



Embodied conversational agents in verbal and non-verbal communication



**Björn Granström & David House
School of Computer Science and
Communication**



KTH - Kungliga tekniska högskolan

Department of Speech, Music and Hearing

Kungliga tekniska högskolan, KTH

Welcome to KTH

- 10000 undergraduate students
- 1500 graduate students
- 3000 staff
- 800 professors and teachers



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

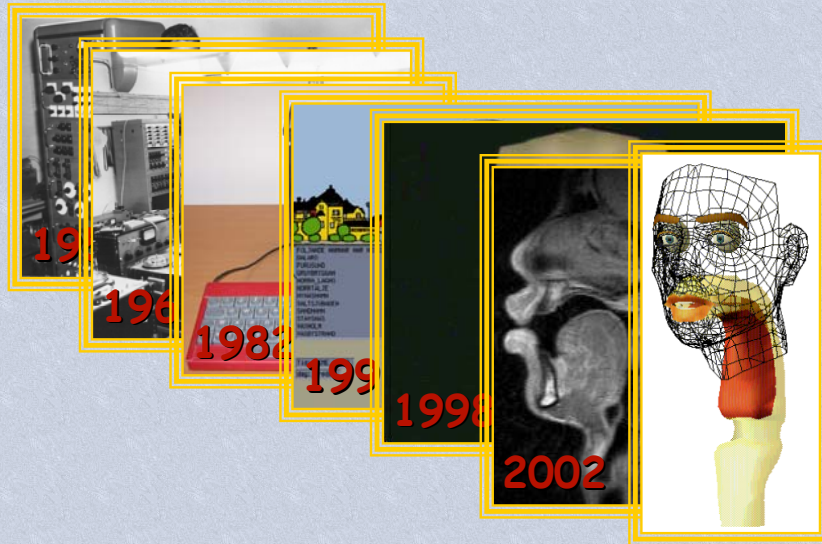


<http://www.speech.kth.se>

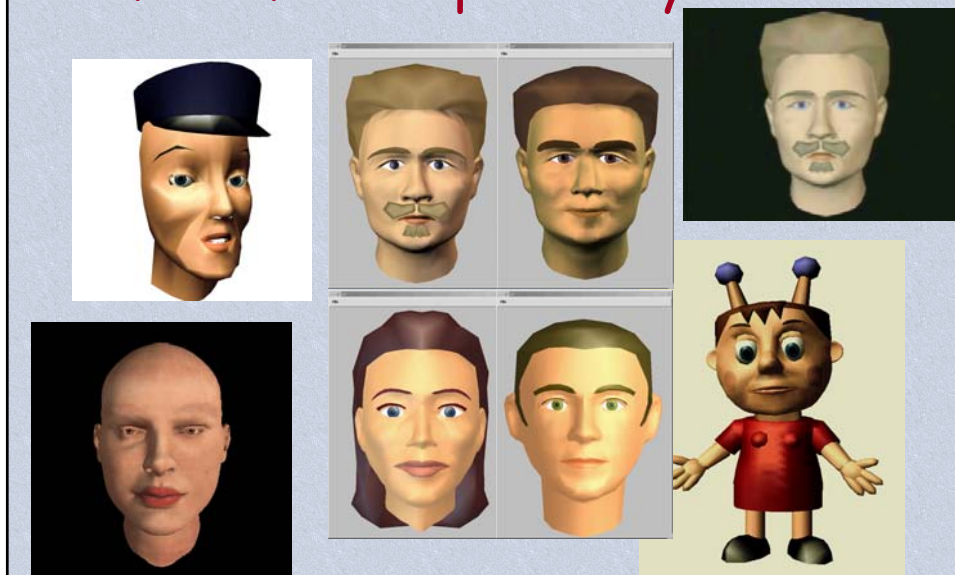
- Headed by Björn Granström & Rolf Carlson
- 25 employees (some part time)
 - including 10 PhD students
 - Multidisciplinary staff
- Activities
 - Education – mostly post grad.
 - National and EU projects



Speech synthesis developments at KTH



Animated talking agents - multimodal speech synthesis



Paradigm transition in human-computer interaction

- Shift from desktop metaphor to person metaphor
- Non-verbal communication as well as spoken dialogue
- Takes advantage of the user's social and communicative skills
- Coherence between vision and audio
- Strive for believability?
- Managing expectations

One application: EU-project Synface

[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 trialed a prototype in the UK.



