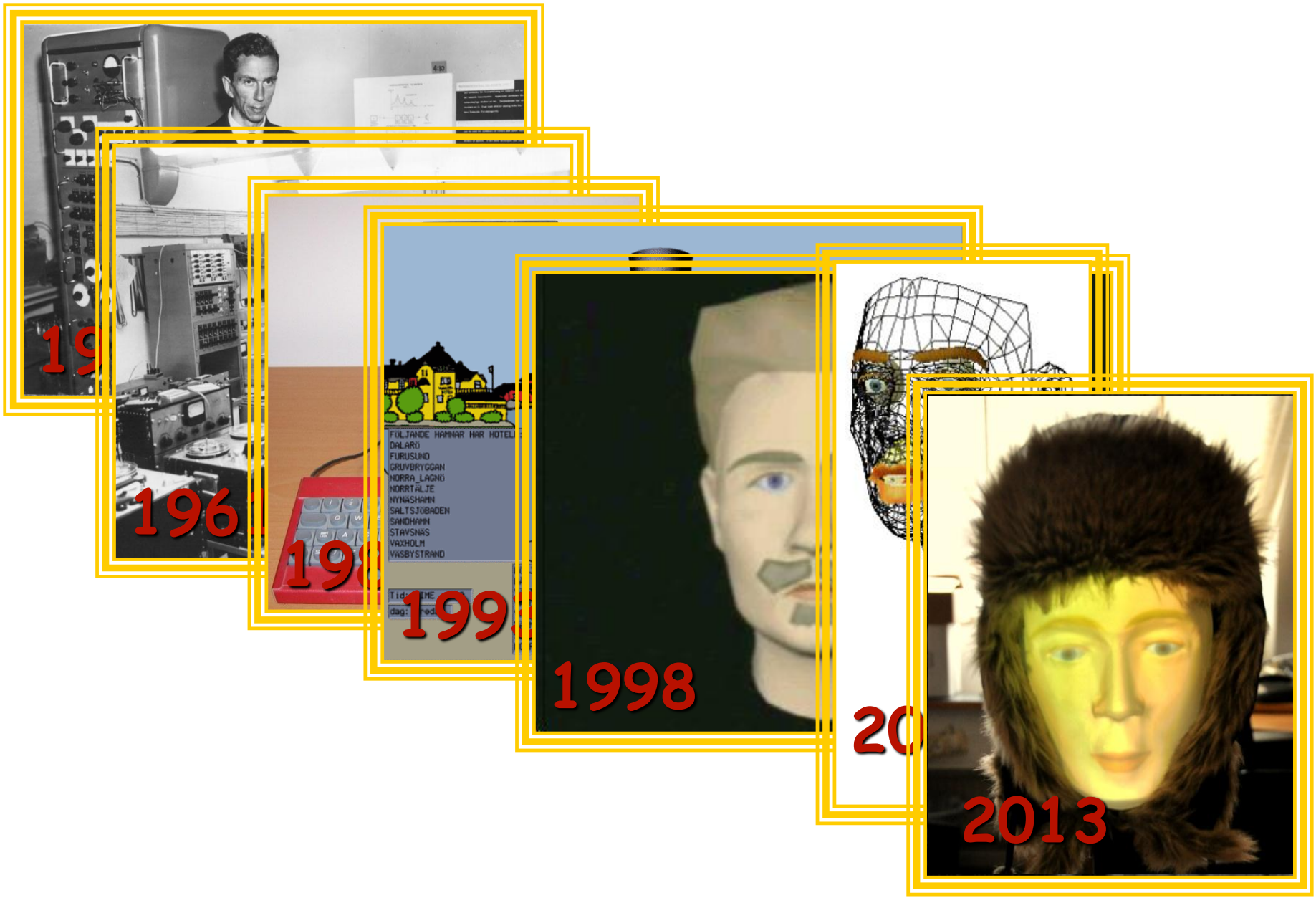
The KTH talking head in space - a vehicle for situated multi-party interaction

Björn Granström-CTT, KTH

Centre for Speech Technology

KTH VETENSKAP OCH KONST

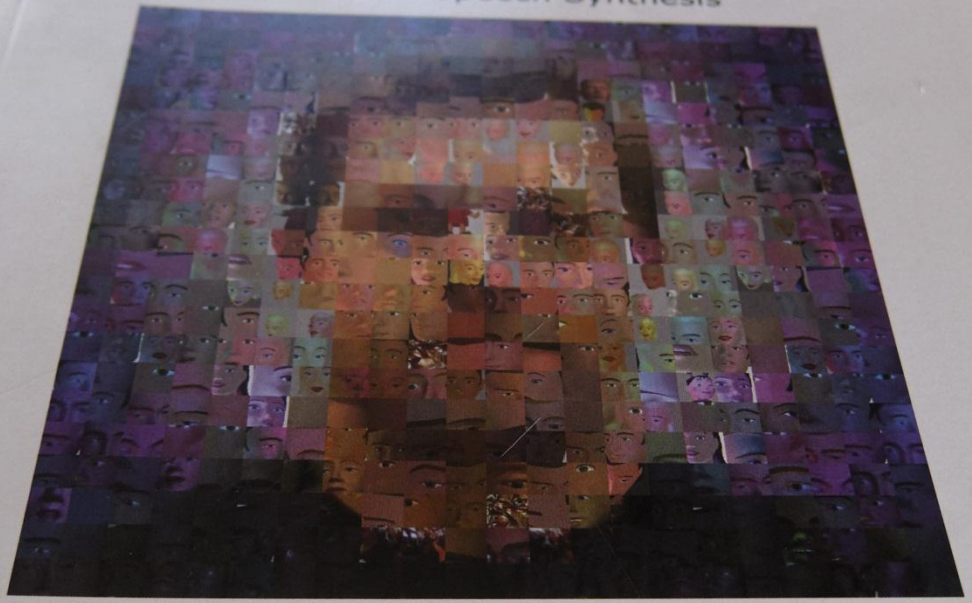
Speech synthesis developments at KTH



KTH

Talking Heads

Models and Applications for Multimodal Speech Synthesis



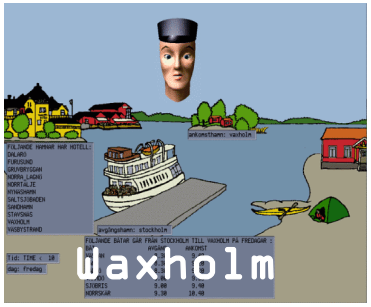
Jonas Beskow

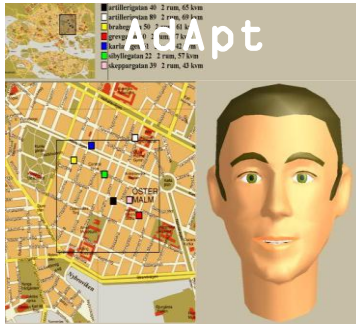
Doctoral Dissertation
Stockholm 2003

KTH interactional data collections

- Human-human

- Simulated human-computer





- Human-computer

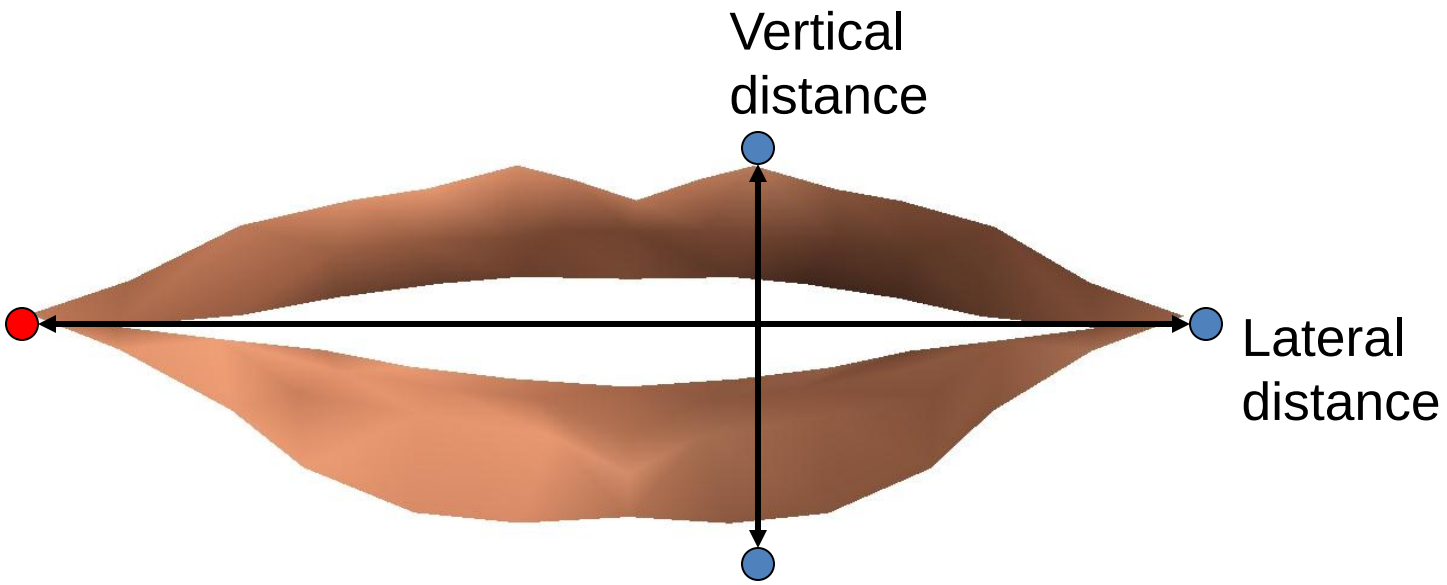
Collection of audio-visual databases: interactive spontaneous dialogues

Eliciting technique: information seeking scenario

Focus on the speaker who has the role of information giver

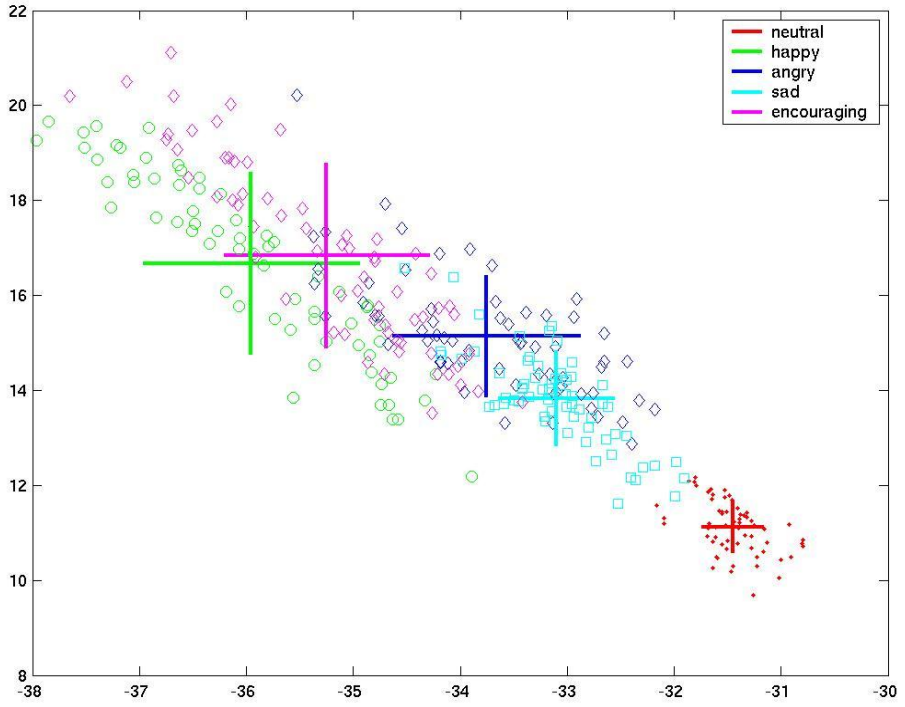
The speaker seats facing 4 infrared cameras, a digital video-camera, a microphone The other person is only video recorded.

