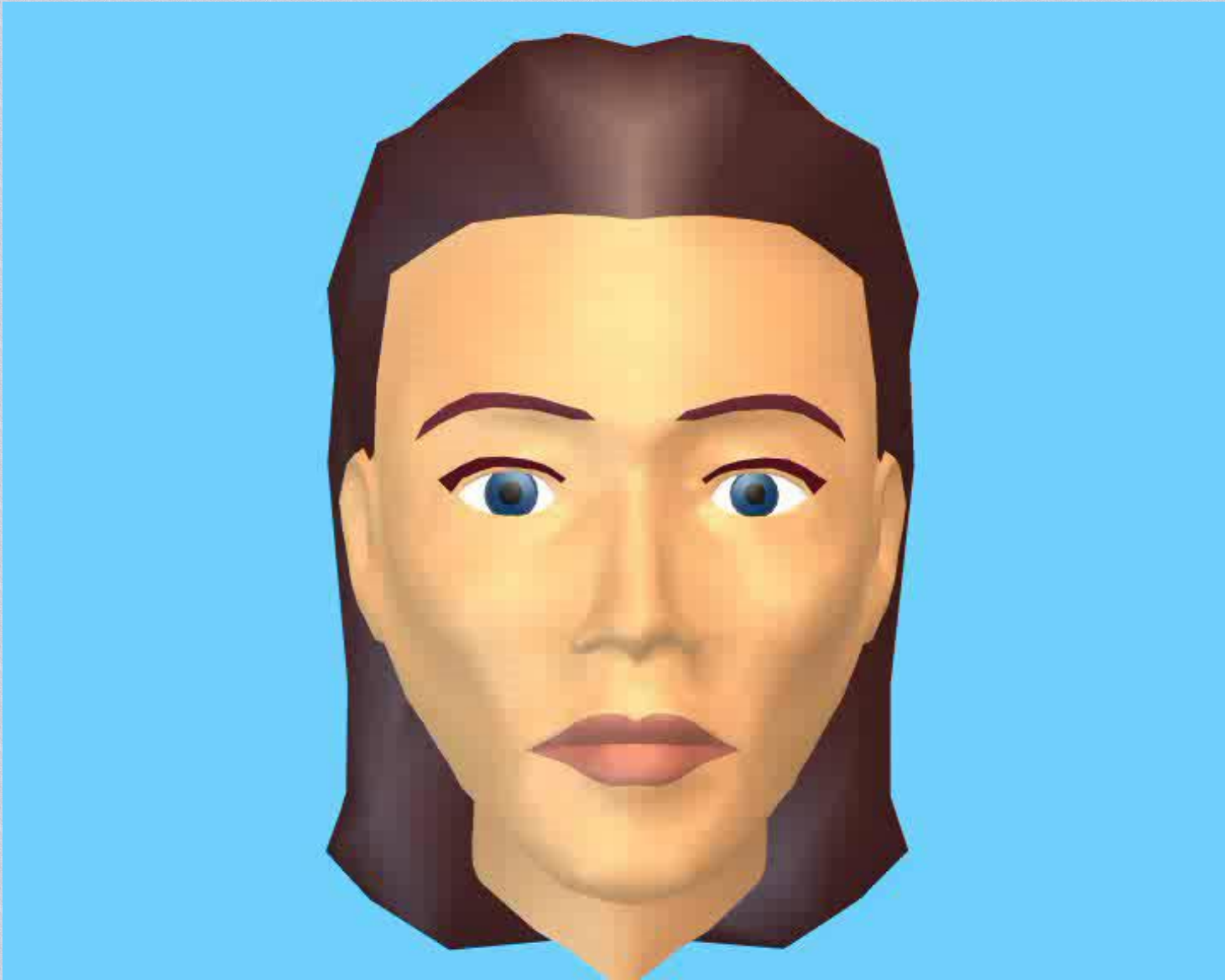
European scientists will use the trial results to tweak the device before it is made available in coming years.