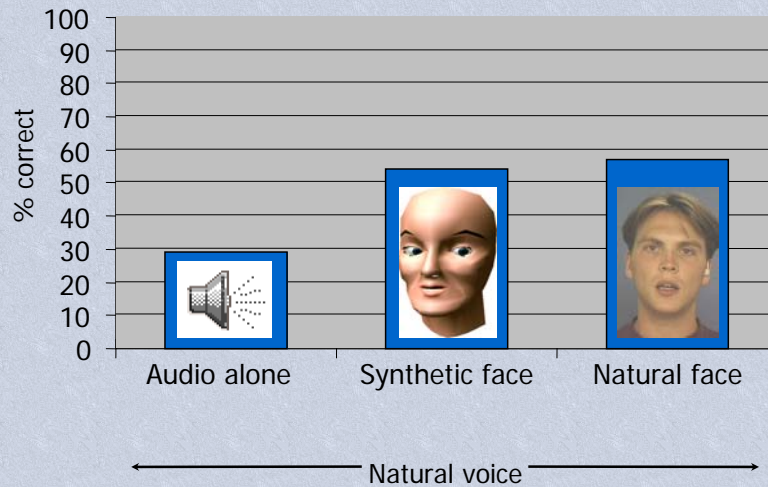
Phone conversations are animated

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

www.speech.kth.se/synface & www.synface.com

Face gives increase in intelligibility!
 Results for VCV-words (hearing impaired subjects)
 Used in lip reading aid for telephone developed in EU/SynFace project

Beta version for Skype available from <http://www.synface.com/> <http://www.synface.com/>



Better than a natural face?

| | Synthetic face | | | | | Natural face | | | | |
|--------------------|----------------|------|-----|------|------|--------------|------|------|------|------|
| | bil | labd | den | pal | vel | bil | labd | den | pal | vel |
| bilabial | 100 | | | | | 96,3 | | 2,5 | 1,3 | |
| labiodental | | 96,3 | 3,7 | | | | 92,6 | 5,6 | 1,9 | |
| dental | | | 3,0 | 78,0 | 5,5 | 13,4 | | 85,8 | 7,4 | 6,8 |
| palatal | | | | 9,9 | 70,4 | 19,8 | | 1,2 | 17,3 | 71,6 |
| velar | | | | 4,9 | 16,0 | 79,0 | | | 2,5 | 25,0 |

Key challenges in developing interactive talking agents with communicative and emotional capabilities

- How to obtain data?
- How to model it and its interaction with speech?
- How to exploit it in dialogue systems?

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



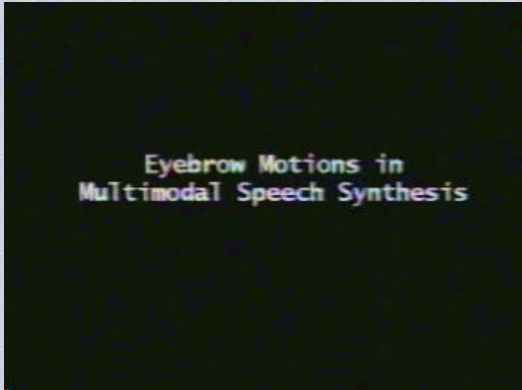
SONY SDR



Tool for testing hypotheses in multimodal communication

- Our bias: communication is multi-modal
- Traditionally prosodic functions are signaled by "gestures", perceived by "eye and ear"
- This concerns both body and face gestures
- Preliminary hypothesis: $F_0 \sim$ eyebrow height - e.g. Cavé et al. (1996)
- Easy to put to a test with multimodal speech synthesis

Eyebrow vs intonation



Eyebrow Motions in
Multimodal Speech Synthesis

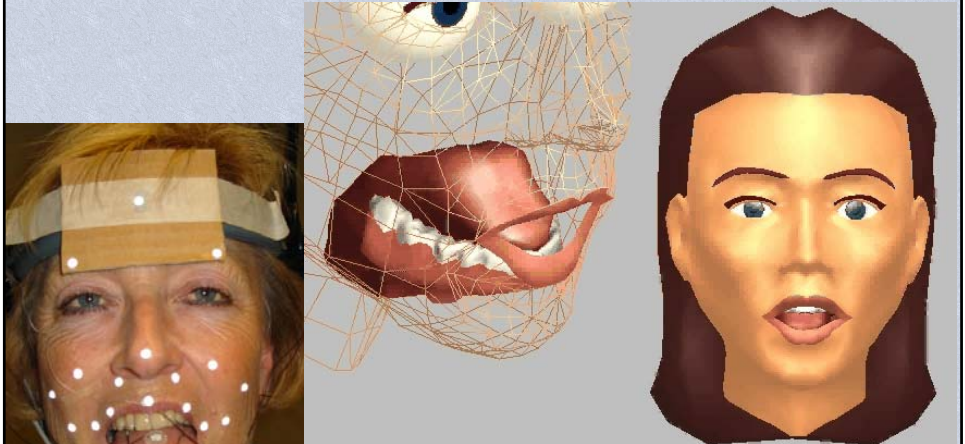
- 1 No eyebrow motion
- 2 Eyebrow motion controlled by the fundamental frequency of the voice
- 3 Eyebrow motion at focal accents +
- 4 Eyebrow motion at the first focal accent +

