Humanoid Robot Heads

Past, Present, and Future

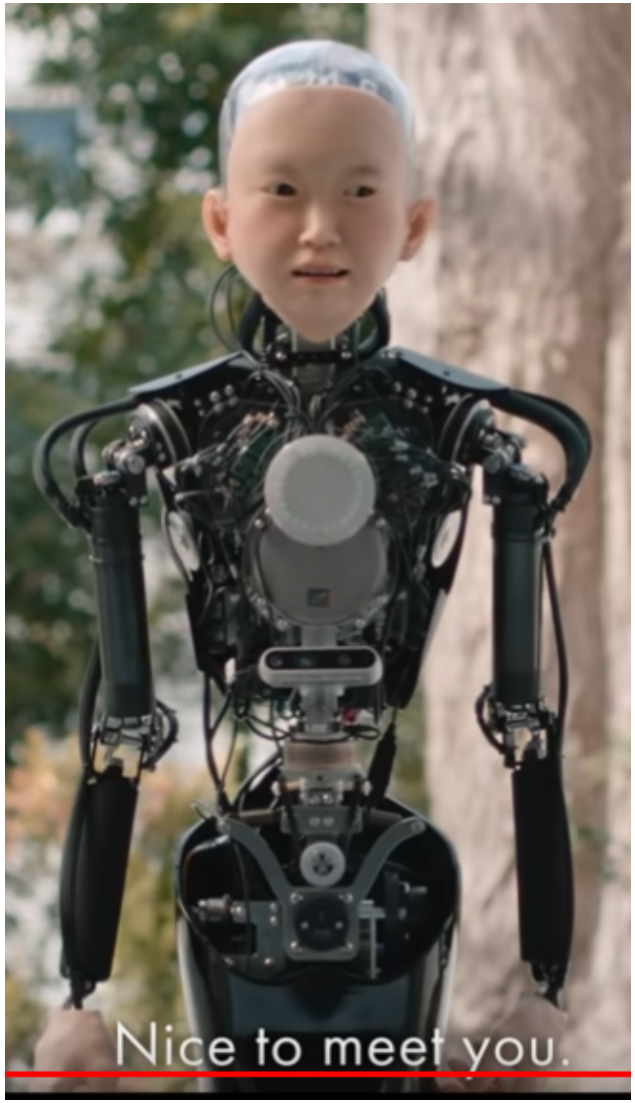
Norman Hendrich

University of Hamburg, Informatics Department
Vogt-Koelln-Str. 30, D-22527 Hamburg, Germany
hendrich@informatik.uni-hamburg.de

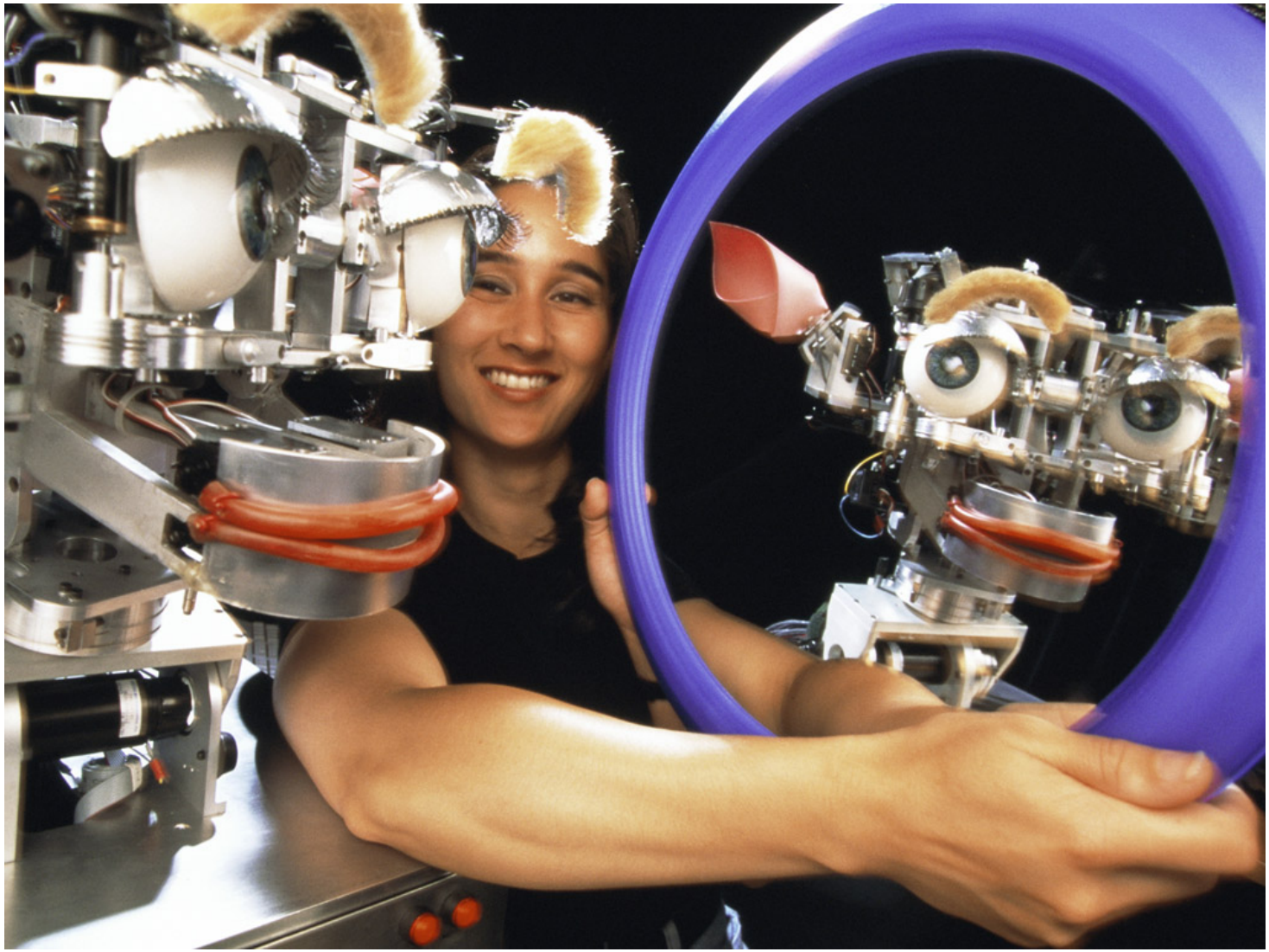
01 November 2022

Outline

- Motivation
- Robot Heads for Social Robotics
- Ultra-realisitic Robot Heads
- What can a Robot Head Do?
- USST Robot Head
- Future Work



How it began: Kismet (Breazeal 1993–2000)



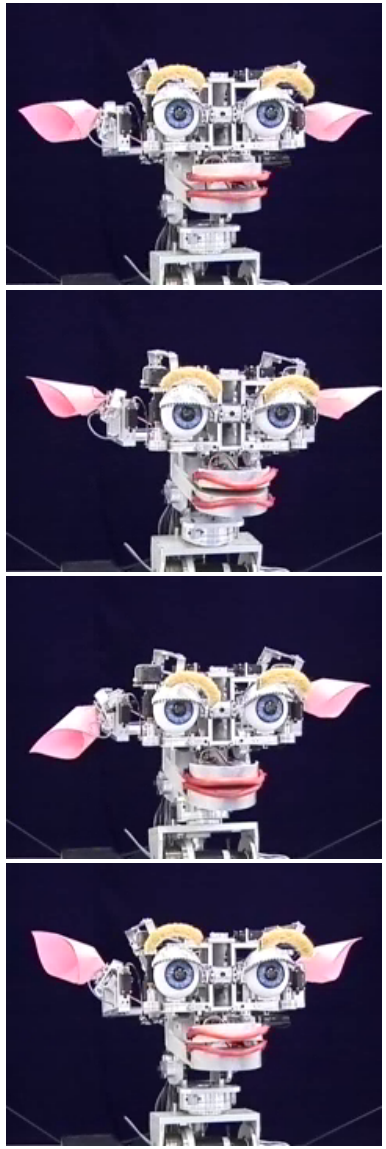
Cynthia Breazeal, Designing Socialiable Robots, MIT Press, 2002

Kismet: The first “Social Robot”

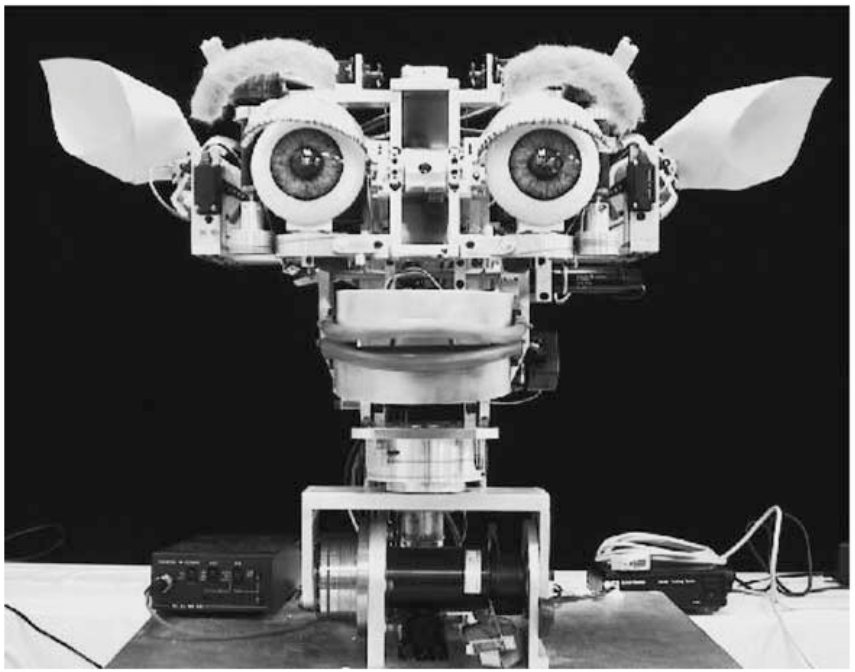
- pan-tilt robot head
- forward/retract head shift
- pan-tilt eyes with eyelids
- expressive mouth, eyebrows, ears
- cameras, microphone, and speaker

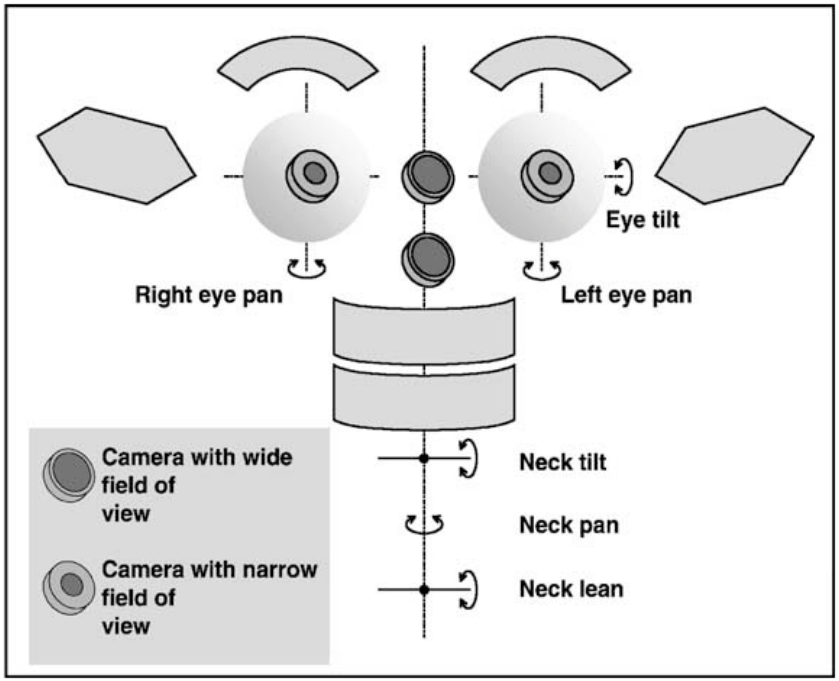
- human-like gaze (e.g. saccades)
- no speech, but expressive audio
- behavior and motivation system

- intentionally, no face skin

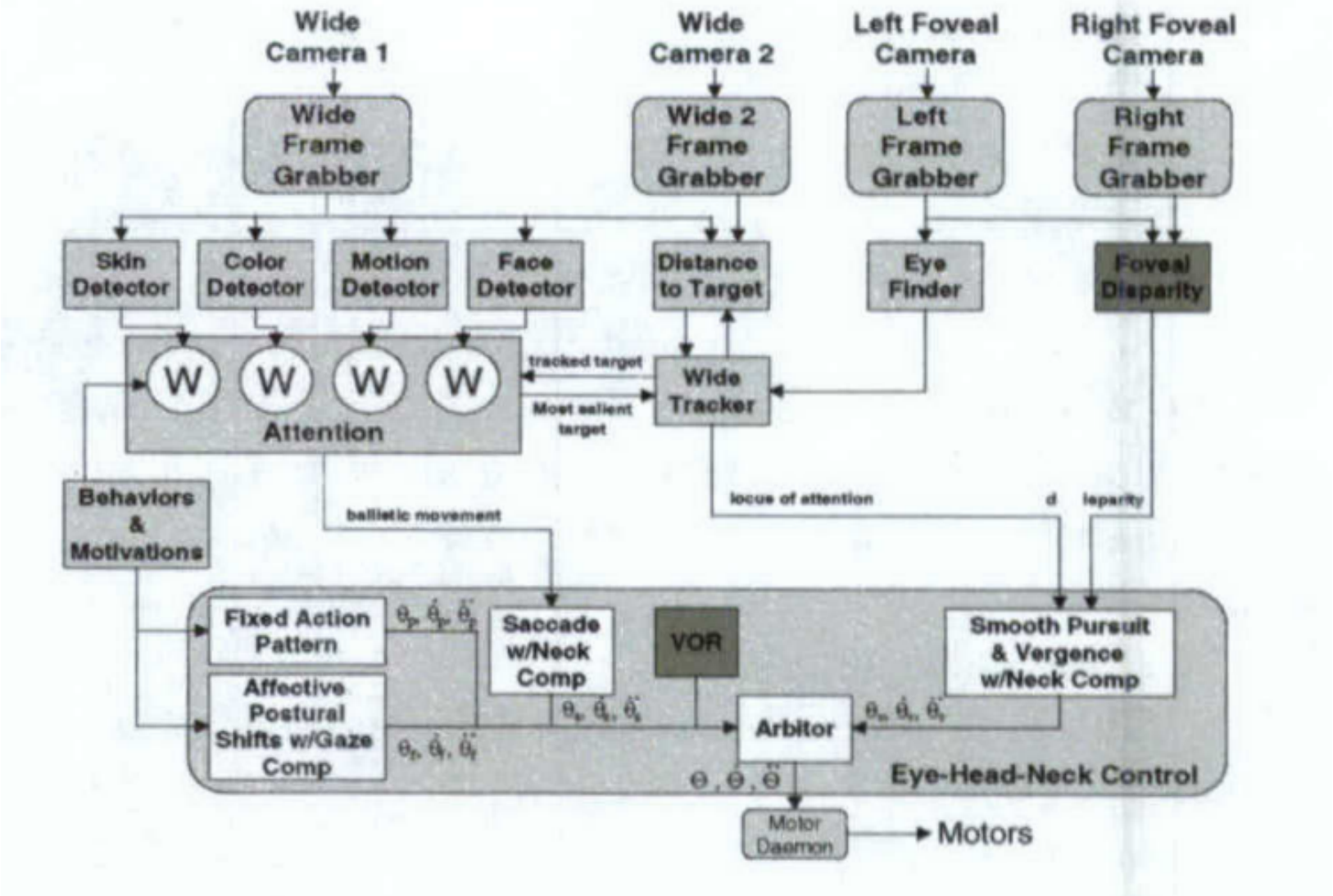


Kismet: Hardware Overview



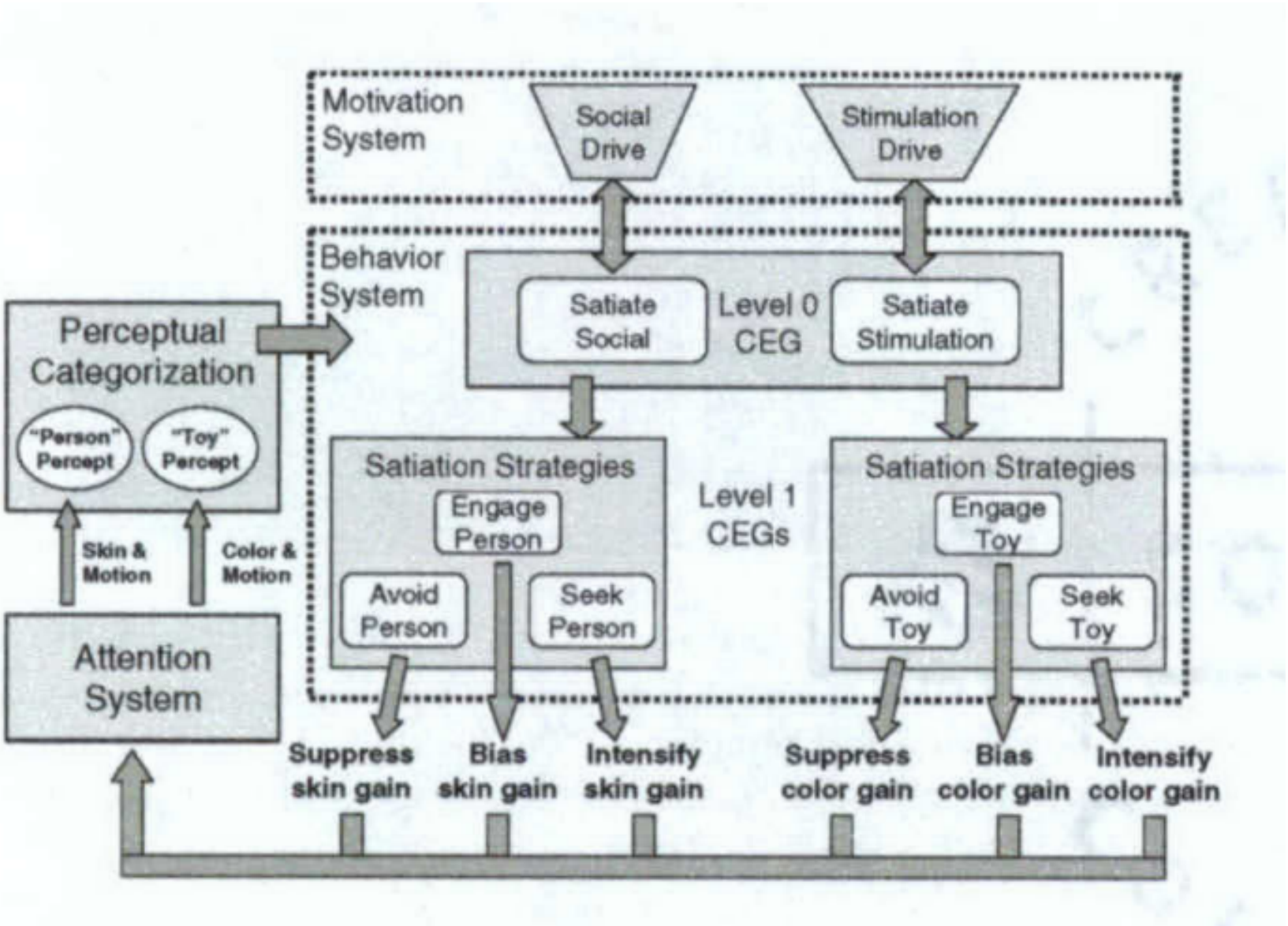


Kismet: Vision- and Motion-Software Architecture



one head, total 15 computers, 4 operating systems (speech system not shown here)

Kismet: Behaviour System



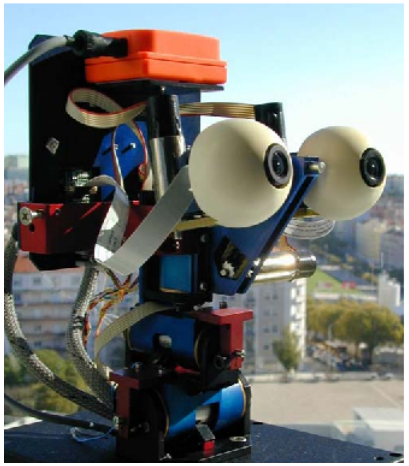
Motivations, attention, novelty + habituation

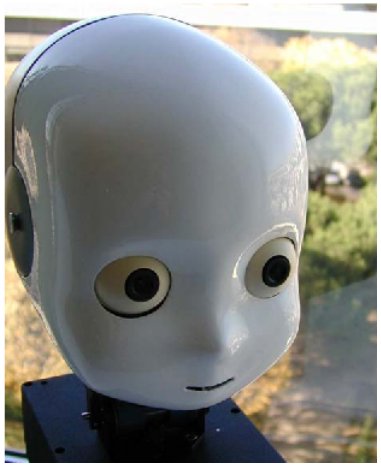
iCub Head (Metta 2006)

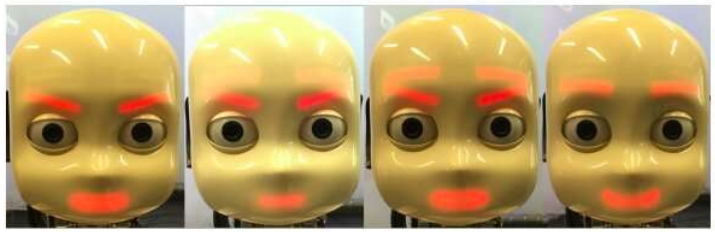
- size of 4-year child
- rigid face mesh

- 3-DOF neck + IMU
- 3-DOF eyeballs (i.e. vergence)
- 1-DOF eyelids (version 2)
- LEDs for 7 basic emotions

- YARP middleware
- price 40.000 EUR





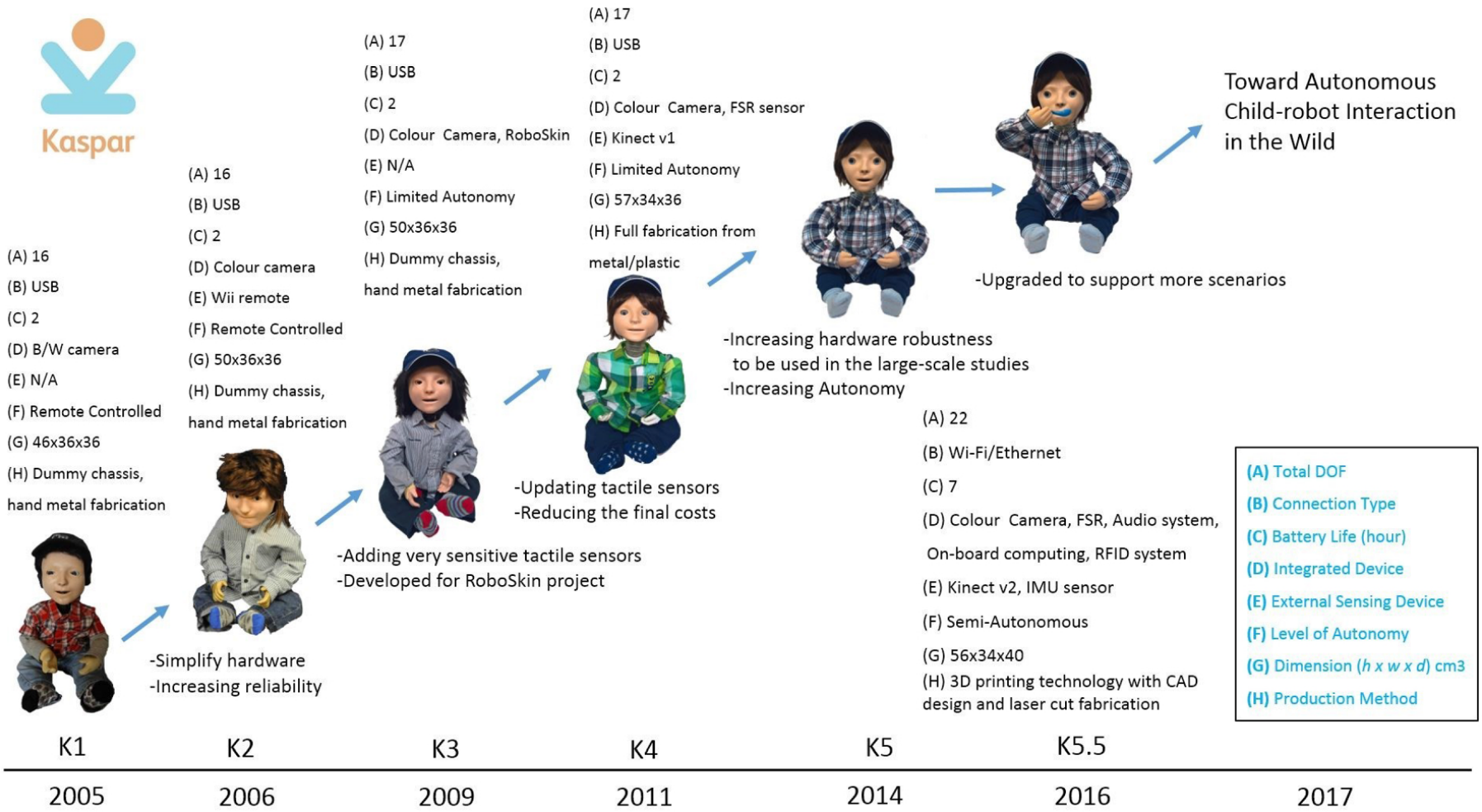


Beira et al., Design of the Robot-cub (iCub) Head, ICRA 2006, www.iit.it/web/icub/products/product-catalog

Kaspar (Dautenhahn 2006)

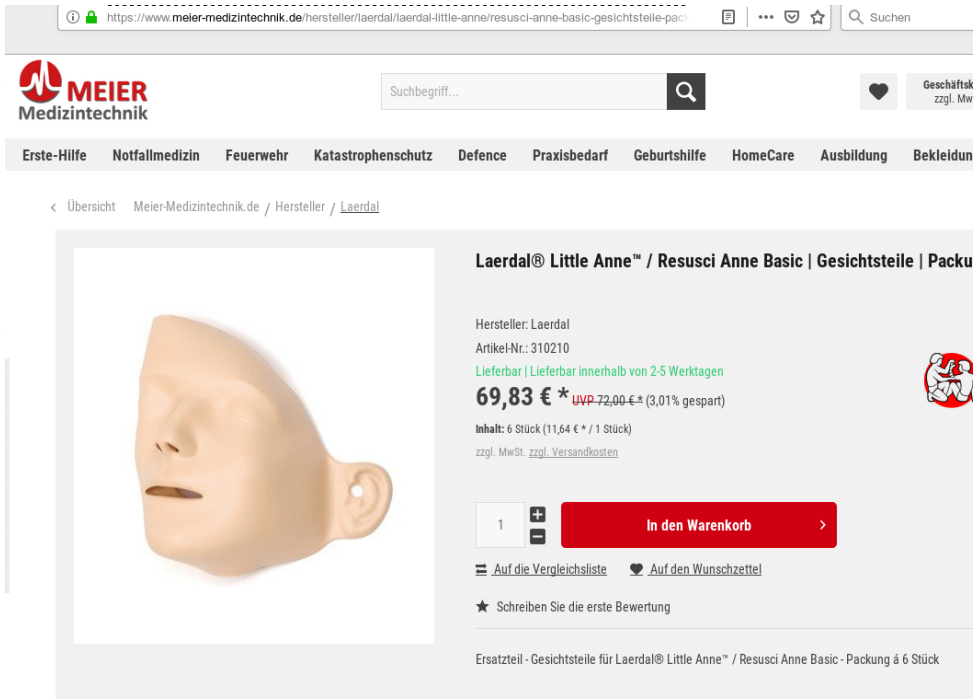
Low-cost design, portable, head with flexible skin, 16-DOF, targeting HRI-studies with autistic children
M. Blow, K. Dautenhand et al., The art of designing robot faces: ACM conference on HRI, 2006

Kaspar: Evolution 2005–2019



Wood et al., Developing Kaspar: A Humanoid Robot for Children with Autism, IJSR 2019

Leardal Resusci Anne

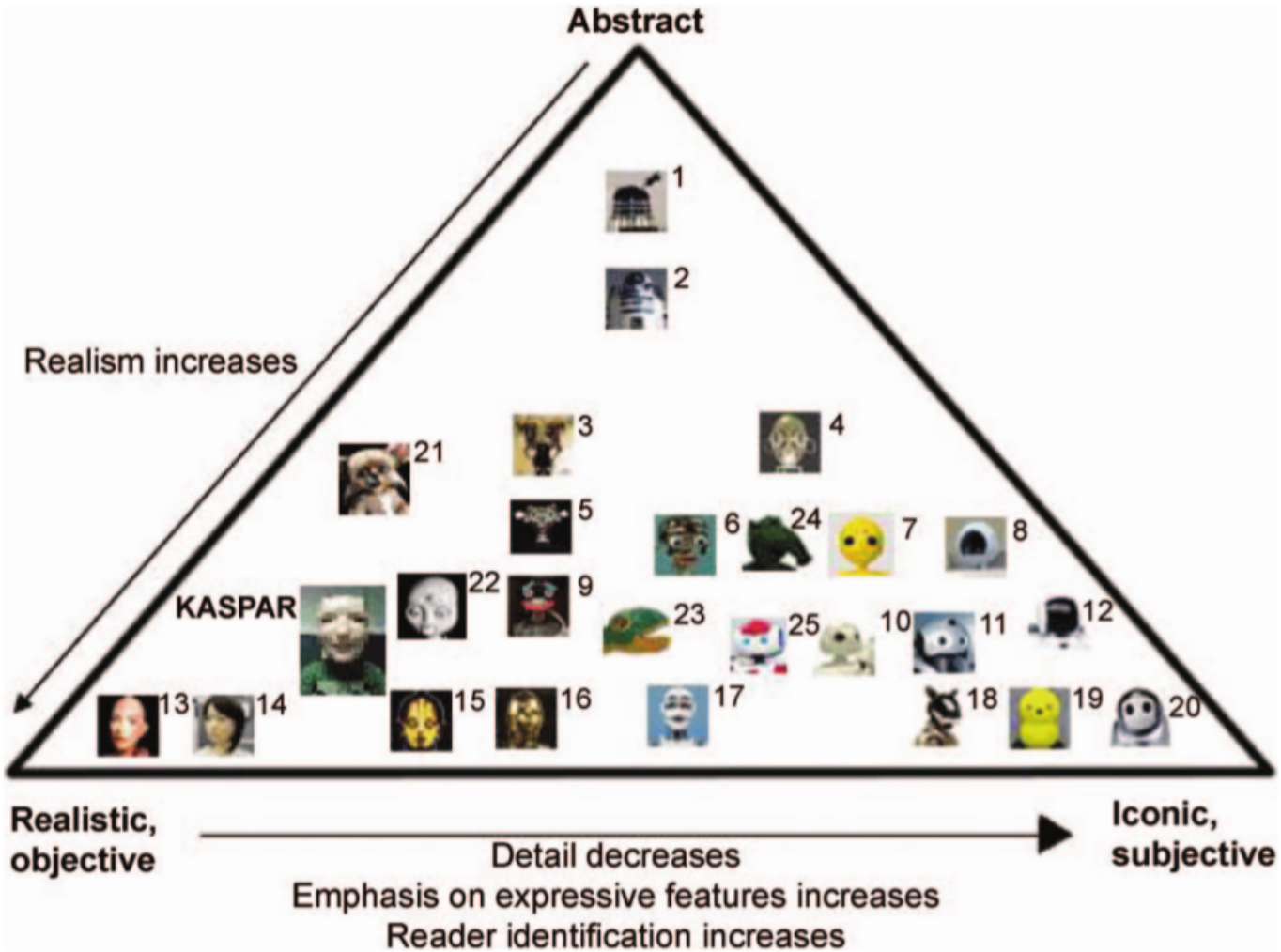


- key hardware part for Kaspar: affordable silicon face
- perhaps we should order a few of these?