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/>

Demonstration video from EU

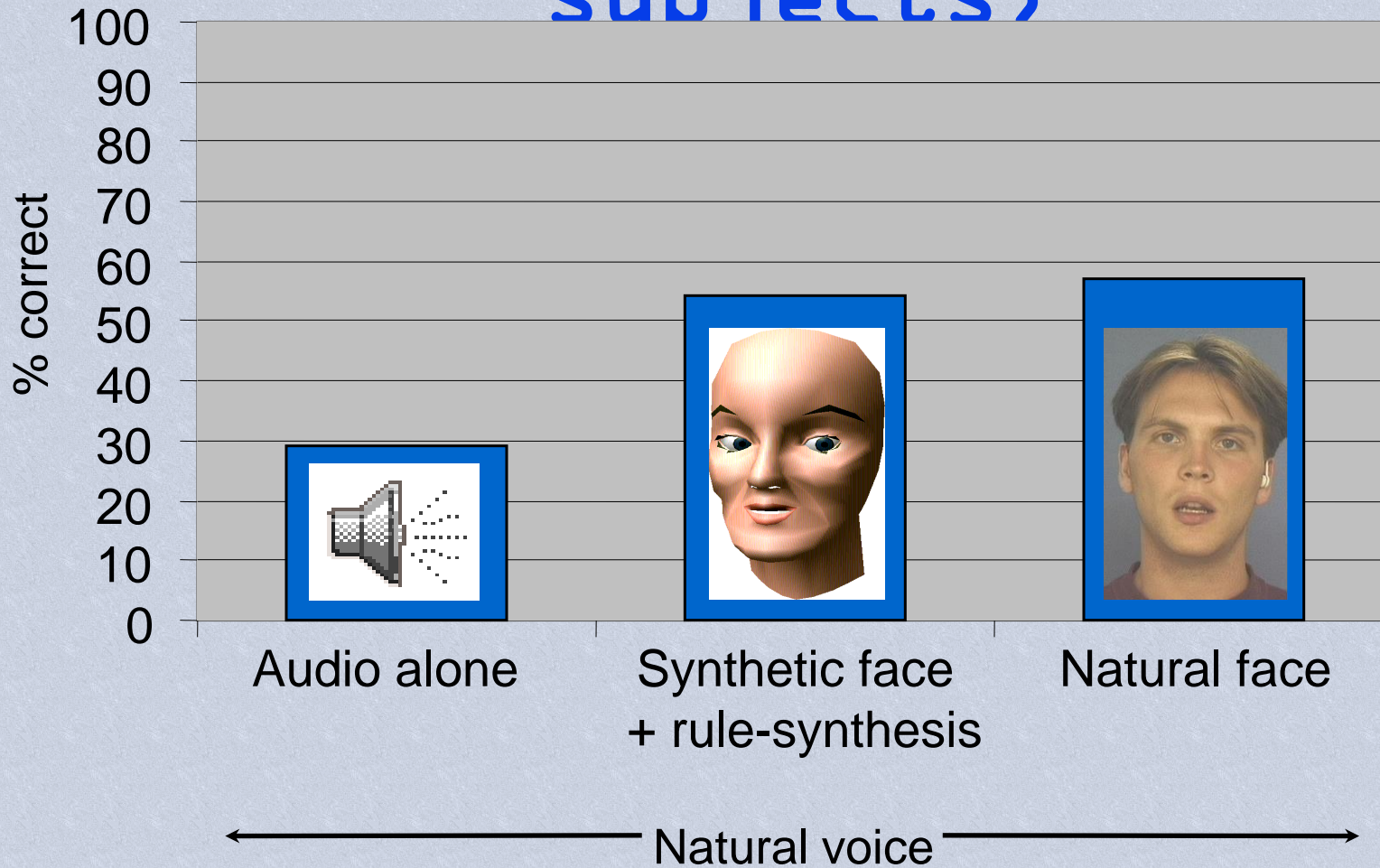




Formal intelligibility test

- Material: VCV (symmetric vowel context)
 - 2 vowels: /**u**, **a**/
 - 17 consonants:
/p, b, m, f, v, t, d, n, s, l, r, k, g, ŋ, ʃ, 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"