Measurement points for lip coarticulation analysis



The expressive mouth

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

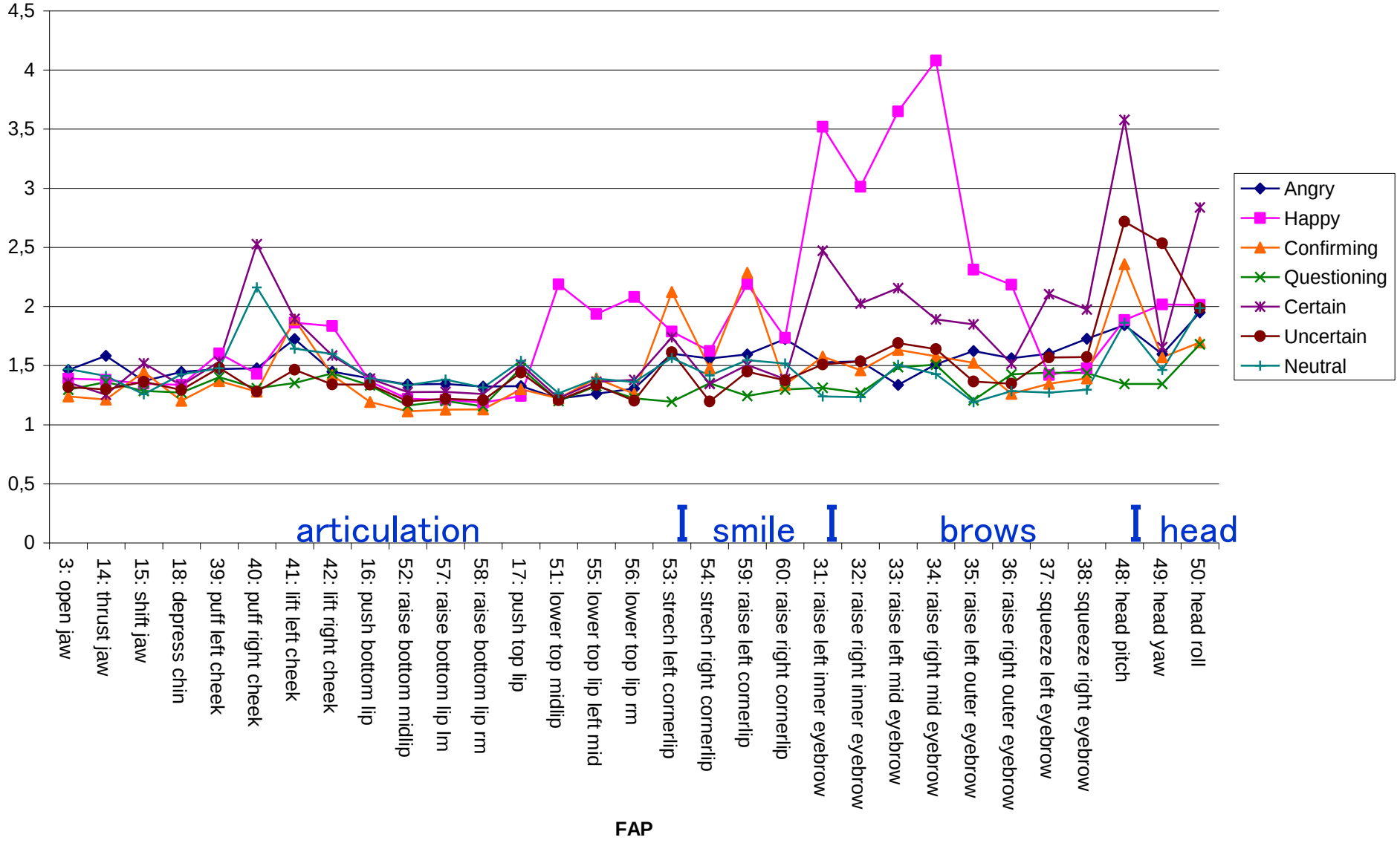


"left mouth corner"

Visual prosody
Word focus vs. articulation

- Simple 3-word sentences with systematically varied focus
- Pronounced in seven different expressive modes
- Lab/acted speech
- Motion capture of points on face mapped on facial animation parameters (FAP)
- Focal motion quotient (FMQ) calculated, as variation in FAP over the focal word divided by the (mean) variation in FAP of the same word in non-focal position)
- If FMQ=1, focus does not affect that point on the face

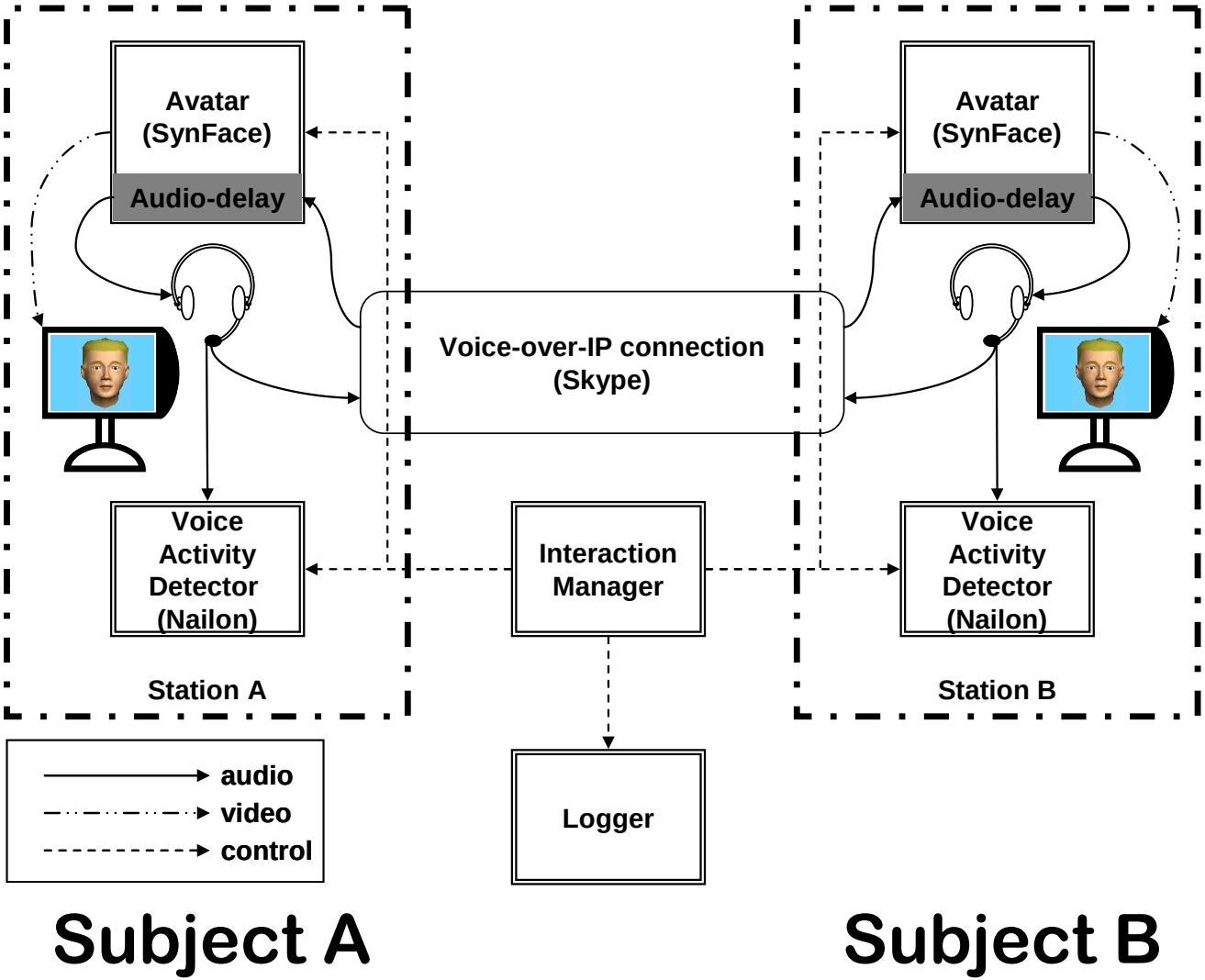
The focal motion quotient, FMQ, averaged across all sentences, for several expressive modes



Turntaking in mediated audio-visual human-human communication

- 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?

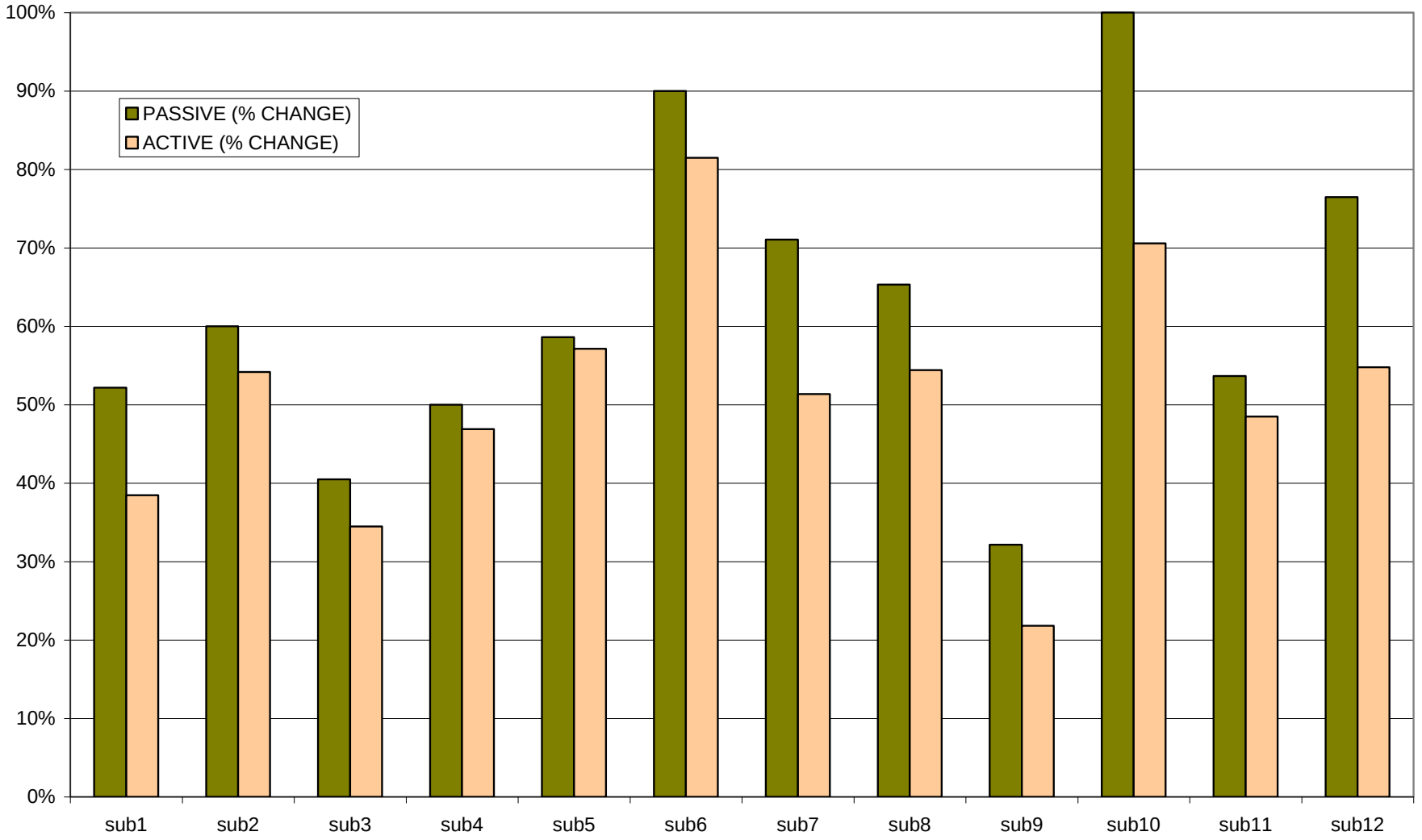
Manipulating turntaking in mediated human-human communication (experimental set-up)