Medical training, mouth-to-mouth breathing and intubation, exchangable face masks, used for Kaspar

Social Robots Spectrum (Dautenhahn 2009)

Applied Bionics and Biomechanics



Flobi (CITEC 2010)

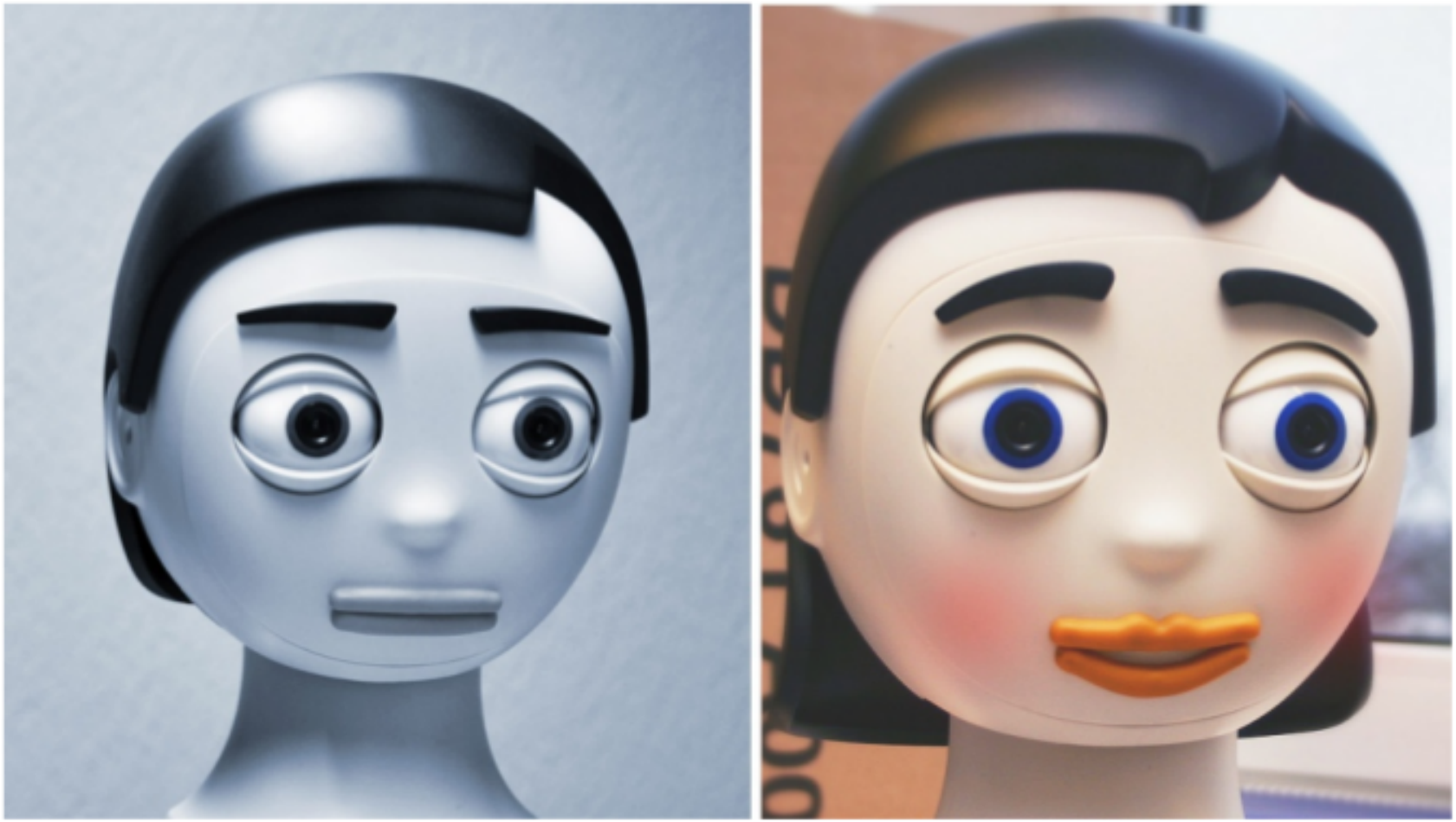
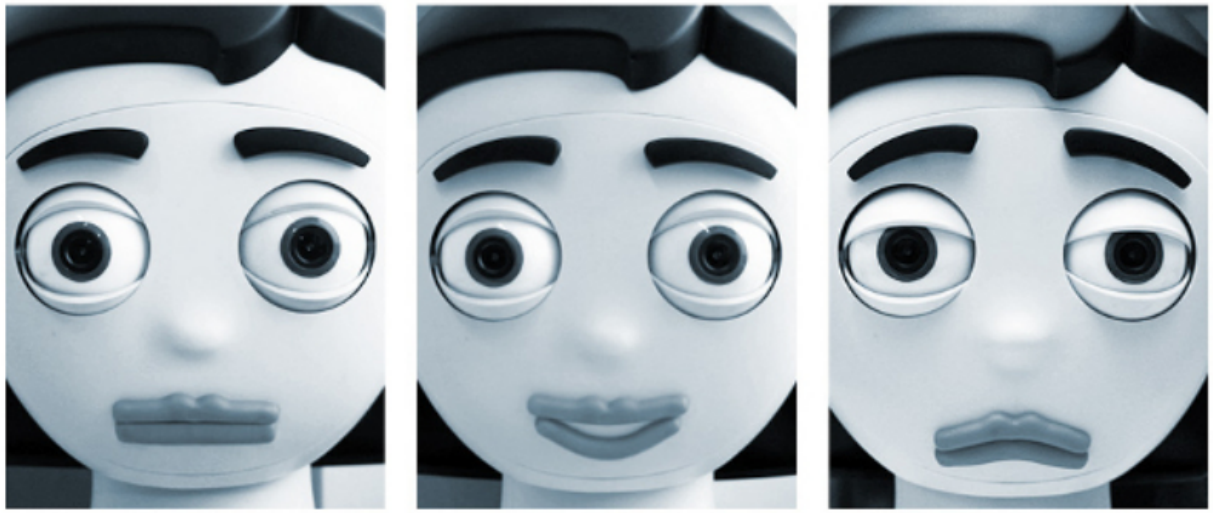


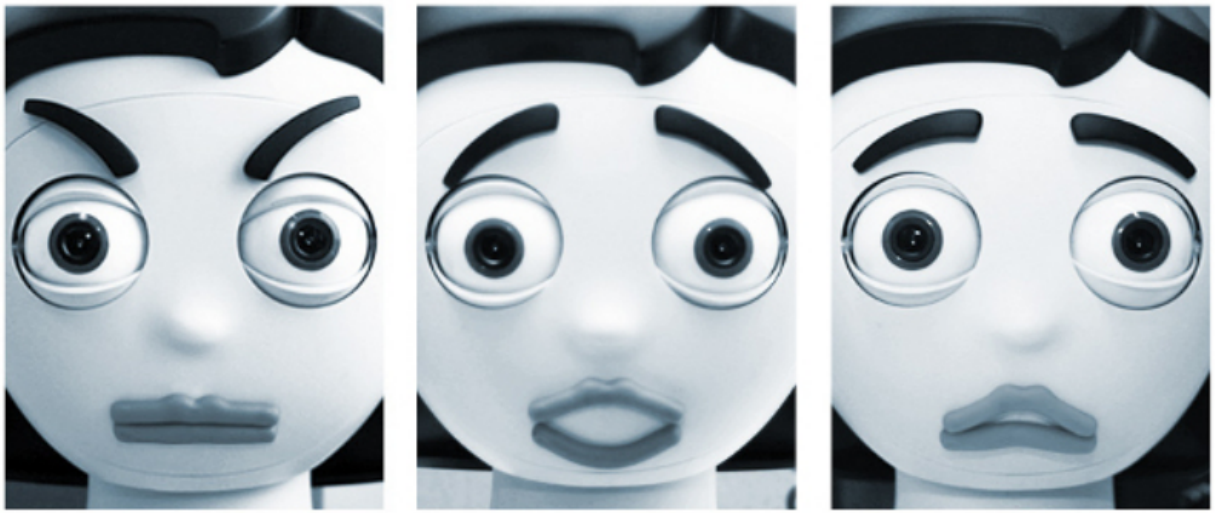
Fig. 1. Flobi male-neutral (left) and female-smiling (right) configurations.

modular head, face+neck+hair+mouth can be swapped at runtime (magnetic)
Lütkebohle et al. (Wachsmuth), The Bielefeld Anthropomorphic Robot Head 'Flobi', ICRA 2010

Flobi: Six basic emotions



Neutral | Happiness | Sadness



Anger | Surprise | Fear

Gruselkabinett (from Flobi paper 2010)

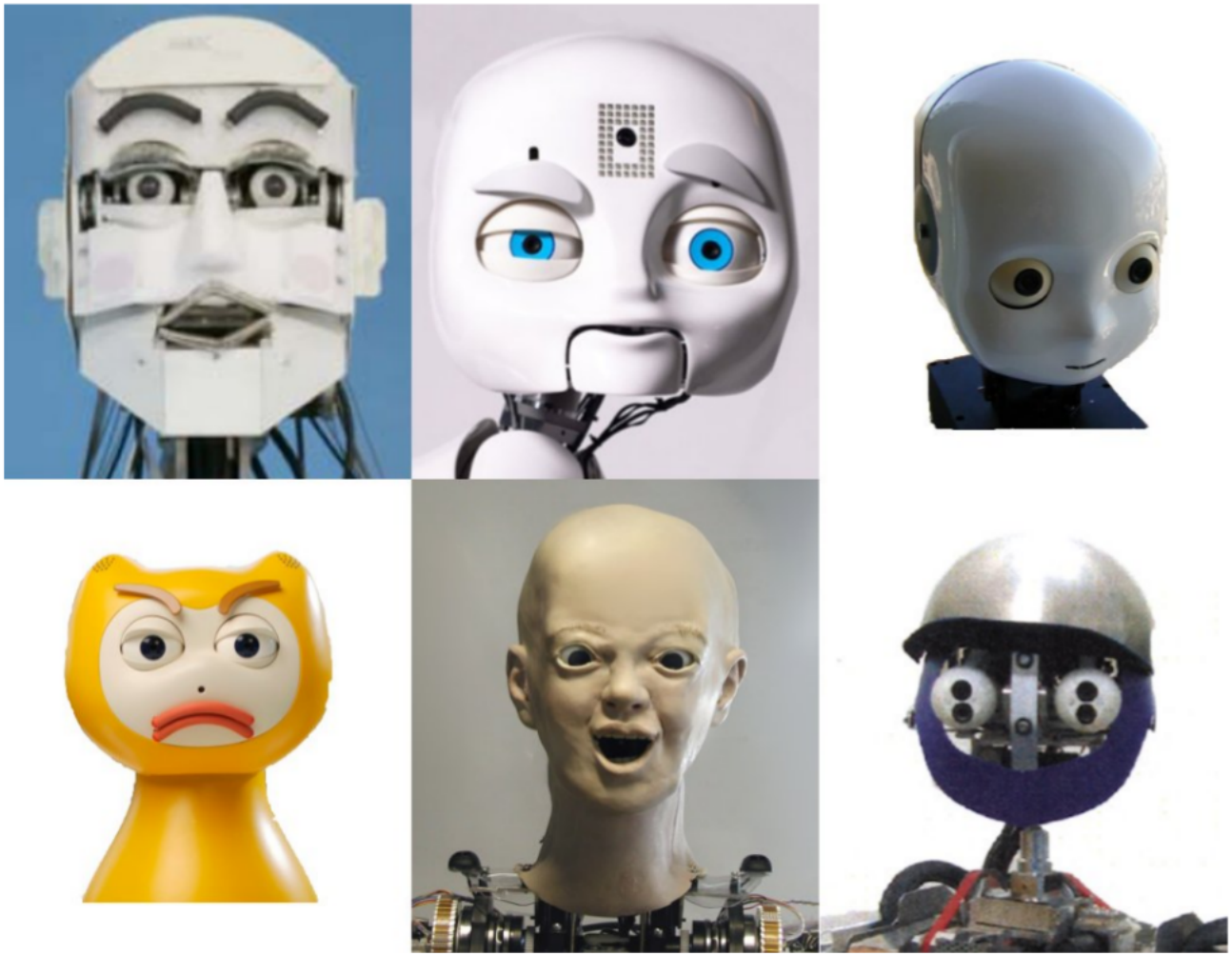


Fig. 2. Heads for comparison. From left to right and top to bottom: WE-4RII, Nexi, iCub, iCat, BARTHOC, Karlsruhe Humanoid Head.

Head Comparison

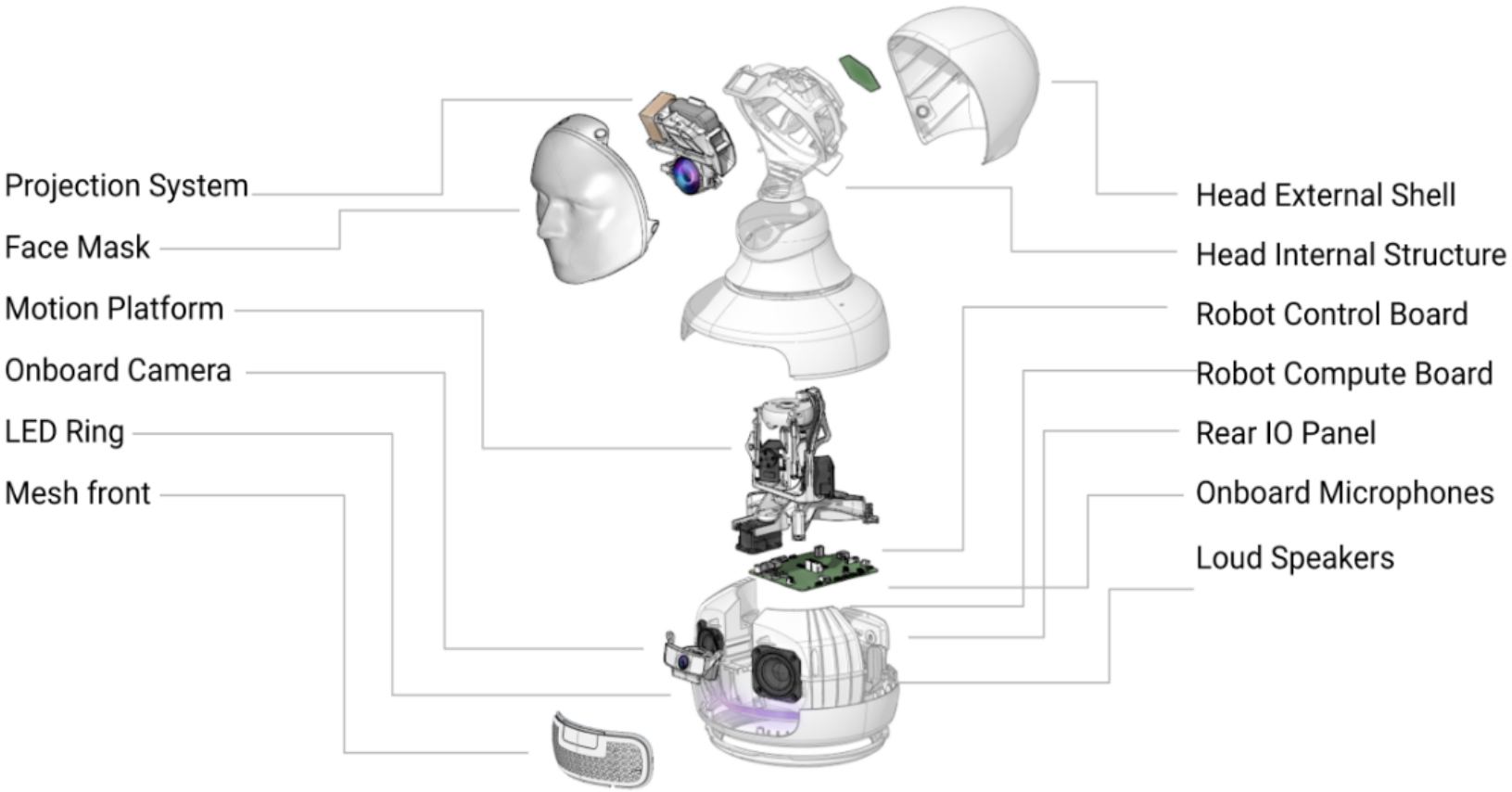
	WE-4RII	iCub	Nexi	ICAT	BARTHOC	KHH
Breadth (cm)	18.6	15.2	~24[1]	18	14	
Eye DoF	3	3		0	3	3
Neck DoF	4	3		2	4	4
Eyebrow DoF	8	LED	4	2	2	n/a
Eyelid DoF	6	n/a	4	4	2	n/a
Mouth DoF	5	LED	1	2	3	n/a
Stereo Vision	yes	yes	yes	no	yes	2x
Stereo Audio	yes	yes		yes	no	6ch
Gyroscope	yes	yes		no	no	yes
Eye pan (°/s)		180				
Eye pan range		90				
Apperance	A	A	A	Z	A	T
Tech visible	yes	no	yes	no	no	yes

TABLE I

COMPARISON OF ANTHROPOMORPHIC ROBOT HEADS.

(A - ANTHROPOMORPHIC, Z - ZOOMORPHIC, T - TECHNOMORPHIC)

Furhat Social Robot



furhat.com/

This makes navel different from other social robots

Computing power

Communication

Eye contact

Liveliness

Autonomy

Privacy

Eye contact is the foundation of all social interaction, which only a few social robots besides **navel** have mastered.

navel has eyes that no other robot has! Special 3D optics are mounted above the displays, giving navel real three-dimensional eyeballs. Because with eyes that are only shown on a planar display, no real eye contact is possible.

Because **navel** uses its camera to recognise where the eyes of its conversation partner are, navel can look exactly into their eyes. And because a static stare is unpleasant, navel has natural eye movements that continuously change the focus, including gaze aversion.

Navel robotics website, navelrobotics.com/en/home-en-2/

Navel: Head Expressions and Emotions

Navel robotics website

QTrobot

https://luxai.com/product/qtrobot-for-autism-home-education/

Suchen

LuxAI

Academia Schools Parents Autism Curriculum Resources Buy

QTrobot for Home

QTrobot is an expressive social robot for parents
social, emotional and communication skills.
QTrobot provides a friendly and effective setup t
life skills at home. The QTrobot for home pack
educational curricula with hundreds of coherent
calls with our specialists.

For more information about QTrobot for home vis

Your purchase includes:

30 day no questions asked money back guara

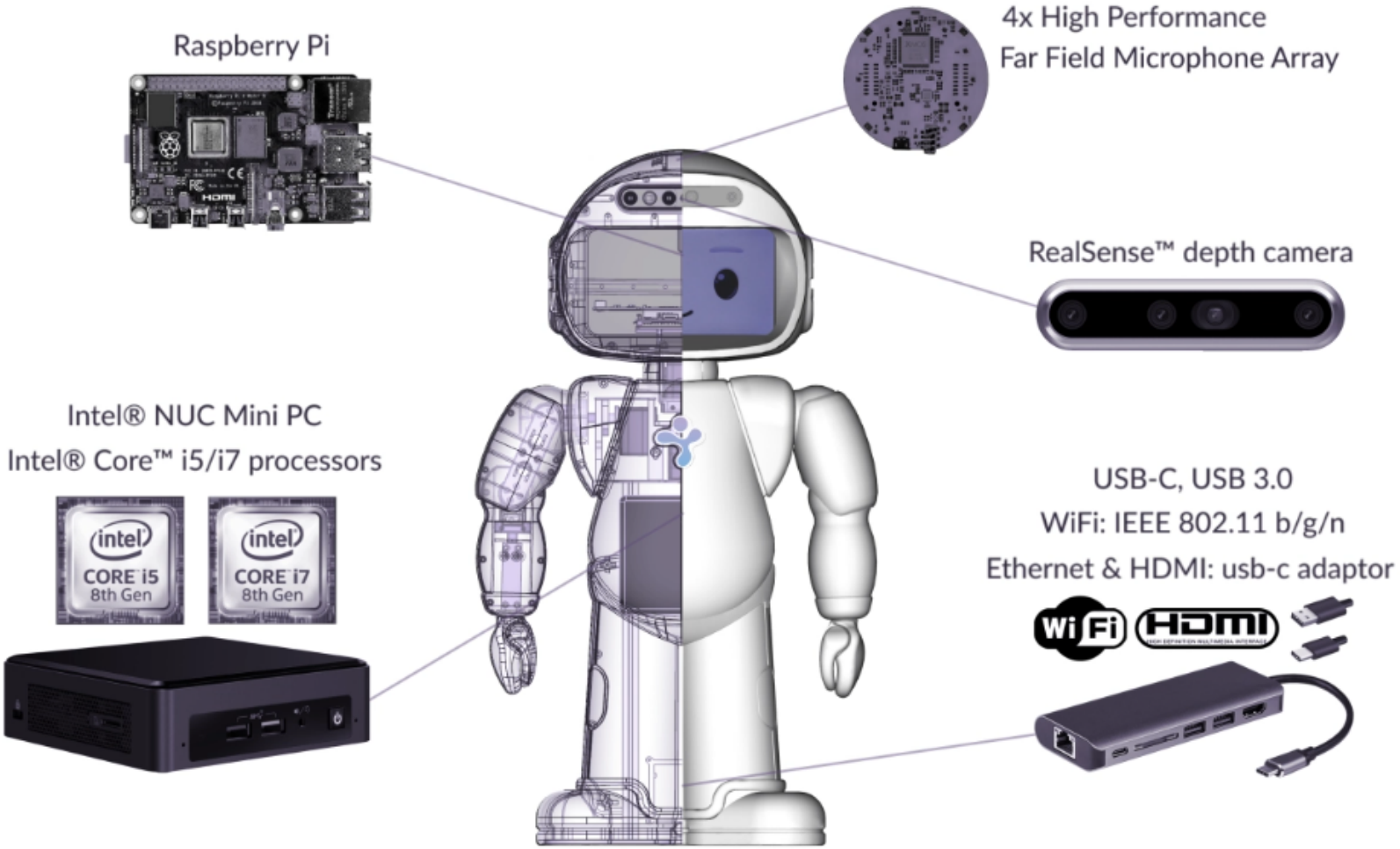
Monthly support calls with LuxAI SEN special

QTrobot Teaching Platform

2.356,20 €

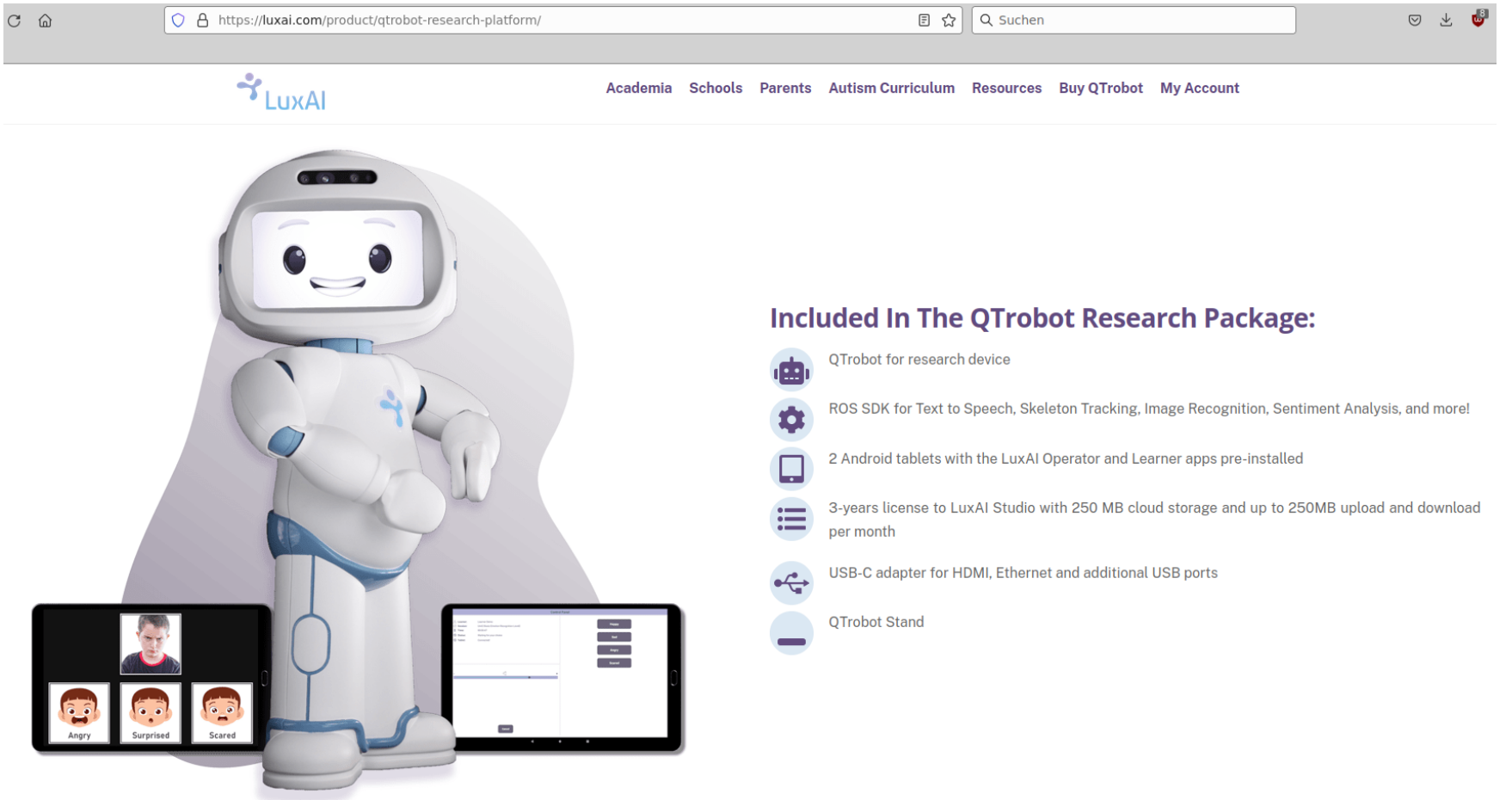
Software Subscription

QTrobot hardware



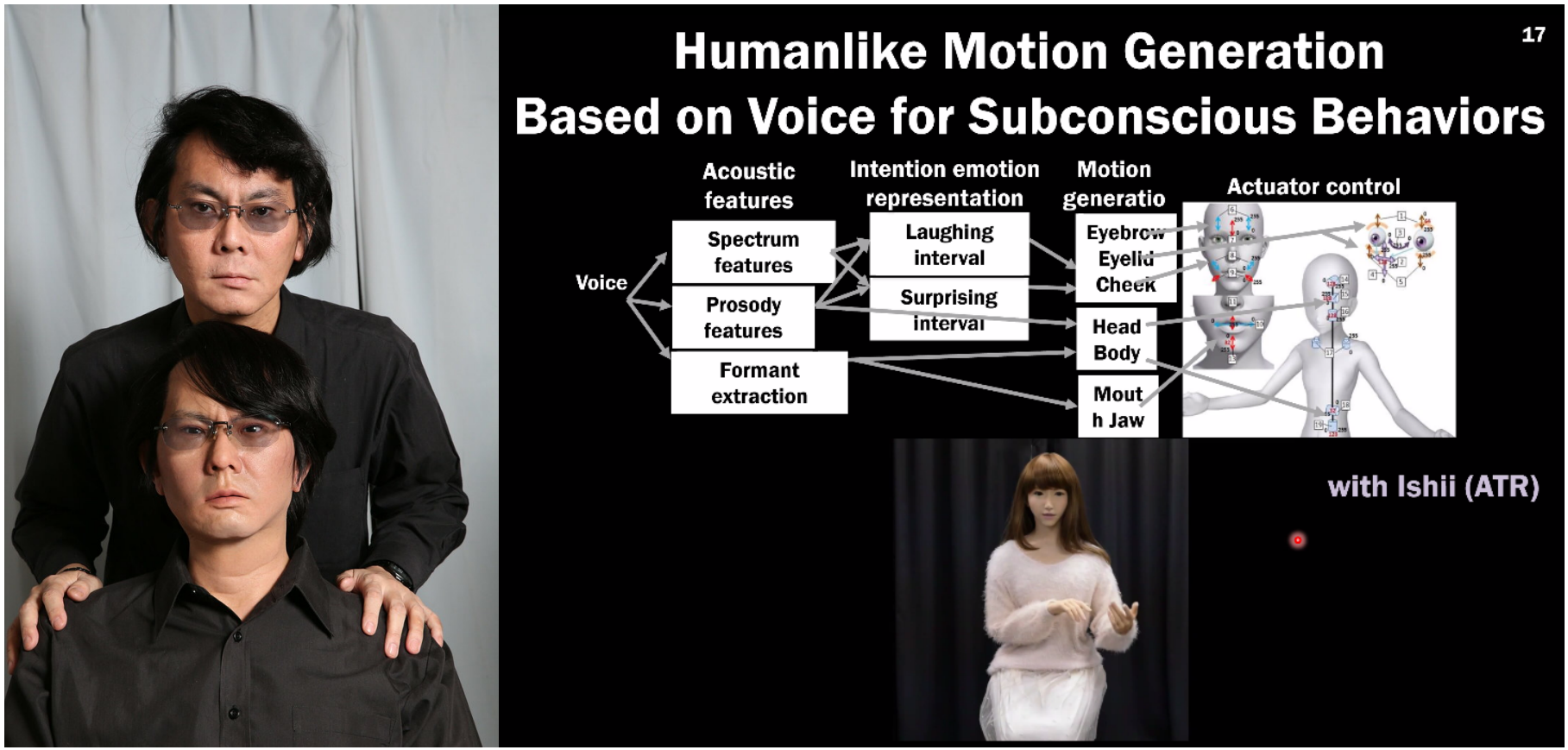
https://luxai.com/humanoid-social-robot-for-research-and-teaching/

QTrobot research package



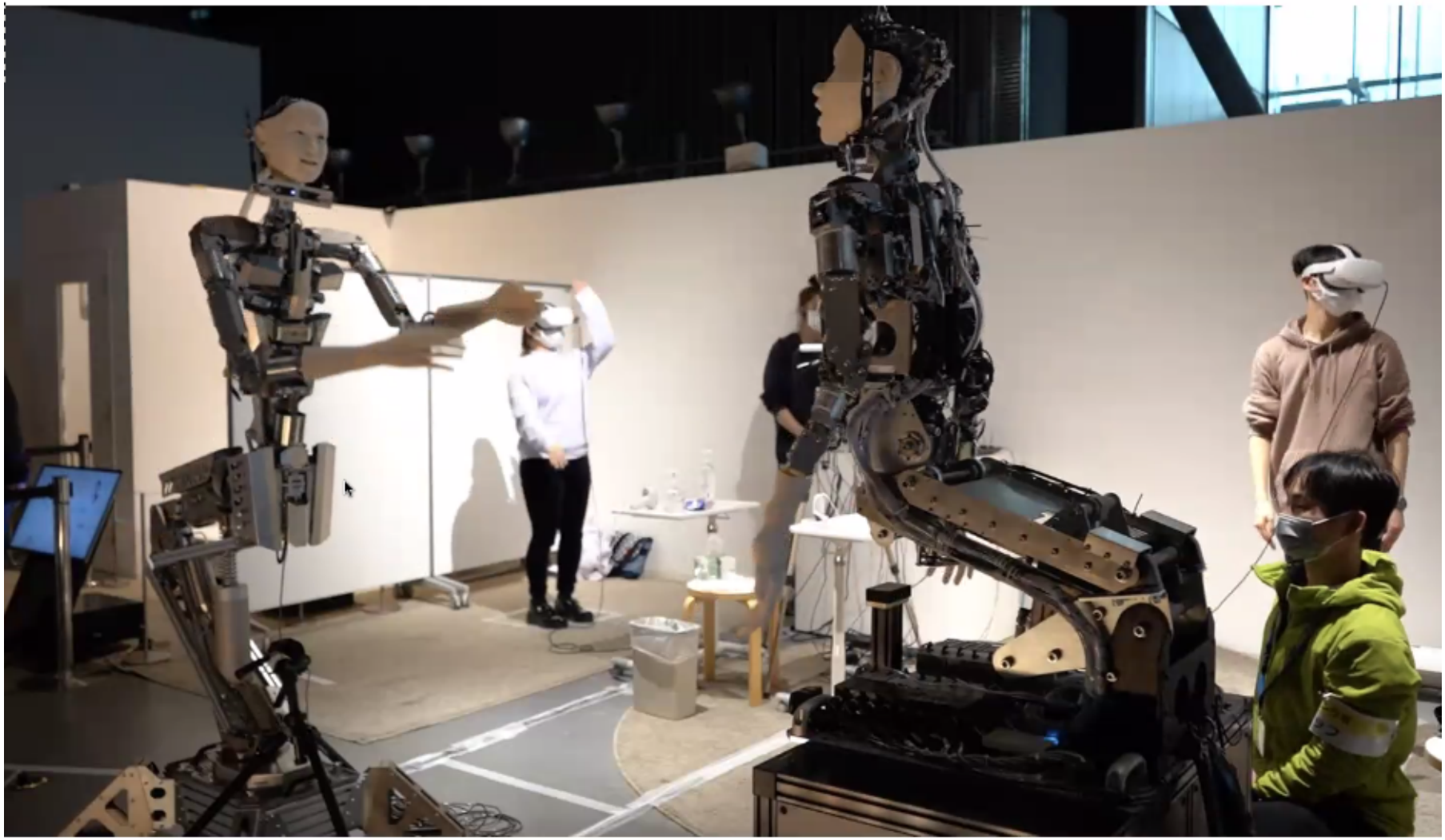
https://luxai.com/humanoid-social-robot-for-research-and-teaching/