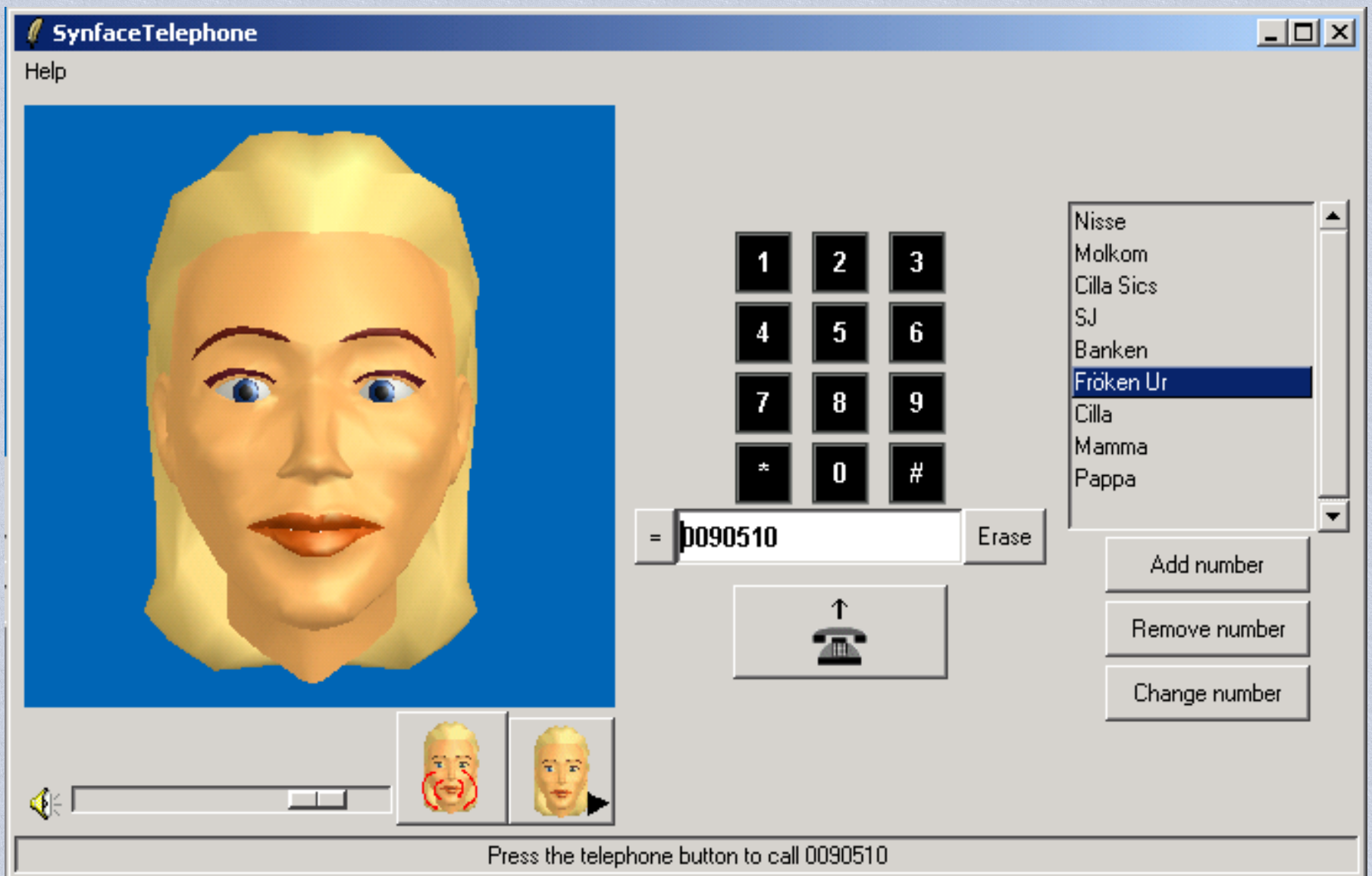
- Great variation in human speakers due to for example
 - Speaking rate
 - Extent of articulatory movements (the hypo – hyper dimension)
 - Anatomy, facial hair
 - Light, distance, viewing angle...