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

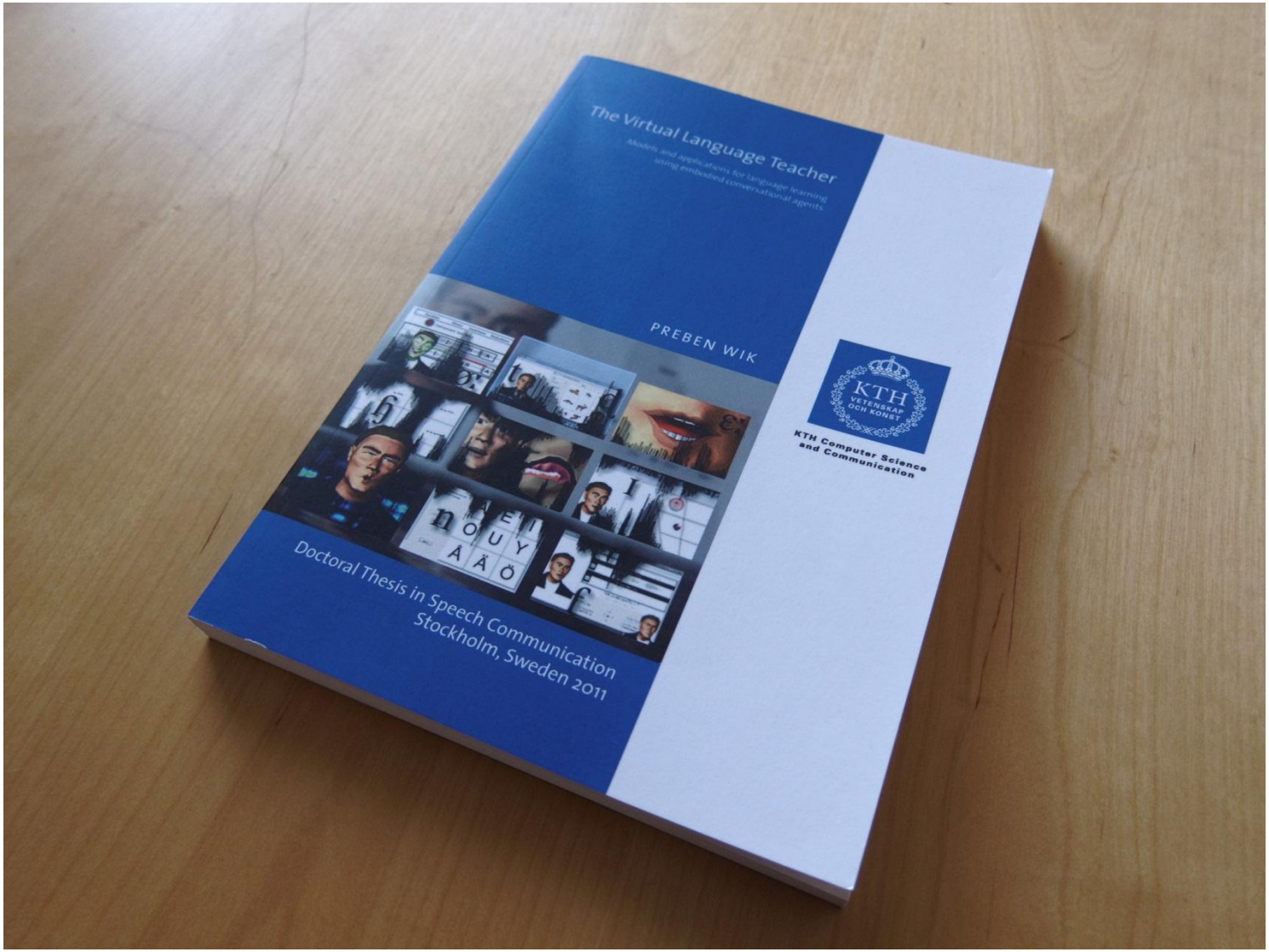
Percentage of contributions followed by a change of turn for twelve subjects represented by passive vs. active avatars