Geminoid + Eve + Erica (Ishiguro 2009)



Geminoid initially teleoperated only (including voice), some autonomy since www.geminoid.jp/projects/kibans/resources-j.html

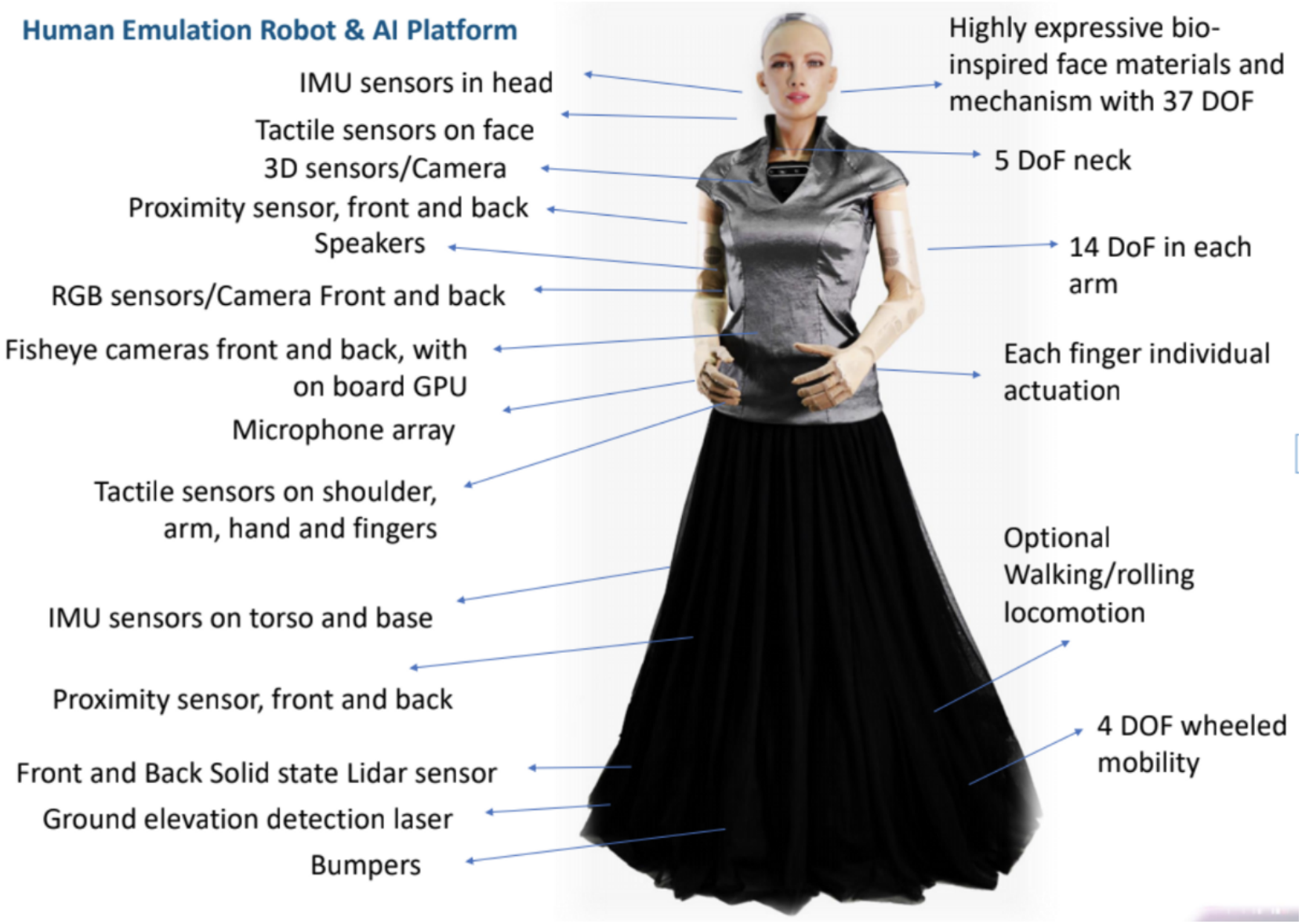
Alter3 (Ikegami 2018)



presented at

- Tokyo Future Science Museum (Tokyo, March 2021)

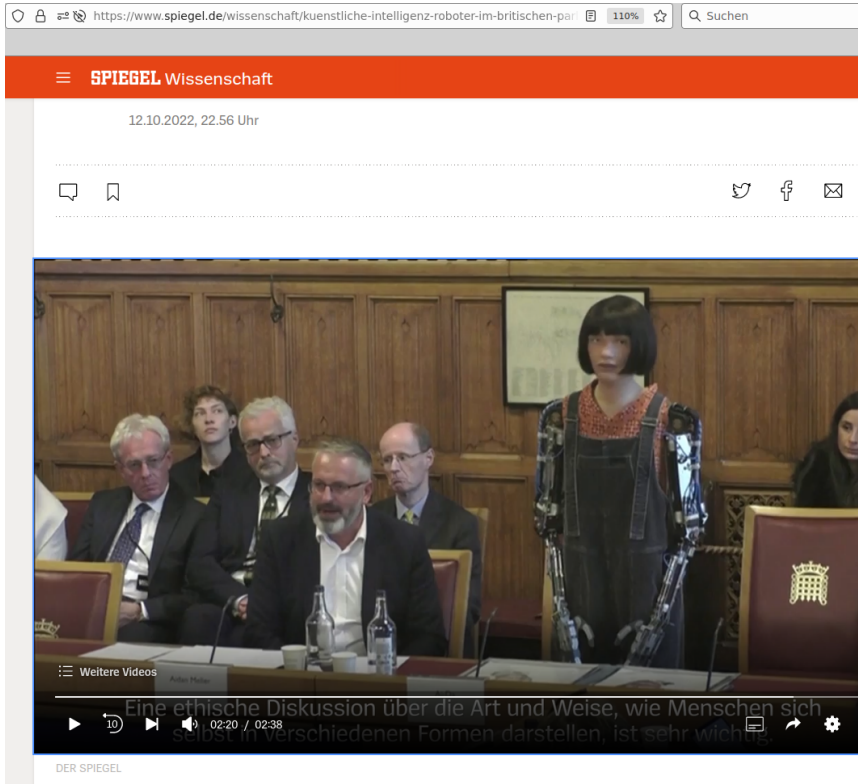
Sophia (Hanson Robotics, 2017)



Sophia (Oktoberfest, 2019)

DIA 2019 Munich, www.youtube.com/watch?v=Y0HkIG2x4FU

Ai-Da (U Oxford, 2022)



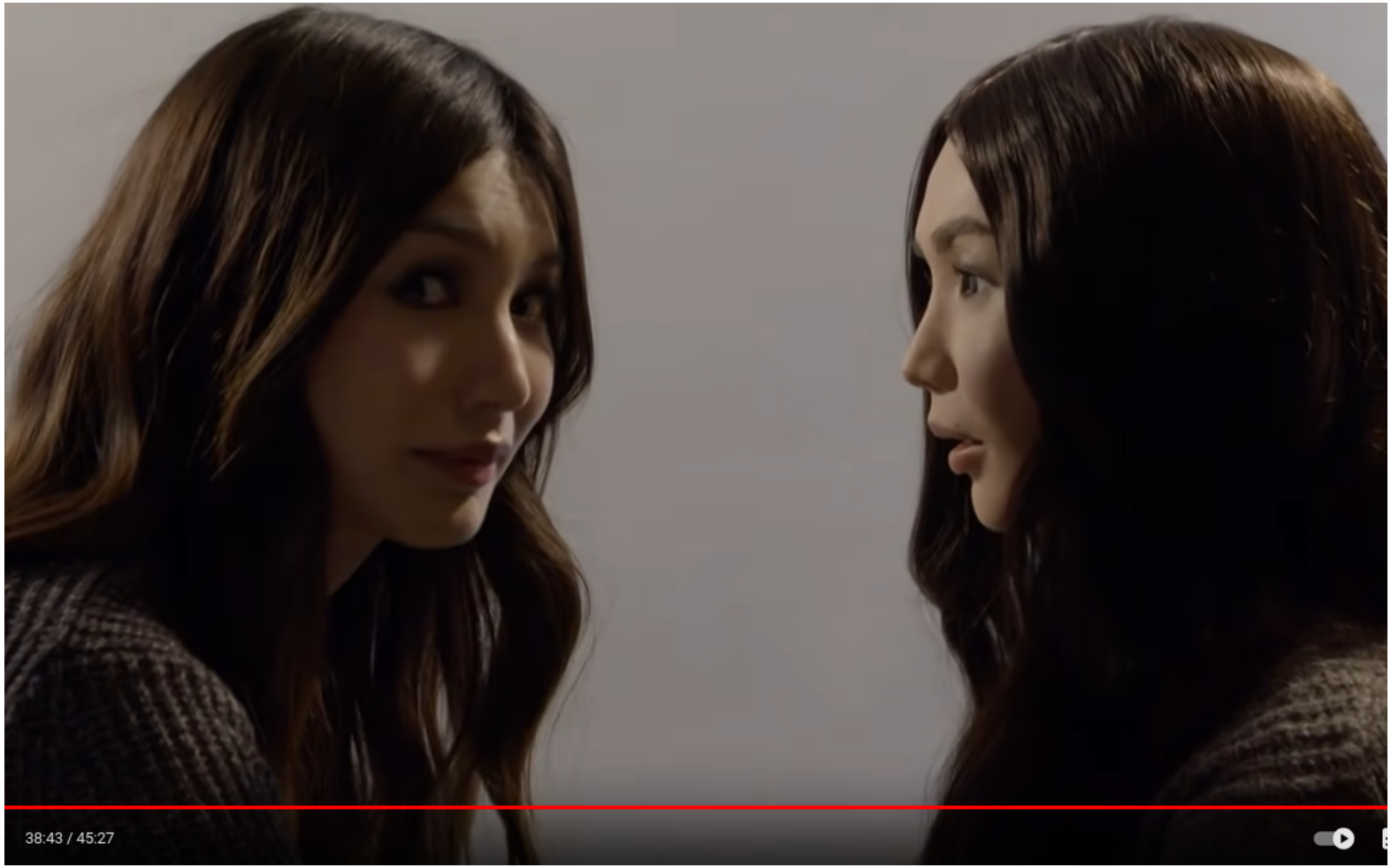
Researcher Lucy Seal (left) and creator Aidan Meller (right) will unveil Ai-Da's work at Oxford University

An exhibition of art created by a humanoid AI robot is set to open at the University of Oxford.

The robot, called Ai-Da after the mathematician Ada Lovelace, uses a robotic arm and a pencil to draw what it sees with a camera in its eye.

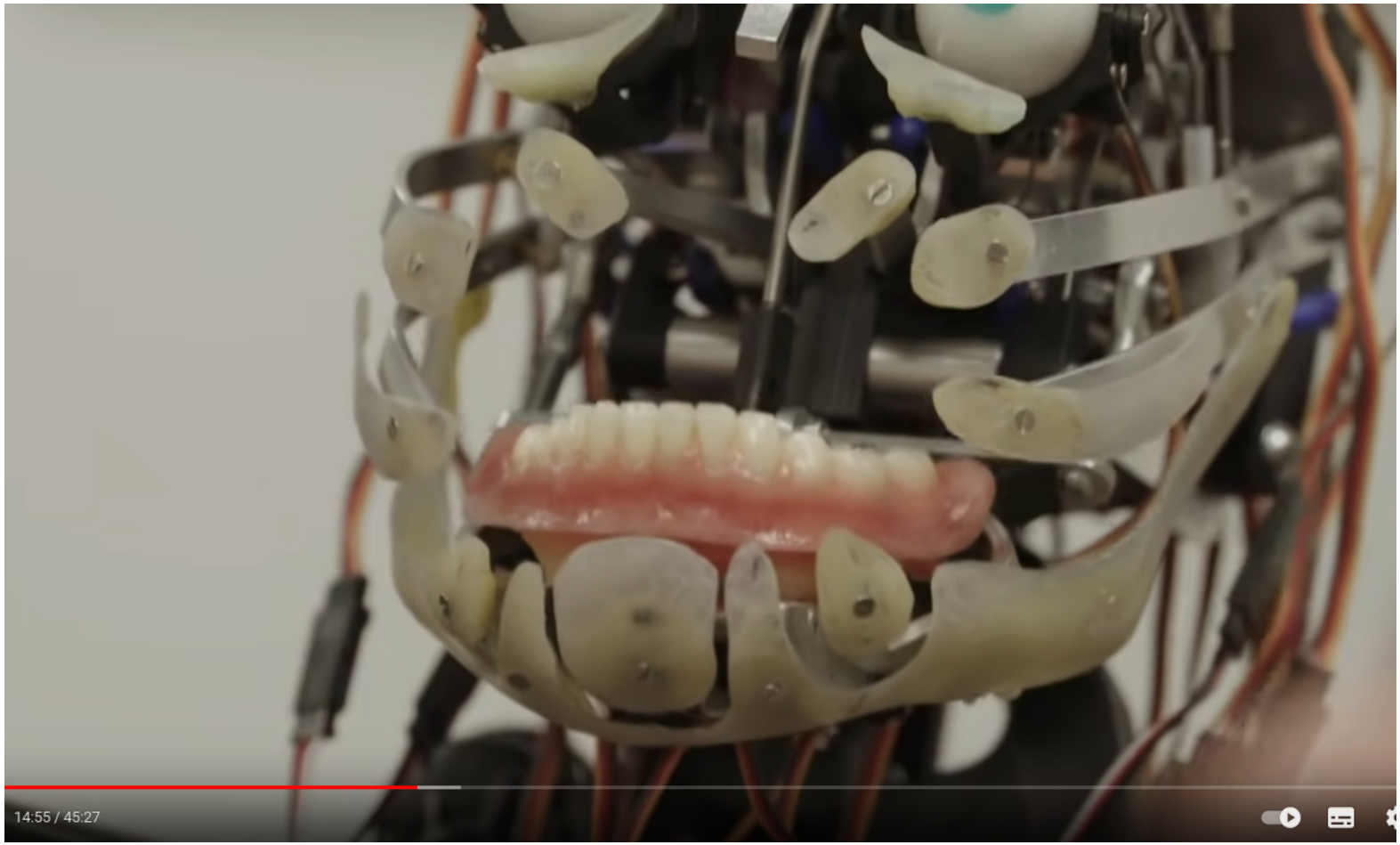
Spiegel Online, The Guardian

Gemma (Engineered Arts, 2021)



www.youtube.com/watch?v=bC_DZIwevIl — The Rising World of Building AI Human Clones

Gemma: Face Actuators



www.youtube.com/watch?v=bC_DZIwevIl — The Rising World of Building AI Human Clones

Gemma: Mask-Making

www.youtube.com/watch?v=bC_DZlwevIl — The Rising World of Building AI Human Clones

What can a Robot Head do?

- visual and audio and touch sensing
- uni-modal and multi-modal
- visual object detection, tracking
- sound-source localization, separation
- visual/audio integration: ego-sphere

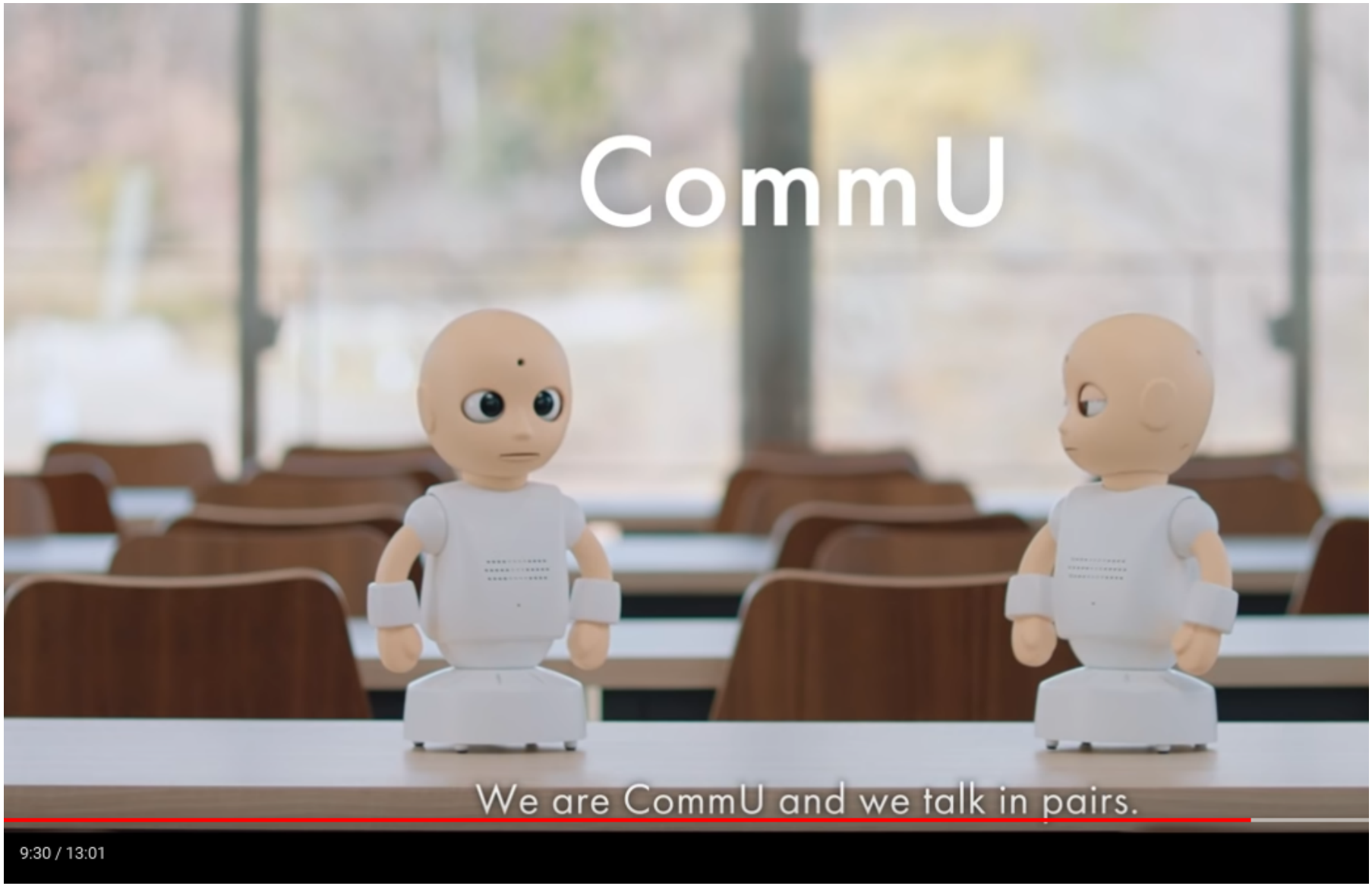
- eye motions
- facial expressions and emotions

- expressive voice
- speech dialog system
- long-term goal: human-like interactions

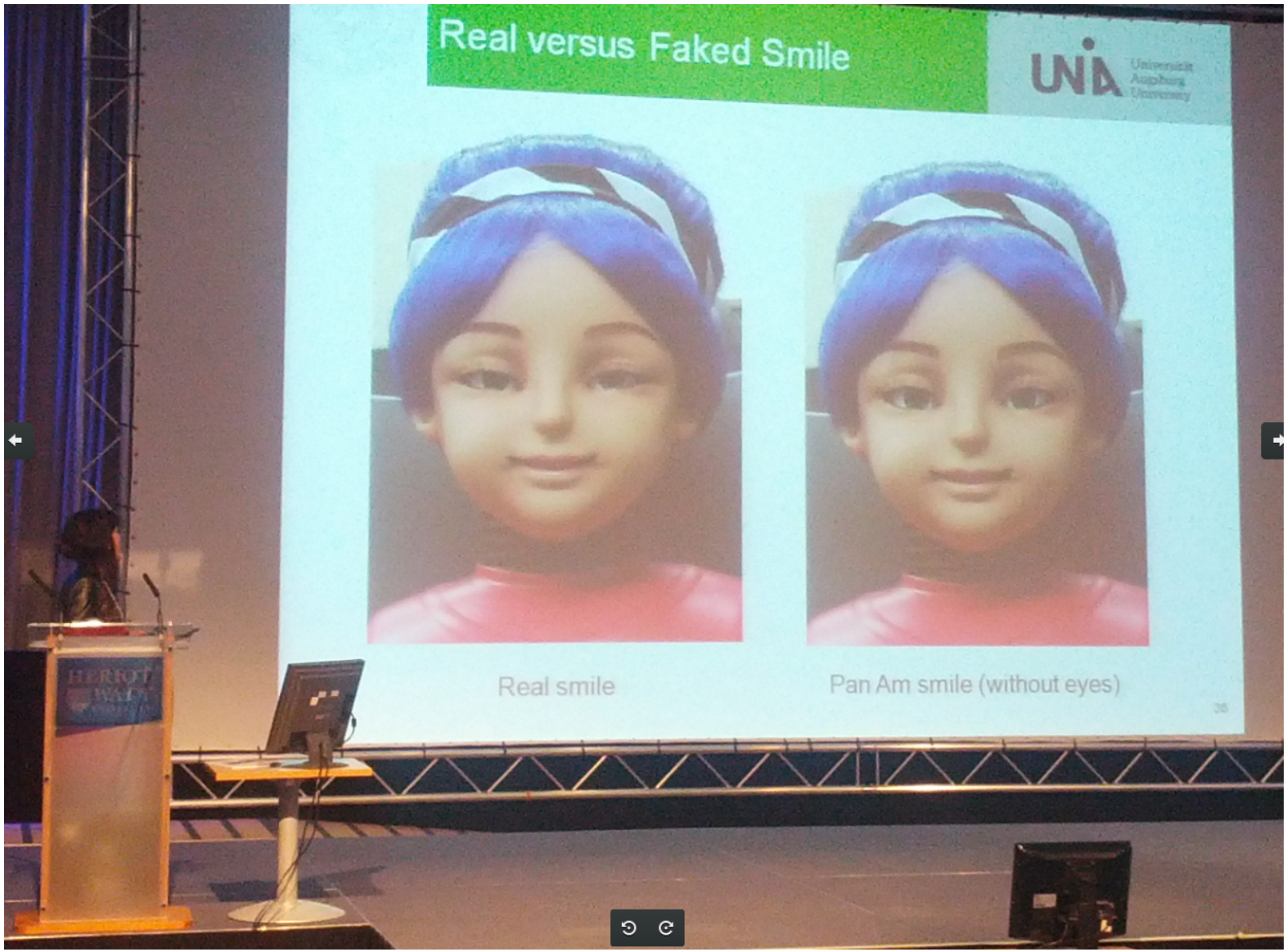
USST Robot Head 2021 (Ula)

Amazon Echo Dot 2022 (Alexa)

Example: CommU food recommendations

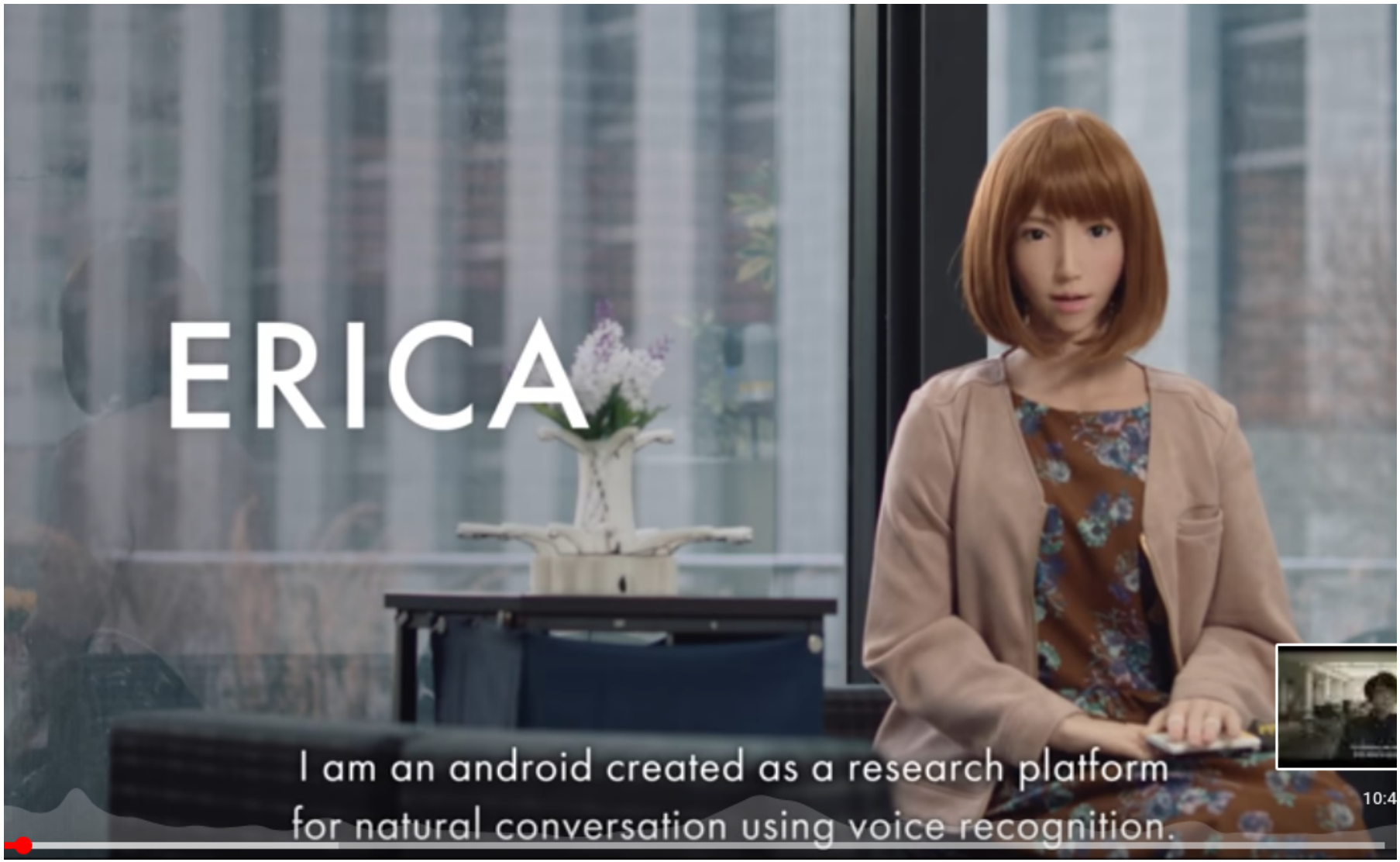


Facial Expressions: Pan-Am Smile



Zeca toy robot, University of Augsburg, ROMAN Workshop 2016

Example: ERICA job interview

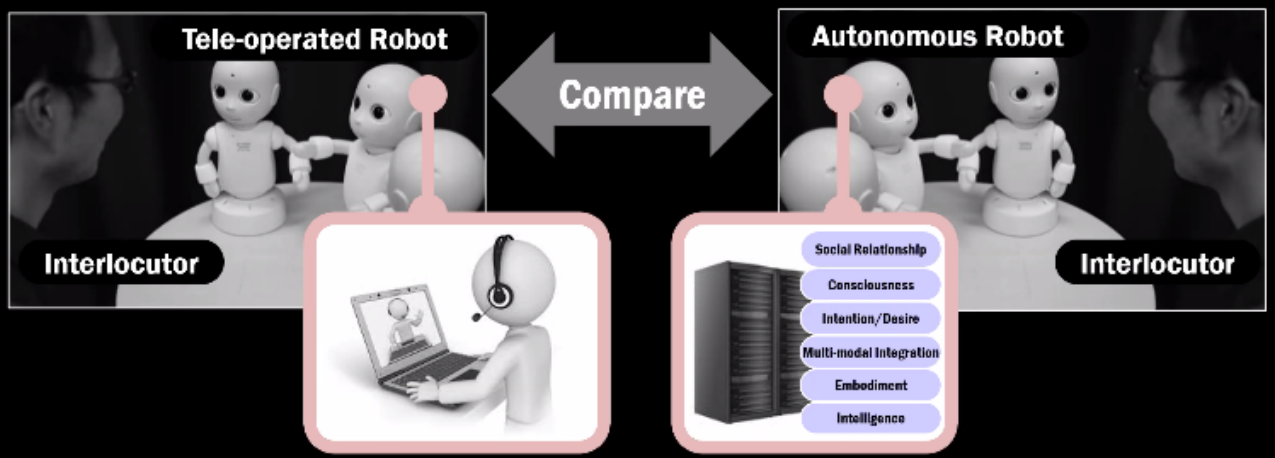


www.youtube.com/watch?v=j1h1KOeCHjg - JST ERATO IsHIGURO Symbiotic Human-Robot Interaction project

Multimodal Turing Test (Teleoperation)

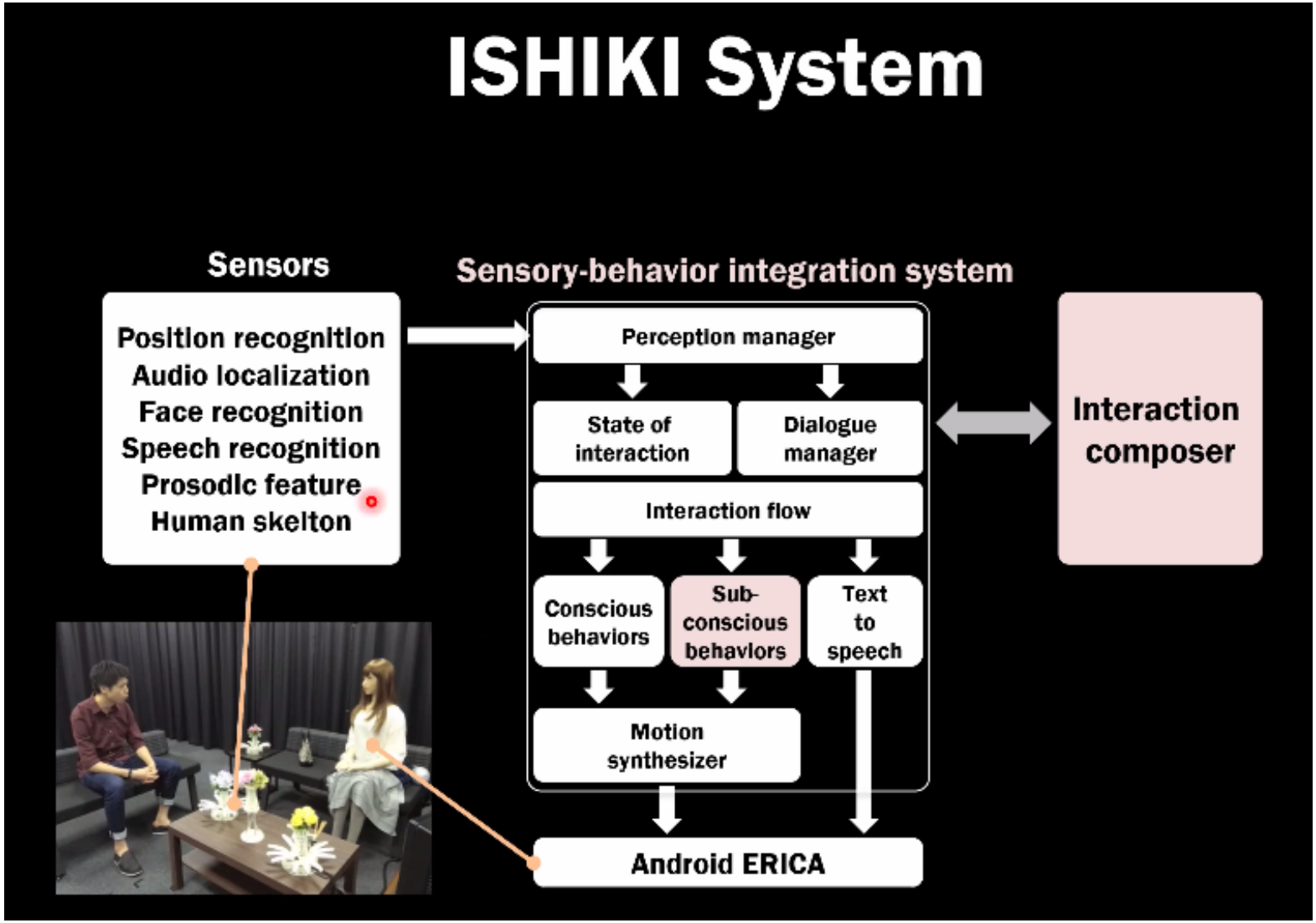
Multimodal Turing Test (MTT) as a scientific and engineering goal

- MTT is to compare between a robot manipulated by human operators and an autonomous robot controlled by developed technology.
- One of the important challenges in intelligent robotics is to pass MTT.
- It evaluates the total humanlikeness through all of the modalities.
- It evaluates the social acceptance as a member of our society.



