VISION-BASED SPEECH IDENTIFICATION AND HUMAN FACE TRACKING ON REEM-C

04/01/2022

Pranav Barot (20616032)

Chandrasekar Elankannan (20958948)

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

Contents

Purpose

- Responding to speech is one of the most important requirements in human-robot interaction

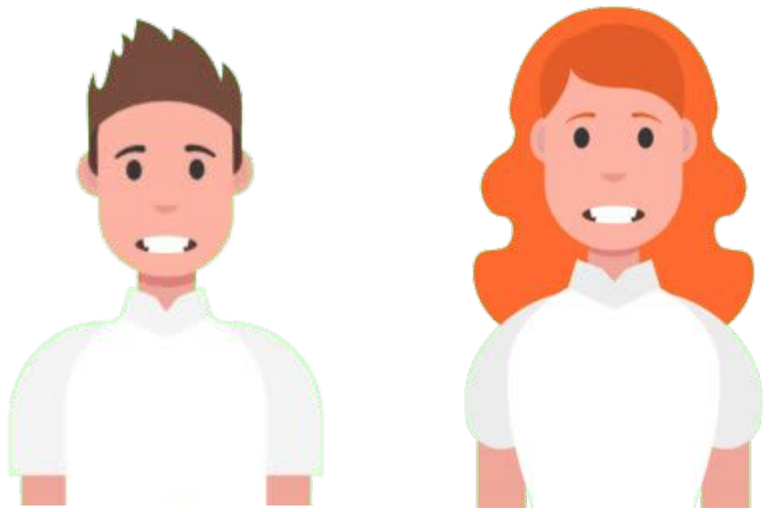
- Providing attention to human speech improves human-robot interaction

- Thus, head tracking and speech identification is one of the key methods to improve human-robot interaction