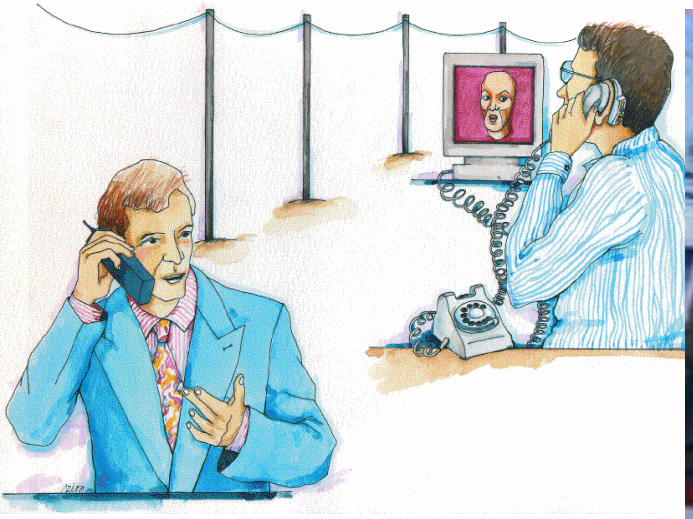
Summary of the three early studies

- 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

Applications in dialogue systems, virtual tutors and rehabilitation



Visual impact on intelligibility. Early application in the TELEFACE project

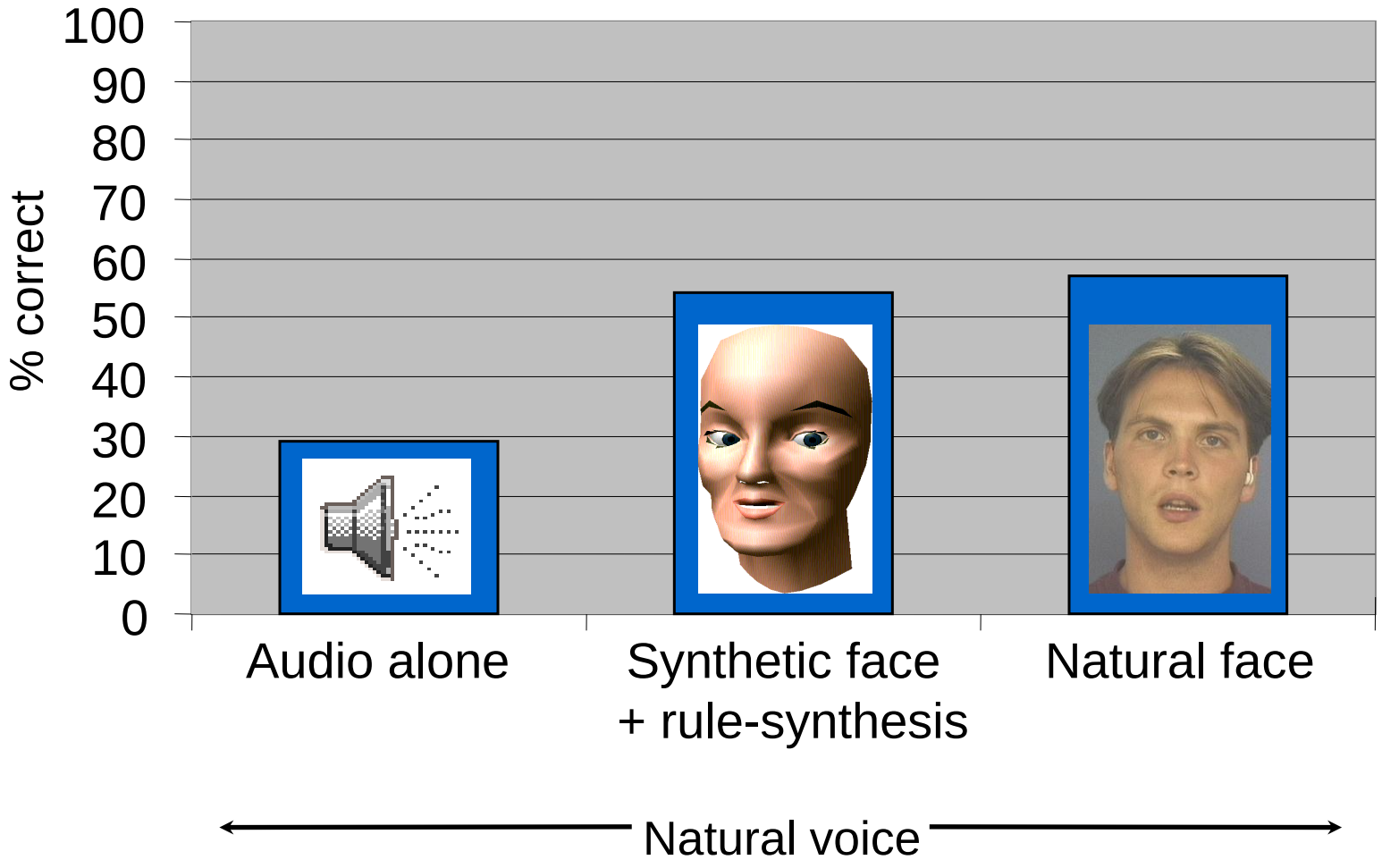


Multi-modal speech communication for hearing impaired persons

Concept video

Continued in EU projects SYNFACE and Hearing at Home,

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

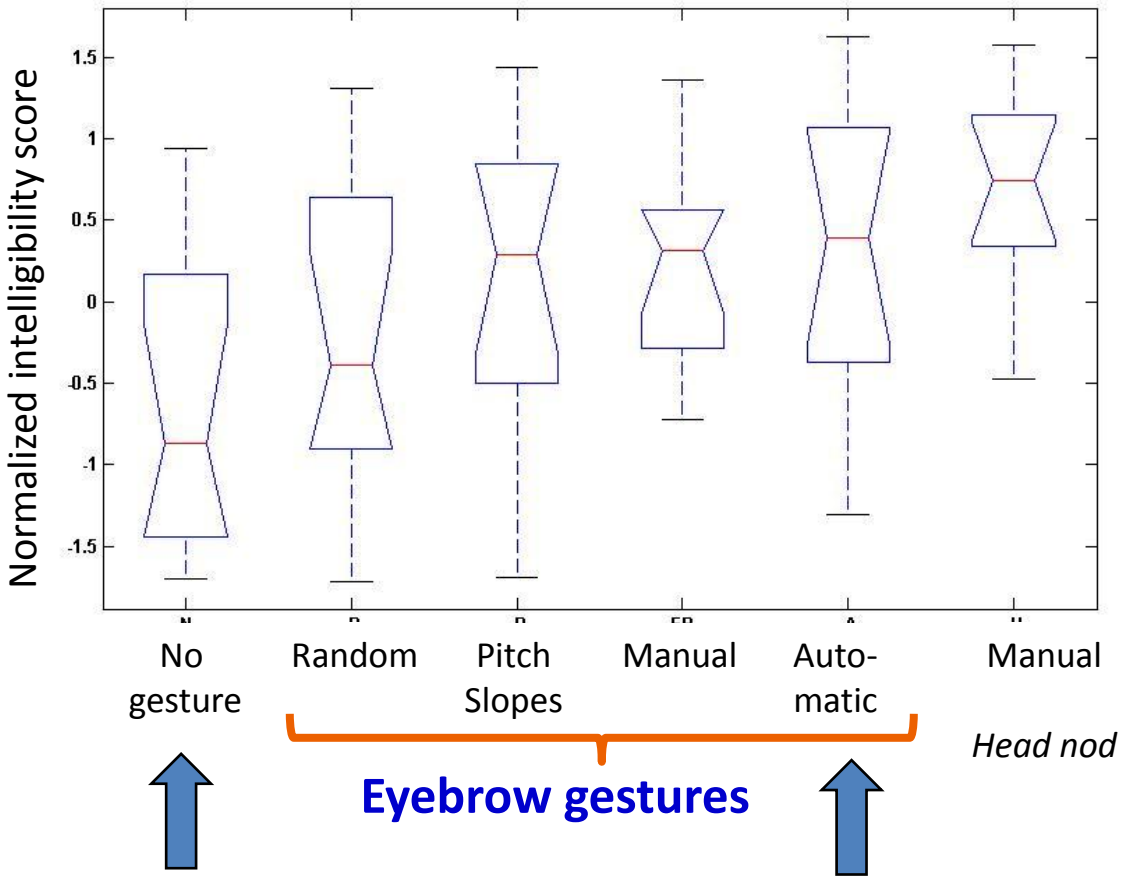
Synthetic face

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

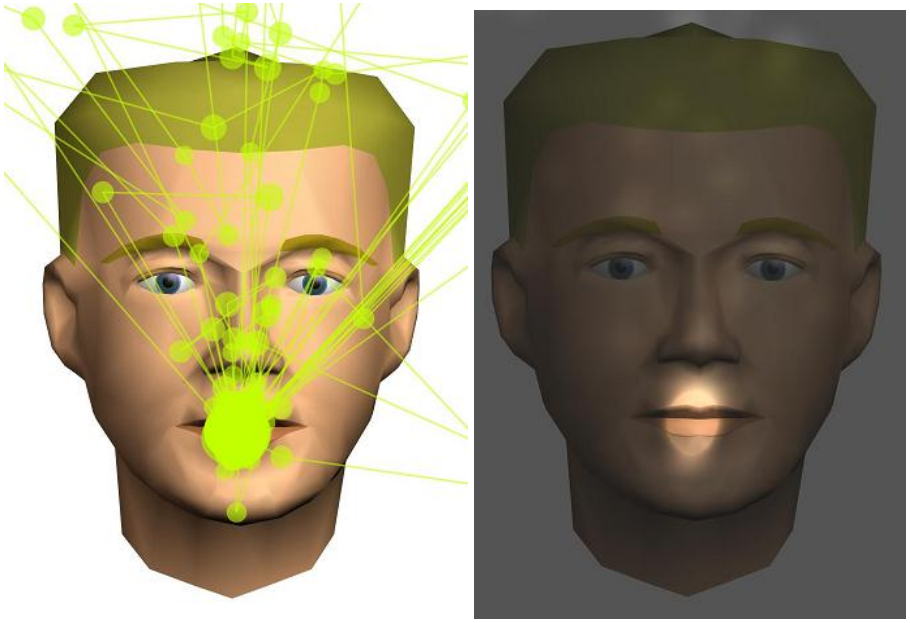
Expanded use: Reading talking books with SynFace and TOBII Eye-tracker (EU/HaH project)

Effect of non-verbal information
Addition of prominence gestures

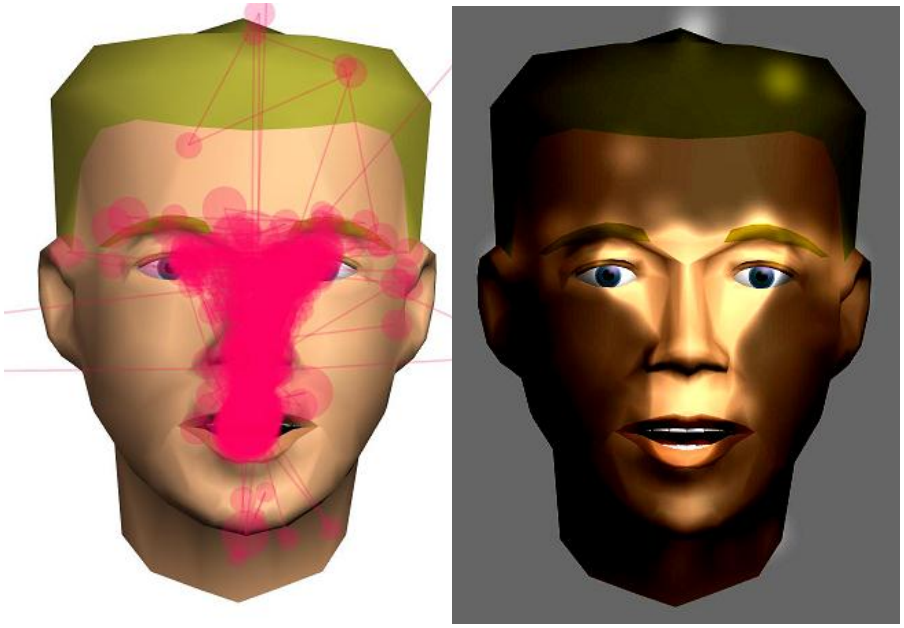


Significant intelligibility increase for gestures (eyebrows & head nod) based on pitch slopes, automatic prominence detection and manual annotation

Eye-tracking results



Without prominence gestures



With gestures

MonAMI

Mainstreaming on Ambient Intelligence

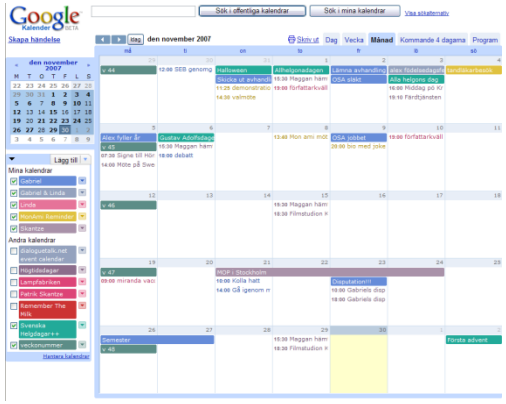
LARGE IP – 4 years

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

Google Calendar interfaces

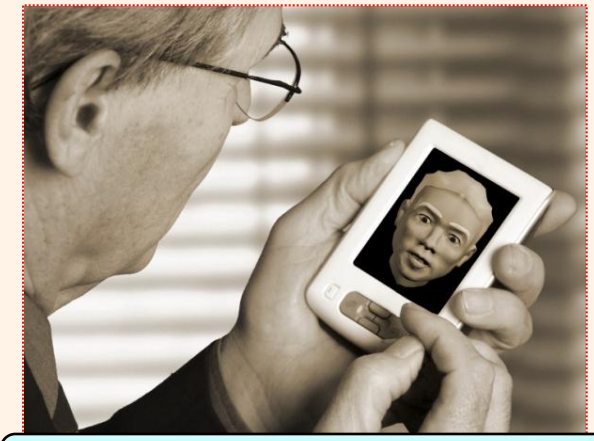


SMS notifications

Web GUI

Google Calendar

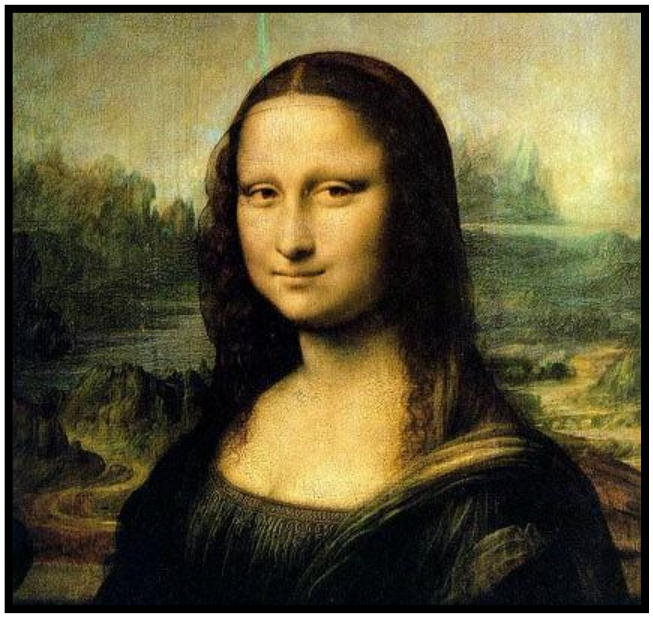
Handwriting



ECA

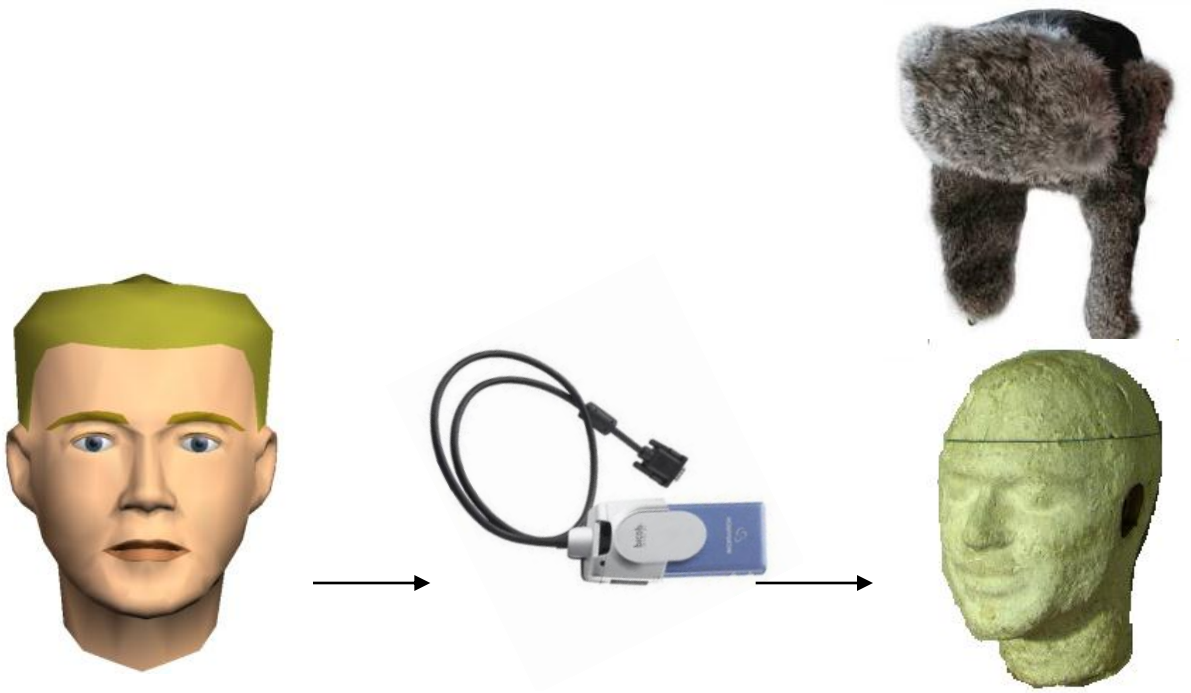
Tutoring session in the MonAMI domain

The Mona Lisa Effect



- The Mona Lisa gaze effect is when all viewers perceive the same gaze direction, independently of where they are seated.
- Good when all should see the same scene
- Also applies to avatars presented on flat screens
- Confusing in situated multi-party interaction