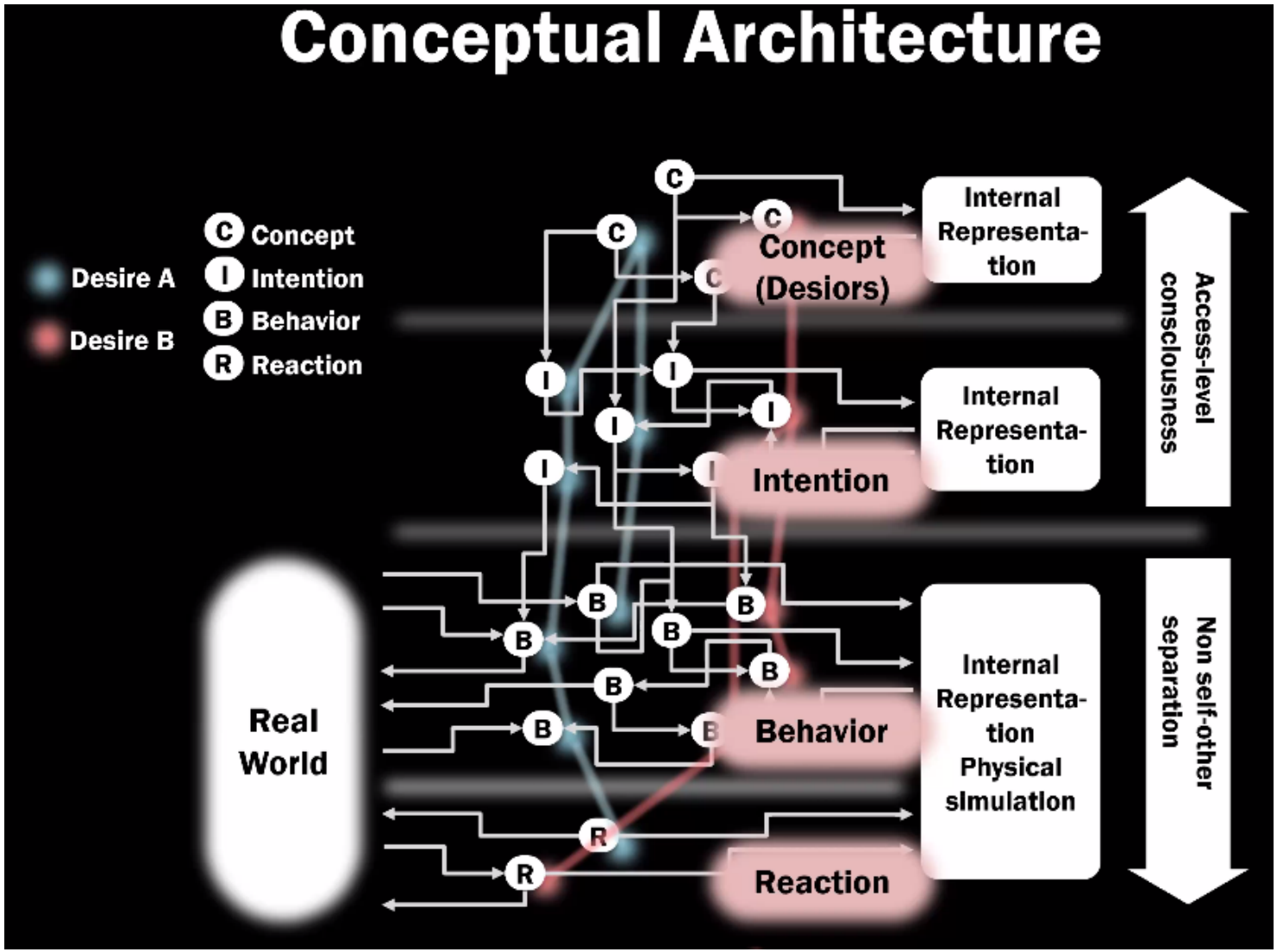
Ishiguro, Embodied Intelligence Workshop, 2021

Ishiki System



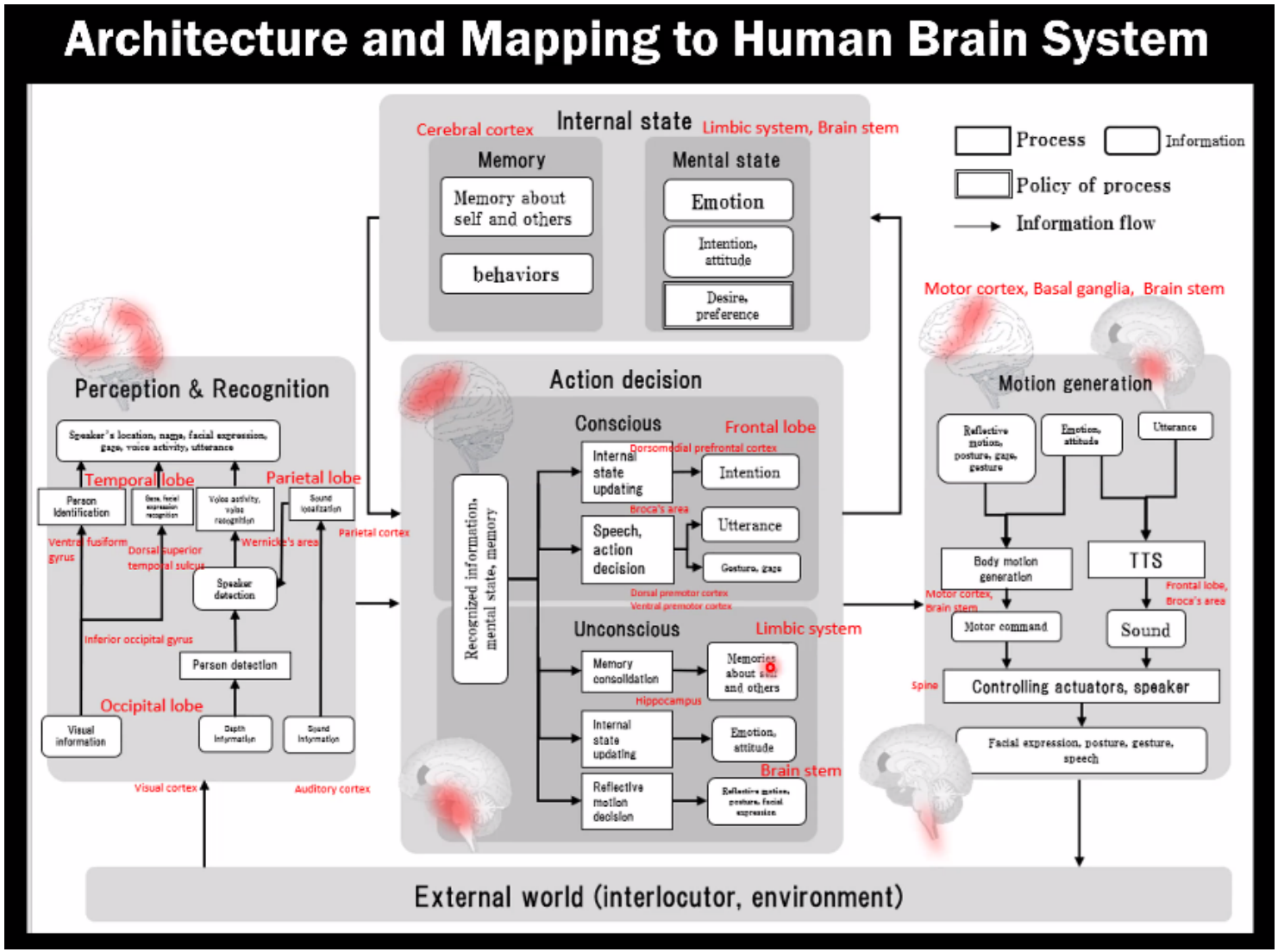
Ishiguro, Embodied Intelligence Workshop, 2021

Ishiki System: Concepts and Architecture



Ishiguro, Embodied Intelligence Workshop, 2021

Ishiki System: Mapping to the Human Brain



Ishiguro, Embodied Intelligence Workshop, 2021

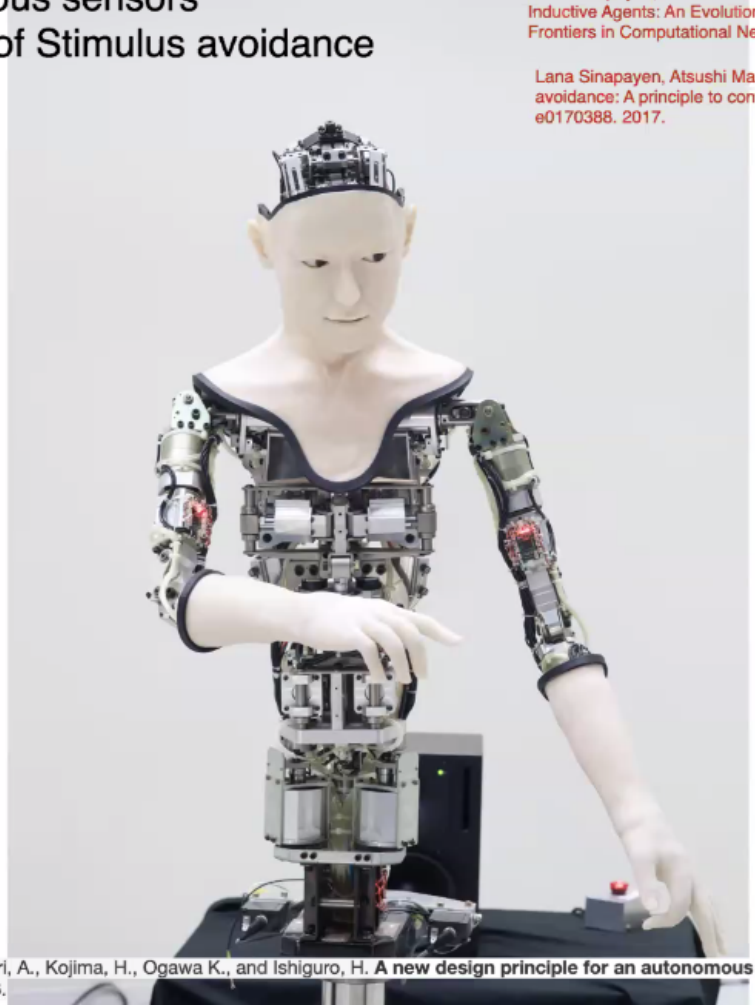
Takashi Ikegami: Alter 1

1. Coupled chaotic oscillators
2. Autonomous sensors
3. Principle of Stimulus avoidance

Norihiro Maruyama, Mizuki Oka, Takashi Ikegami: Creating Space-Time Affordances via an Autonomous Sensor Network, The 2013 IEEE Symposium on Artificial Life,pp.67-73, 2013.

Lana Sinapayen, Atsushi Masumori and Takashi Ikegami. Reactive, Proactive, and Inductive Agents: An Evolutionary Path for Biological and Artificial Spiking Networks. Frontiers in Computational Neuroscience, 2019 13(88)

Lana Sinapayen, Atsushi Masumori, Takashi Ikegami: Learning by stimulation avoidance: A principle to control spiking neural networks dynamics., PLoS ONE,12(2) e0170388. 2017.



Alter1

Doi, I., Ikegami, T., Masumori, A., Kojima, H., Ogawa K., and Ishiguro, H. **A new design principle for an autonomous robot**, *14th European Conference on Artificial Life(ECAL2017)*, pp.490-466.

Takashi Ikegami, Embodied Intelligence Workshop, 2021

Takashi Ikegami: Evolution of Alter

The evolution of Alter's systems

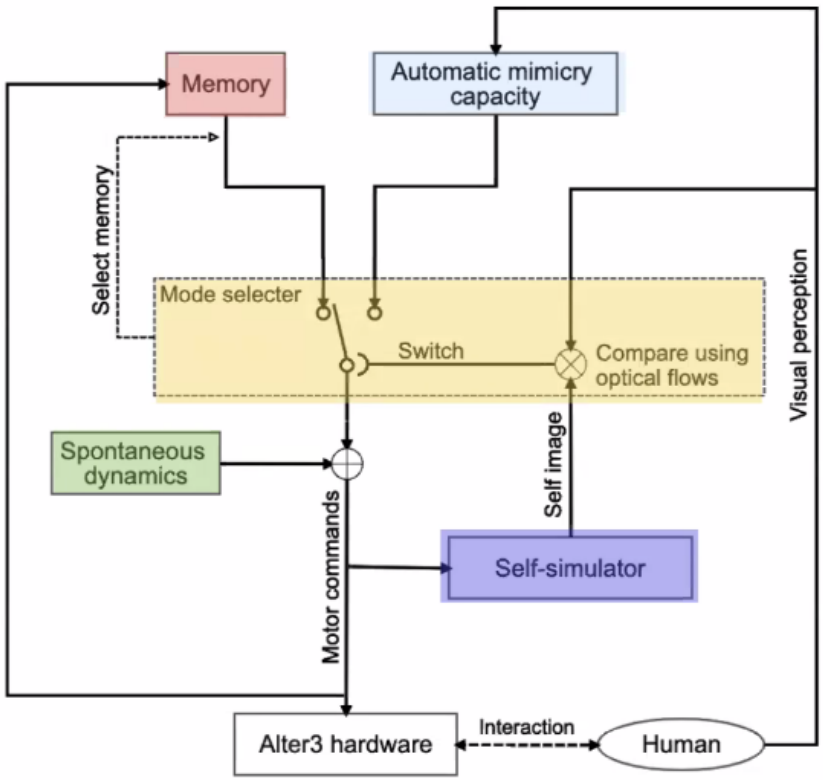
	IR	T3	PSA	Self-I	Mem	Eye	Imi	AL
Alter1 2016	○	○	○	x	x	x	x	x
Alter2 2018	x	○	○	x	x	x	x	○
Alter3 2019	x	x	○	○	○	○	○	x

1. Built-in Self-simulator (self-awareness) .
2. Awake (open-eyes) and Dream (Closed-eyes) mode.
3. Memory selection and variations.

Otto Roessler, *An Artificial Cognitive Map System, BioSystems, 13 (1981) pp.203-209.*

From a tutorial of Peter Dayan, Geoffrey Hinton and Radford M.Neal and Richard Zemel, *The Helmholtz Machine, Neural Computation 7 (1995) 889-904.*

Switching Mimicry and Memory Based Behavior

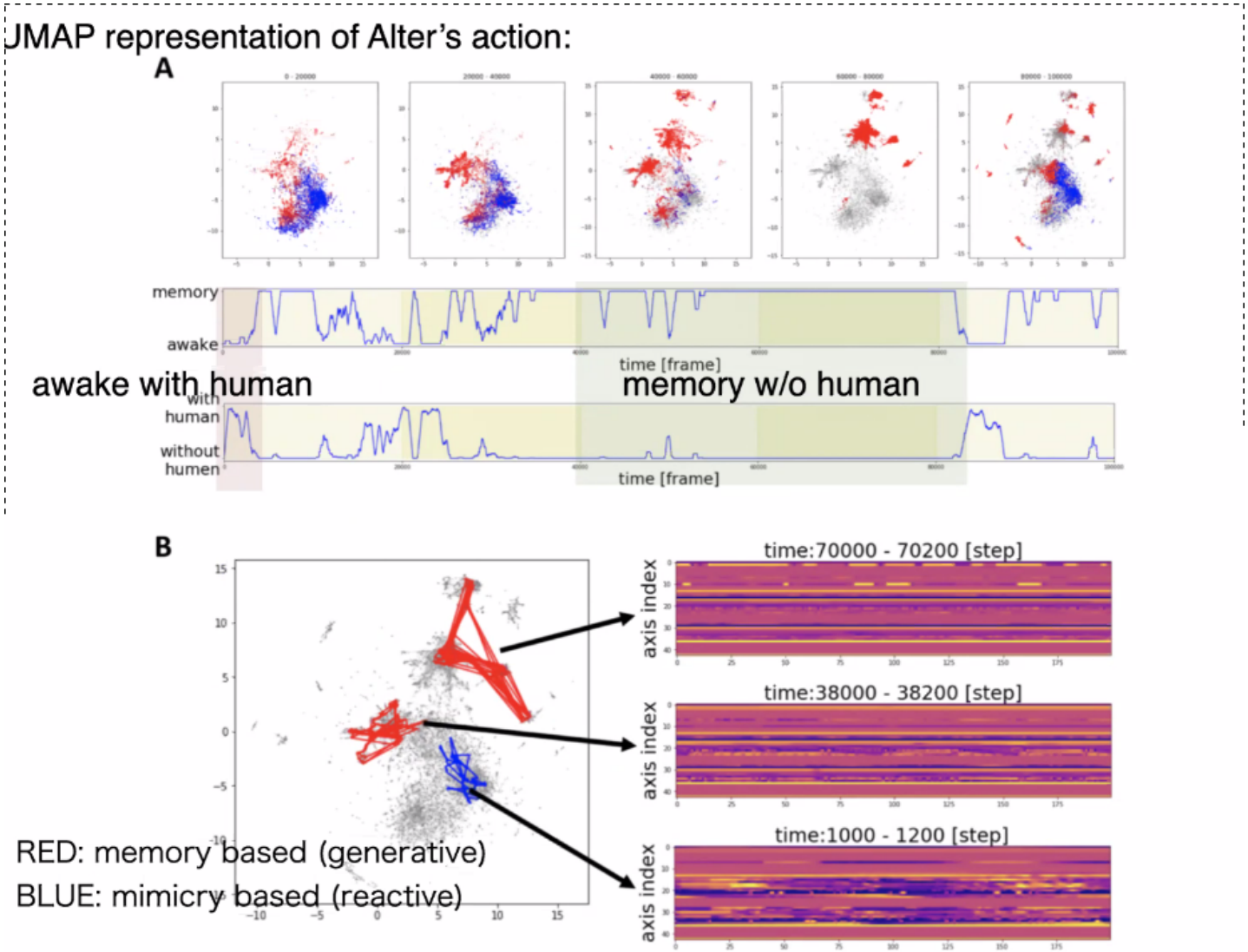


Atsushi Masumori, Norihiro Maruyama, Takashi Ikegami: Personogenesis Through Imitating Human Behavior in a Humanoid Robot "Alter3". Frontiers Robotics AI 7: 532375 (2020)

Lana Sinapayen, Atsushi Masumori and Takashi Ikegami. Reactive, Proactive, and Inductive Agents: An Evolutionary Path for Biological and Artificial Spiking Networks. Frontiers in Computational Neuroscience, 2019 13(88)

Ikegami, Embodied Intelligence Workshop 2021

Switching Mimicry and Memory Based Behavior



Imitation Learning of Facial Expressions

Chen at al., ICRA 2021

Imitation Learning of Facial Expressions

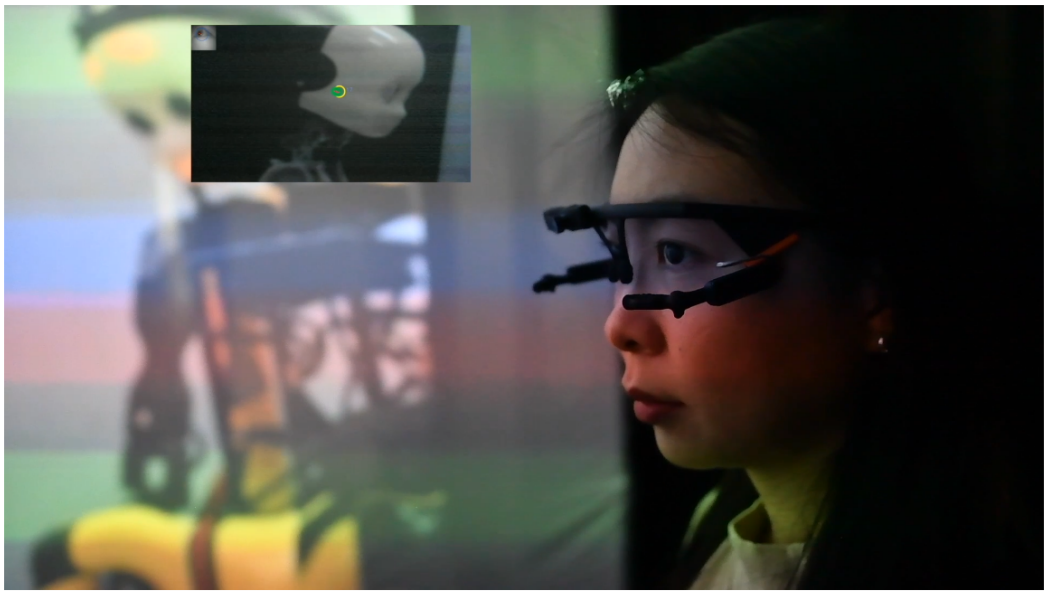
Chen at al., ICRA 2021

Eye-Tracking for Learning Visual Attention

SFB TRR 169

A5 – Current work
social attention between human and robot

Task B: Eye Tracking Data Collection



(Fu, Abawi*, Strahl, & Wermter, 2022, RO-MAN workshop)*

Fu et al., CML summerschool 2022

Lip-Motion Detection and Speaker Tracking

Fu et al., CML summerschool 2022

ULA Head (USST + UHAM 2020)

Actuators

10 RC-servos

- neck: yaw, pitch, roll
- lower jaw: chewing

left and right eye each:

- eyeballs: pan, tilt
- eyelid: open/close

facial expressions:

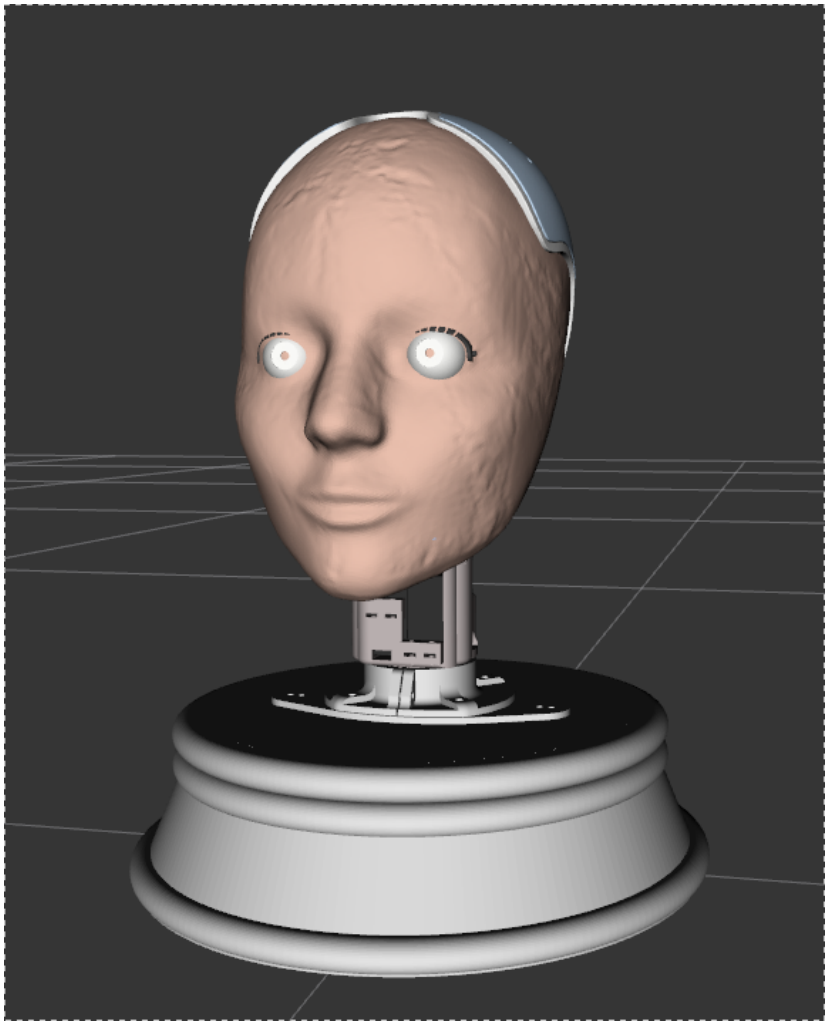
- mouth corners
- eyebrows

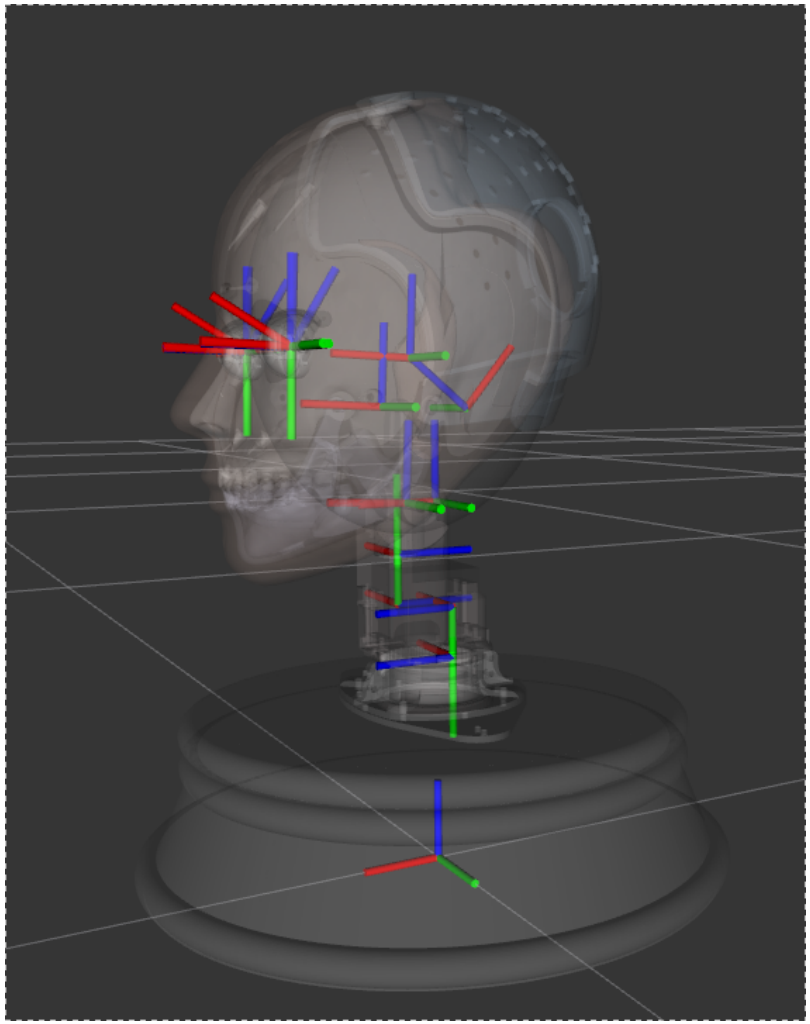
- loudspeaker

Sensors

- 2 webcams
- 4-microphone array
- IMU
- tactile skin?