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

How to obtain data?
Combination of inside and
outside registrations
Qualisys and EMA
(Movetrack)



Example of resynthesis



Could be used for e.g.
pronunciation training

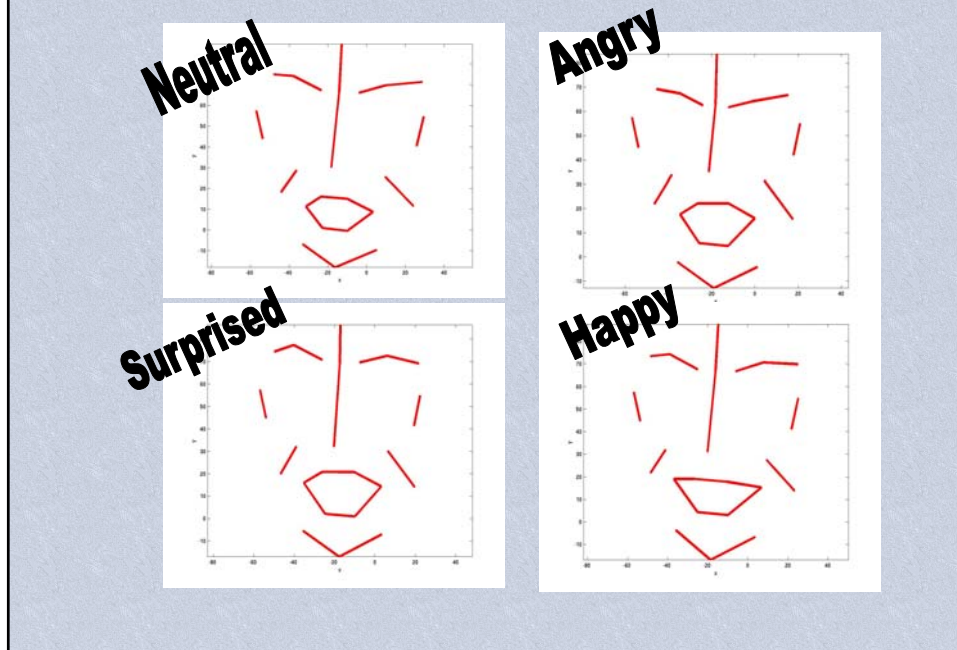
Prompted read speech databases

- Expressive modes (acted):
 - Happy, angry, sad, surprised, afraid, disgusted
 - confirming, questioning, certain, uncertain, encouraging and neutral
- 39 short, content neutral sentences with three possible focal accent positions each, e.g.
 - *Båten seglade förbi* (The boat sailed by)
 - *Dom flyttade möblerna* (They moved the furniture)
- Nonsense words (VCV, VCCV, CVC)
- Digits

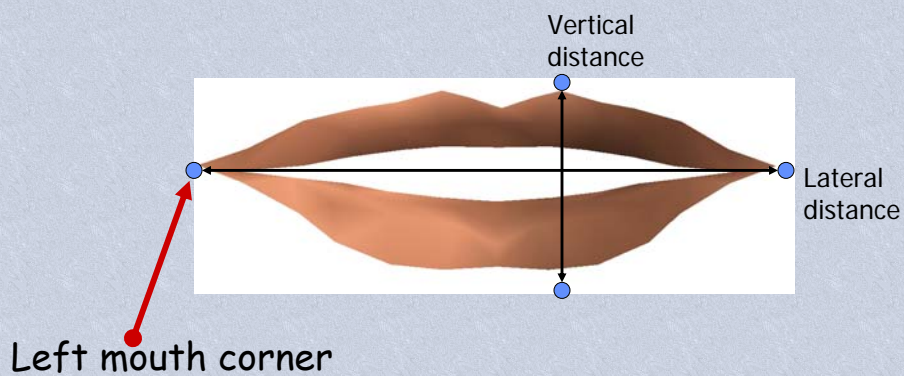
Interactions: emotion and articulation (resynthesis) (from AV speech database - EU/PF_STAR project)