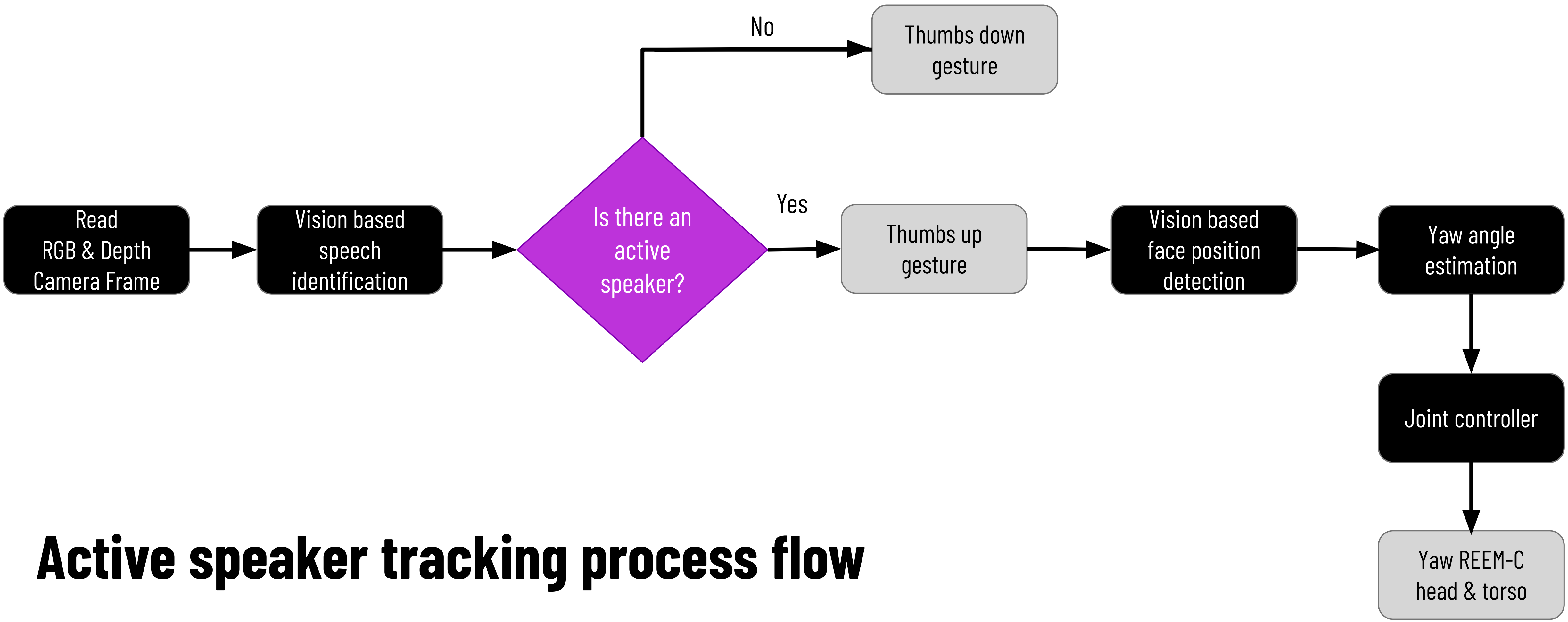
Objective

- To detect the active human speaker using computer vision
- To track the head of the active human speaker by controlling head and torso of the humanoid robot
- Signal a thumbs-up/down gesture for presence/absence of active human speaker

SCENARIO 1:ONE PERSON



SCENARIO 2:TWO PERSONS



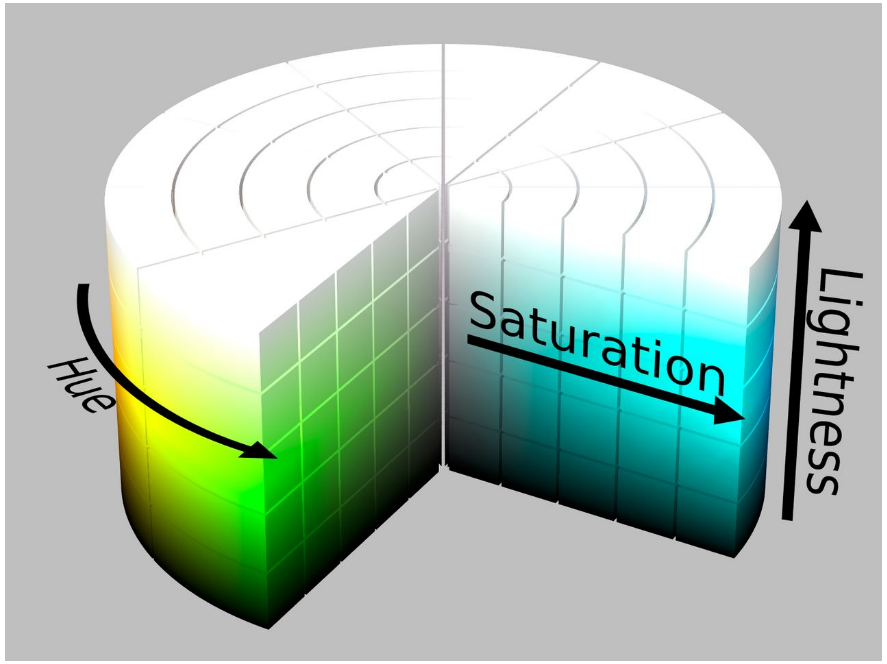
Active speaker tracking process flow

Vision based speech identification

Speech Identification → Face position estimation → Thumbs up /down gesture → Head,Torso Control

- Consider features that may indicate speech… mouth height? Color intensities?
- Refer to [1], [2] for ideas
- Iteration + hypothesizing leads to 4 unique features:

  1. Mouth height
  2. Area of mouth
  3. Lightness in HSL space
  4. Average depth of mouth region (Intel RealSense)

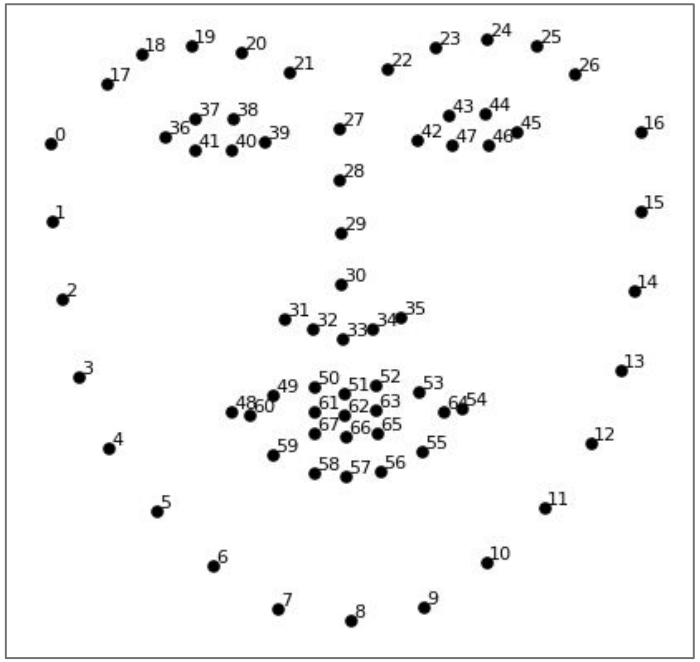


Vision based speech identification

Speech Identification → Face position estimation → Thumbs up /down gesture → Head,Torso Control

Feature crafting:

- Mouth height: given points 61,67,62,66,63,65, $\frac{\sum_1^3(mheight_i)}{3}$
  normalized by height of face (points 27, 8), **mouth_height**



- Area of mouth: Size of binary mask drawn by points 60-67 normalized by the lower face (2-14 + 64-48), **mouth_area**

- Lightness: Mean value of binary mask in lightness channel, **mean_lightness**, normalized by lower face

- Depth: Mean value of binary mask in aligned depth channel, **mean_depth**, normalized by lower face

Vision based speech identification

Speech Identification → Face position estimation → Thumbs up /down gesture → Head,Torso Control

Classification:

- Frame energy based thresholding, given 10 images per frame
- Behaviour: height, area and depth increase, lightness decreases
- Establish threshold via first 5 images for the speaker ⟶ update every 5 frames (50 images)