Proposal: 3D Displays for Avatars



Animated face

Laser projector

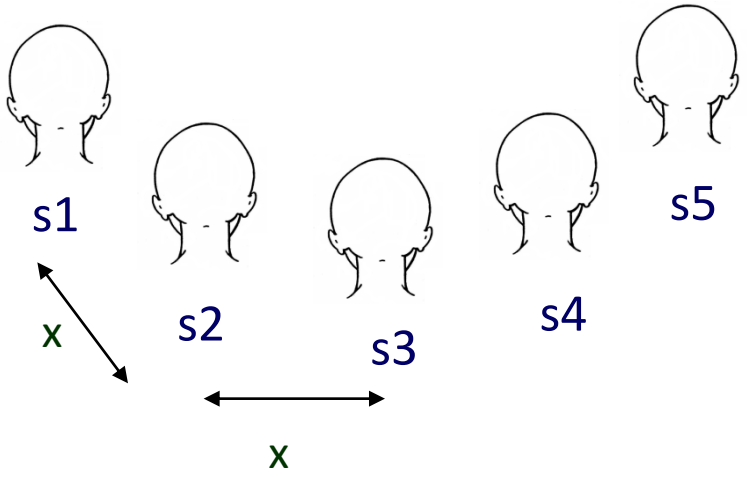
Static head model

Investigating Gaze Accuracy with 3D Projected Avatars

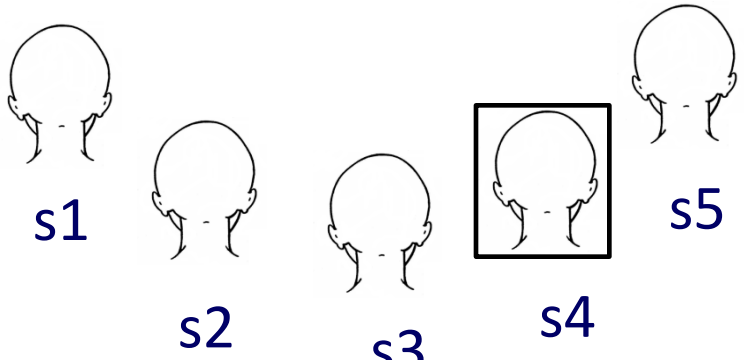
Experiment setup

- Two display methods:
  - Projection on 2D surface
  - Projection on 3D head model

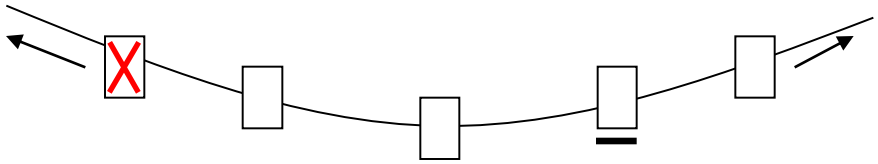
- 5 Subjects, seated at equal distance from each other. with equal distance to the display.
- Gaze varied in 20 angles.
- Five runs with each subject in all positions.
- Task: Decide who is being gazed at.



Question: On whom is he looking?

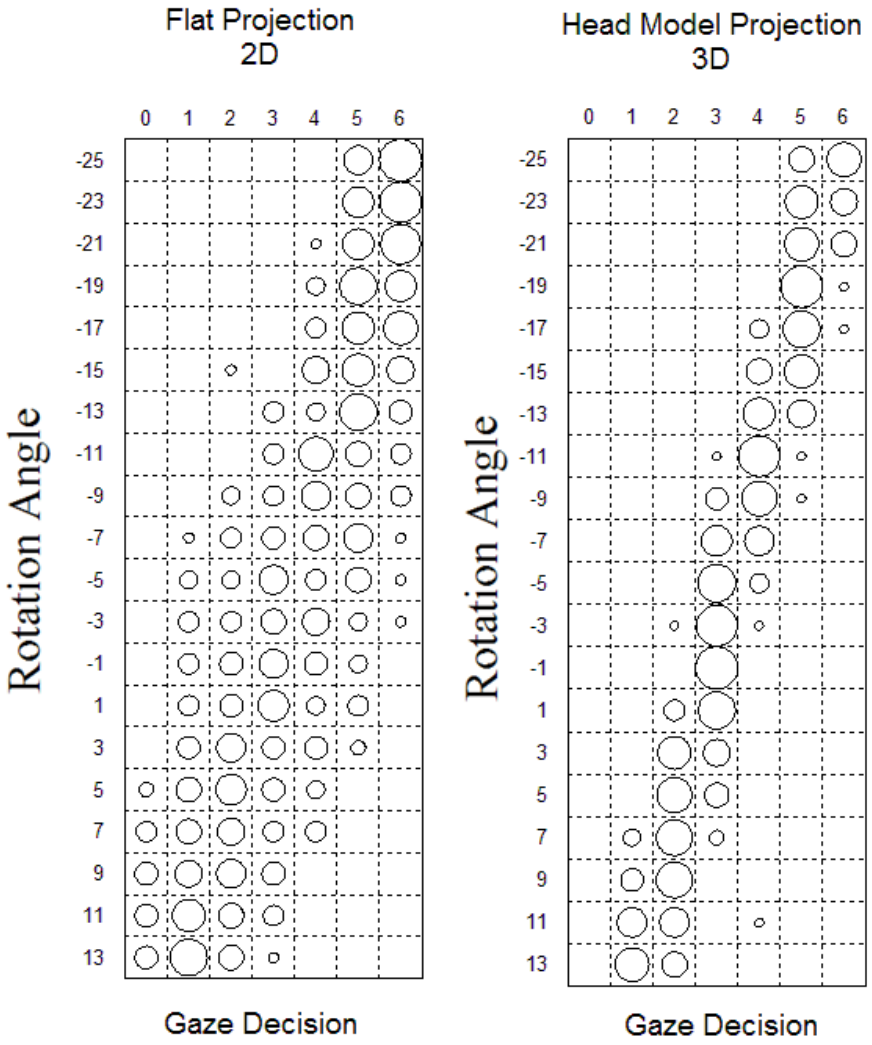


Answer sheet:



Decision encoding: 0 1 2 3 4 5 6

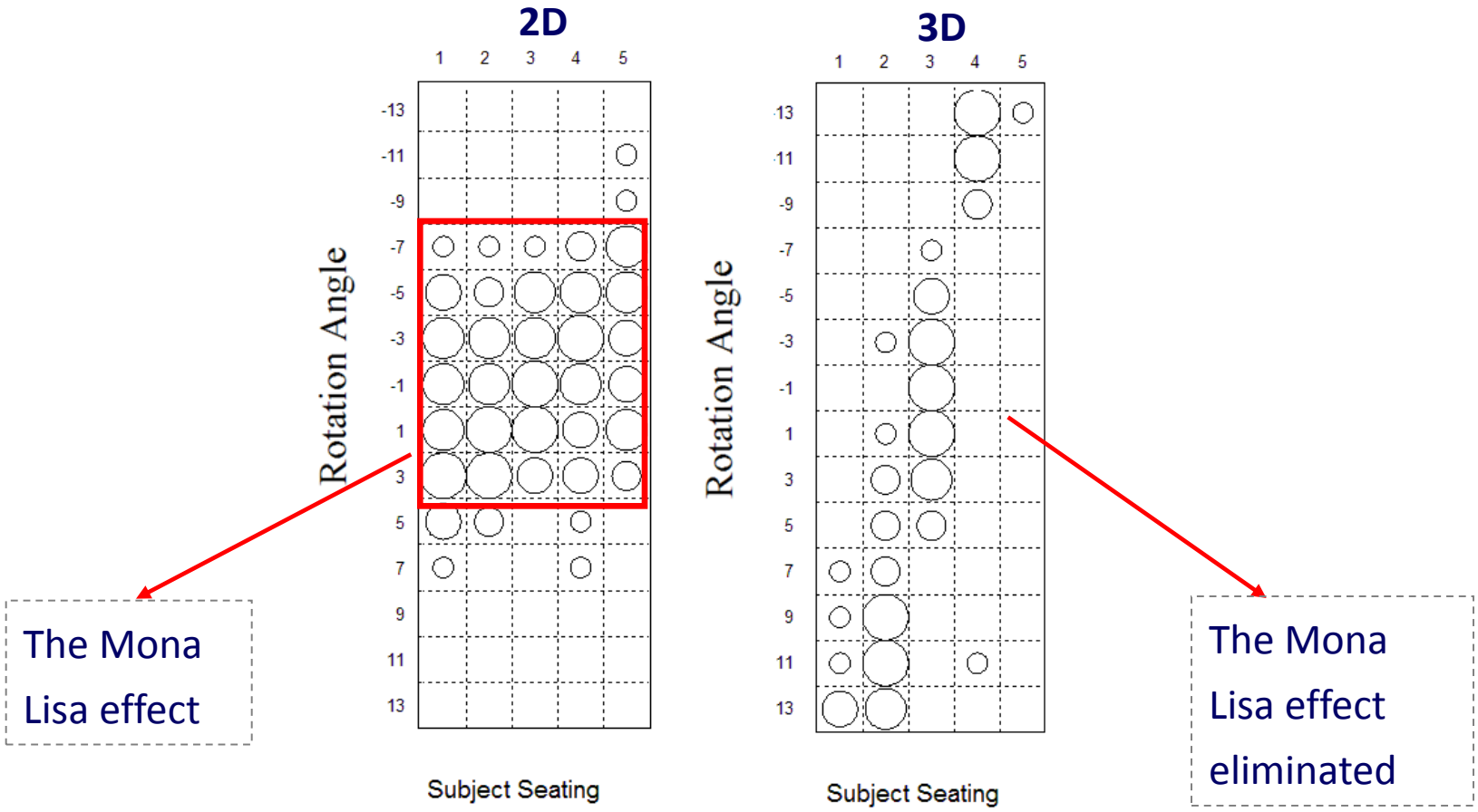
Results: Gaze Accuracy



Raw Data Plot

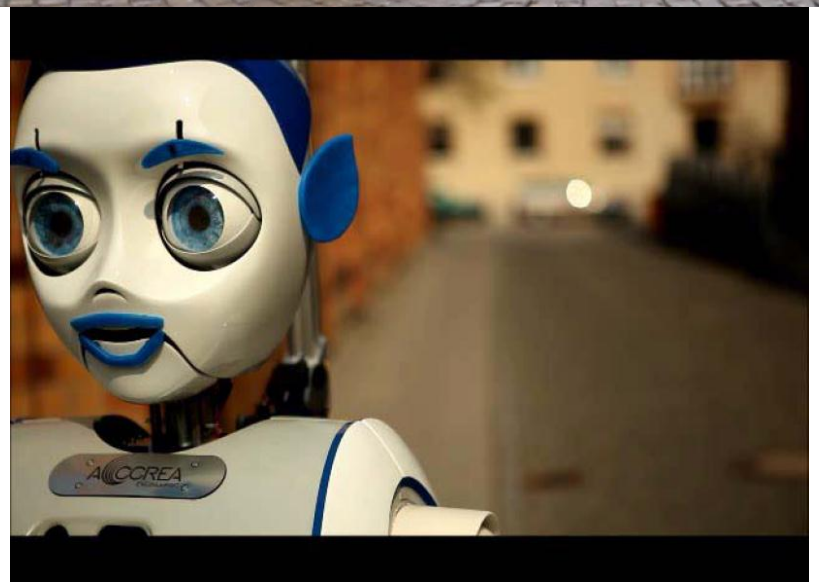
The Mona Lisa gaze effect in the data

Only looking at data when eye-contact is reported



The EU/IURO project 2010-2013

The goal of the Interactive Urban Robot (IURO) project is to develop and implement methods and technologies enabling robots to navigate and interact in densely populated, unknown human-centred environments and retrieve information from human partners in order to achieve a given navigation or interaction goal.