Connected mean positions for recorded points



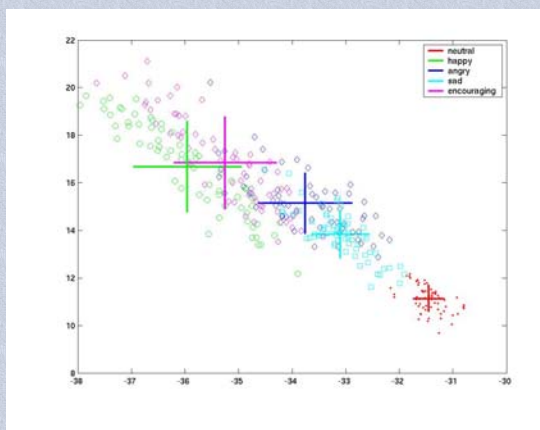
Measurement points for lip coarticulation analysis



The expressive mouth

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

"left mouth corner"



Focal accent examples from the database

Focal accent on:

Båten

seglade

förbi

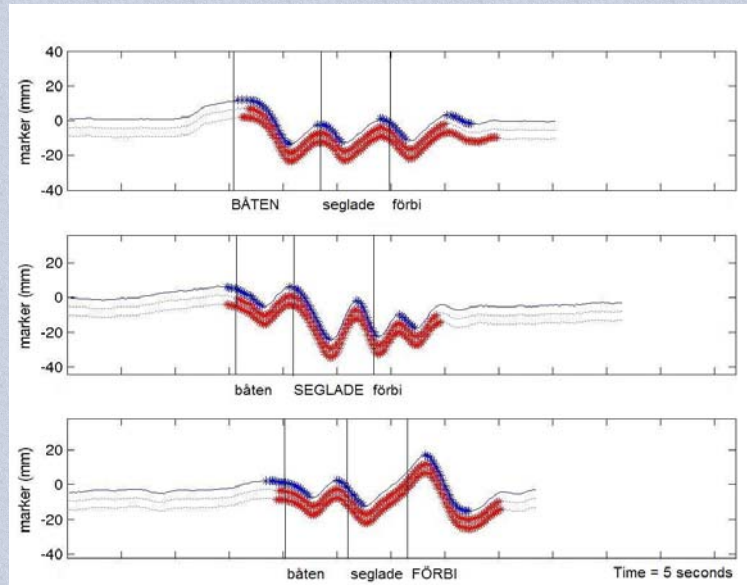
Happy



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

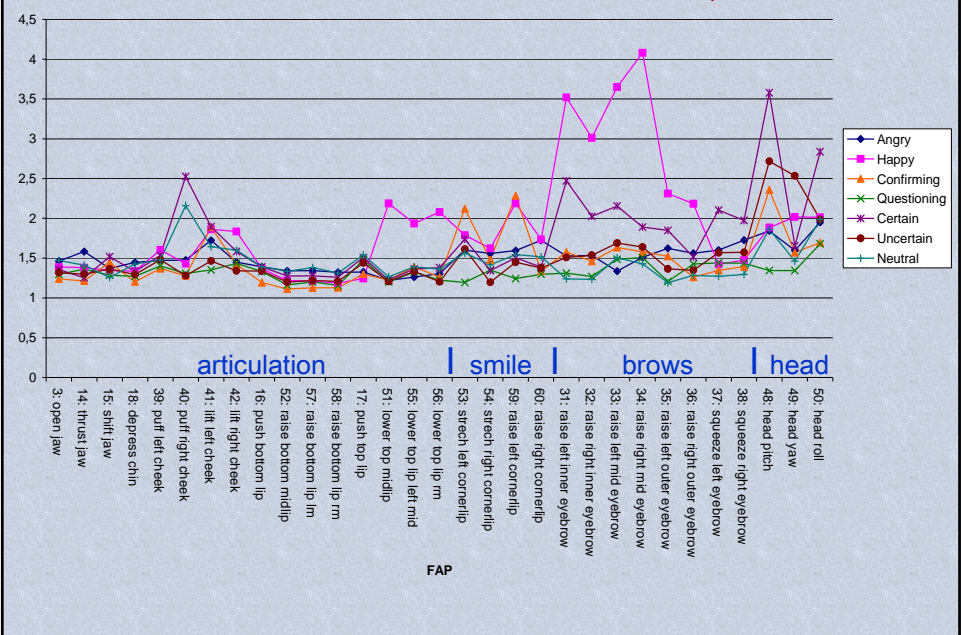
MPEG-4 Facial Animation Parameter (FAP)

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.

Beskow, Granström & House (2006)

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



Datadriven facial synthesis with MPEG4 model



Happy



Angry



Sad



Surprised

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 FO analysis.
- The Hummer is not patent pending!

Collection of audio-visual databases also for 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.