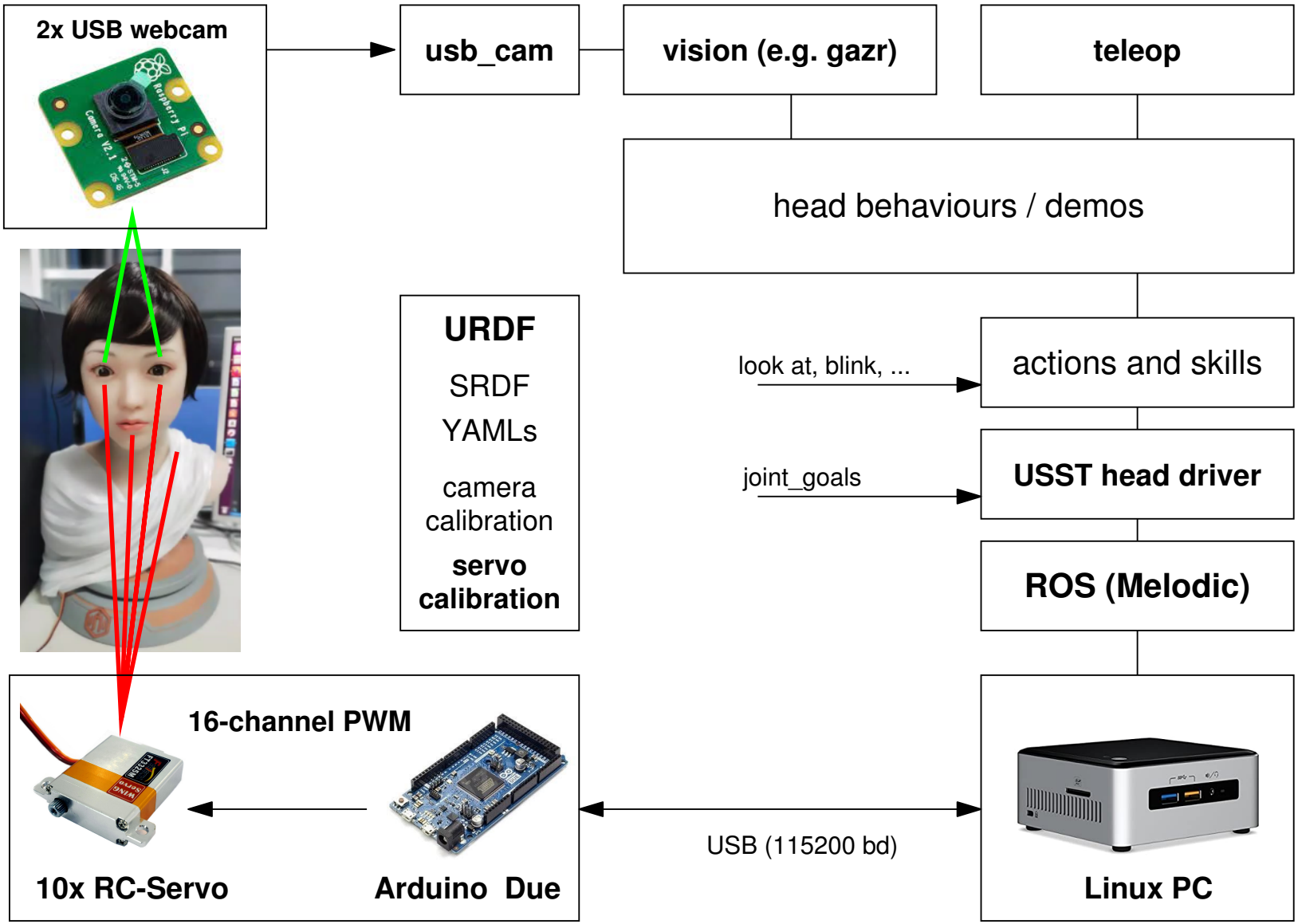
URDF Model





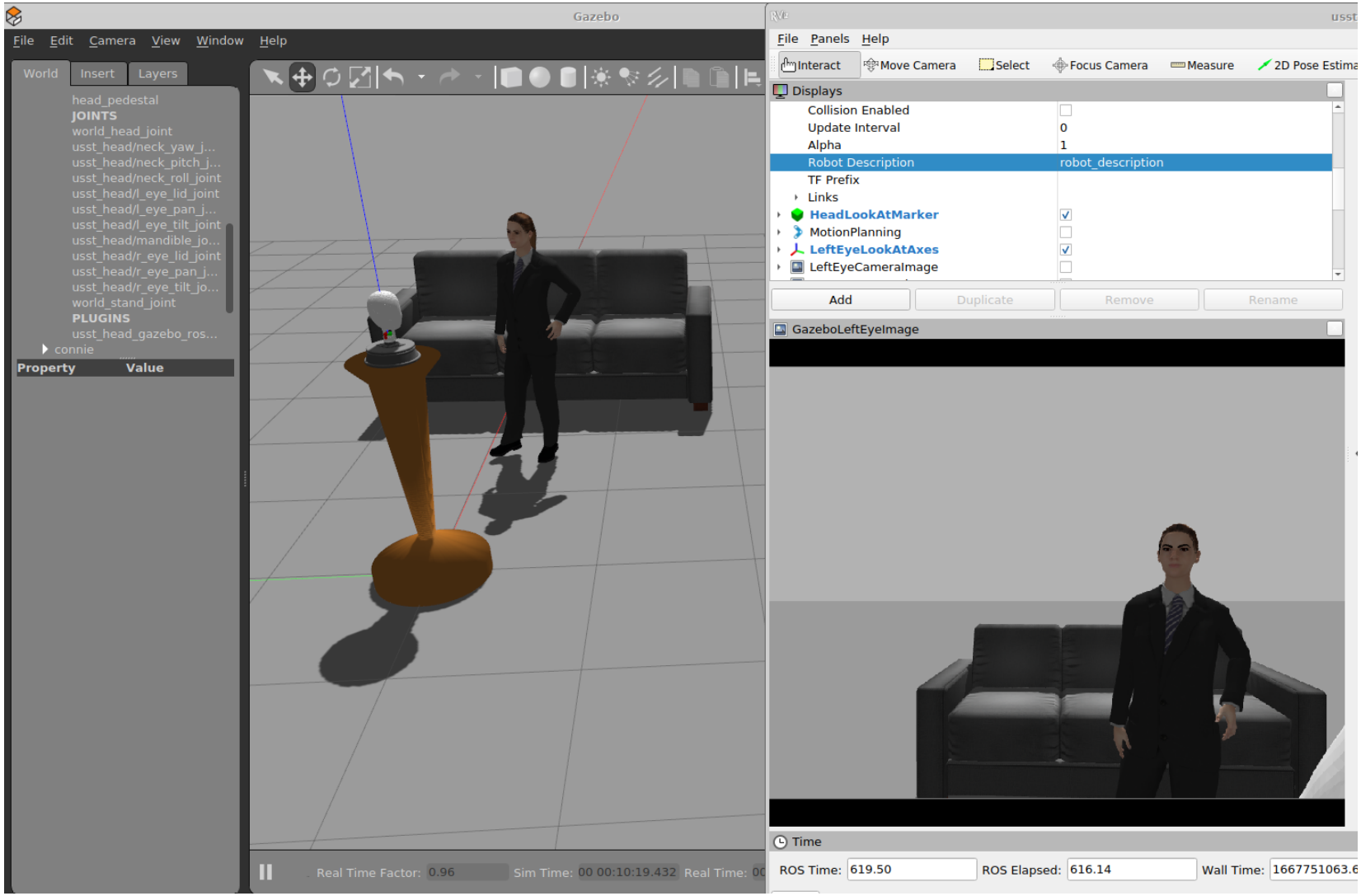
Fixed face mesh. 10-DOF: 3x neck pan+tilt, 1x yaw, 4x eyes, 2x eyelids.

ROS Drivers



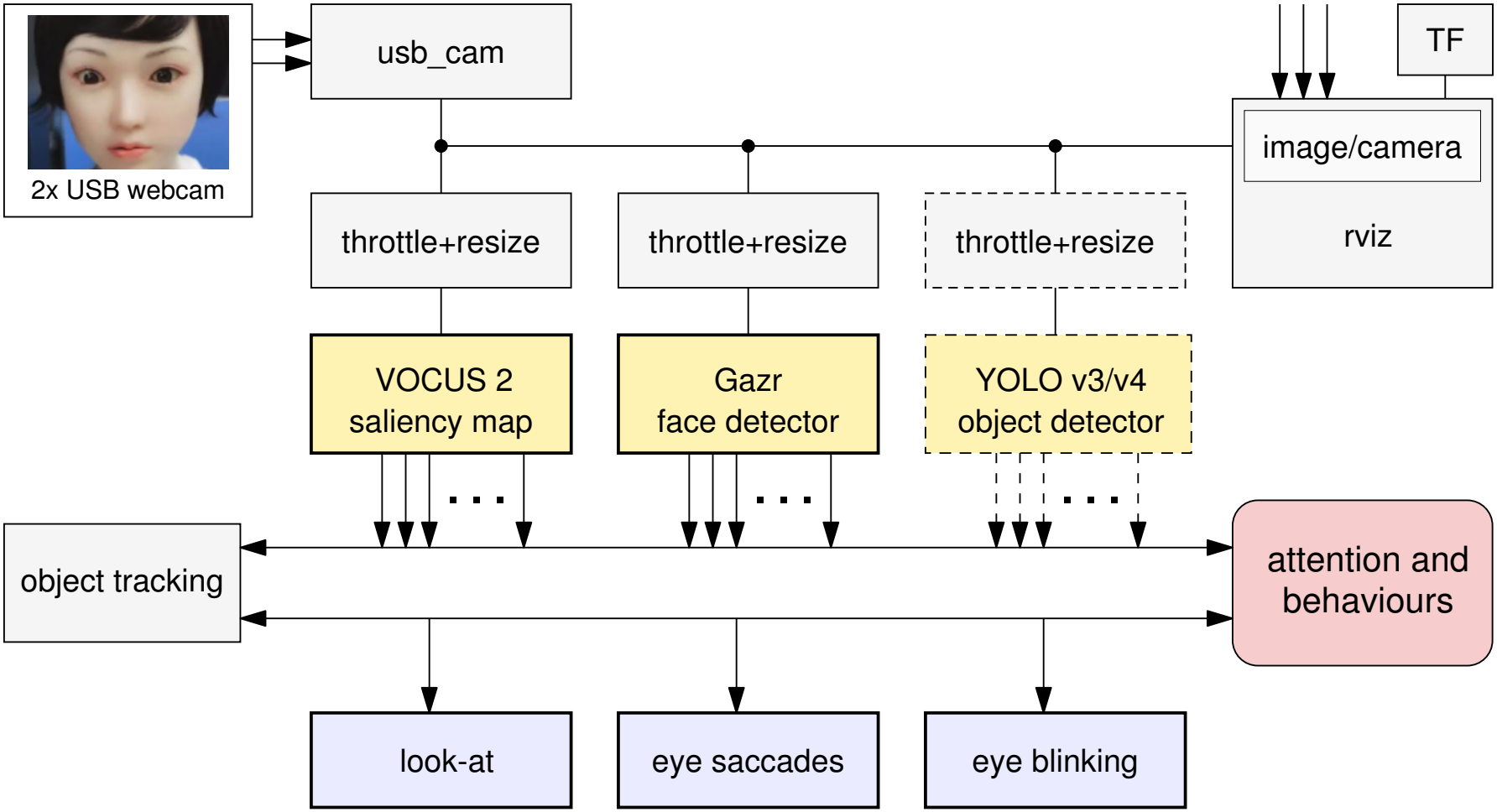
git.mafiasi.de:hendrich/usst_head_ros.git + git.mafiasi.de:hendrich/usst_head_arduino.git

Gazebo simulation

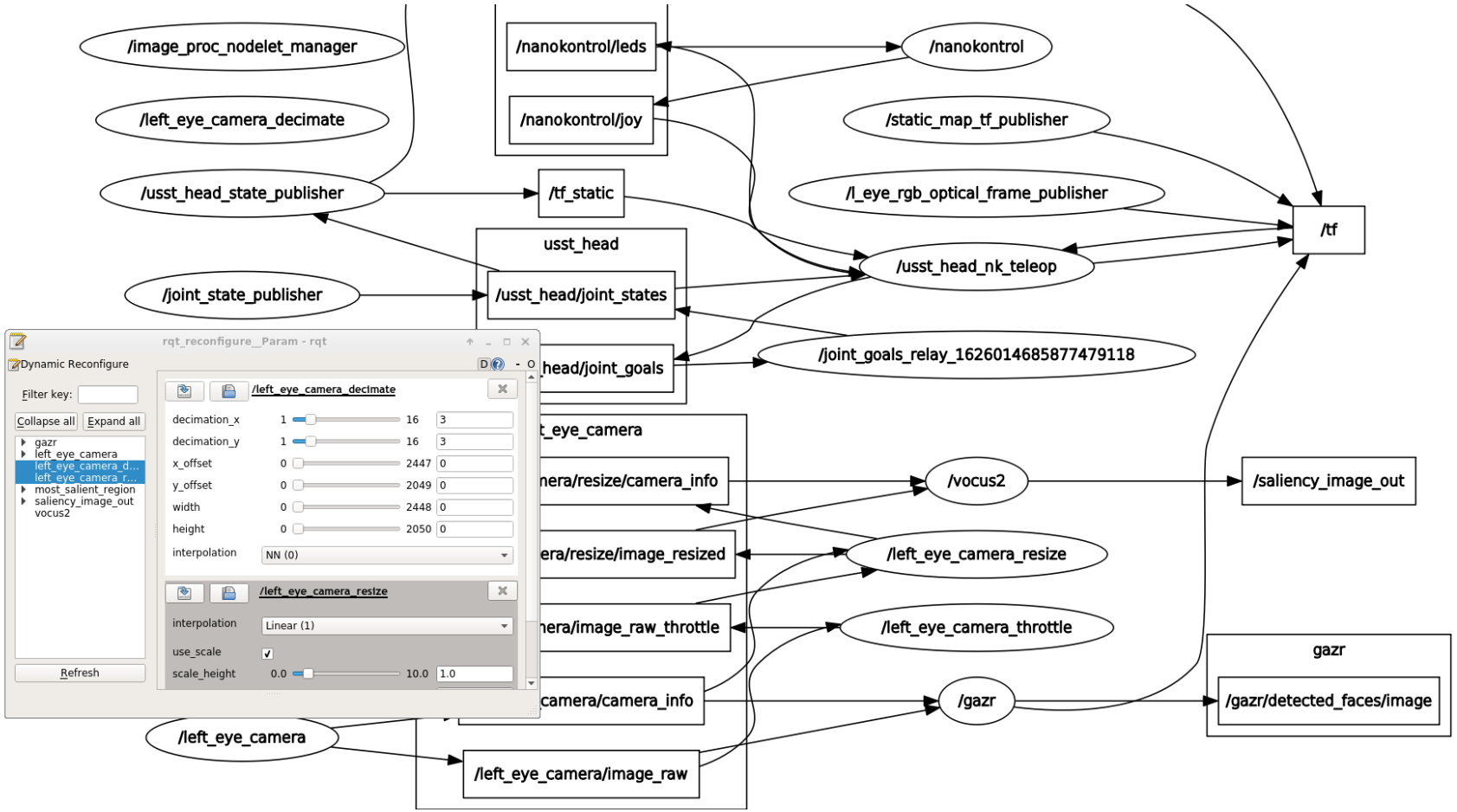


USST head + position controllers + cameras, furniture, sliding humans

Vision Pipeline

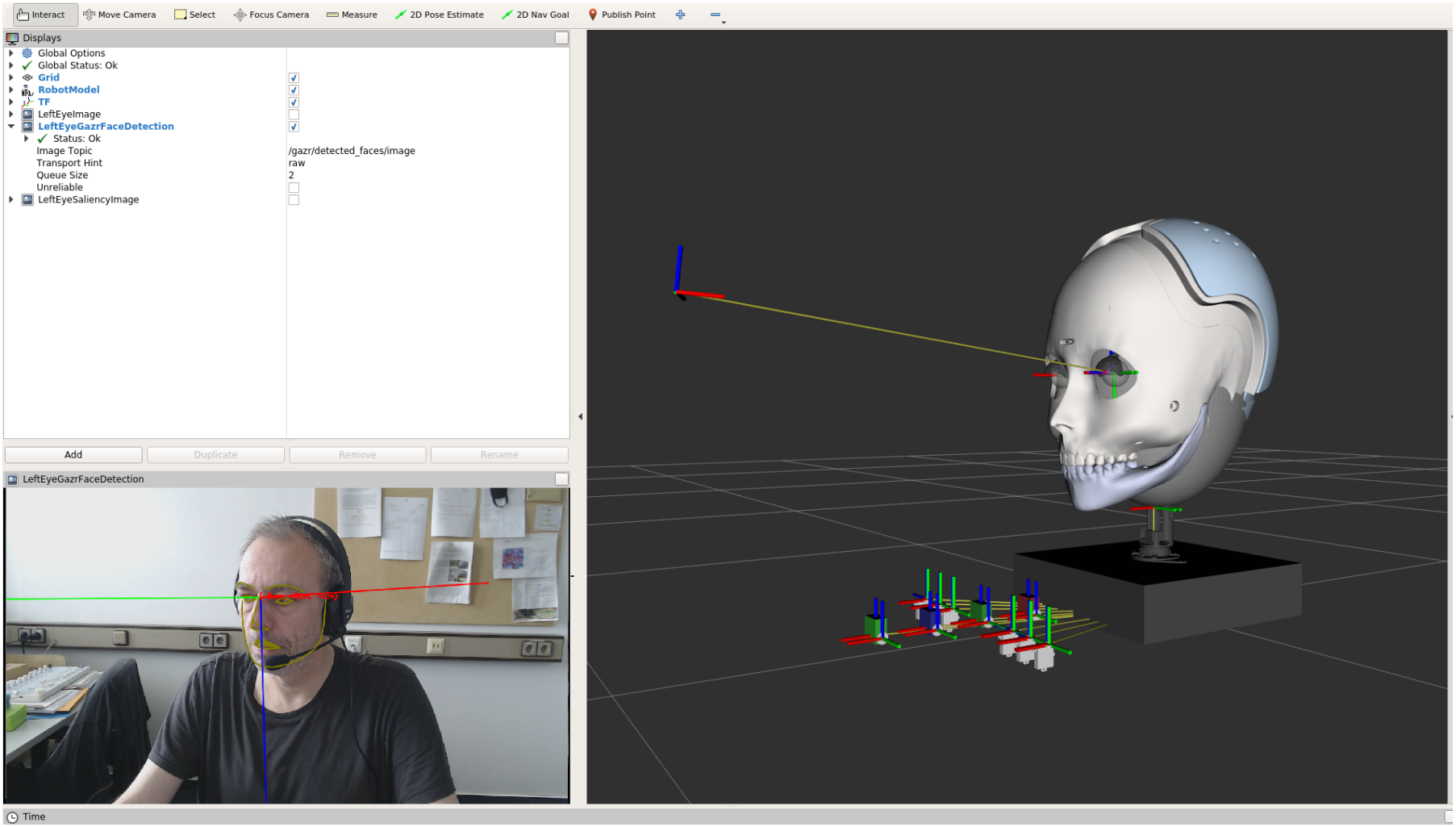


Vision pipeline: ROS nodes

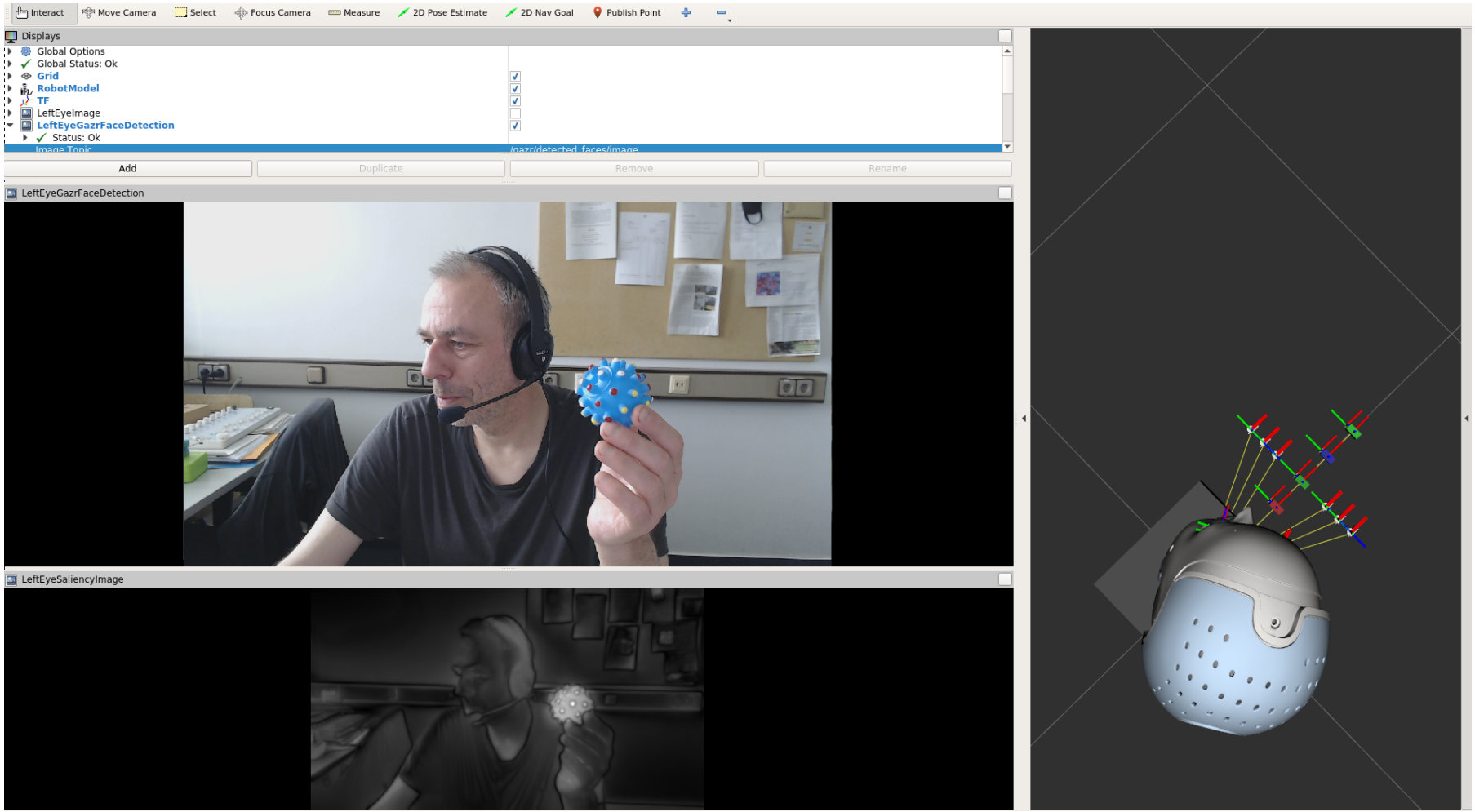


usb_cam@60Hz + throttle + ../image_proc/crop_decimate + ../image_proc/resize + rqt_reconfigure
pipeline 60% CPU + gazr 100% CPU + vocus 400% CPU

Gazr



Face detector (eyes, nose, mouth), outputs human head position and orientation)



Coronavirus dogtoy the most salient object ever :-)

Usage / Demos

```
% prepare workspace, clone dependencies
cd ~catkin_ws/src
git clone git:mafiasi.de:hendrich/usst_head_ros.git
git clone ... (ros-*-desktop-full ros-*-usb-cam tams-gazr vocus ...)
catkin_make

roslaunch usst_head_ros usst_head_urdf.launch          (URDF model in rviz)
roslaunch usst_head_ros usst_head_plotjuggler.launch   (not a lot to plot here)
...

roslaunch usst_head_ros usst_head_bringup.launch       (Arduino driver + urdf + rvi
roslaunch usst_head_ros usst_head_nanokontrol.launch   (NK teleop)
roslaunch usst_head_ros vocus_usbcam.launch            (Vocus saliency tracking)
roslaunch usst_head_ros gazr_usbcam.launch             (Gazr person tracking)
...
```

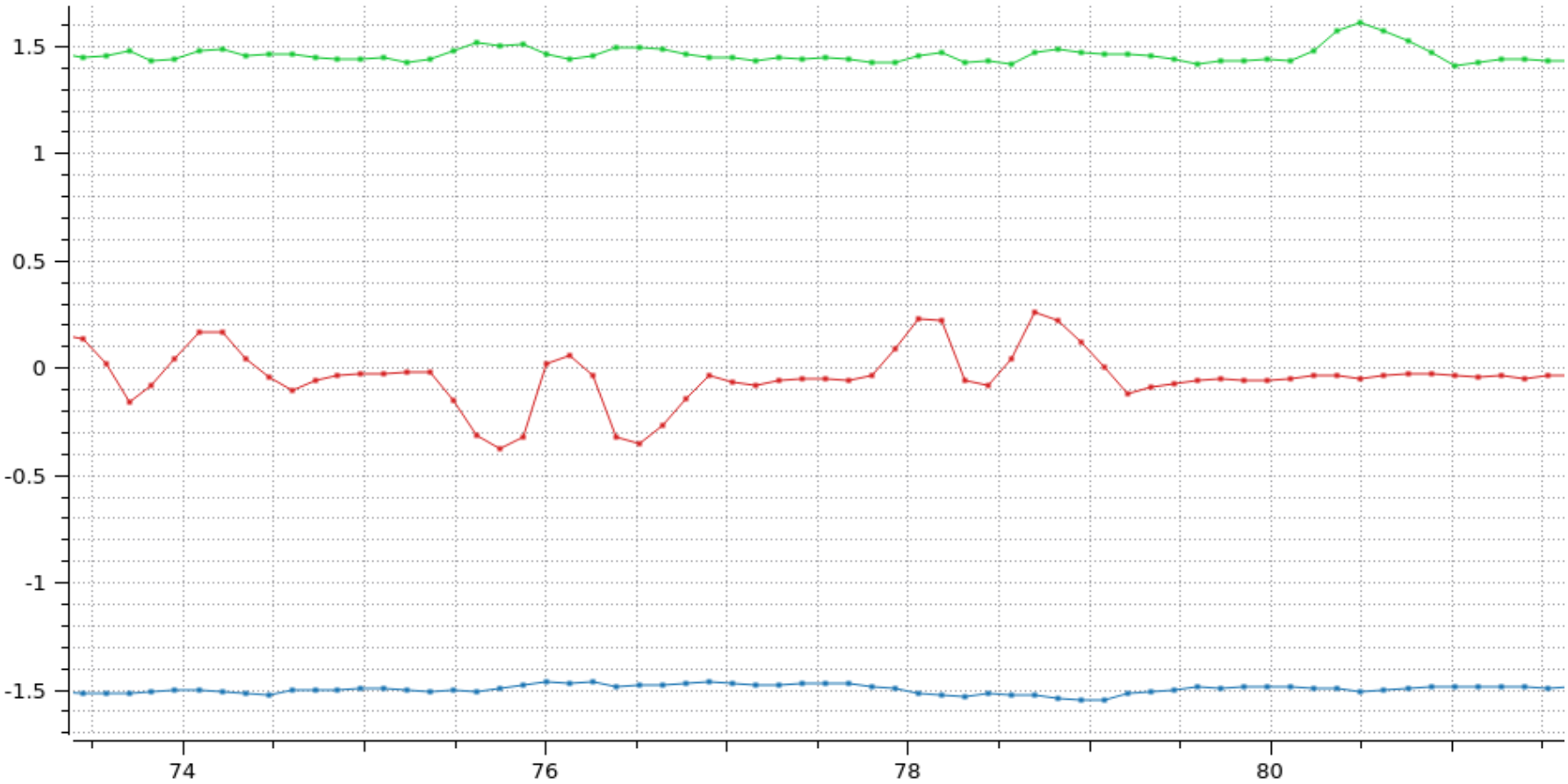
Neck motions: head shaking and nodding

- simple joint-level controller
- sine-wave motions
- user-defined bias, amplitude, frequency
  $\Phi(t) = \Phi_0 + a \cdot \sin(2\pi \cdot t/T + \omega_0)$

- set via node params or dynamic-reconfigure
- initial values based on Gazr tracking

- triggered by teleop
- triggered or inhibited by behaviours/speech

Neck motions: example from Gazr tracking



Gazr face poses converted to Euler angles, tracking around 8 Hz only. Head shaking to the left ($t = 76$), then right($t = 78.25$), then nodding ($t = 80.25$). Amplitudes: 0.45, 0.35, 0.2, period: 400 ms

Four types of eye-motions

1. vergence (focusing near objects)
2. head motion compensation (image stabilization)
3. object tracking (keeping objects in fovea)
4. saccades (scanning the scene)

Eyelid motions:

- eyelid wide open (surprise, staring)
- eyelid closing reflex (protecting the eye)
- eyelid blinking (cleaning the eye)
- winking (social interaction)
- eyelid closing (getting sleepy)

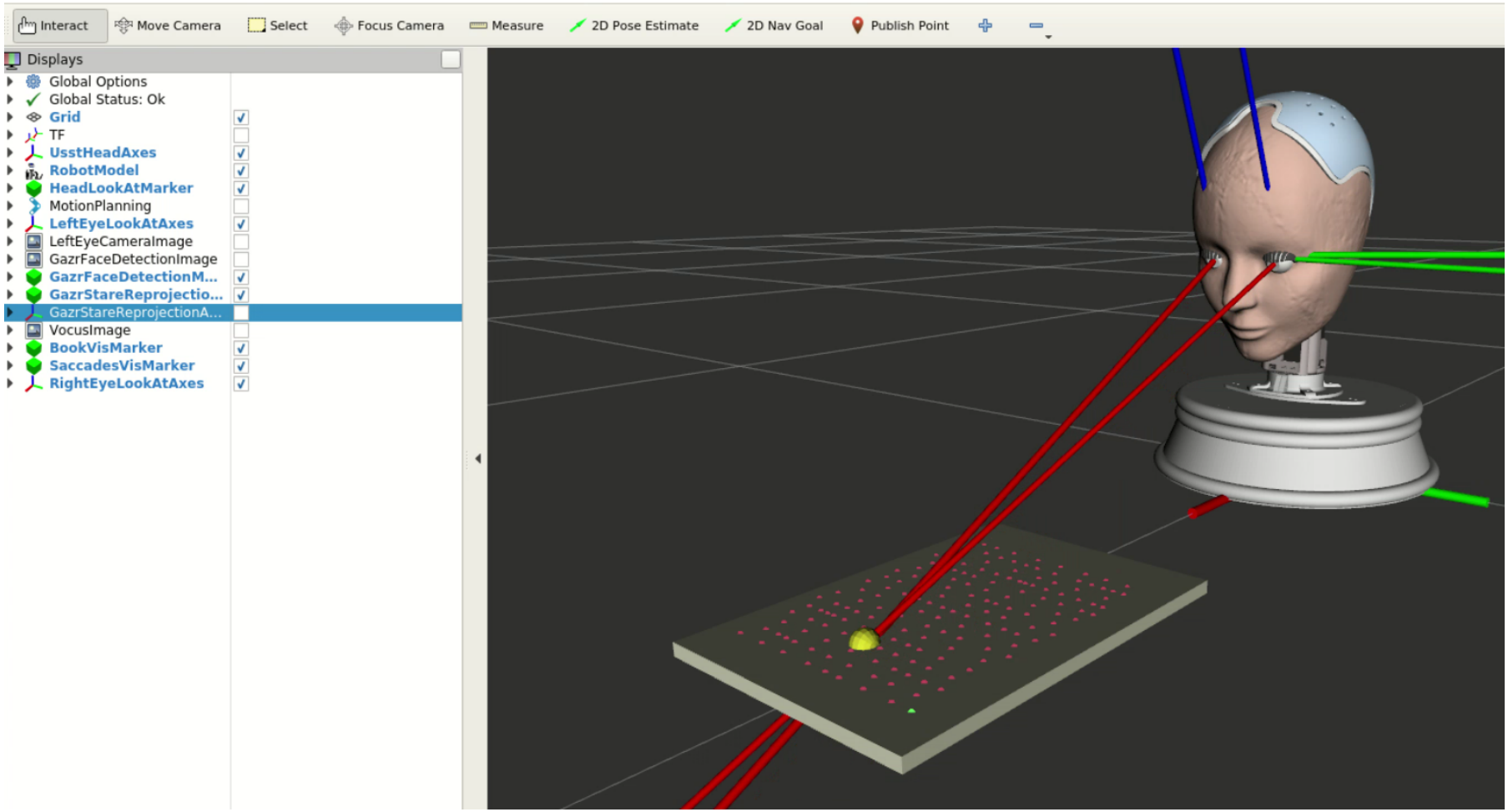
Eye Motions: 1. Vergence

- adjust pan-angle of both eyes
- to look precisely at a given target object
- so that the object is mapped to the fovea

- most social robots do it
- important social clue: where is my partner looking at? which distance?