Speech is identified in the frame if all 4 conditions are True

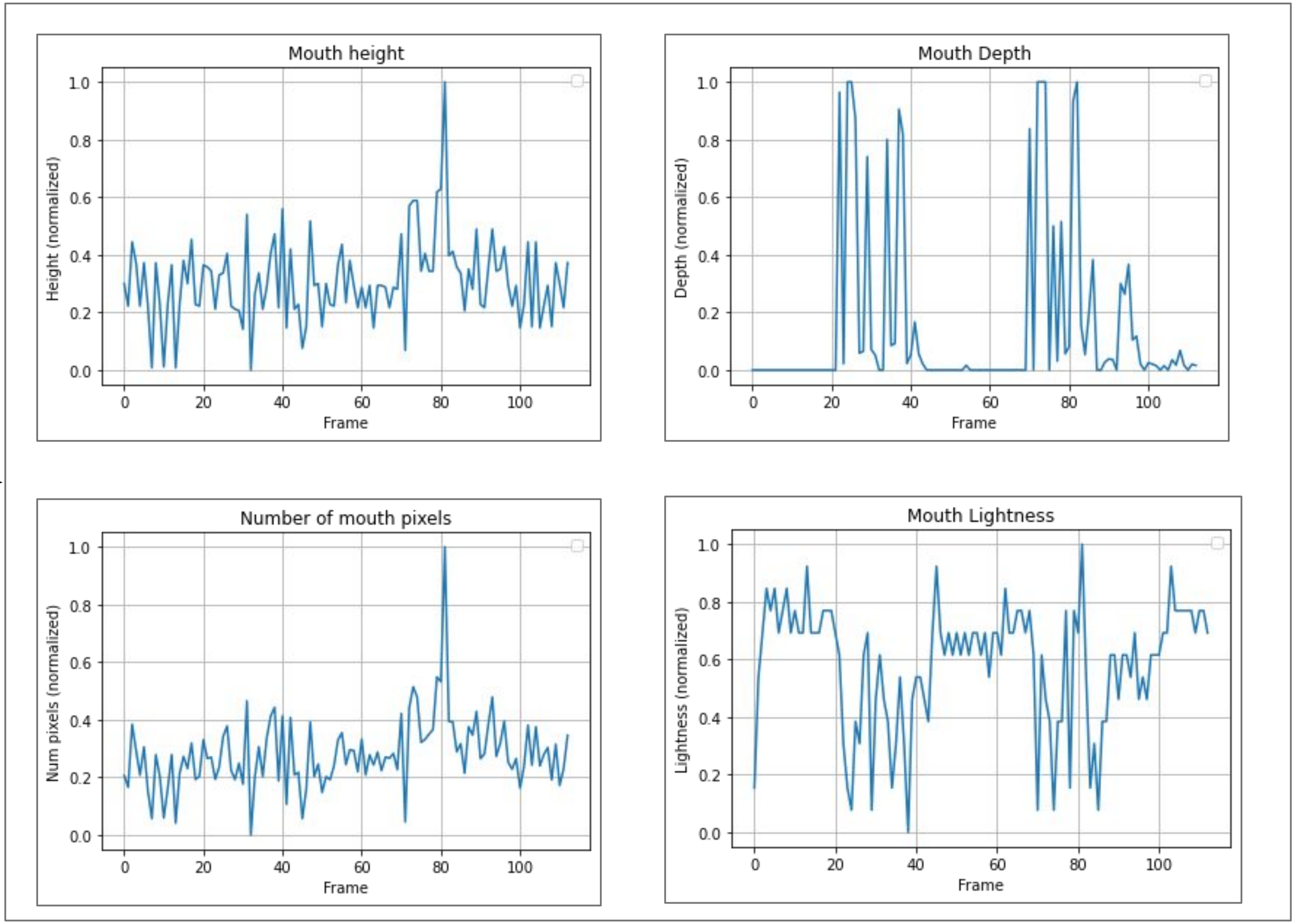
Metric	**Expected Behaviour**	**Tuned threshold (ratio to base values)**
Mouth Height	>	1.4
Area	>	1.1
Lightness	<	0.8
Depth	>	1.002

Vision based speech identification

Speech Identification → Face position estimation → Thumbs up /down gesture → Head,Torso Control

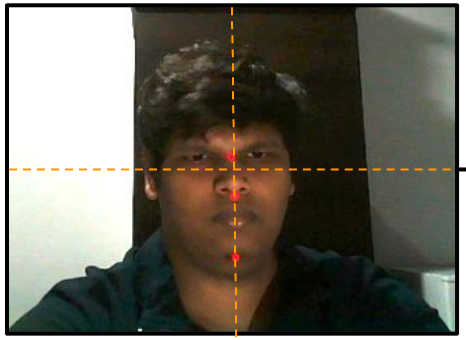
RGB frame

Binary mask (mouth)