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

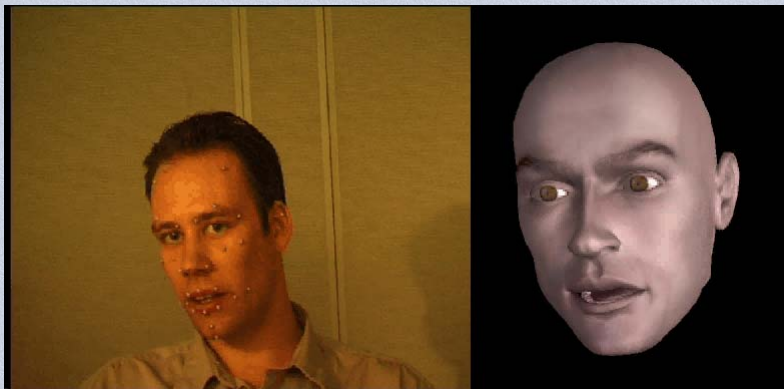
VISPP summerschool, Aug 2008



SPONTAL conversational database

- Free conversation in pairs
- Gender: same, different
- Partners: known, unknown
- Audio, video and motion capture
- 120 recordings*30 minutes
- Will be available for research

Recording and model



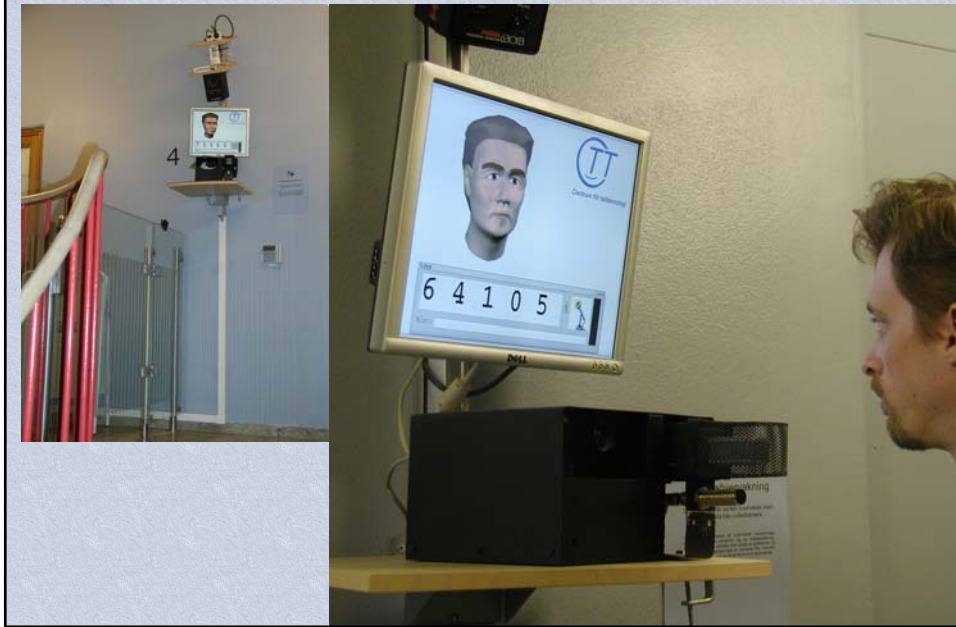
Conversation with agent



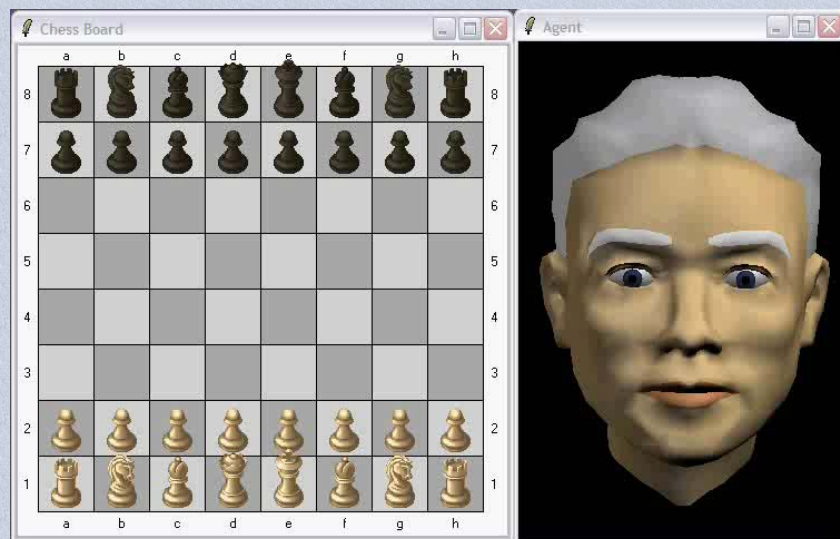
Some applications

- Per (St. Peter) our gatekeeper
- Chess
- Hearing at Home - lip reading support
- MonAMI - Innovative interfaces
- Language learning