COOPERATION

TUM ETH

UNIVERSITÄT SALZBURG

KTH VETENSKAP OCH KONST

ACCREA ENGINEERING

IURO Interactive Urban Robot

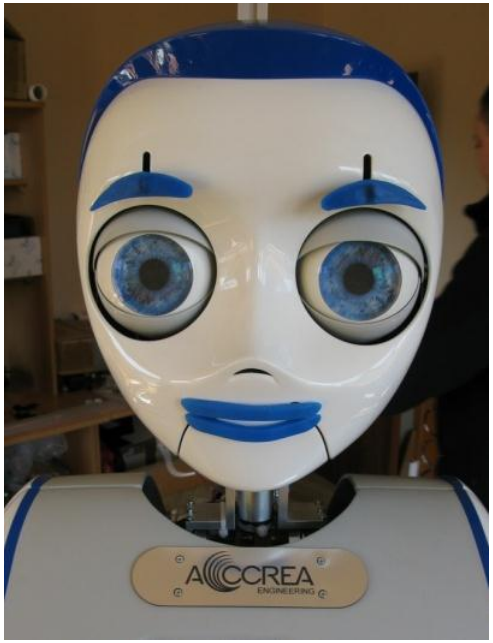
IURO Research Issues at KTH

- How can a robot extract information from people?
  - How can we encourage people to speak?
  - How can we interpret spoken language into route graphs?
  - How can we deal with partial understanding?
- How can the robot handle multi-party dialogue?
  - Multiparty turn-taking, competition for the floor
  - Gaze, head pose and turn-taking gestures

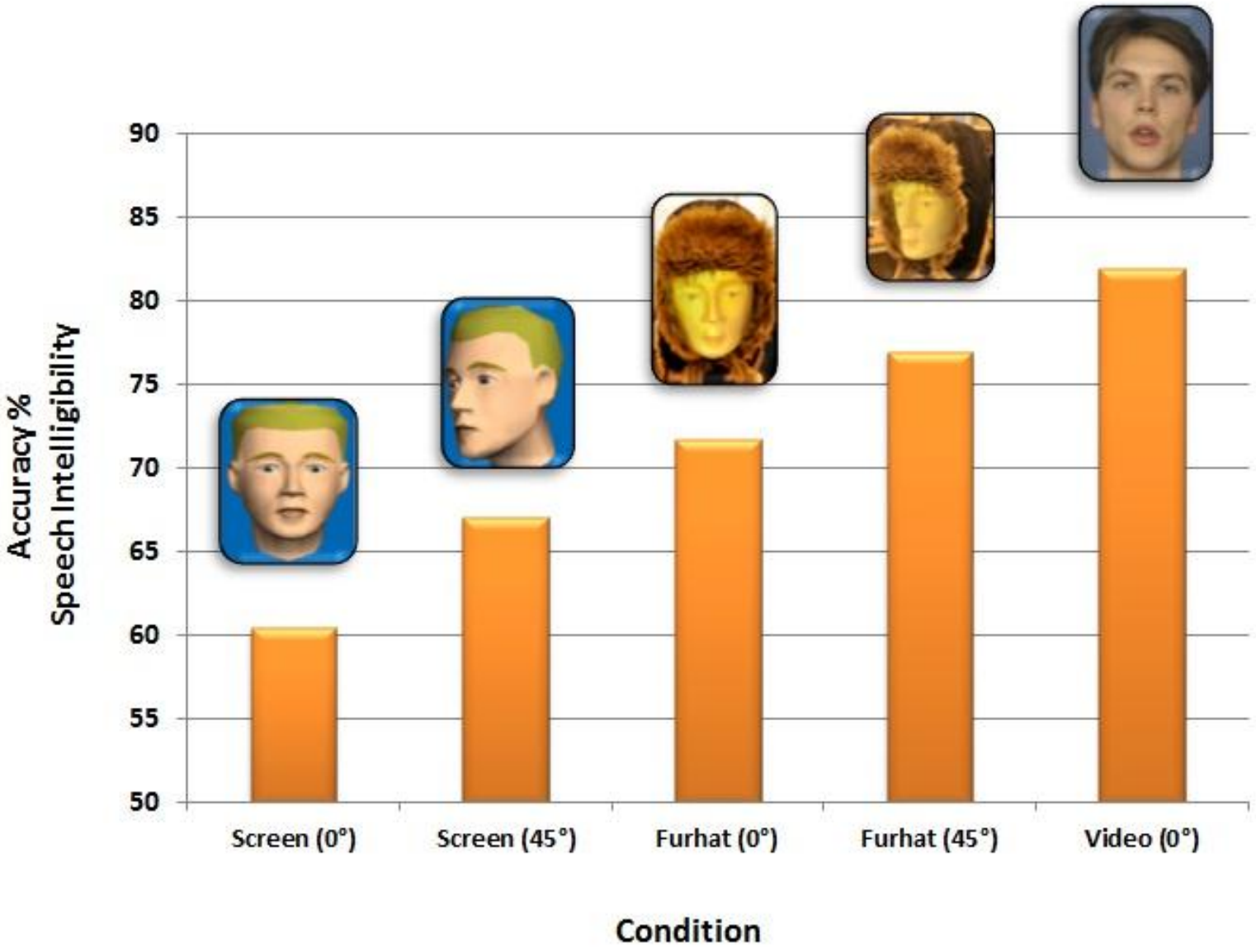
The IURO project investigates two head solutions

A back-projected face

A mechatronic face



Developing the back-projected robot head FurHat



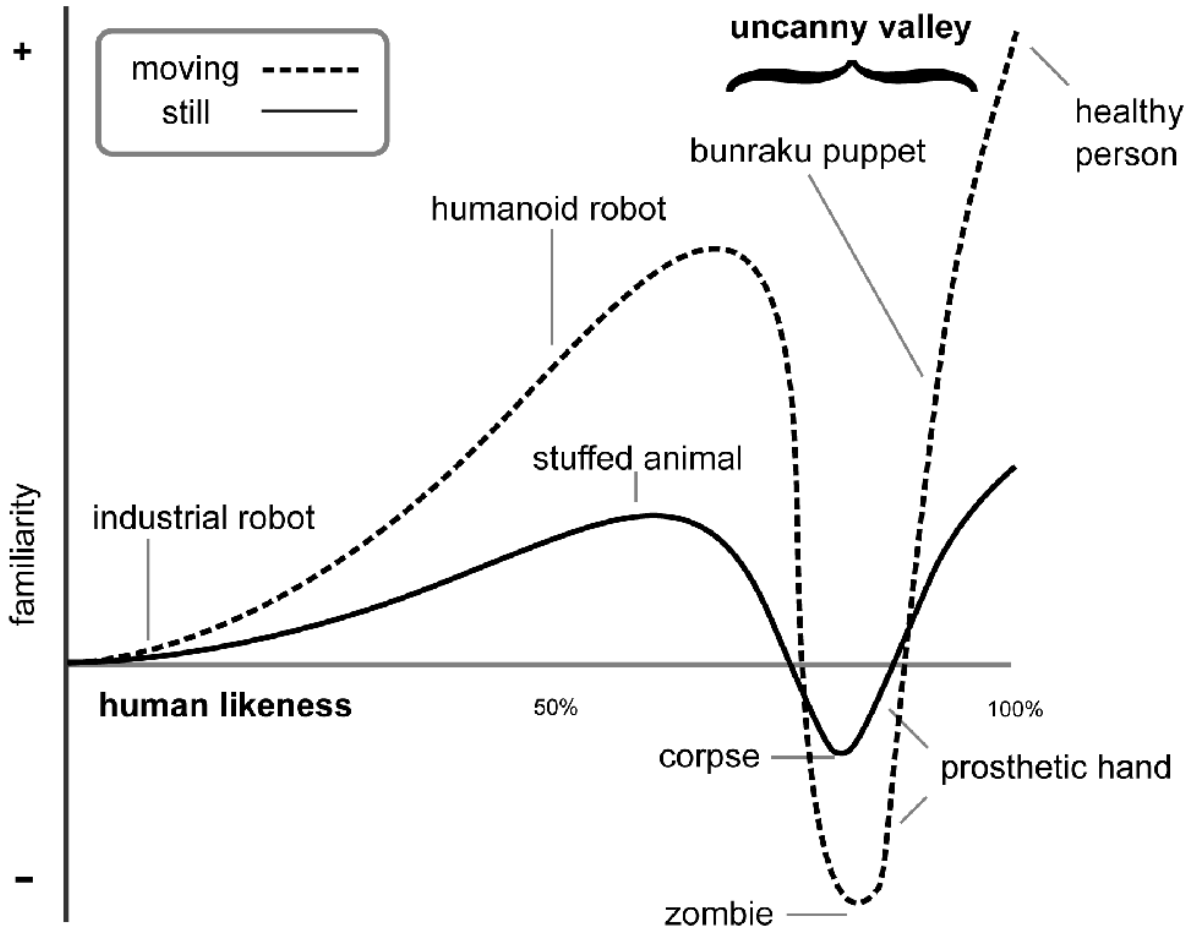
Robot Head for
Notes in Computer

FurHat's usage of head and gaze direction

- *Indicate idle/no contact* by looking down
- *Establish contact* by turning head and gaze to newly arrived
- Put newly arrived on hold while in conversation by giving the new visitor a brief glance while saying “please wait a while!”
- Deflecting understanding problems by turning head and gaze towards the other user and ask "What do you think about that?"
- *Pose open question* to both visitors by directing the head straight ahead, then switching the gaze between them
- ”Pointing” with gaze

Multi-party dialogue test at the London Science Museum:

Did we experience the "uncanny valley"?



Remember what happened to the answering machine.