- my URDF/simuluation supports this
- but ULA-head has coupled eye-pan
- implemented using Bio-IK look-at goals

Eye vergence: Reading a book



bio-ik: 2x look-at goal, minimum-displacement, random delays

Eye Motions: 2. Head-motion compensation

- “rolling shutter” retina
- most animals stabilize the eyes during head turns
- reducing motion blur on the fovea
- eye-pan is reset when joint-limits are reached
- also triggered by inner ear (IMU)

- again, using Bio-IK look-at goals
- want velocity-controlled servos for best performance

- simple self experiment:
  - move hand in front of the eyes: motion blur
  - move head while looking at the hand: sharp image

Eye Motions: 3. Object tracking

- tries to keep target object in fovea
- combination of eye pan/tilt and neck motions
- similar to head motion-compensation
- again, eyes ”reset” when near joint limits

- implemented by Bio-IK look-at goals
- again, want velocity-controlled servos for best performance

Eye Motions: Bio-IK with Look-At Goals

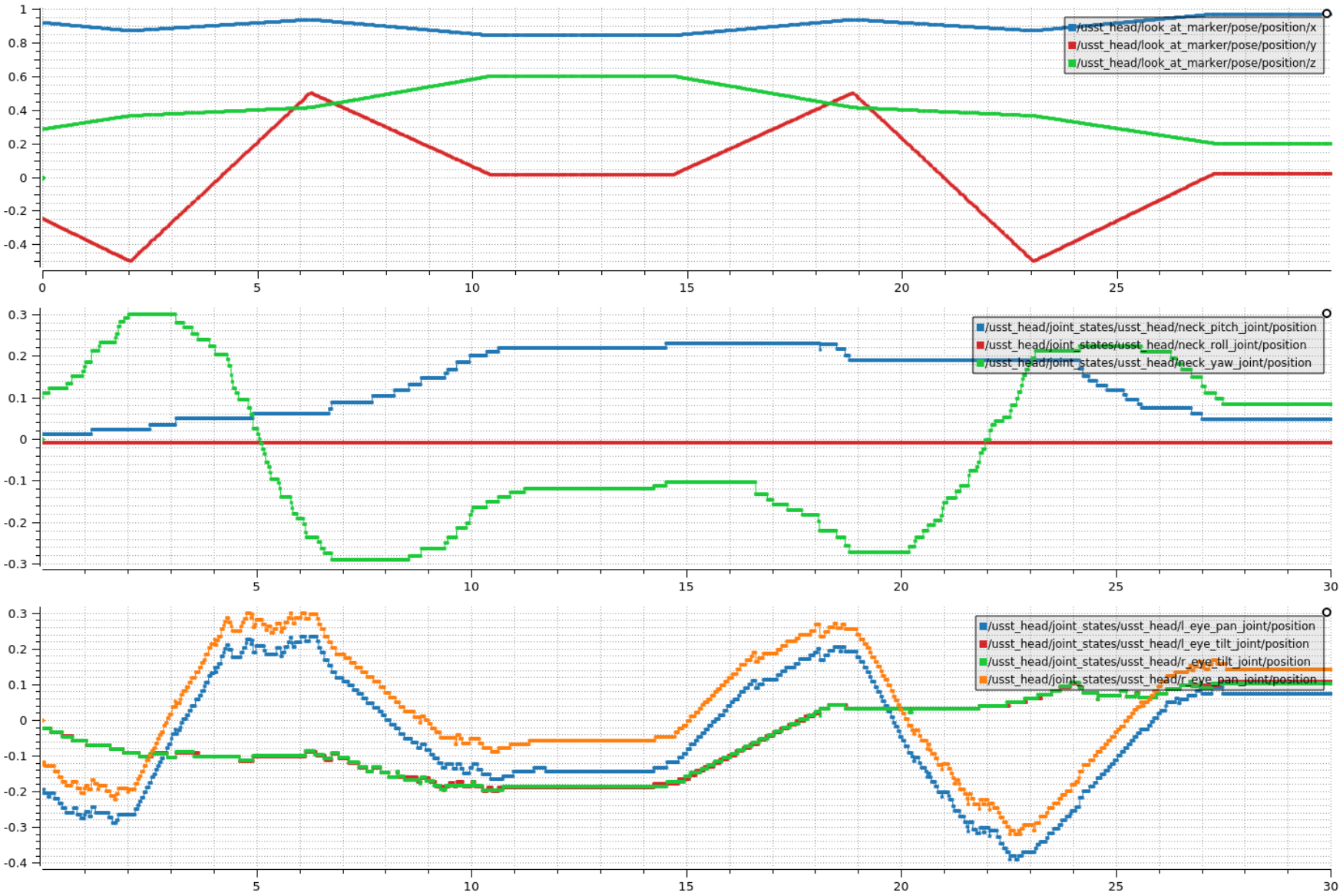
```
request = bio_ik_msgs.msg.IKRequest()
request.group_name = "head_all_group"           # neck and eyes and lids
request.timeout.secs = 0.005
request.approximate = True

l_look_at_goal = bio_ik_msgs.msg.LookAtGoal()  # left eye
l_look_at_goal.link_name = "usst_head/l_eye"   # link from urdf model
l_look_at_goal.weight = 2.0
l_look_at_goal.axis.x = 1.0                    # x points forward y=z=0
l_look_at_goal.target.x = target.x             # same for y and z
request.look_at_goals.append( l_look_at_goal )
...                                            # repeat for right eye

keep_neck_yaw_goal = bio_ik_msgs.msg.JointVariableGoal()
keep_neck_yaw_goal.variable_name = "usst_head/neck_yaw_joint"
keep_neck_yaw_goal.weight = 0.1
keep_neck_yaw_goal.variable_position = joint_state.position[0]
request.joint_variable_goals.append( keep_neck_yaw_goal )
...                                            # repeat for pitch roll

ik_server = rospy.ServiceProxy( "/bio_ik/get_bio_ik", bio_ik_msgs.srv.GetIK)
response  = ik_server( request )
angles    = response.ik_response.solution.joint_state
```

Head motion compensation: Face tracking



top: target (x,y,z) middle: neck (y,p,r) bottom: eye (pan,tilt)

Eye Motions: 4. Ballistic Saccades

- idle eye motions are not smooth at all
- very rapid (”ballistic”) motions
- followed by short fixation of the scene
- 3..5 saccades per second, or
- about 150.000 *Saccades* per day
- motions different for stimuli-driven and memory-driven

- fixation points not randomly distributed
- e.g. eyes + nose + mouth + face boundary, but not chin ...
- typical patterns well researched
- but high inter-person variances
- different saccade patterns for autism, ...

Saccades: Looking at a face

```
if (detected_face):
    select saccade target position
    (25% l-eye, 25% r-eye, 10% nose, 20% mouth, 20% random)
    execute saccade motion
    publish VisualMarkerArray on top of gazr face pose

else:
    select salient object
    ...
```

Eye Motions: blinking

- simple control process
- stationary states: sleeping - awake - wide-open
- random blinking with controllable timing
- deliberate (one eye-) winking

- enable/inhibit via single `activation` topic
- e.g. disabled when danger detected

Eye Motions: want depth from head turning

- eyes (cameras) mounted in front of rotation axes
- neck motions generate disparity images
- try to recover depth information from multiple images

- also used by (one-eyed) people (and animals)

- already mentioned by Brooks in 1993, probably by others
- but I don't remember any implementation - why?
- needs high precision eye+neck pan angles

e.g., Newcombe and Davison, Live Dense Reconstruction with a Single Moving Camera, CVPR 2010

Object tracking / Neck-motion compensation

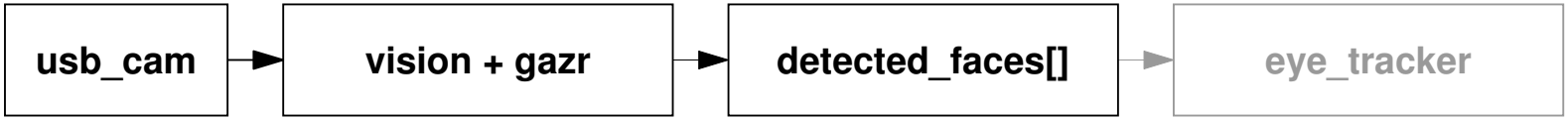
- target direction from vision pipeline
  (Vocus, Yolo, Gazr)

- create bio-ik look-at goals
- add some head-motion “momentum” term
- add MinimalDisplacementGoal

- RC-servo hardware limitation:
  - servos are position control only
  - low servo update rate (50 Hz)
  - servos not synchronized
  - so, slightly shaky

  - replace with bus-servos (e.g. Dynamixel/Feetech)
  - or direct-drive motors

Staring contest...




```
process stare_detection:
  foreach( face: detected_faces ):
    if have_eye_tracker:
      d = |reproject_eye_direction-my_eye|
    else:
      d = |reproject_gazr_direction-my_eye|
    if (d < threshold): // somebody staring
      stare_duration[ face ] += delta
    else:
      stare_duration[ face ] = 0
```

```
process stare_handler:
  t,challenger = find_max( stare_duration )
  if (t > threshold2): // long stare
    if looks_dangerous( challenger ):
      bashful_look_away( challenger );
    else:
      stare_back( challenger );
  elif (t > threshold1): // first stare
    look_at( challenger );
  else:
    relaunch_previous_tasks();
```

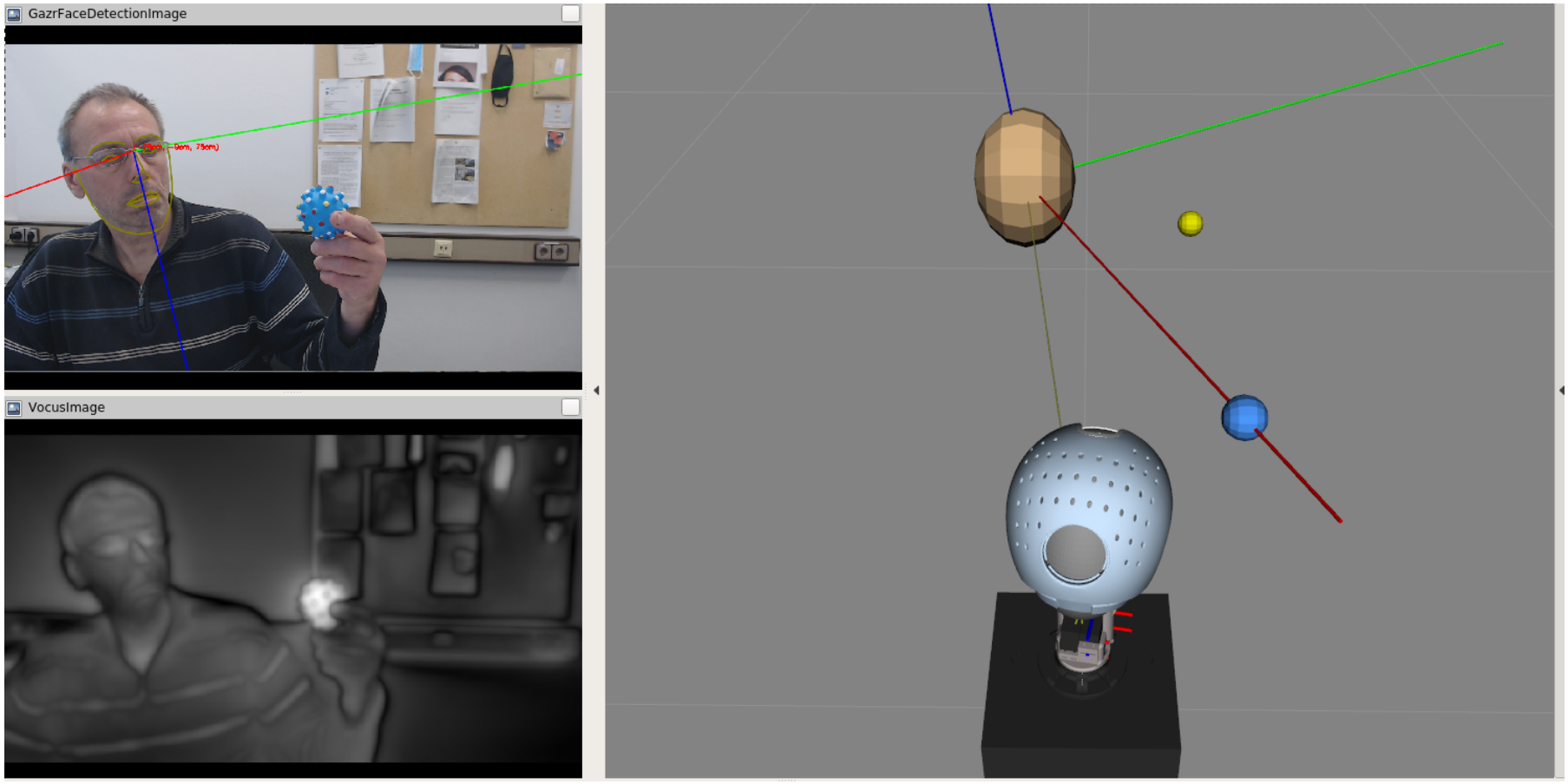
```
function bashful_look_away( face ):
```

```
function look_at( face ):
```

```
function ...()
```

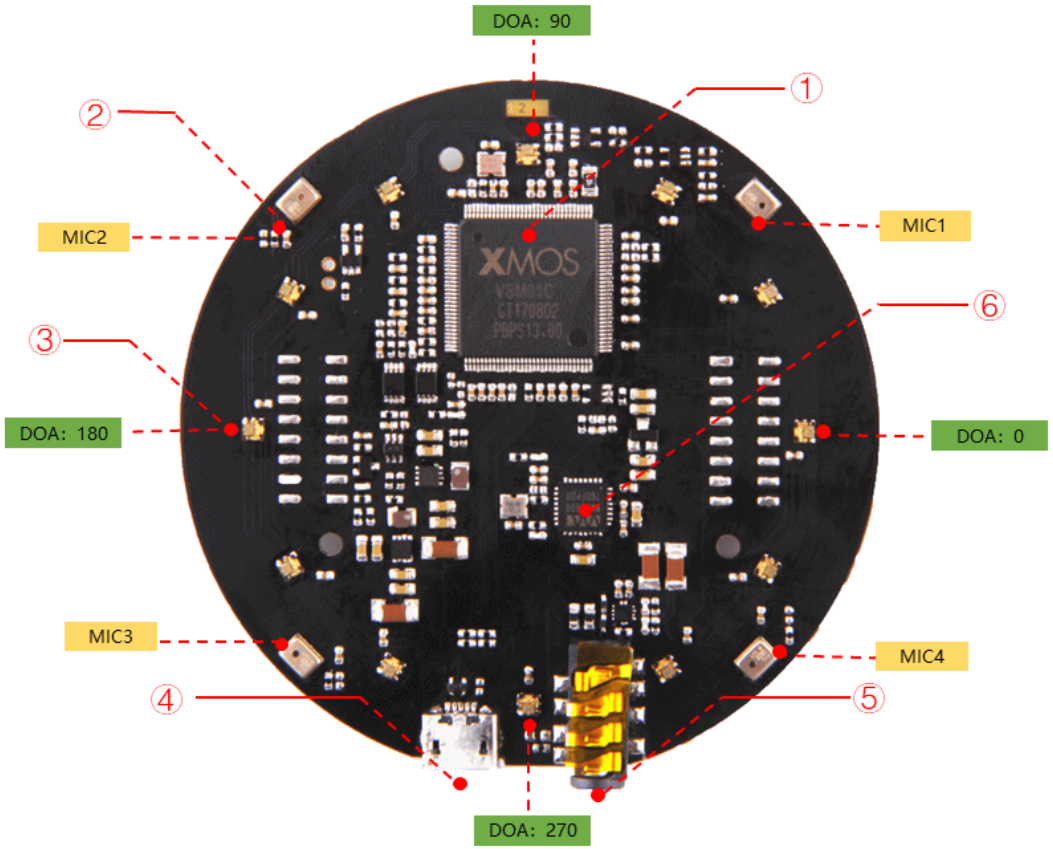
```
function stare_back( challenger ):
  save_current_eye_tasks();
  disable_eye_blinking();
  raise_eyebrows();
  look_at( challenger );
```

Staring detection (gaze direction reprojection)



yellow: look-at target, brown: gazr face, blue: human gaze reprojection

ReSpeaker 4-Mic Array

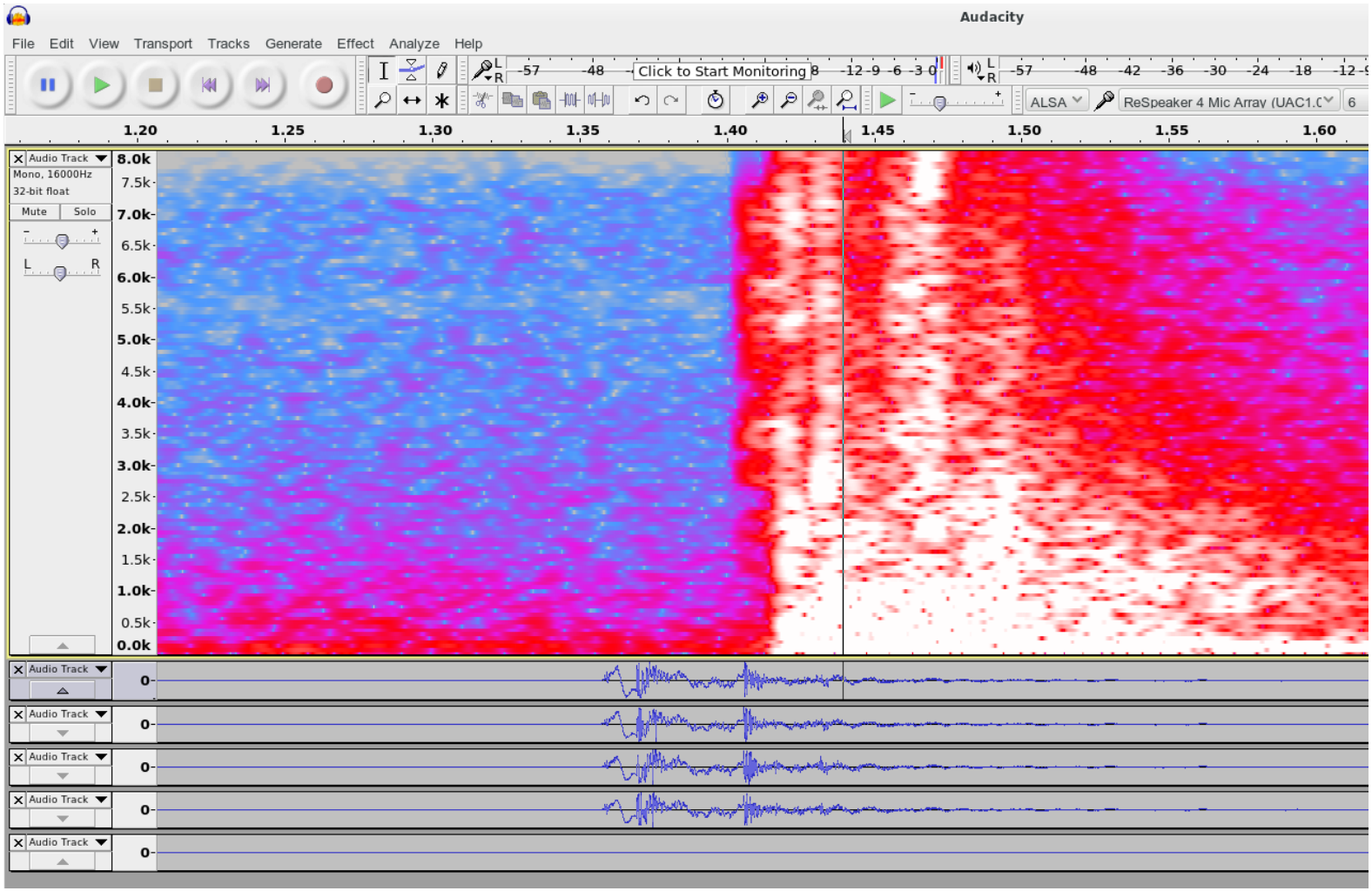


Key Benefits

- Far-field Voice Capture
- Support USB Audio Class 1.0 (UAC 1.0)
- Four Microphones Array
- 12 Programmable RGB LED Indicators
- Speech Algorithms and Features
  - Voice Activity Detection
  - Direction of Arrival
  - Beamforming
  - Noise Suppression
  - De-reverberation
  - Acoustic Echo Cancellation

https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/

ReSpeaker + Audacity



channel 1: combined voice (50 ms delay), channels 2-5: microphones, channel 6: headphone output

ReSpeaker + ROS

- ROS node, realtime audio for all microphone channels
  - channels 2,3,4,5: raw microphone data
  - channel 1: speech (merge channels, normalize volume)
  - publishes sound source localization (direction)
- interfaces to Python SpeechRecognition
- depends on (broken) catkin_virtualenv: remove

```
roslaunch respeaker_ros respeaker.launch
rostopic echo /sound_direction      # Result of DoA
rostopic echo /sound_localization   # Result of DoA as Pose
rostopic echo /is_speeching         # Result of VAD
rostopic echo /audio                # Raw audio
rostopic echo /speech_audio         # Audio data while speeching
```

github.com/furushchev/respeaker_ros
python -m pip install pyusb pixel_ring
vi CMakeLists.txt (remove python_virtualenv) catkin_make

Python SpeechRecognition (3.8.1)

A Python library for speech recognition, with support for several engines and APIs, online and offline:

- CMU Sphinx (works offline)
- Snowboy Hotword Detection (works offline)

- Google Speech Recognition
- Google Cloud Speech, IBM Speech to Text, Microsoft Bing Voice Reconition, Wit.ai, Houndify API

- Demo (using Google test key, could be revoked at any time)

pip install SpeechRecognition
python -m speech_recognition
pypi.org/project/SpeechRecognition

ROS: Text-to-Speech

Mozilla TTS?

- choice of language and vocoder models
- one deep-network trained female voice
- very good speech quality
- but long sentences truncated

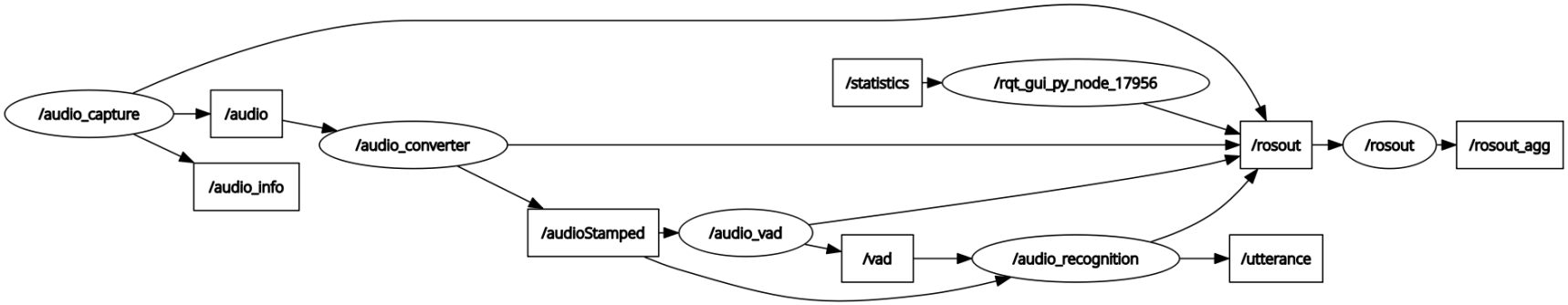
Acapela

- already used at IROS-19 competition (PAL Tiago)
- supports many languages
- choice of different English voices

other systems?

https://www.acapela-group.com/demos/, https://github.com/mozilla/TTS

Speech Recognition




```
cd src/tbd_audio_stack/tbd_audio_recognition_deepspeech && mkdir models && cd models
wget github.com/mozilla/DeepSpeech/releases/download/v0.8.2/deepspeech-0.8.2-models.pbmm
wget github.com/mozilla/DeepSpeech/releases/download/v0.8.2/deepspeech-0.8.2-models.scorer
catkin build -DPYTHON_VERSION=3
roslaunch tbd_audio_recognition_deepspeech run_recognition.launch
rostopic echo /utterance
```

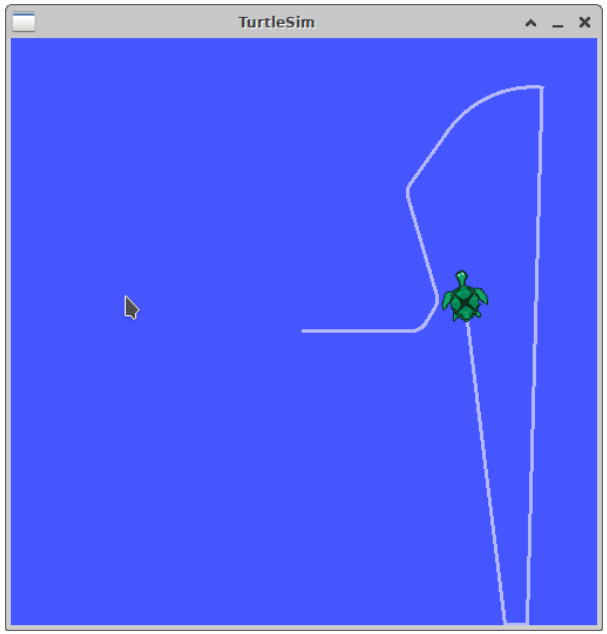
```
header: seq: 41 stamp: secs: 1660751696 ...
text: "bring me the red apple"
confidence: -30.367855072021484
word_list:
  - bring
  - me
  - the
  - red
  - apple
timing_list: [17, 33, 52, 61, 80]
```

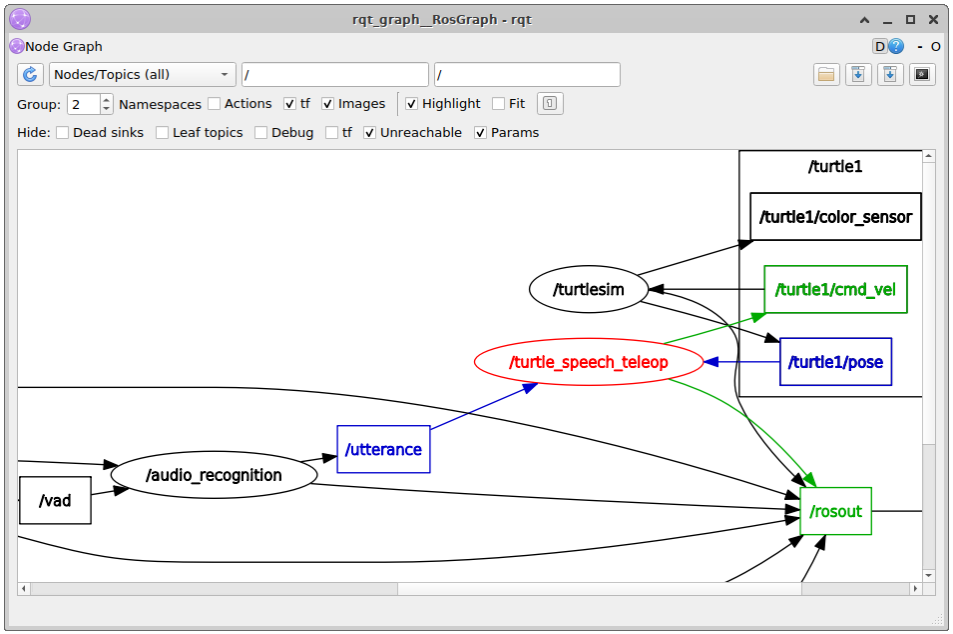
```
header: seq: 42 stamp: secs: 1660751708 ...
text: "university of hamburg"
confidence: -20.71257781982422
word_list:
  - university
  - of
  - hamburg
timing_list: [18, 59, 77]

...
```

Mozilla Deep Speech, https://github.com/mozilla/DeepSpeech, https://github.com/CMU-TBD/tbd_audio_stack

Keyword Recognition: Turtlesim speech teleop






```
roslaunch tams_deep_speech turtlesim_speech_teleop
"start left straight left ... (watchdog)
 right walk slower right stop ... (watchdog)
 right right right jump ... (watchdog)
 left left bla left run ..."

"left" -> { left, lift, let, it, ...}
"reverse" -> { reverse, rivers, ...}
```

git.crossmodal-learning.org/norman.hendrich/tams_deep_speech

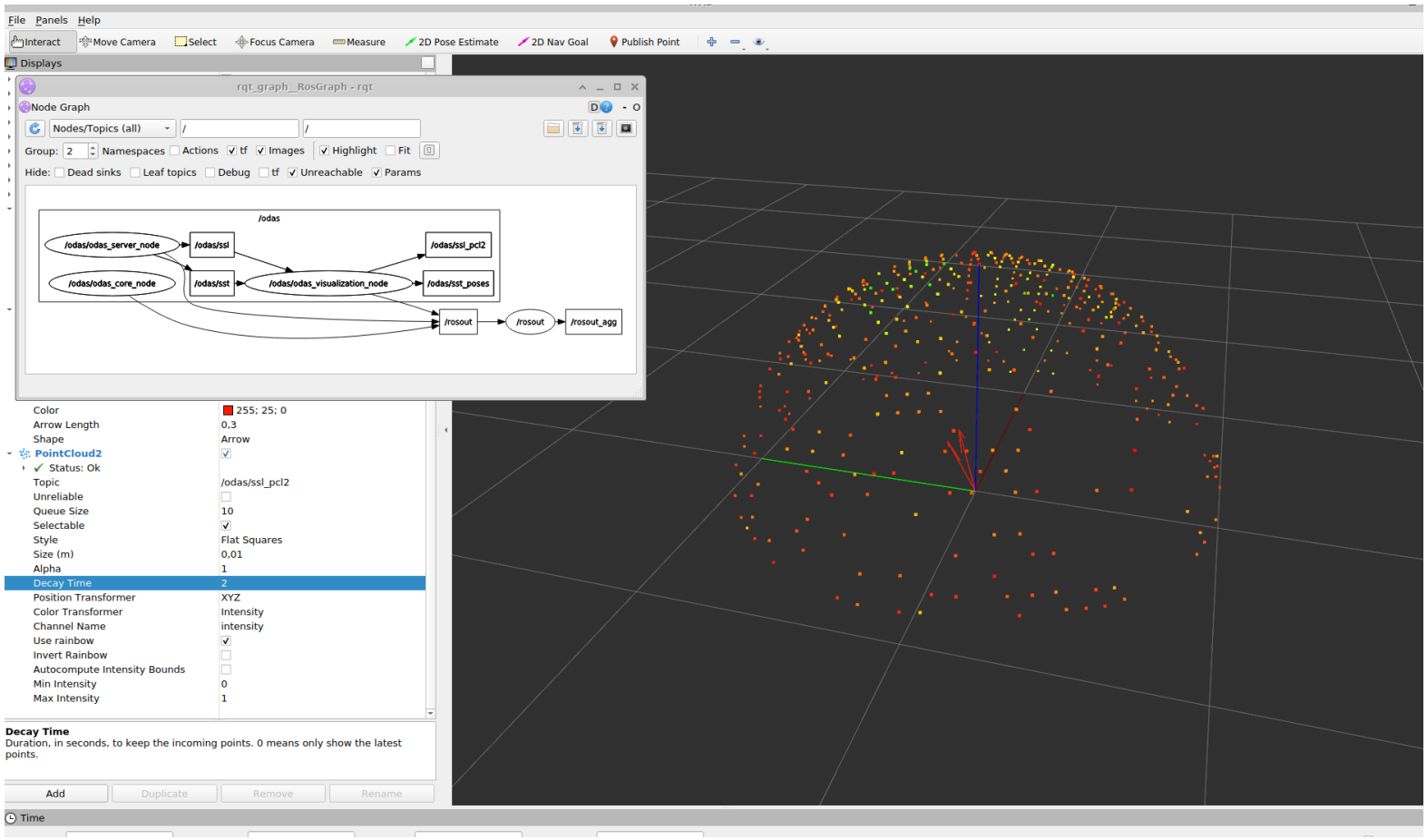
Soundex, Metaphone, and Levenshtein

vanilla keyword recognition not robust enough

- phonetic similarity (Soundex, Metaphone)
  - soundex example: 'robert', 'rupert' $\rightarrow$ 'R163'
- fuzzy word matching (Levenshtein distance)
  - number of steps to reach one string from another
  - addition, deletion, modification of single character
  - 'kitten' $\rightarrow$ 'sitten' $\rightarrow$ 'sittin' $\rightarrow$ 'sitting'
  - computationally expensive for long words / large distances
- ...

en.wikipedia.org/wiki/Soundex, en.wikipedia.org/wiki/Levenshtein_distance, pip install phonetics cologne_phonetics

ODAS: Sound Source Localization and Tracking



(rviz+ODAS+ROS with respeaker 4-microphone input)

Open Embedded Audio System github.com/introlab/odas, github.com/introlab/odas_ros

What is missing? User Interaction Example

multimodal user-interaction

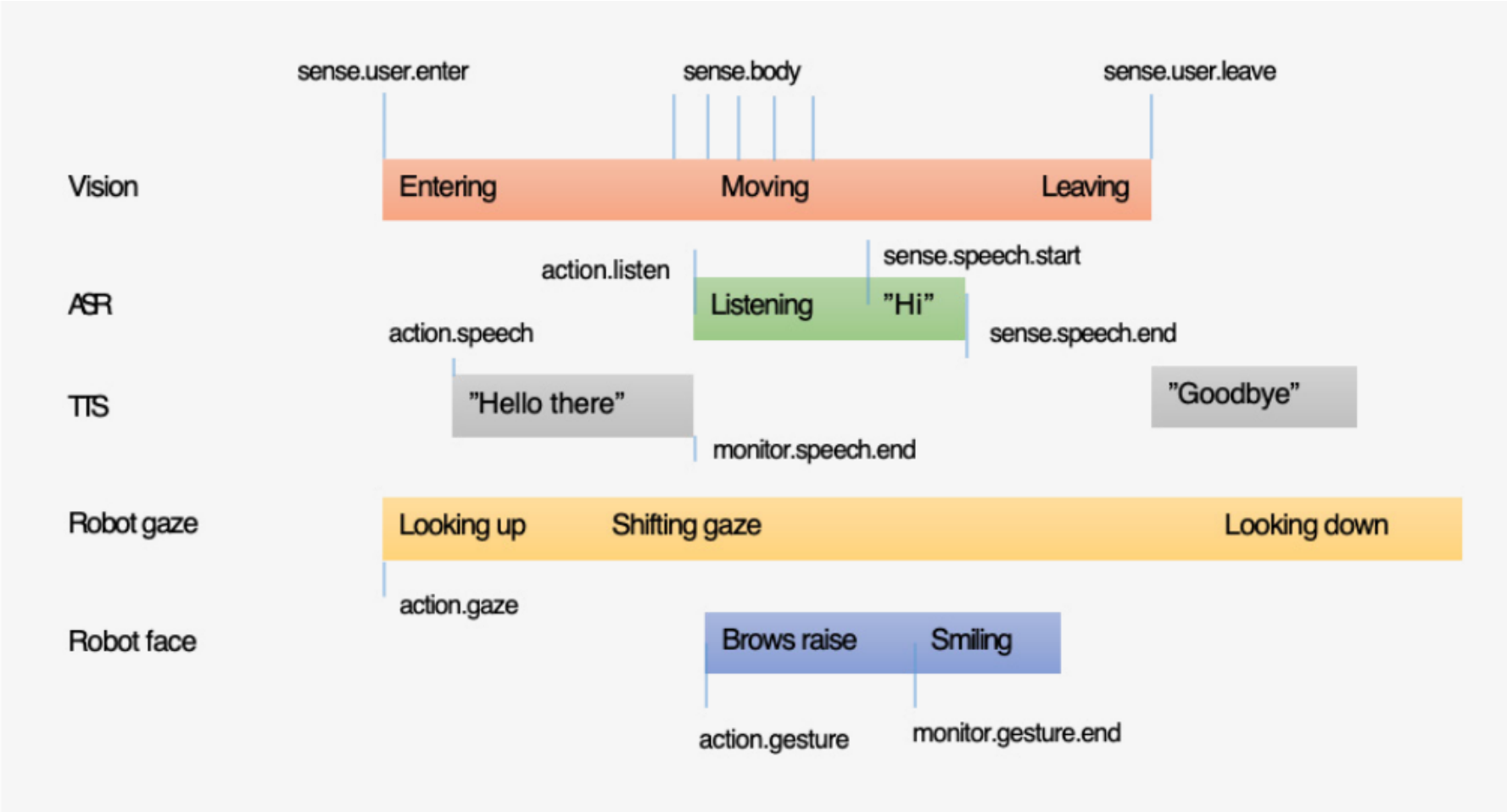


Figure 1: Different events occurring in different modules of the system, illustrated on a timeline.

furhat.com, Furhat Robot Manual

Future work: What? Who?