Per - the CTT gatekeeper

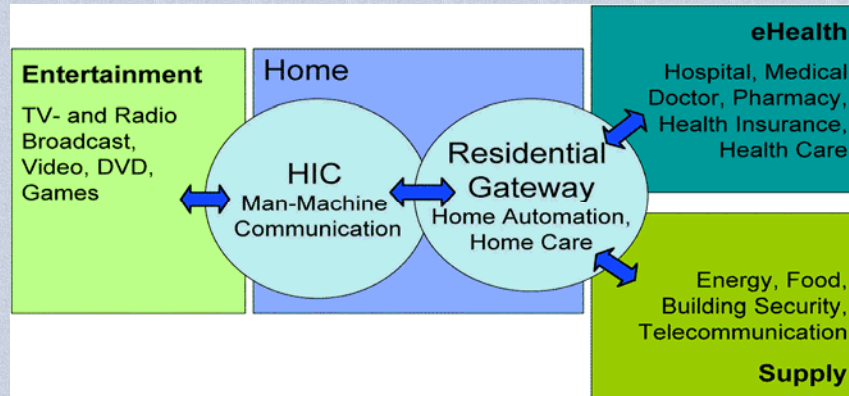


The CTT chess player



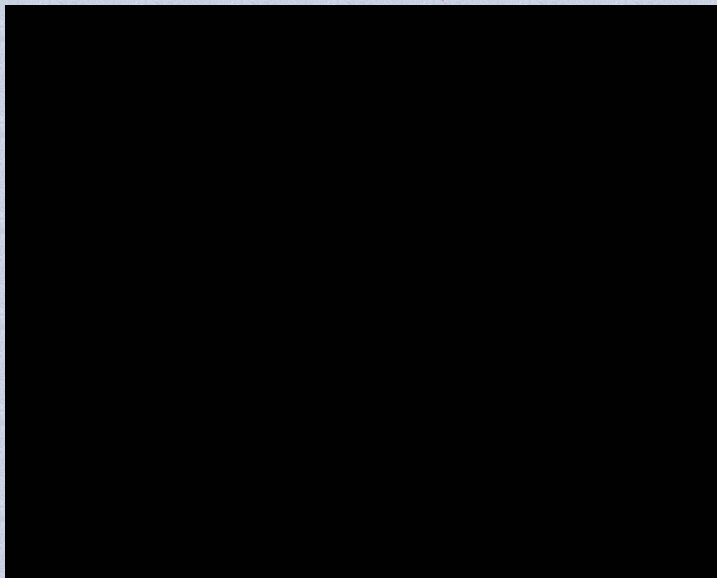
Hearing at Home

Goal: Barrier free access for the hearing impaired to IT solutions used at their homes in future: Information, Communication, HomeCare Applications, Assistive Technologies

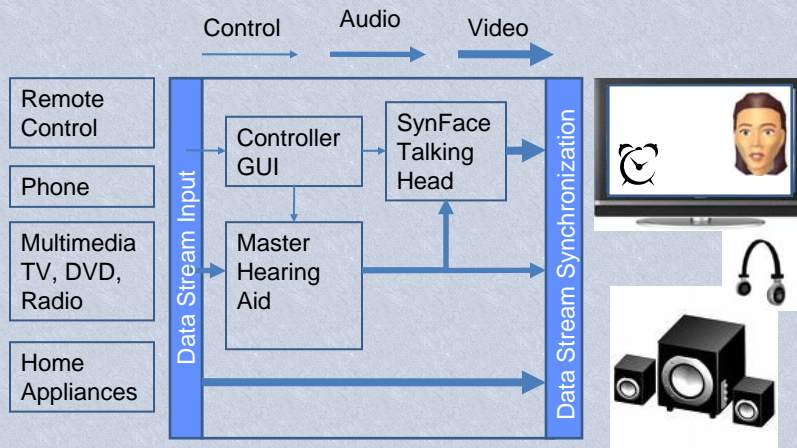


<http://www.hearing-at-home.eu/>

Includes SynFace



Hearing at Home



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

Animated speaker for TV

CHIL
Center for Human-Computer Interaction

Computers in the
Human Interaction
Loop

Simplify your daily life!