Bringing the avatar to life

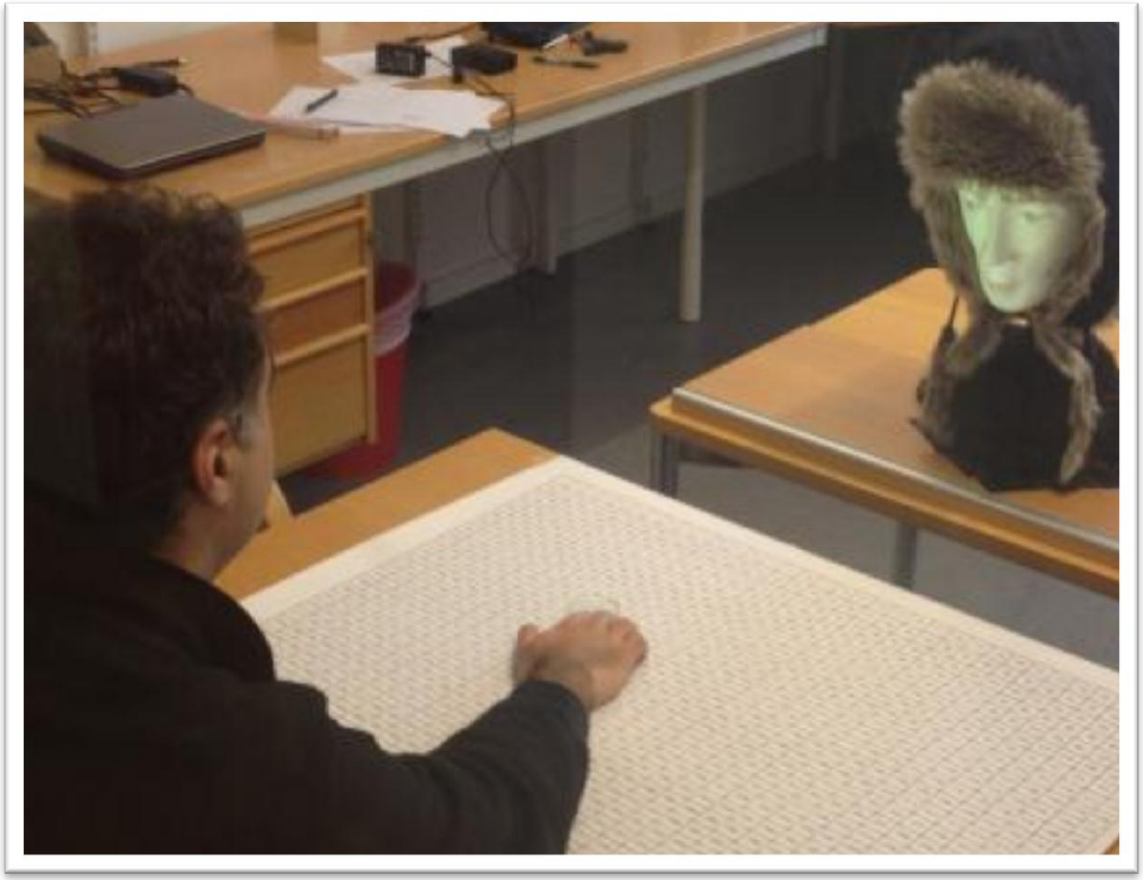
Studies and developments in facial communication for virtual agents and robots

SAMER AL MOUBAYED

KTH Computer Science and Communication

Doctoral Thesis in Speech Technology and Communication
Stockholm, Sweden 2012

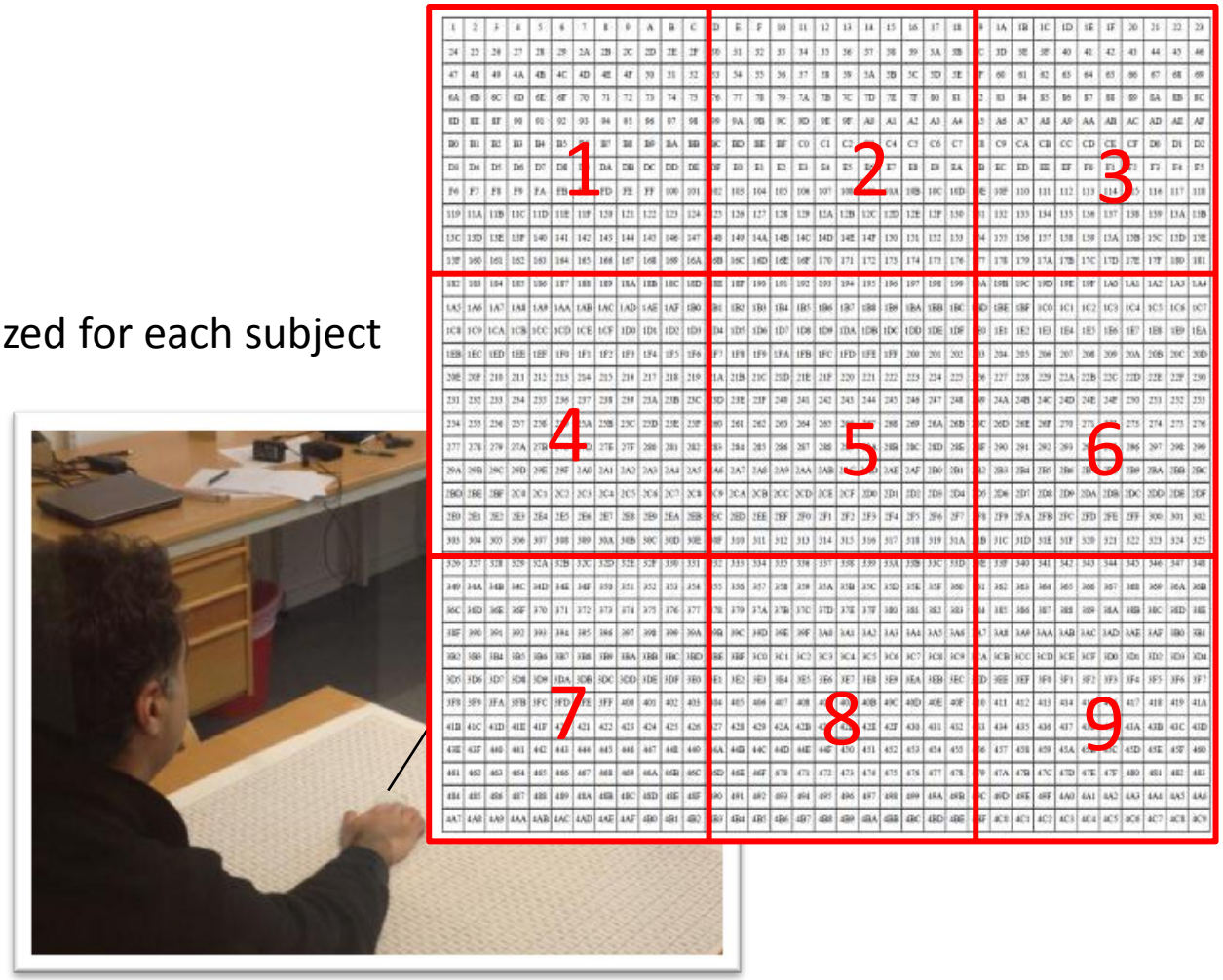
Gaze and head pose for pointing



Experimental setting: Human-agent collaborative (task solving) scenario

Task: indicate where the agent is "pointing"

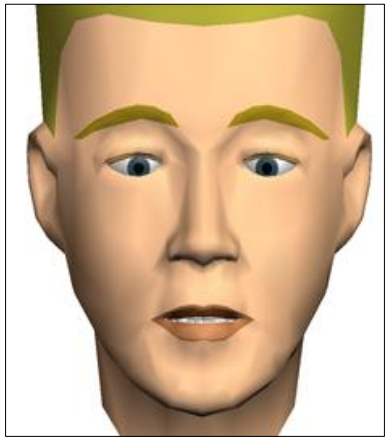
- 18 subjects
- 9 gaze targets per condition
- Conditions * targets randomized for each subject



Gaze vector was estimated using a highly detailed geometric model

Human vs. agent conditions

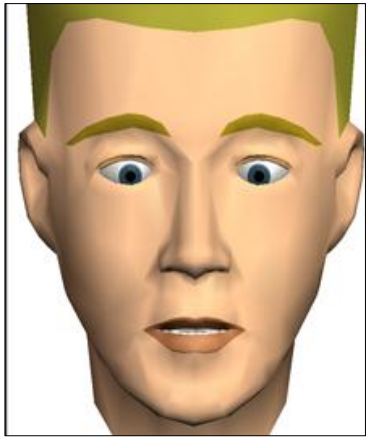
Human-human condition



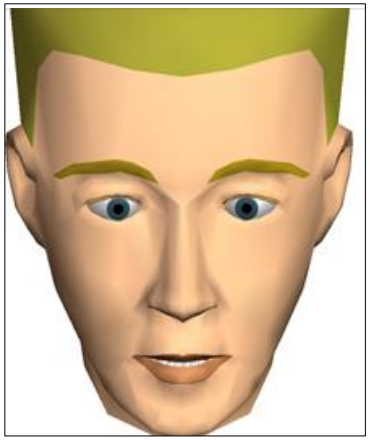
The *normal* condition (baseline)

- Static head, moving eyes
- Eyelids follow the eyeballs vertically
- Subject sitting in front of the head

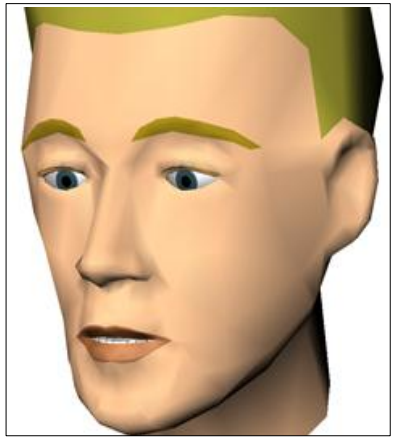
Three more, "anomal", conditions



Eyelids condition
Static Eyelids
Static head
Moving eyes

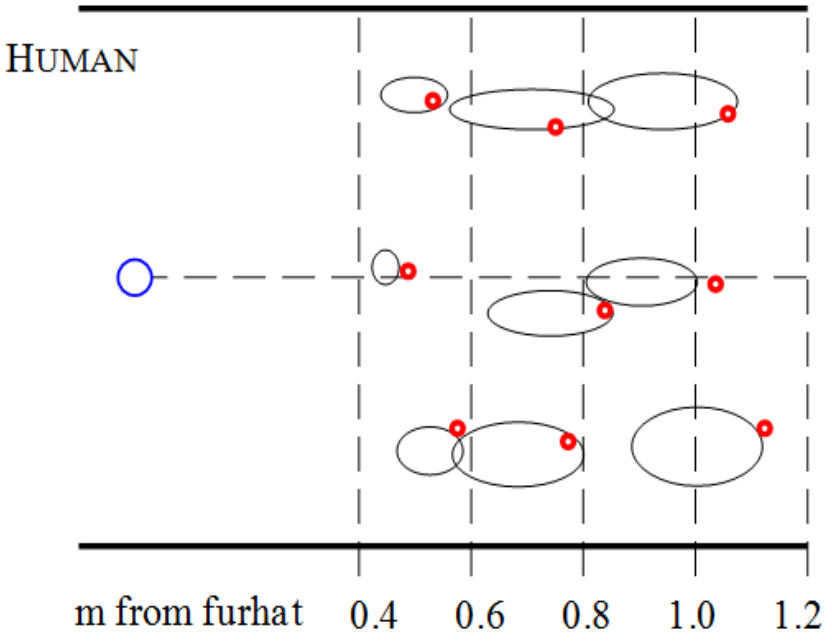


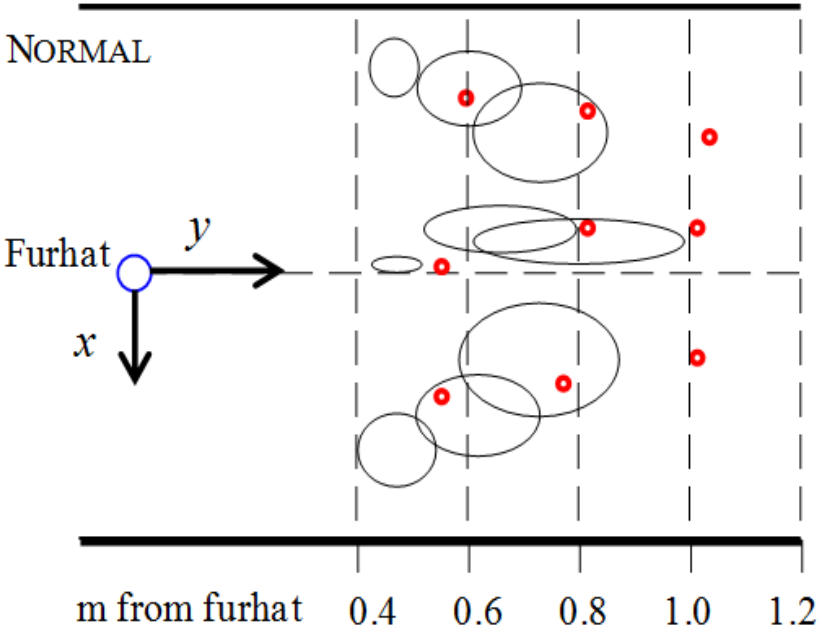
Neck condition
Neck moving
Eyes frontal



Side condition
Subject seated at 45 degrees from the head.