Hypo to hyper articulation

Pilot study stimuli



The Synface telephone



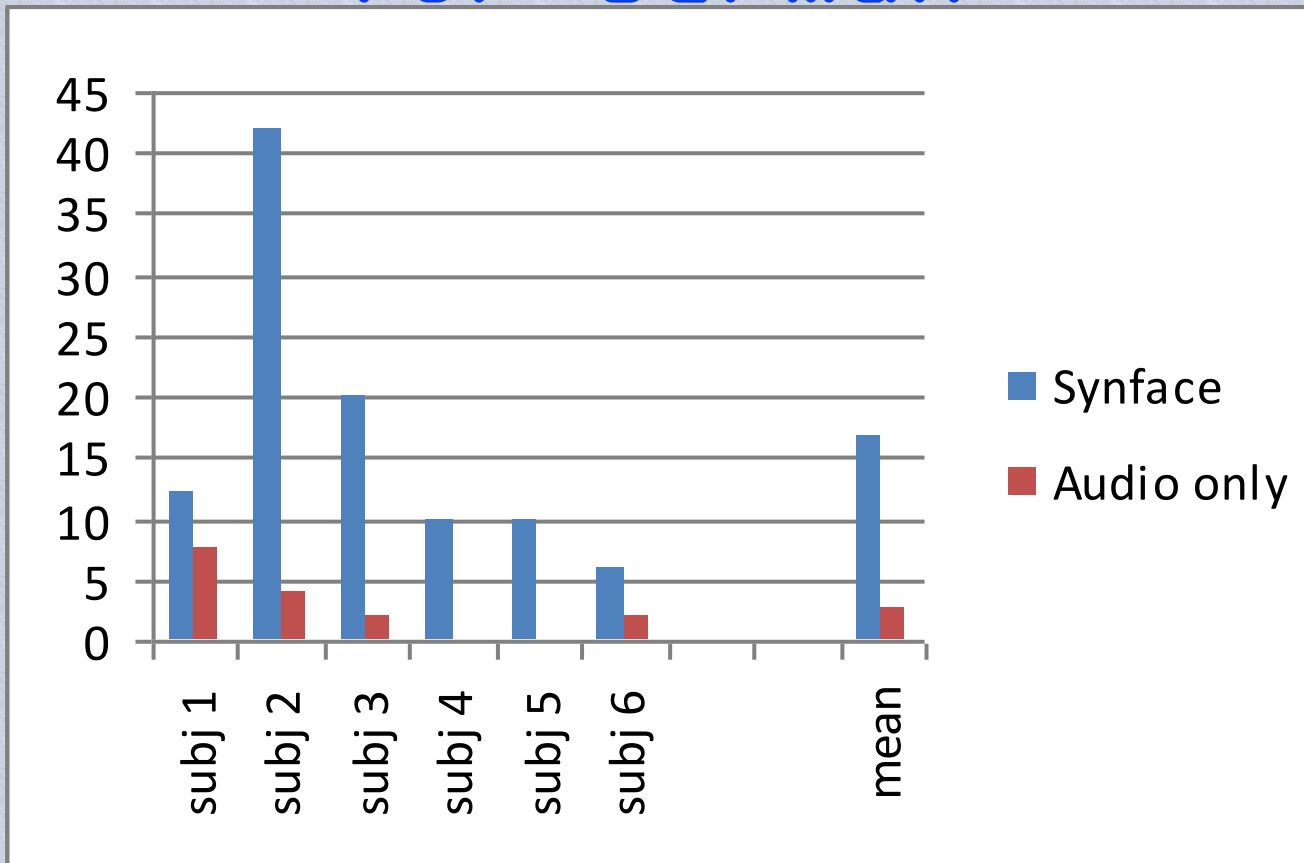
Real user tests



SYNFACE: Recent tests/results

- Tested on hearing impaired users in England, Holland and Sweden
- Most were positive
 - Believe that SYNFACE helps (80%)
 - Think that SYNFACE is a useful product (80%)
- Desire a more realistic face (79%)

SynFace: preliminary results for German





[Home](#) · [Technology](#) · [Demo](#) · [About us](#) · [Account](#)



[Download Trial Version >>](#)

World leading technique for synthesized talking face derived from speech

SynFace AB sells the world leading and award winning software EyePhone. EyePhone transforms the sound in the spoken language (phonemes) to facial and lip movement that is visualized on a synthesized face in real time.

EyePhone helps the hard of hearing people

EyePhone enables for hard of hearing people that need lip reading to interpret speech to now have telephone conversations. Simply install EyePhone on your computer and use IP-telephone. When the opposite party speaks you will get support in the conversation as you simultaneously can read the lips on the synthesized talking face.

Other areas that can use EyePhone

There are many other uses for EyePhone such as on-line games, public information systems and language studies.

News

2006-09-25

Free EyePhone download now available for testing

2006-09-21

SynFace is invited by European Commission, Directorate-General for Research Information and Communication Unit to participate at the exhibition "Today is the Future - 07" on March 7-18 in Brussels

2006-09-19

Meet SynFace at ID-dagarna Oktober 11-13 at Factory Nacka Strand

2006-08-17

Pål Ljungberger is elected CEO and Per Junesand is elected Chairman of the Board for Synface AB at extra shareholders meeting.