ist
Information Society

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*

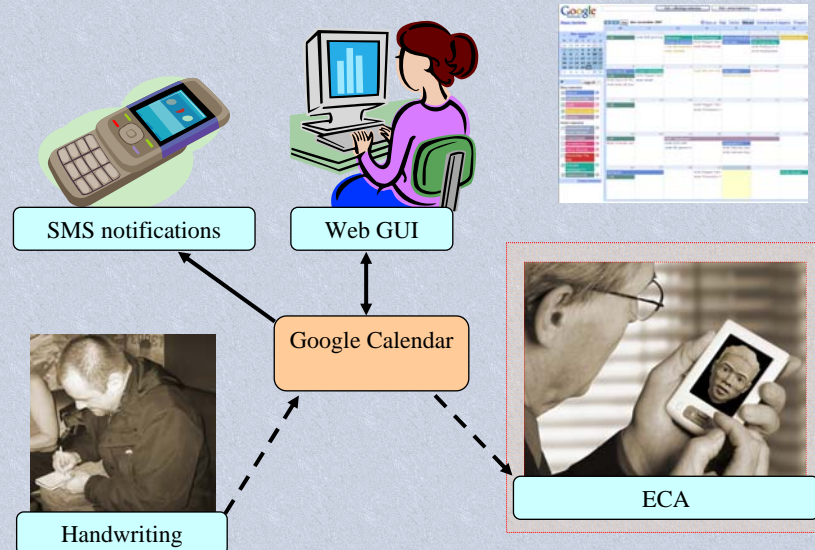


<http://www.monami.info/>

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



MonAMI reminder application

MonAMI
Mainstreaming on Ambient Intelligence

The ECA Reminder



Language learning



Also for conversational training...



Three AV perception experiments

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

Prominence due to eyebrow rise

5 content words: "När pappa fiskar stör piper Putte"

When dad is fishing sturgeon, Putte is whimpering

Task: "which word is most prominent"

(identical acoustics - varied location of eyebrow movement)

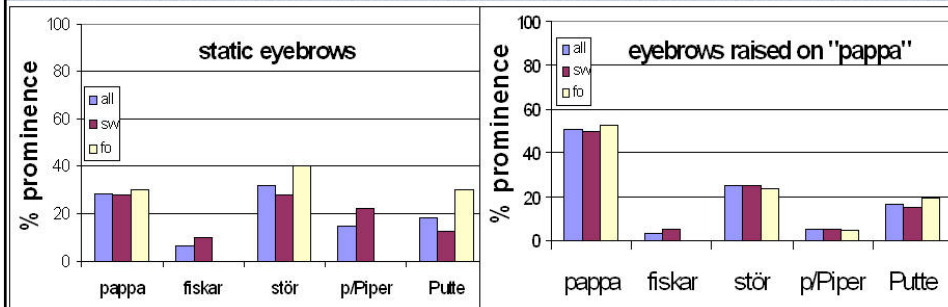


No eyebrow movement (neutral)