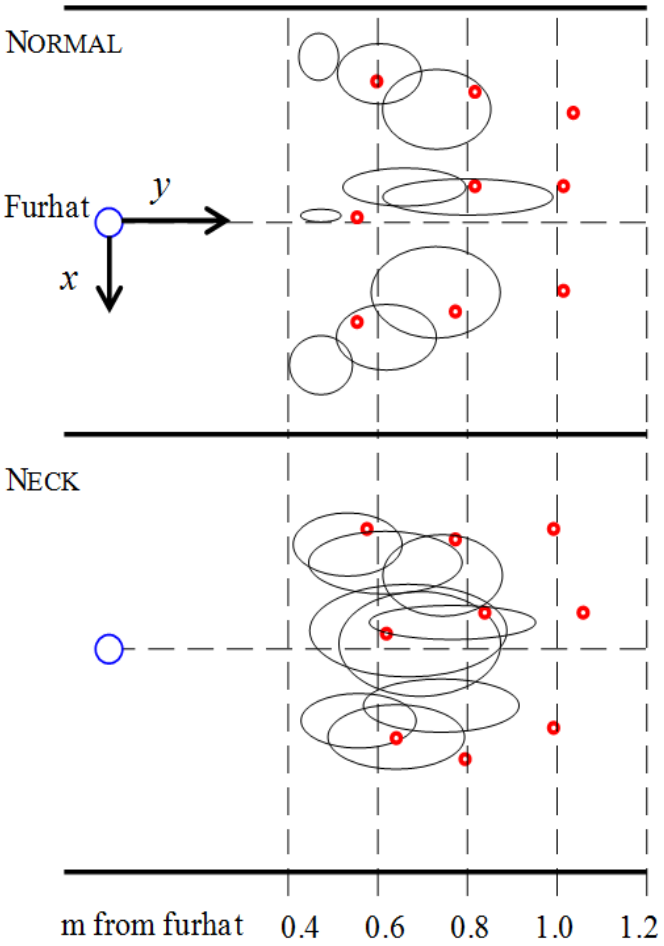
Results





dots: intended target points
ellipses: distributions of answers on x and y

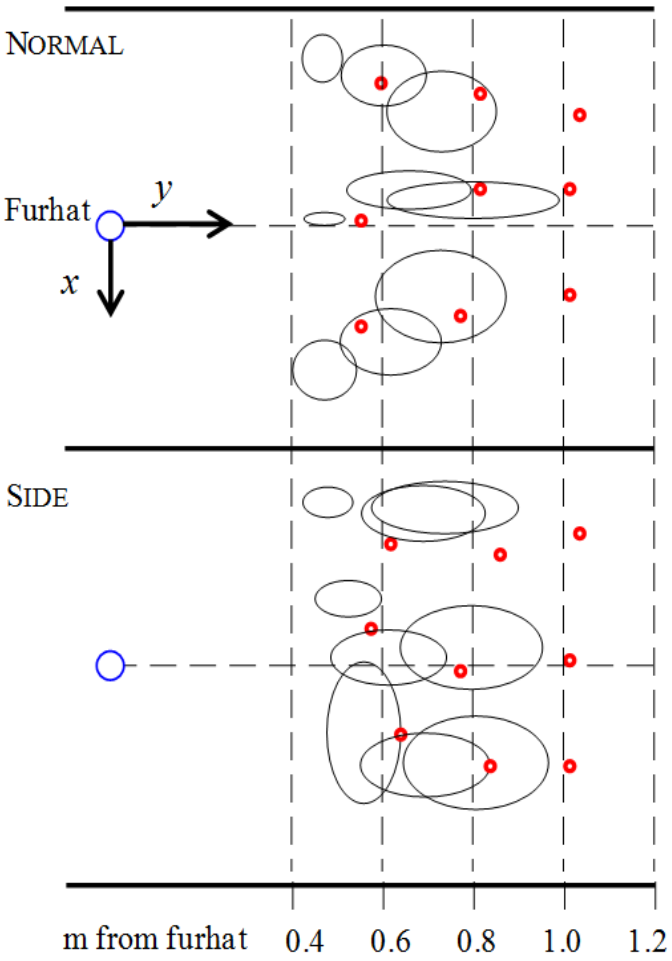
Results



Eyelids: Perceived target is 1.47 degrees lower.

Neck: Resolution very bad on the *y* axis (5,86). Displacement on *x* (7.1) and *y* (2.3).

Results

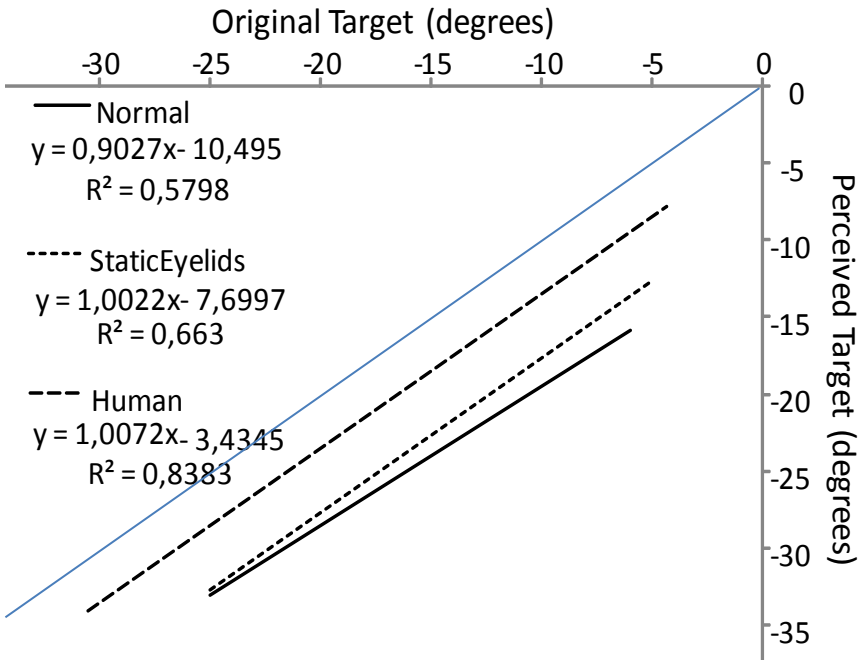


Eyelids: Perceived target is 1.47 degrees lower.

Neck: Resolution very bad on the *y* axis (5,86).
Displacement on *x* (7.1) and *y* (2.3).

Side: resolution lower on *x*
Displacement on *x*

The setup serves as a gaze calibration method



Psychometric functions for three of the conditions on the *y*-axis.

Experiment on "pointing" and interaction
Map task – gaze and turntaking

Mapping FurHat to the IURO mechatronic head
Tällberg forum domain

Possible non-robotic applications in mediated communication

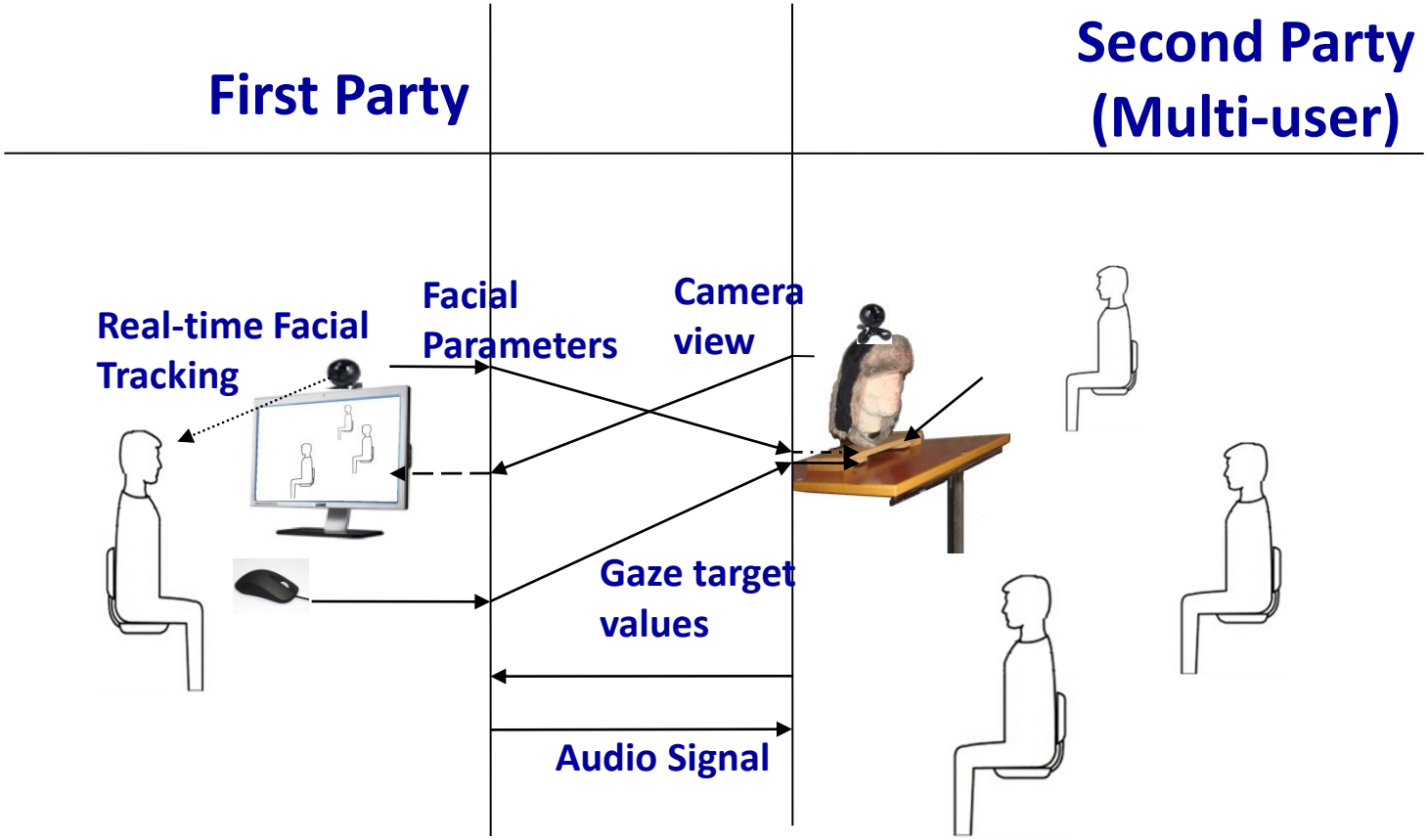
Standard video conferencing setup

No shared access to environment

No ability to have exclusive eye-contact (the Mona Lisa effect)

User always looks at the video instead of the camera (no shared attention)

Proposal: use a 3D head for the remote subject



Tele-presence on one side

In summary:

- Something Olde, Something New, Something Borrowed, Something Blue……

- Due thanks go to all my co-workers at KTH, in particular:
  Samer Al Moubayed
  Jonas Beskow
  Jens Edlund
  Joakim Gustafson
  Anna Hjalmarsson
  Gabriel Skantze

Furhat at our Christmas party

The End

Questions?

Find me on facebook

www.facebook.com/furhat.svensson