- flexible skin (German)

- ros4hri (Norman)
- realistic eye-motions (?)
- speech and dialog system (?)

- duplicate existing demos? (e.g. iCub @ WTM)
- re-implement cognitive architectures? (e.g. Ishiguro's stuff)
- learning from observing humans?

- your ideas here…
- BSc and MSc theses?

ROS4HRI: “social signal processing”

- recent initiative
- implement HRI basics for ROS1

- 18-DOF human URDF model (sticks+spheres / ragdoll)

- face/body/voice identifiers
- person IDs and tracking
- based on OpenPose/OpenFace

- gaze and group interaction
- age, gender, emotions
- based on Intel OpenVINO

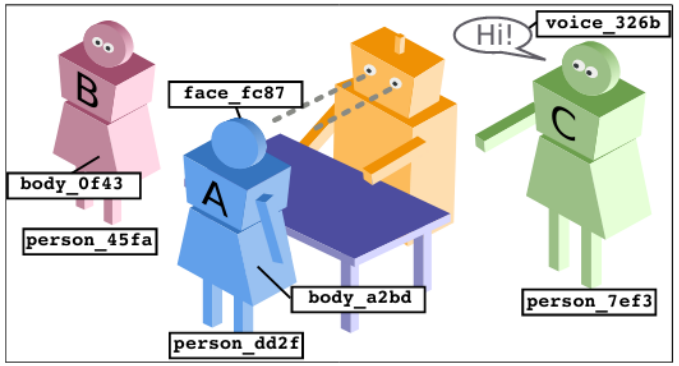
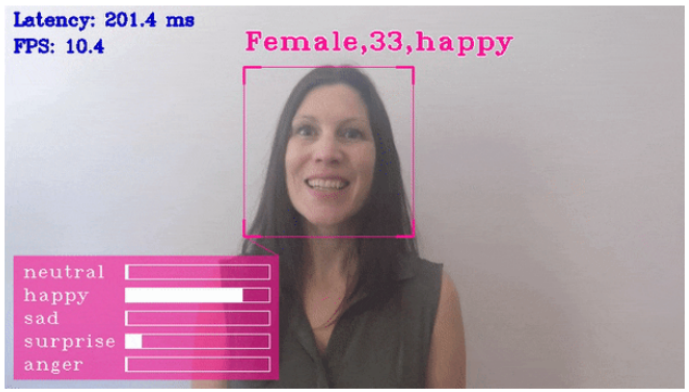


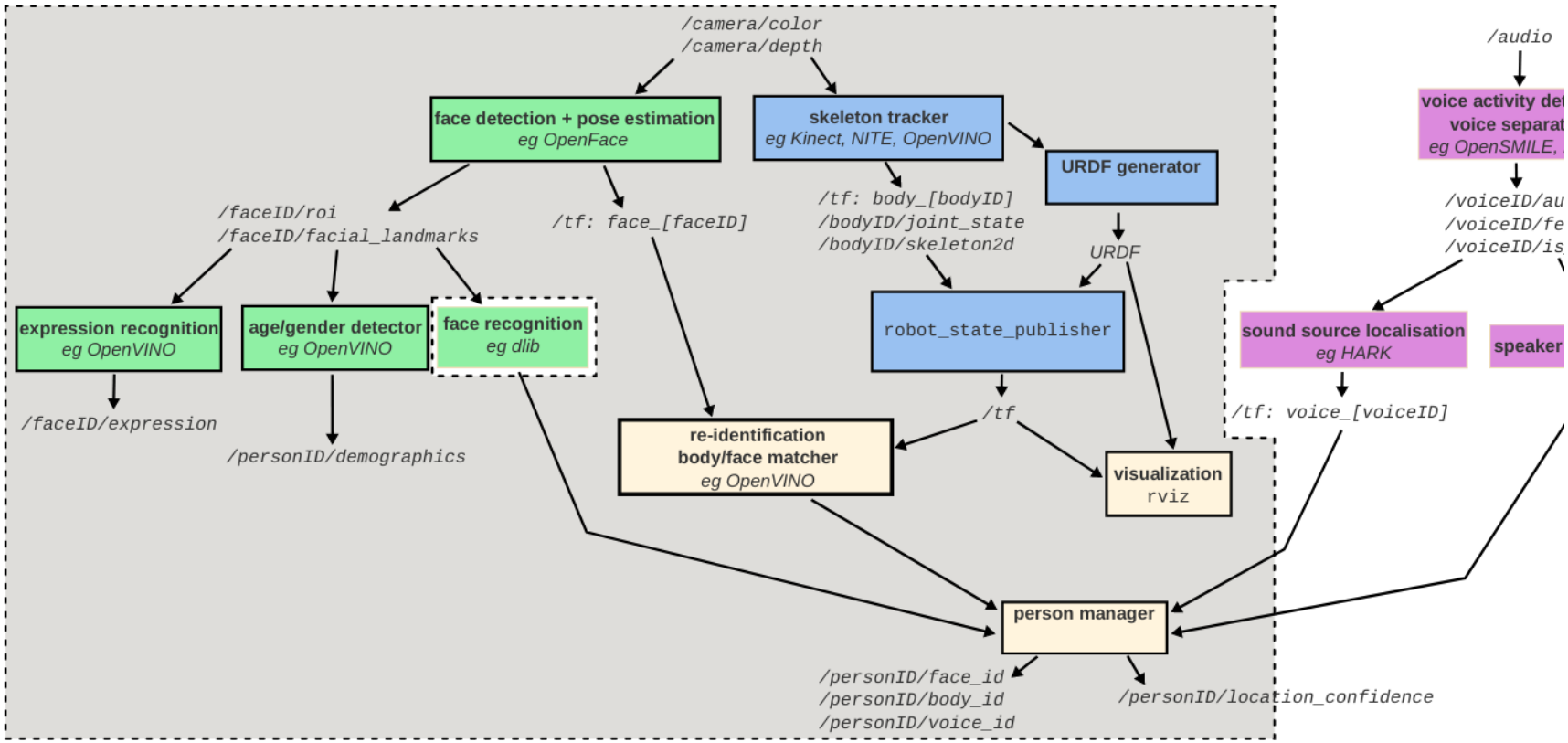
Fig. 1: In this situation: A is facing the robot: A gets a unique `faceID`, a unique `bodyID`, and a unique `personID`; B’s body is visible to the robot, but not the face: B only gets a `bodyID` and `personID`; C is not seen, but heard: C gets a `voiceID` and a `personID`.

Interactive Face Detection C++ Demo



Y. Mohamed and S. Lemaignan, ROS for Human-Robot Interaction, IROS 2021, github.com/ros4hri
github.com/CMU-Perceptual-Computing-Lab/openpose
github.com/openvinotoolkit/openvino

ROS4HRI Pipeline



- note: no audio yet, no ego-sphere or similar

Y. Mohamed and S. Lemaignan, ROS for Human-Robot Interaction, IROS 2021, github.com/ros4hri

OpenVINO: pre-trained networks and demos

https://github.com/openvinotoolkit/open_model_zoo/tree/master/demos 110% Suchen

README.md

- Crossroad Camera C++ Demo - Person Detection followed by the Person Attributes Recognition and Person Reidentification Retail, supports images/video and camera inputs.
- Deblurring Python* Demo - Demo for deblurring the input images.
- Face Detection MTCNN Python* Demo - The demo demonstrates how to run MTCNN face detection model to detect faces on images.
- Face Detection MTCNN C++ G-API* Demo - The demo demonstrates how to run MTCNN face detection model to detect faces on images. G-API version.
- Face Recognition Python* Demo - The interactive face recognition demo.
- Formula Recognition Python* Demo - The demo demonstrates how to run Im2latex formula recognition models and recognize latex formulas.
- Gaze Estimation C++ Demo - Face detection followed by gaze estimation, head pose estimation and facial landmarks regression.
- Gaze Estimation C++ G-API* Demo - Face detection followed by gaze estimation, head pose estimation and facial landmarks regression. G-API version.
- Gesture Recognition Python* Demo - Demo application for Gesture Recognition algorithm (e.g. American Sign Language gestures), which classifies gesture actions that are being performed on input video.
- Gesture Recognition C++ G-API* Demo - Demo application for Gesture Recognition algorithm (e.g. American Sign Language gestures), which classifies gesture actions that are being performed on input video. G-API version.
- GPT-2 Text Prediction Python* Demo - GPT-2 text prediction demo.
- Handwritten Text Recognition Python* Demo - The demo demonstrates how to run Handwritten Text Recognition models for Japanese, Simplified Chinese and English.
- Human Pose Estimation C++ Demo - Human pose estimation demo.
- Human Pose Estimation Python* Demo - Human pose estimation demo.

more than 70 ready-to-run networks: github.com/openvinotoolkit/openvino

Multimodal Perception: Sensory-Ego-Sphere?

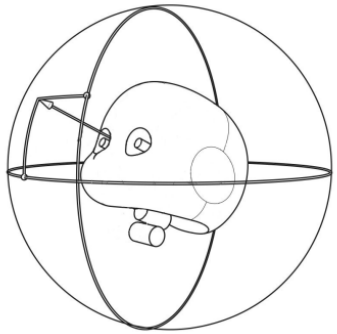


Fig. 1. The iCub ego-sphere: a spherical, torso-based projection surface for egocentric, multimodal saliency information.

- saliency + habituation
- visual+audio+tactile events

- mapped to spherical coordinates
- image-based representation
- controls head-motion and gaze

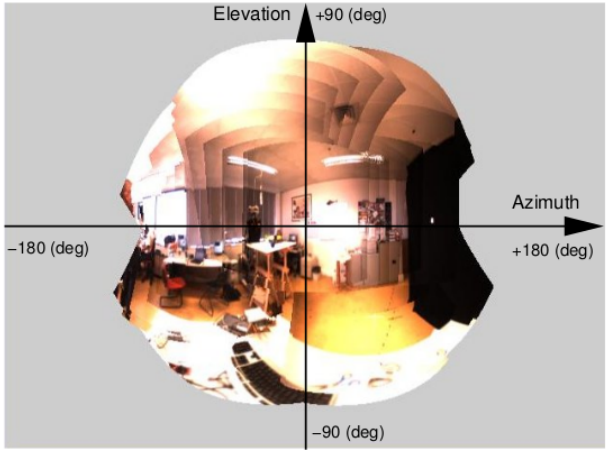


Fig. 6. Experiment V-A: spherical mosaic

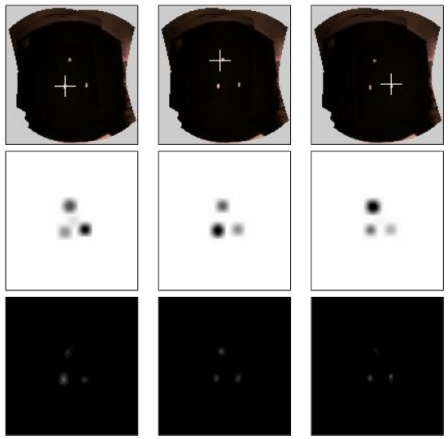
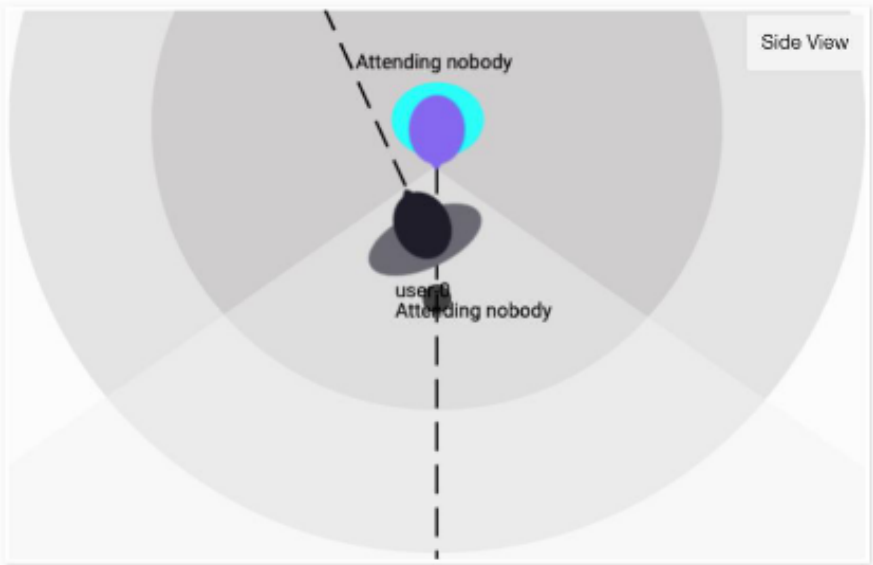


Fig. 7. Experiment V-B (from left to right): Top: selected target locations on the ego-sphere mosaic; middle: corresponding IOR maps (inhibited areas in black); bottom: egocentric saliency map

J. Ruesch, A. Lopes, A. Bernardino, J. Hörnstein, J. Santos-Victor, R. Pfeifer, Multimodal Saliency-based Bottom-Up Attention. A Framework for the Humanoid Robot iCub, ICRA 2008

Compare: Furhat Attention Panel

Your robot can detect and track the presence of users within the robots interaction space[4] using it's onboard camera and the sensory perception features of FurhatOS. The Attention Panel provides an abstract graphical representation of the robots attention model and supports viewing the model from a top down (plan) and side (elevation) view point.



The Attention panel represents your Robot using the blue/purple icon at the top dead center of the representation.

Users that have been detected by the robot are denoted using a gray/black icon and given a sequential identity of the form `User-N`. As the users move about the interaction space the attention panel updates their position dynamically.

furhat.com, Furhat Robot Manual

Human Saccades: Recording Approach

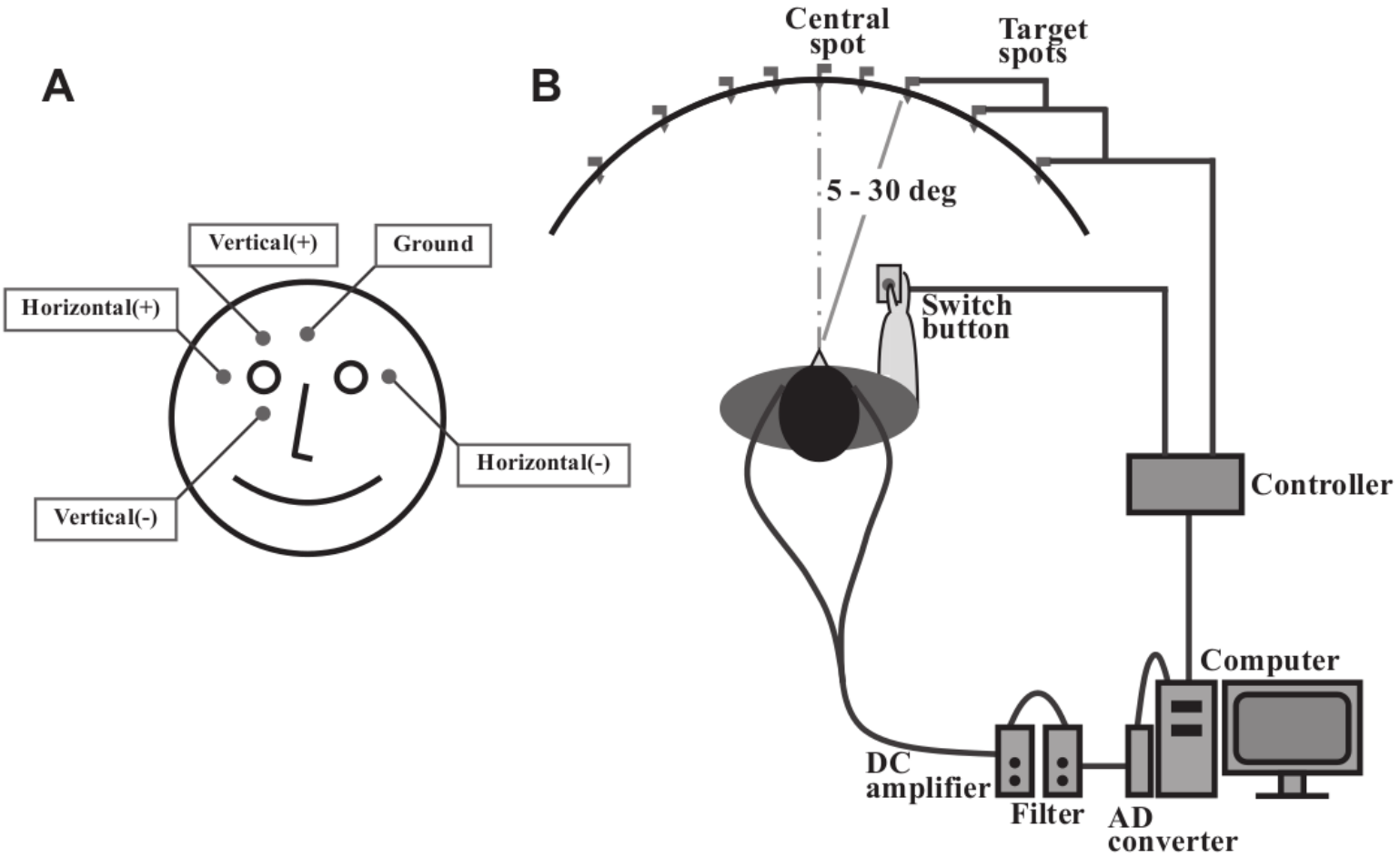


Fig. 1. Typical setup for recording an electrooculogram. **A**. For recording horizontal saccades, electrodes are placed at the bilateral outer canthi, whereas for recording vertical saccades, electrodes are placed above and below one eye. **B**. Subjects are seated in front of a black, concave, dome-shaped screen measuring 90 cm in diameter that contains light-emitting diodes embedded in pinholes, which serve as the fixation points and saccade targets. The subject holds a microswitch button connected to the microcomputer, allowing the subject to initiate and terminate a task trial by pressing and releasing the button. The target point is turned on at a random location 5, 10, 20, or 30 degrees horizontally to the left or right of the central fixation point.

Saccades: Traces

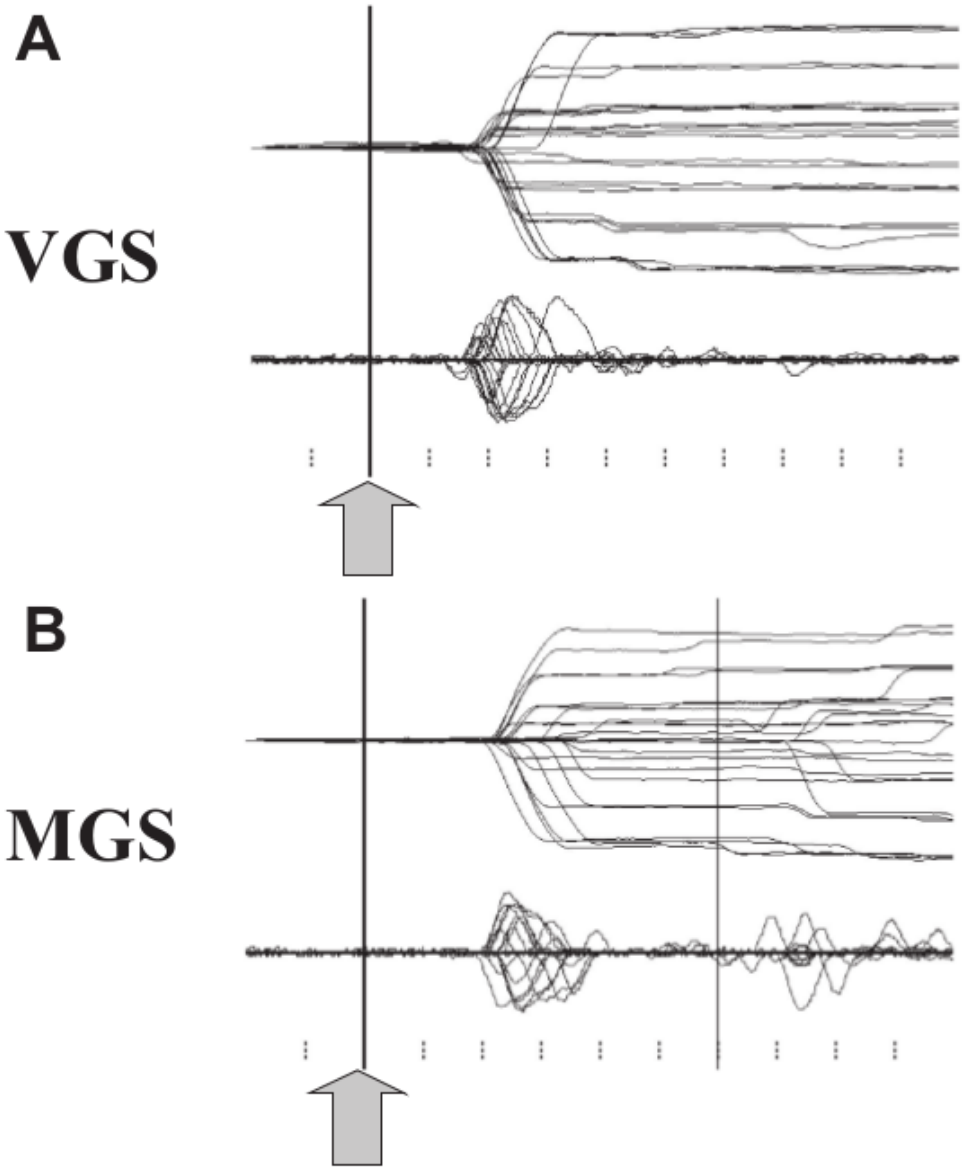


Fig. 3. Example of saccade records in a normal subject. **A**. Saccades records of visually guided saccade (VGS). **B**. Memory-guided saccade (MGS). A total of 25 VGS and MGS traces are superimposed, time-locked to the signal instructing the start of saccades, *i.e.*, presentation of a target (VGS) or offset of the central fixation spot (MGS, shown by arrows and vertical bars). Lower traces in each figure depict the velocity profile of the saccades. The horizontal axis gives the time, and the vertical axis gives the eye position (upper trace) or velocity (lower traces). Ticks below are marked at 100-ms intervals.

Saccades: Velocity Profile

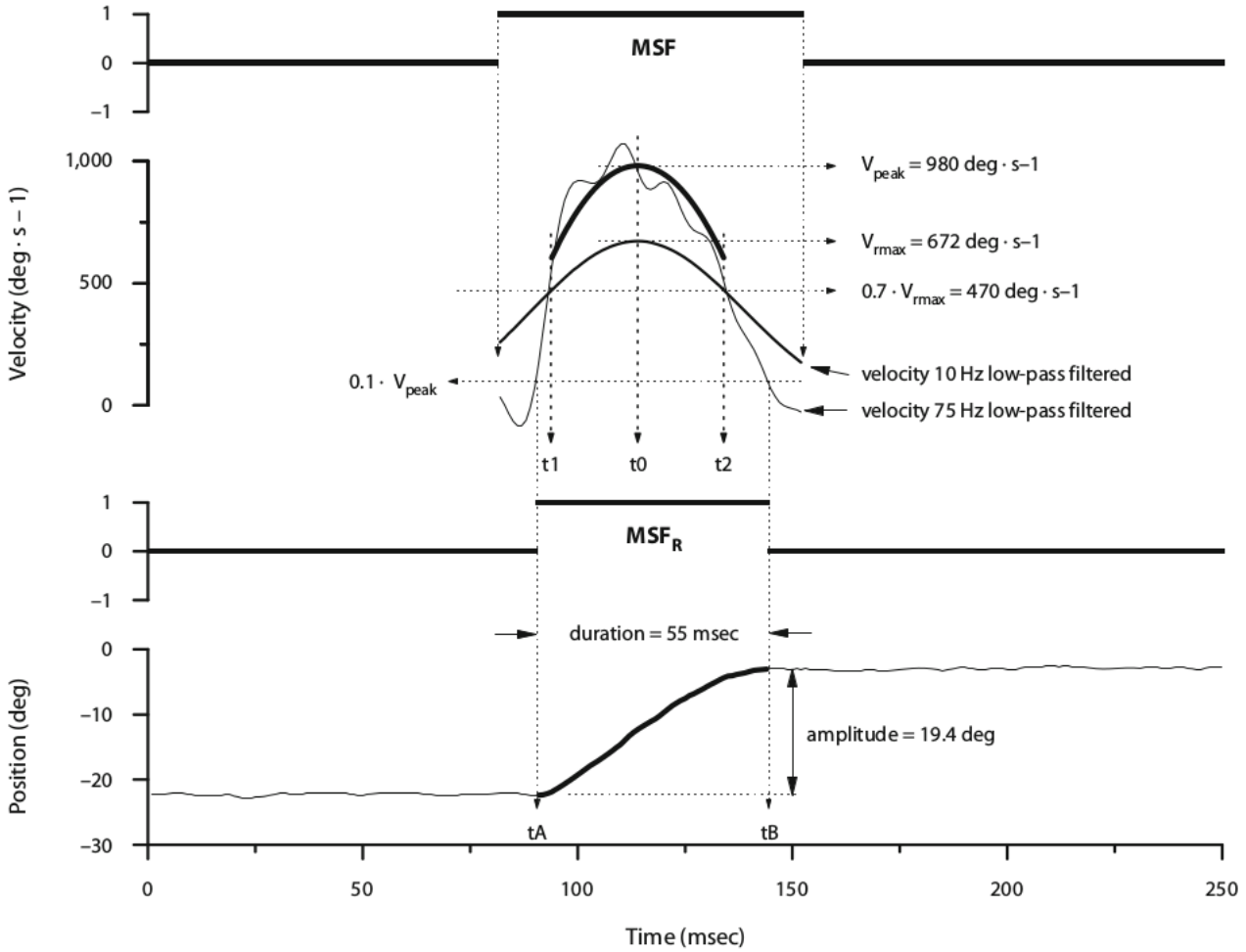


Figure 4. The parameter calculation demonstrated by the saccade from Figure 2A with doubled time resolution. The membership function (MSF) from the detection algorithm is displayed in trace 1. Only data within MSF ≠ 0 are calculated. The 10-Hz low-pass filtered velocity is displayed to find the relative maxima of velocity (V_{rmax}) and t0. The interval t1–t2 is defined by data greater than 0.7 · V_{rmax}. The 75-Hz low-pass filtered noisy velocity is fitted with a second-order function around f0 during the interval t1–t2. The absolute value of the fit represents the maximum velocity (V_{peak}) of the saccade. The beginning and end of the saccade are defined by the samples tA and tB, where the velocity becomes smaller than 0.1 · V_{peak}. Trace 3 displays the revised membership function (MSF_R), and trace 4 shows the eye position before and after the detected saccade (thin line) and during the detected saccade (thick line).