Single Person Case-Binary masking of mouth region

Vision based face tracking



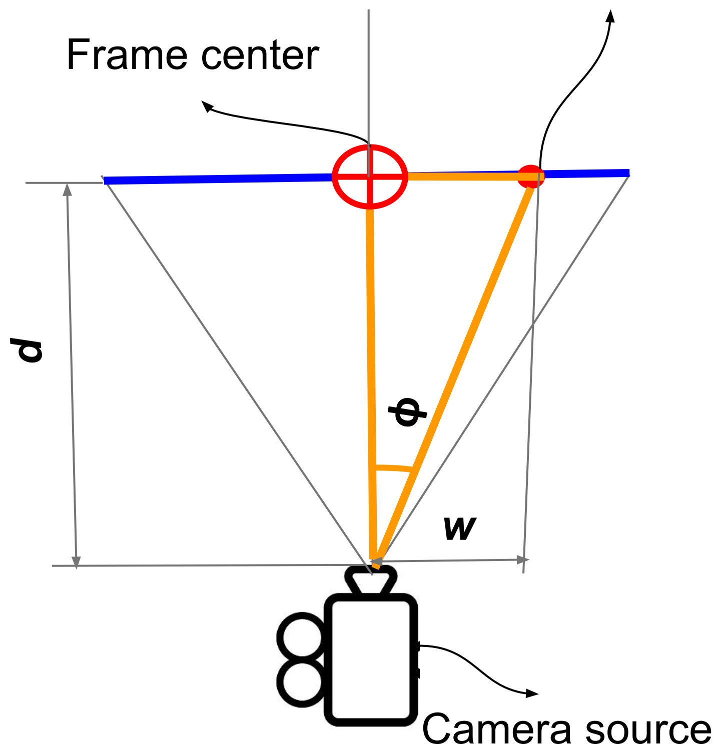
RGB Stream Frame

Depth Stream Frame

Relative position ***(w)*** of face center w.r.t frame center

Depth of face center from camera ***(d)***

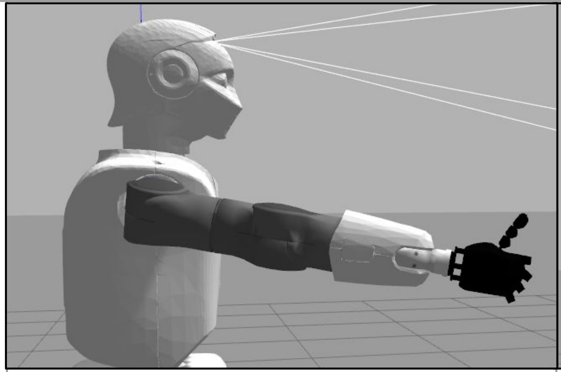
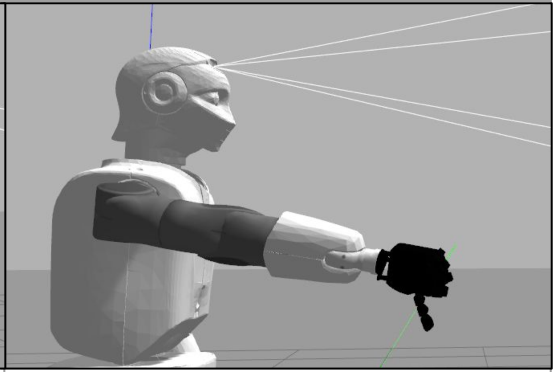
→ Yaw Angle ***(ϕ)*** Estimation