Talking faces for speech reading support in TV



Talking heads in educational applications

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

Unacceptable pronunciation needs to be identified



from a Vinnova video (on CTTs webpages)

Automatic tutor simulation

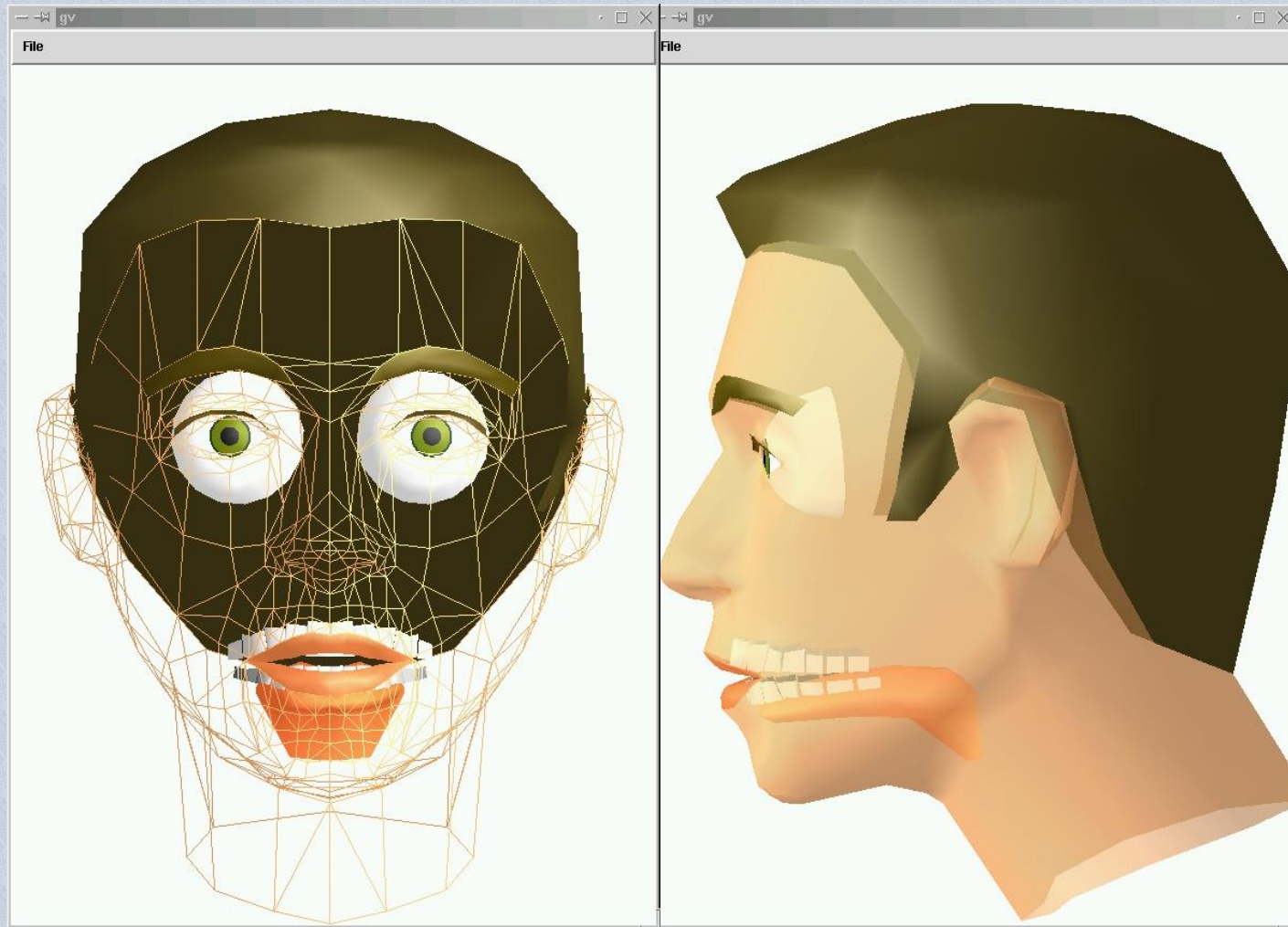


no gestures

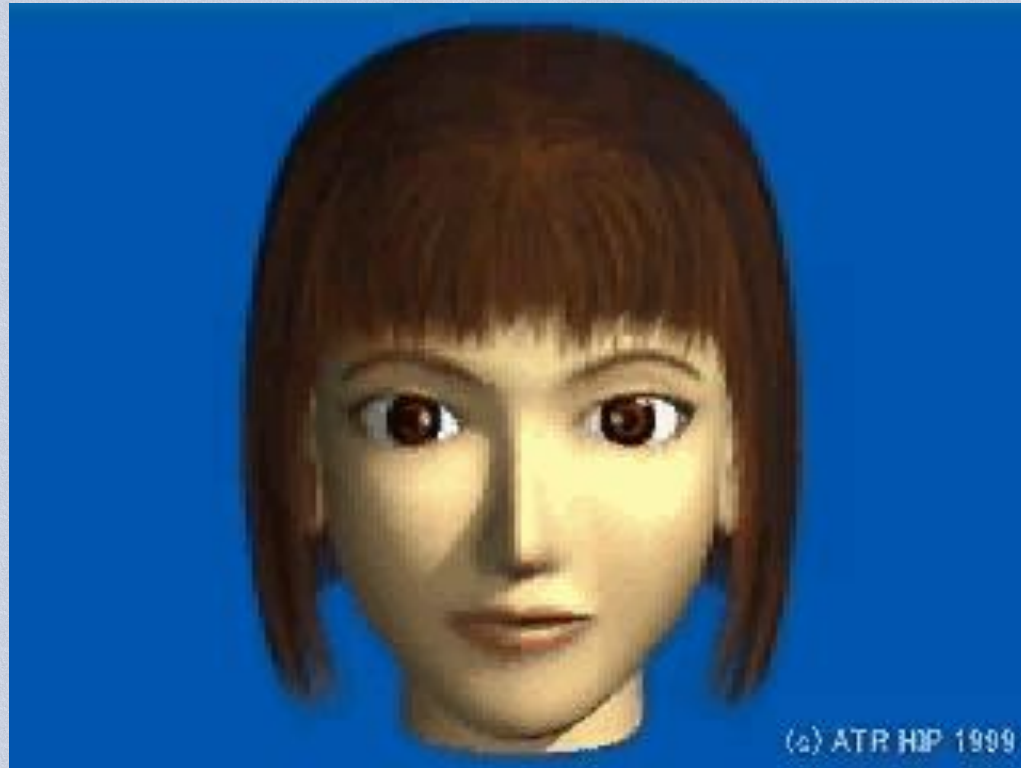


some gestures

Different representations



Reiko Yamada ATR, 1999



National project ARTUR

What?

Automatic articulatory feedback display using face and vocal tract models.

For whom?