Saccades: Neural Mechanism

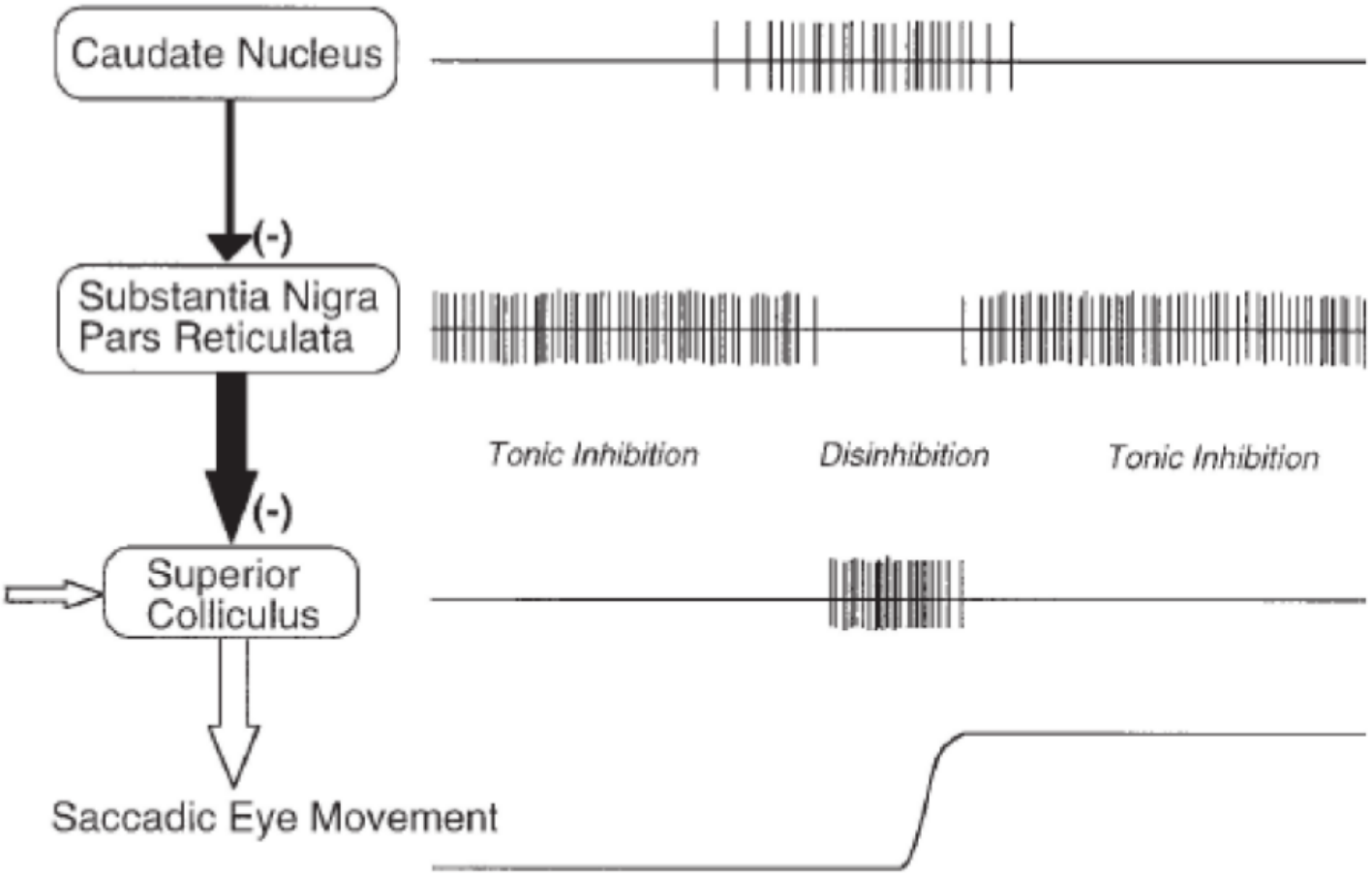


Fig. 5. Neural mechanism for generating memory-guided saccades (MGSs). Hikosaka *et al.* showed that the initiation of an MGS is mediated by cortical commands projected via the caudate nucleus and substantia nigra pars reticulata (SNr) to the superior colliculus (SC). This pathway corresponds to the direct pathway of the basal ganglia (BG). It contains two inhibitory neurons, *i.e.*, from the caudate to the SNr, and then from the SNr to the SC, hence the name double inhibition pathway. A phasic inhibition of the high-frequency firing of the SNr by the caudate nucleus disinhibits the downstream SC, allowing a saccade to occur. Reproduced with permission from Hikosaka *et al.* (2000).[58)]

Saccades: Control (theory)

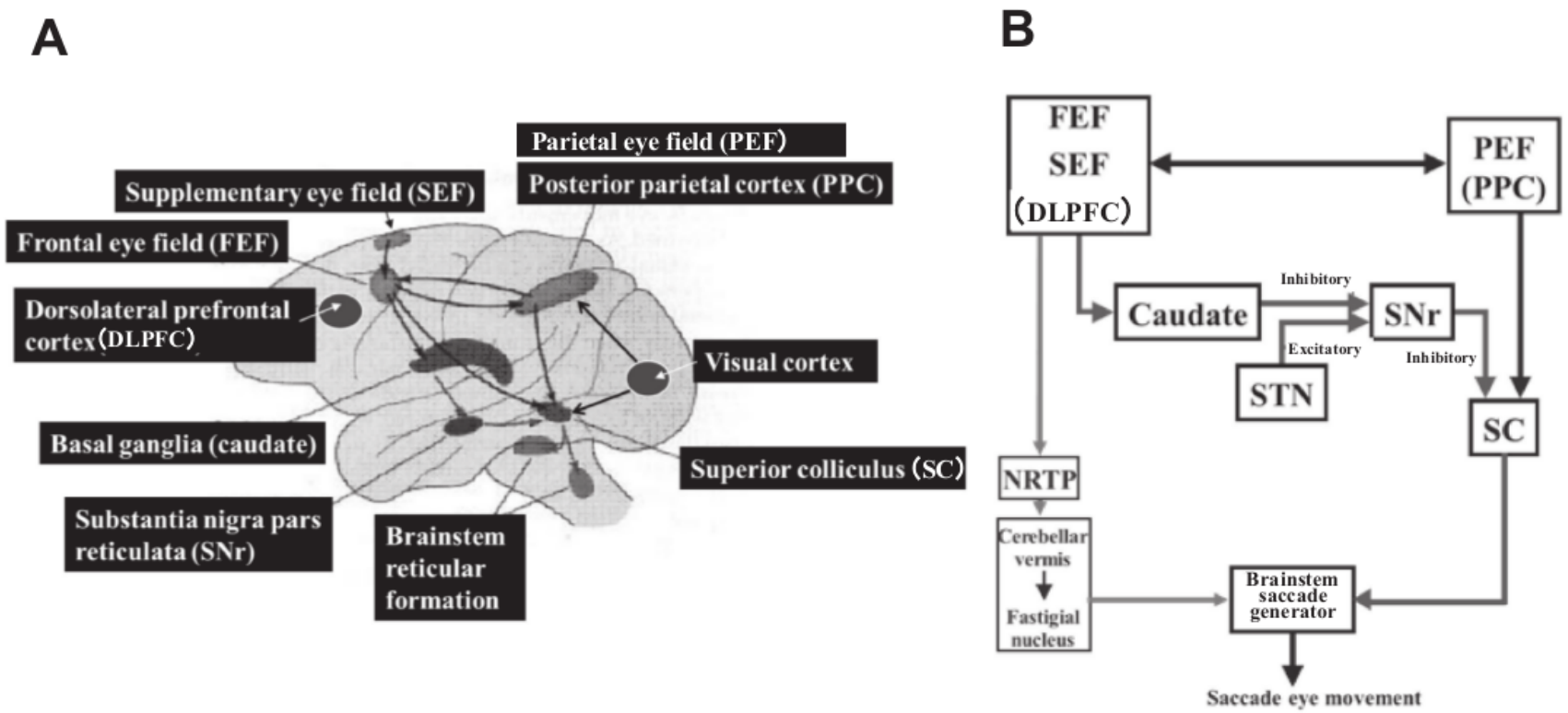
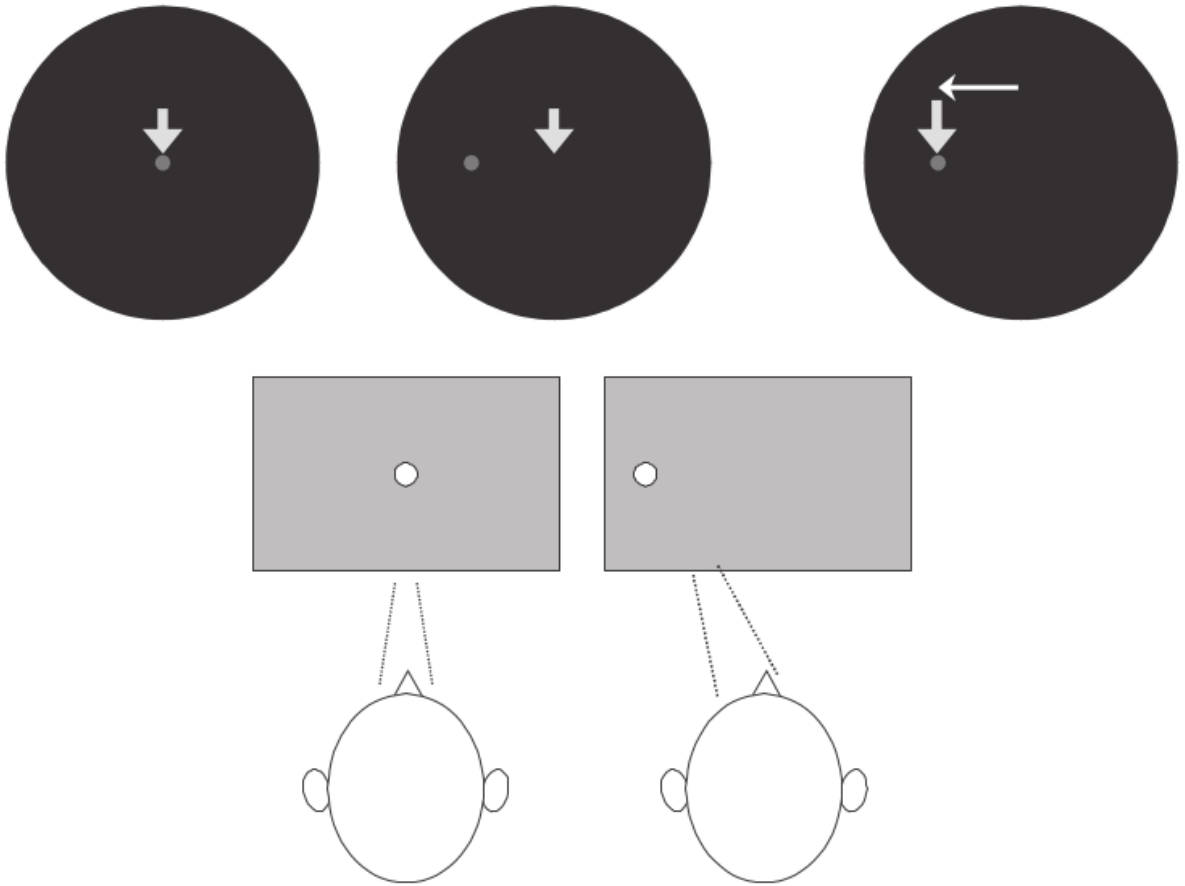


Fig. 4. **A**. Main neural structures for controlling saccades. **B**. Schematic diagram depicting the involvement of the basal ganglia and cerebellum in saccade generation. DLPFC: dorsolateral prefrontal cortex, FEF: frontal eye field, NRTP: nucleus reticularis tegmenti pontis, PEF: parietal eye field, PPC: posterior parietal cortex, SC: superior colliculus, SEF: supplementary eye field, SNr: substantia nigra pars reticulata, STN: subthalamic nucleus.

Saccades: Visually Guided (salient region)

A **Visually guided saccade (VGS)**



Saccades: Memory Guided

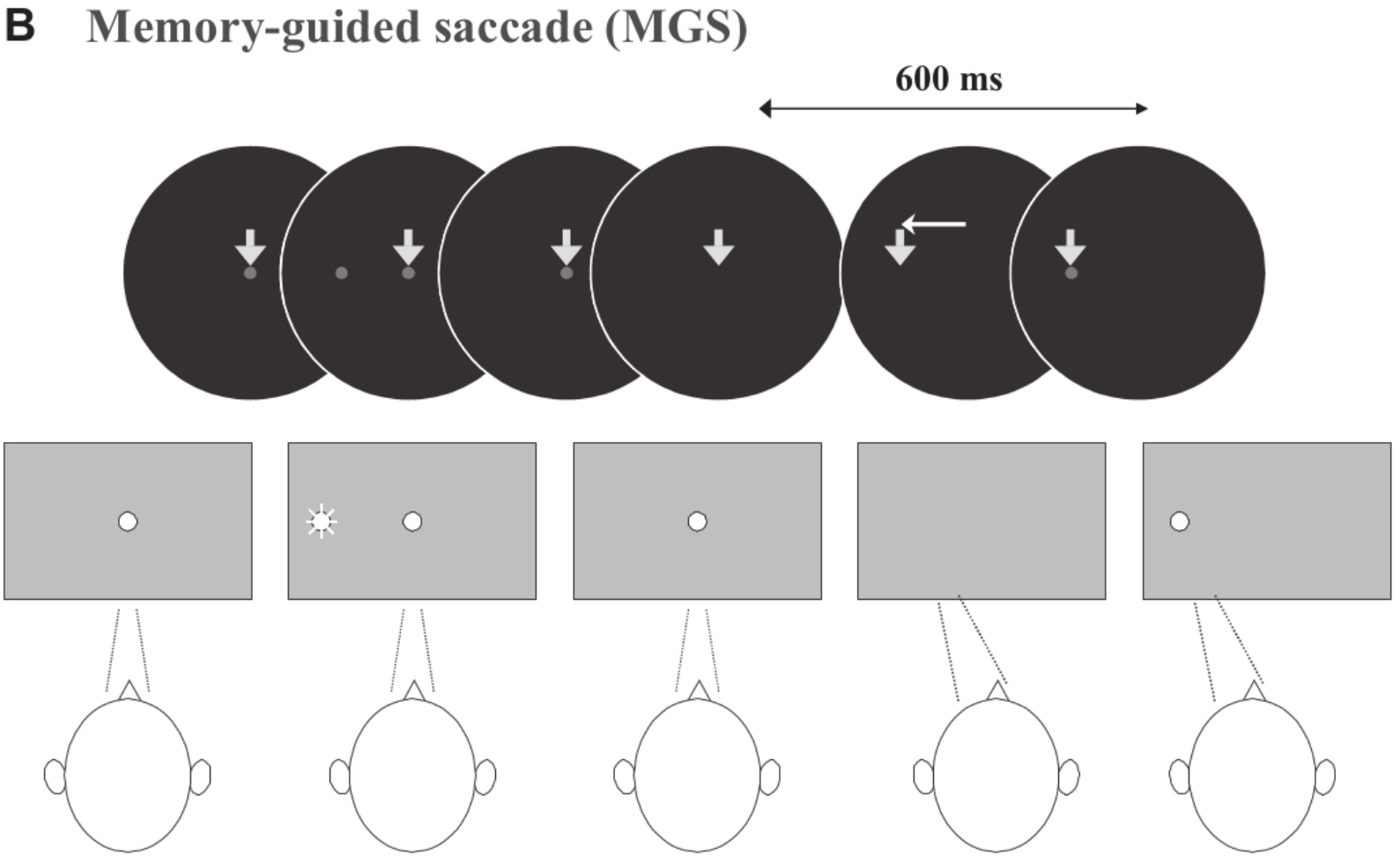
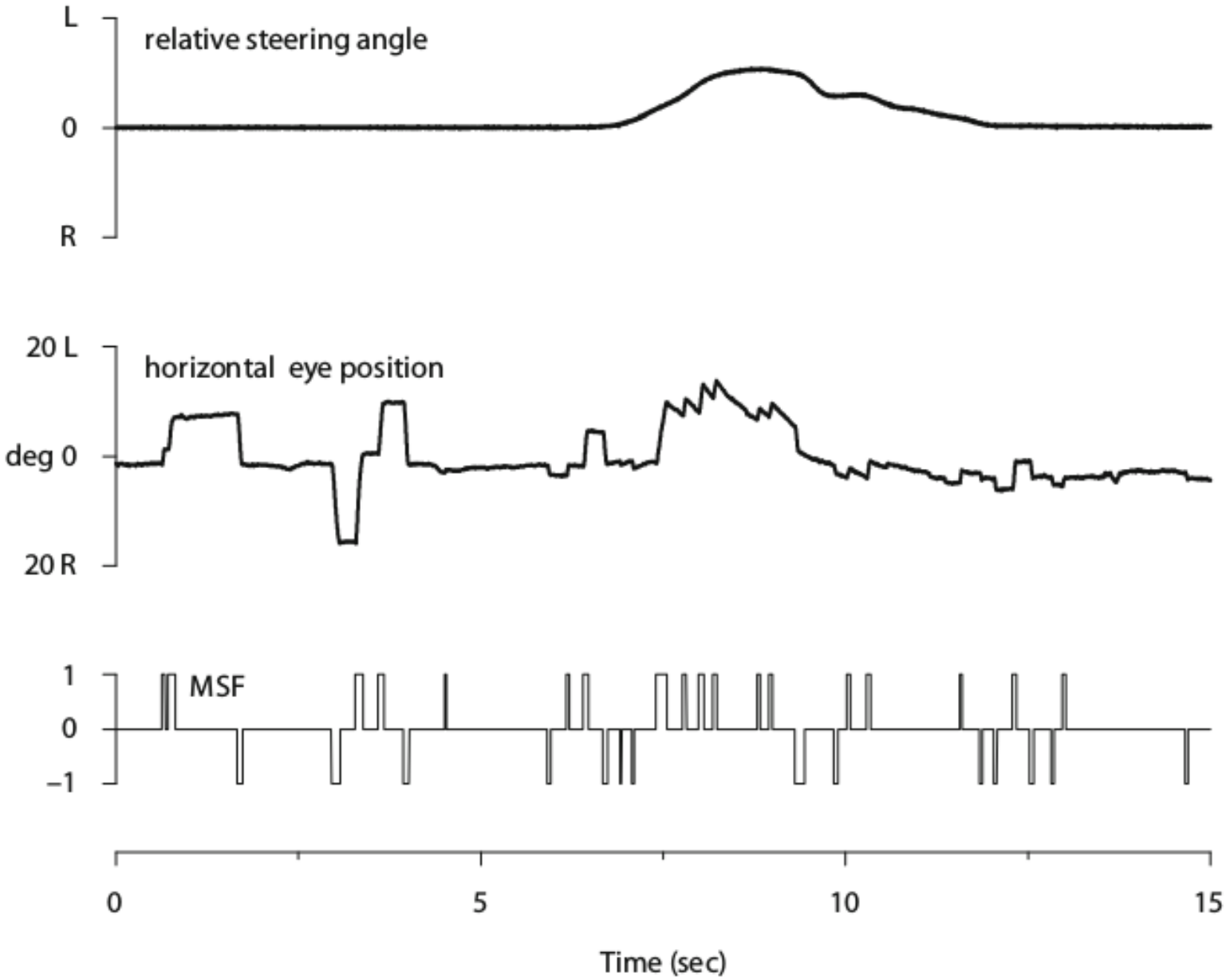
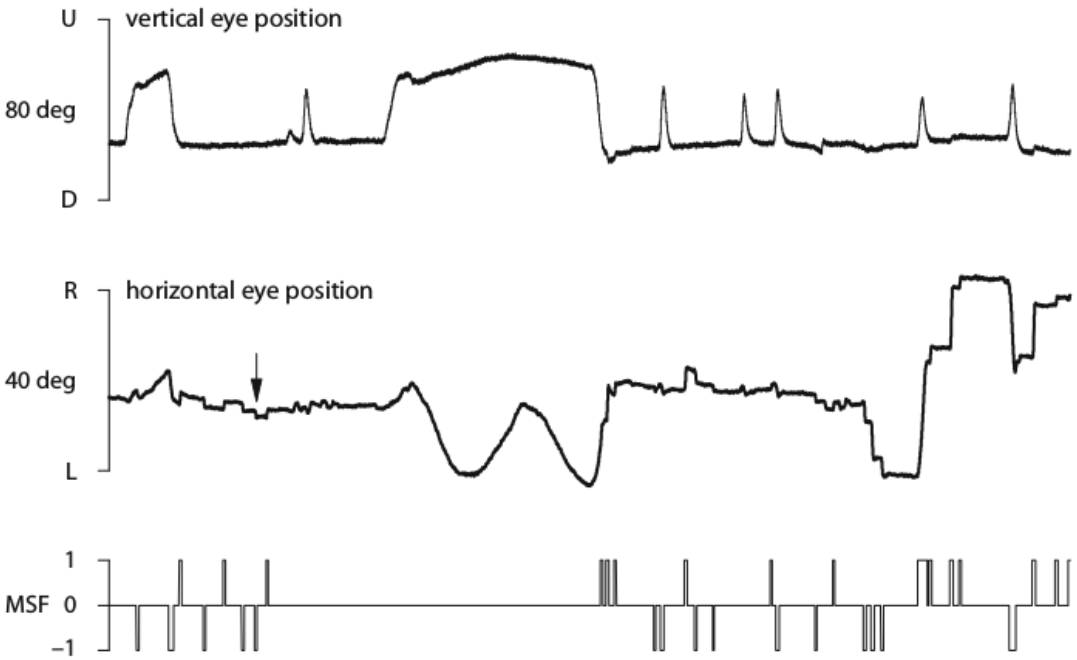


Fig. 2. Oculomotor tasks used in clinical saccade studies. **A**. Visually guided saccade (VGS), **B**. Memory-guided saccade (MGS). Reproduced with permission and modified from Terao *et al.*[96)]

Saccades: Driving a Car



Saccades: Microsleep



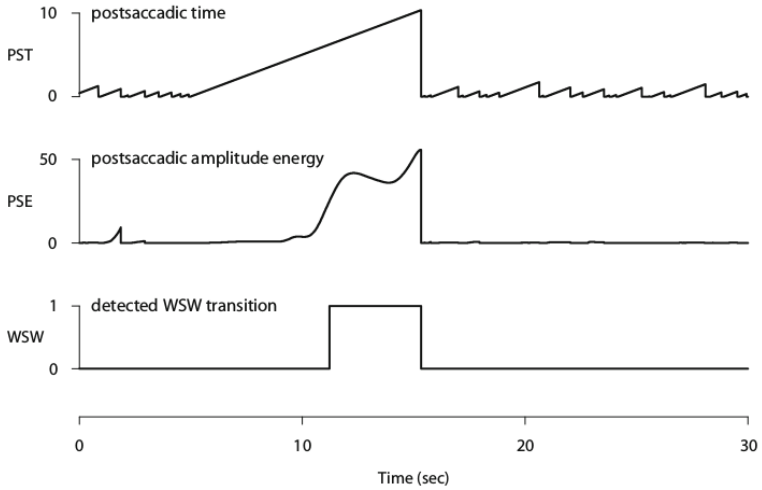


Figure 6. Demonstration of the usefulness of the algorithm by the detection of a wake–sleep–wake (WSW) transition during a microsleep episode on the basis of the analysis of horizontal eye movements (trace 2). The vertical component of eye movements is shown in trace 1. Blinks and lid closures can be discriminated, which is a benefit of the EOG method. Small saccades of approximately 1.5° (arrow in trace 2) are detected in the horizontal component. The membership function MSF is plotted in trace 3, which is the basis for the calculation of the postsaccadic time (PST; trace 4). Postsaccadic amplitude energy (PSE), the squared sum of differences of the actual amplitude and the amplitude at the end of the last saccade divided by the postsaccadic time, is shown in trace 5. The detected microsleep episode is indicated in the bottom trace.

Saccades: Parkinson's Disease

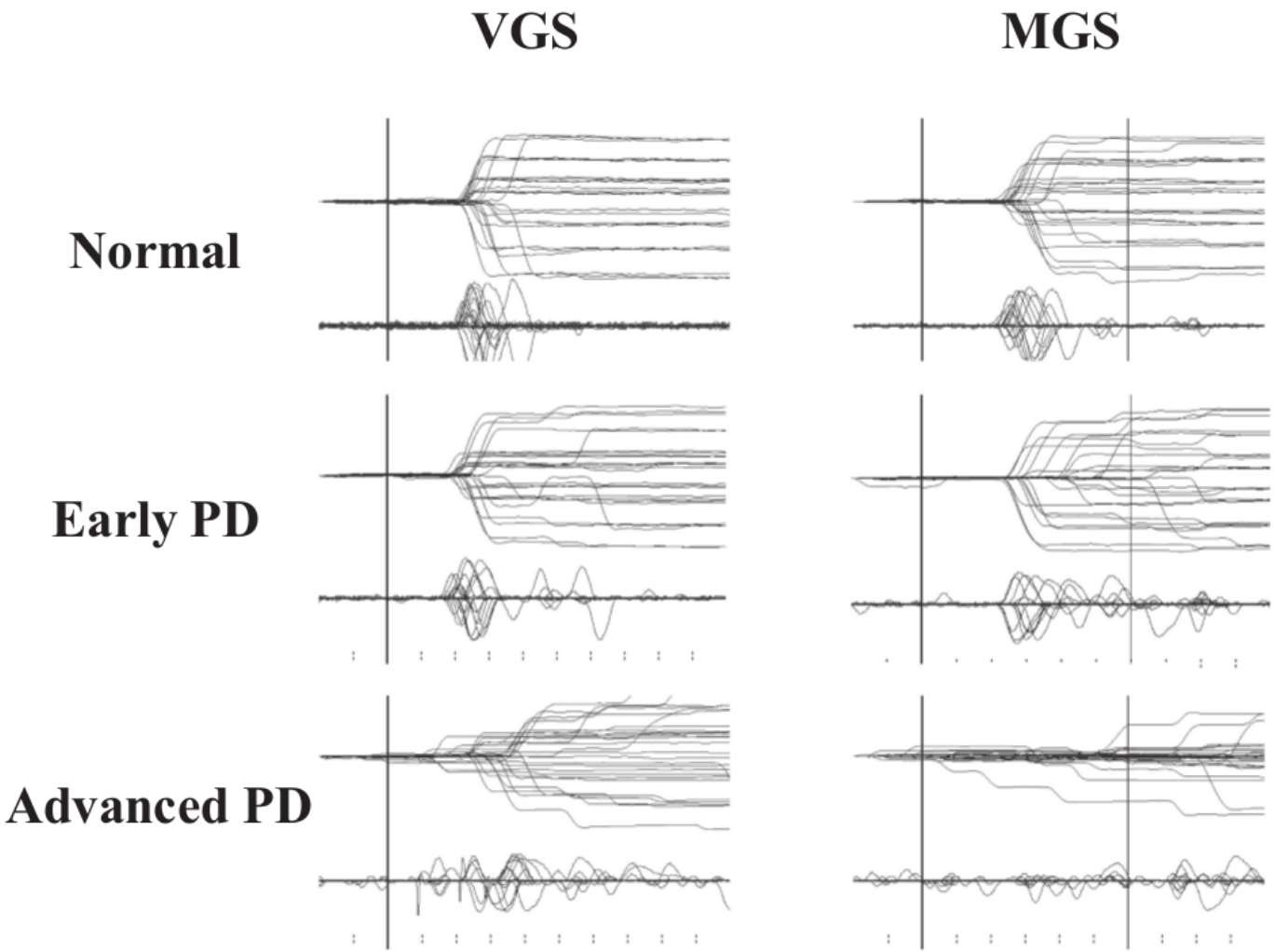


Fig. 7. Saccade records in patients with Parkinson's disease (PD). Saccade records of patients with early (middle traces) and advanced PD (bottom traces) are shown in comparison with those of a normal subject (top traces). Conventions as in Fig. 3. Superimposition of traces for 20–30 trials each. The left half shows a visually guided saccade (VGS), the right half shows memory-guided saccade (MGS). Saccades are characterized by hypometria in both tasks. In addition, MGSs are more affected than VGSs in that the latency is more variable and that failed trials also predominate at an advanced stage. Reproduced with permission from Terao *et al.*[10)]

Want a Dialog System!

- users expect Alexa-style capabilities
- but how to implement this?

- no pretrained manipulation/service-robot agents yet (GPT-3 etc. only talk, little context, no robotics)

- adapt existing dialog manager?
  - Prof. Usbeck told us he had one ready,
  - but never replied which one…

- Rasa?
  - only open-source system I found (so far)
  - mixed bag of ML tools…
  - very complex setup, very restricted dialogs

Rasa Open source conversational AI, https://rasa.com/

What are contextual assistants?

At Rasa, we use the concept of 5 Levels of Assistants to describe the capabilities of AI assistants and show how the technology has evolved over time.

Briefly, these are the definitions:

- **Level 1: Notification Assistants**
  - Capable of sending simple notifications, like a text message, push notificat WhatsApp message.
- **Level 2: FAQ Assistants**
  - Can answer simple questions, like FAQs.
  - The most common type of assistant today
  - Often constructed around a set of rules or a state machine.

Rasa: 5 Levels of Dialog Assistants (cont'd)

- **Level 3: Contextual Assistants**
  - Able to understand the context of the conversation, i.e. what the user has said previously and when/where/how they said it.
  - Capable of understanding and responding to different and unexpected inputs
  - Can learn from previous conversations and improve in accuracy over time
    - Buildable today with Rasa
- **Level 4: Personalised Assistants**
  - The next generation of AI assistants, that will get to know you better over time
  - Theoretical only
- **Level 5: Autonomous Organization of Assistants**
  - AI assistants that know every customer personally
  - Capable of running large parts of a company's operations—from lead generatio to sales, HR, or finance.
  - Long-term vision for the industry

Navel: Dialog strategies

- proxemics, facial expression, gestures
- eye contact, saccades, gaze aversion, blinking
- emotional voice
- greeting, get to know name, farewell,
- games, jokes, chit chat,
- vivid behavior, positive reinforcement, compliments,
- dialog guidance, activating questions,
- online services, weather, menu
- positive psychology questions, nodding

- demo video (’dialog strategies…’)
- targeting first-time users
- not clear, whether memory and learning is implemented

Navel robotics, keywords from promo video…

Furhat: Cardgame Example

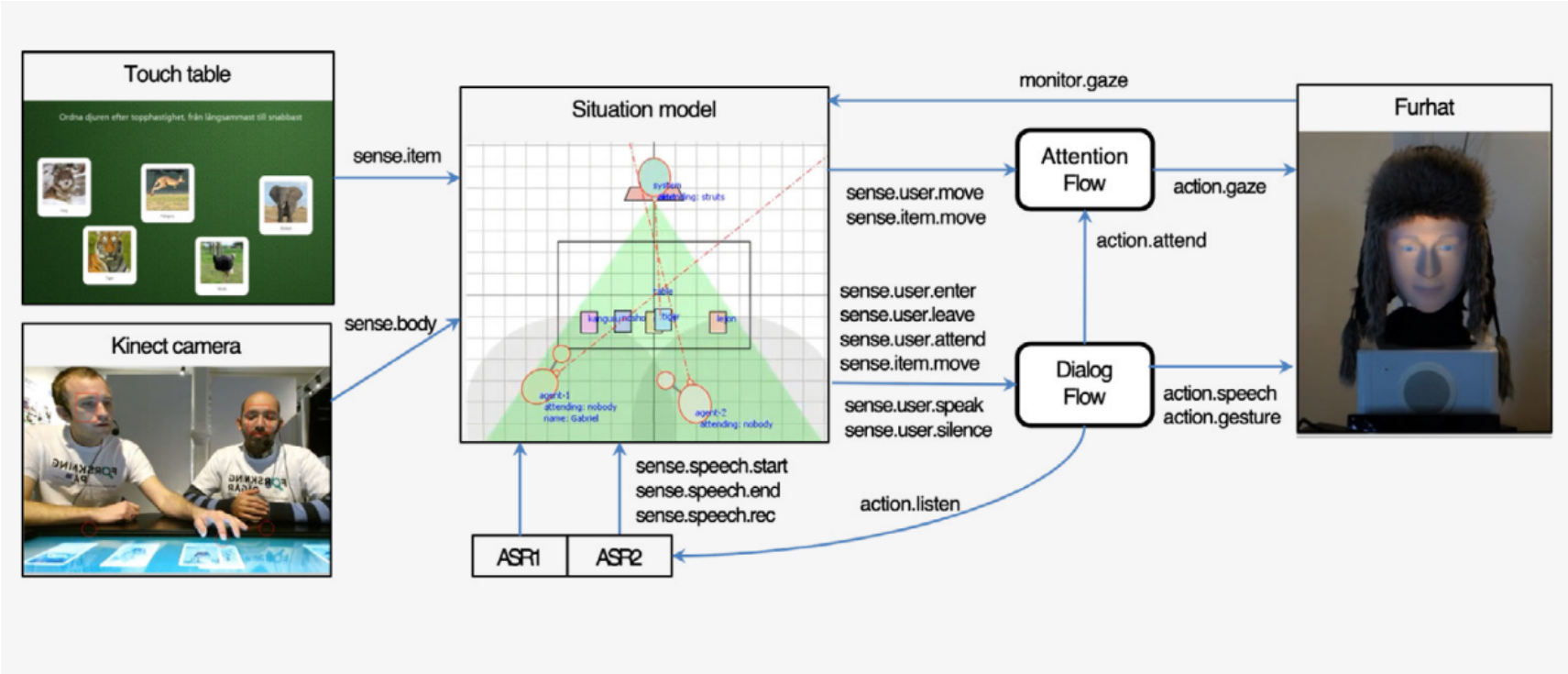


Figure 2: Overview of the different components and some of the events flowing in an example application.

furhat.com/

Take home messages

- robot head market overview
- youtube.com/watch?v=bC_DZIwevIl

- some progress on ULA head
- working on multimodal PR2 HRI demos
  - Neopixel eyes
  - Mozilla TTS + deep_speech
  - ros4hri + openvino

- beware the Uncanny valley :-)

Dog mask