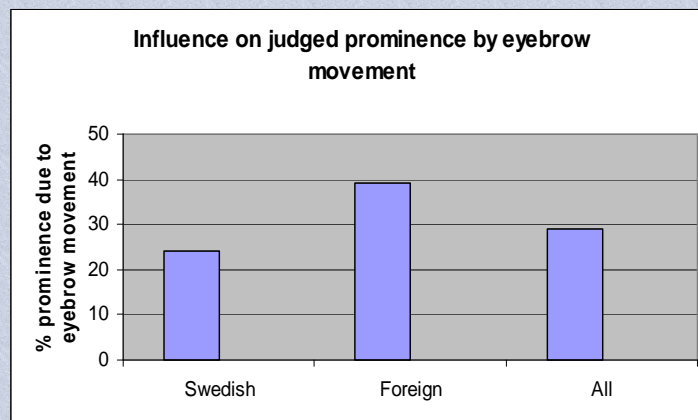


Eyebrow movement

Swedish and foreign subjects



Prominence increase due to eyebrow movement



Cues for feedback - the eavesdropper

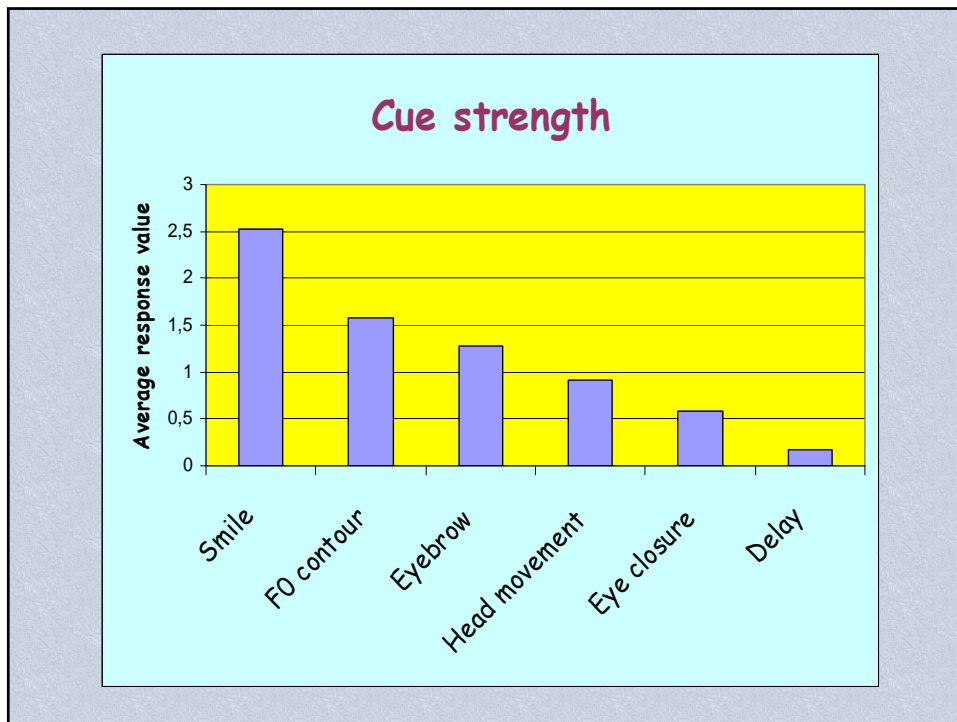
- Mini dialogues (two turns)
- Travel agent application
- subject sees and hears travel agent but only hears customer
- Both visual and acoustic feedback cues
- **Affirmative cues** - agent understands/accepts the request
- **Negative cues** - agent is unsure about the request (seeks confirmation)
- 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 |

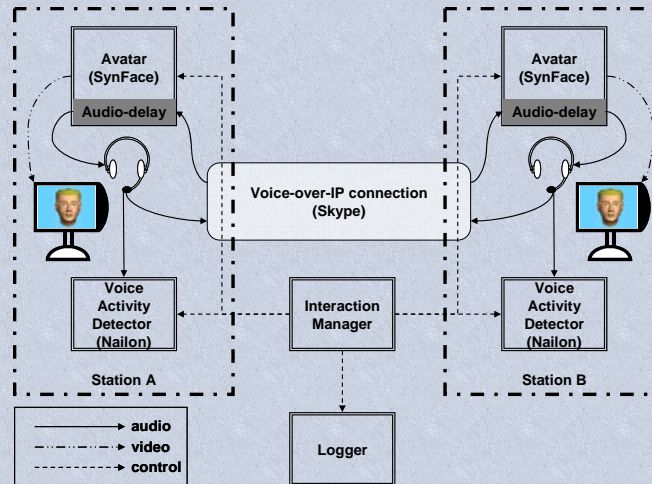


Cues for feedback - the participant

- Based on "SynFace" technology originally 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



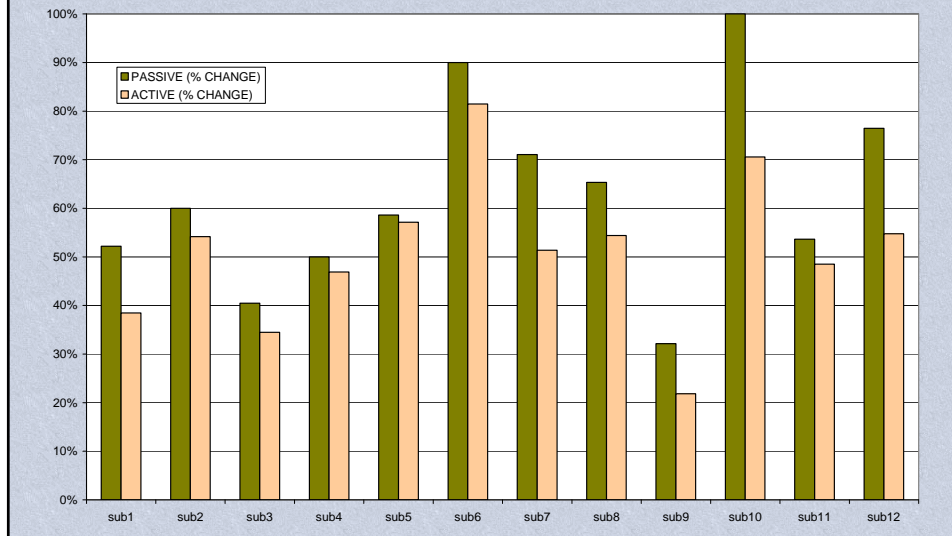
Subject A

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



Summing up

- Emotions more important for (lip) articulation than vowel identity
- The whole face is affected by focal accents, but differently for different kinds of expressive speech
- Visual cues often override audio speech cues
- Interaction behaviour can be manipulated by avatars, useful in e.g. multimodal dialogue systems
- Already useful in several applications

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.