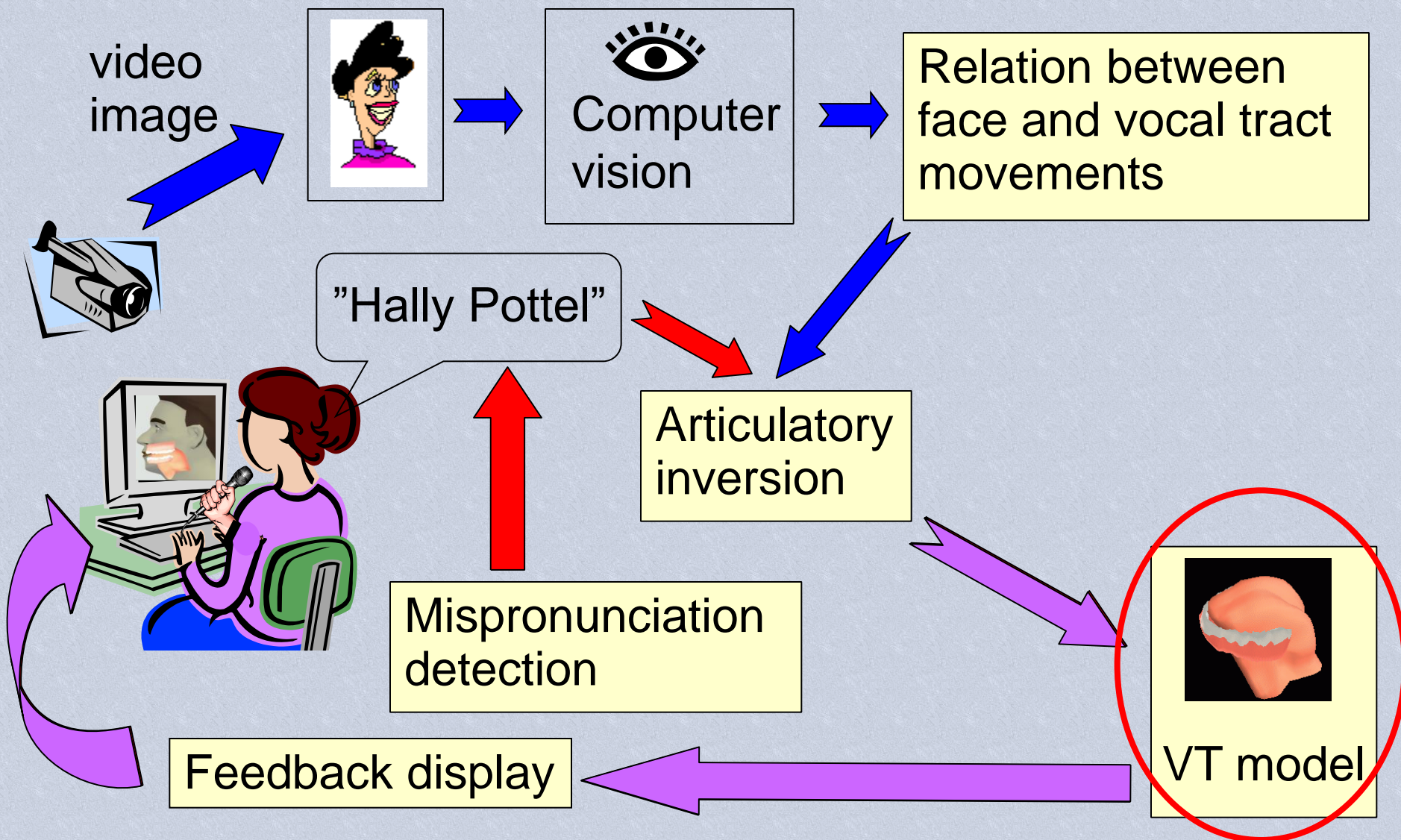
Hearing impaired children, second-language learners, speech therapy patients.

How?

Contrasting the user's articulation with a correct one.



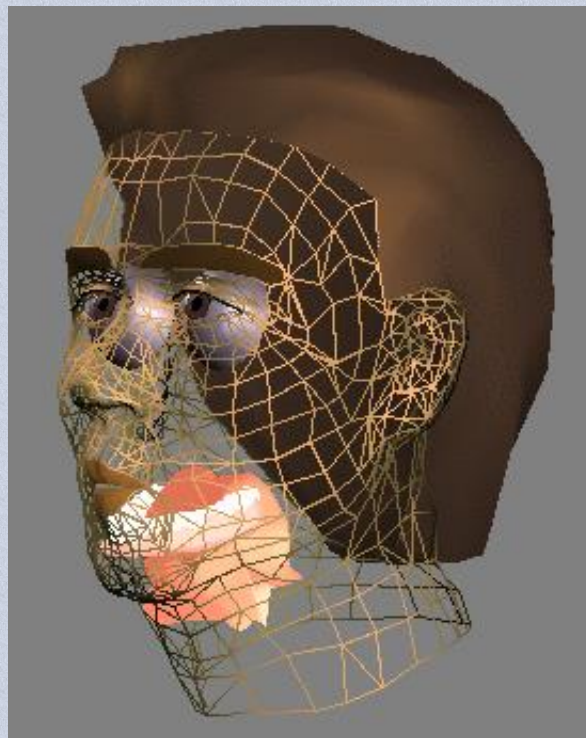
ARTUR: the articulation tutor



3D tongue movements



[s]



[k]



vowels

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

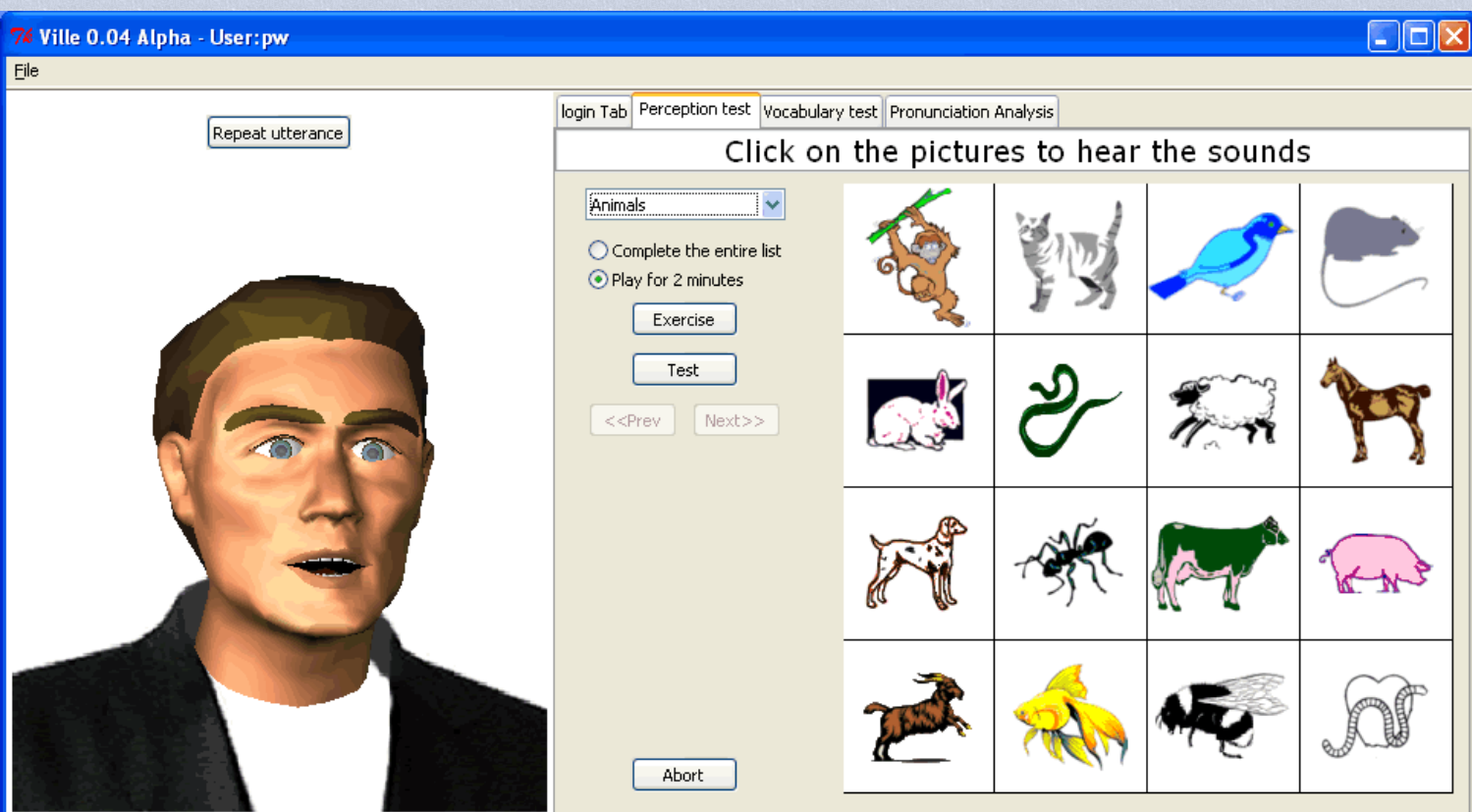
The tongue movements can be presented from different views

CNN video with Arthur



VILLE - Virtual Language Tutor

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



Different types of VILLE 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?

Nordic project – NordPlus Sprog

Using VILLE – CTT Virtual Language Tutor

The screenshot displays the VILLE 0.0402 Alpha software interface. On the left, a 3D-rendered female character is shown with a "Record" button below her. The right side of the interface is divided into several panels:

- Perception test**: A tab at the top right of the analysis area.
- Teacher Spectrogram**: A spectrogram showing the frequency spectrum of the teacher's speech. It includes a "Play" button and a timeline with phonetic labels H, Ä, T, and I:.
- Student Spectrogram**: A spectrogram showing the frequency spectrum of the student's speech. It includes a "Play" button, a "Threshold" slider set to 0.4, an "update" button, and a timeline with phonetic labels H, Ä, T, and I:.
- Student Modified**: A spectrogram showing the modified student speech. It includes a "Play" button, a "Mod" button, and a timeline with phonetic labels H, Ä, T, and I:.

The timeline for all spectrograms is labeled "time" and ranges from 0.2 to 0.8 seconds.