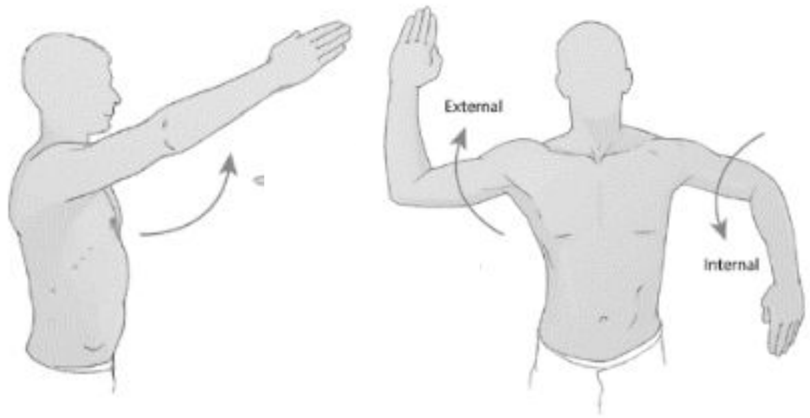


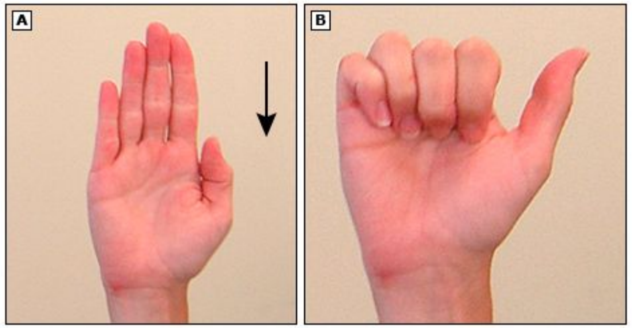
Yaw Angle: $\phi = \arctan(w/d)$

Thumbs up /down gesture

Speech Identification → Face position estimation → **Thumbs up /down gesture** → Head,Torso Control

	Thumbs Up Gesture	Thumbs Down Gesture
Gesture		
Signal	Active speaker detected	No speakers detected
Shoulder flexion	1.47 rads (85°)	1.47 rads (85°)
Shoulder rotation	1.4 rads (80°)	-1.9 rads (-110°)
Elbow rotation	-1.6 rads (-92°)	-1.2 rads (-70°)
Fingers flexion	2.5 rads (143°)	2.5 rads (143°)



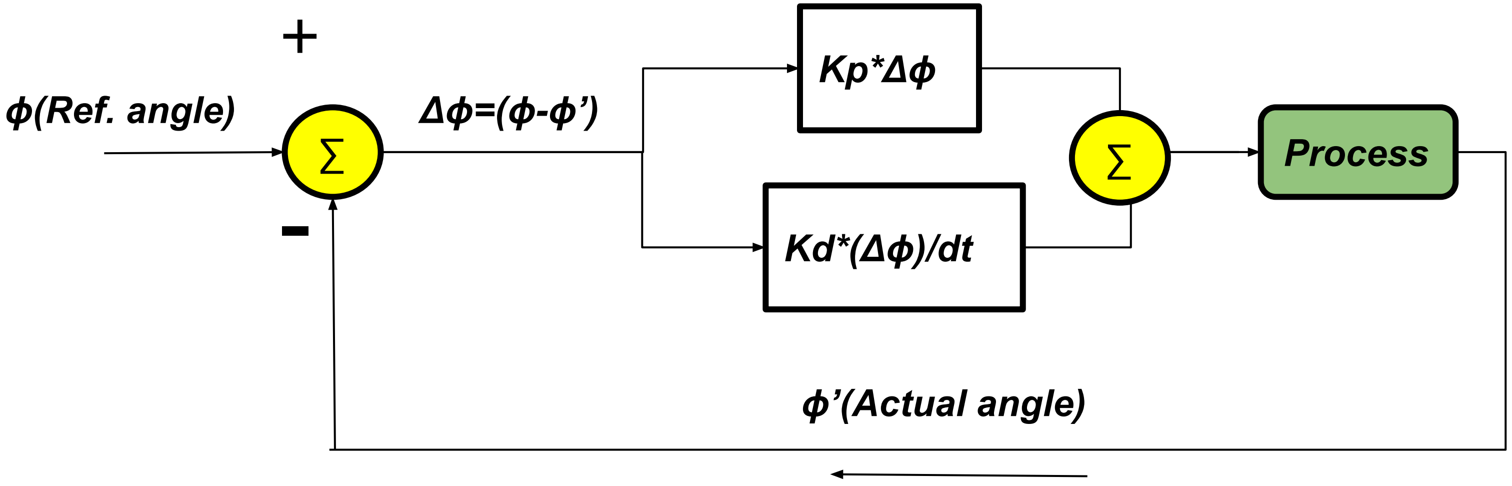


Cross-verified whether arm and hand joint angles are within human joint angle limits [3].
Elbow-shoulder angle =180 degree

Thumbs up/Down gesture demo

Head and Torso Control

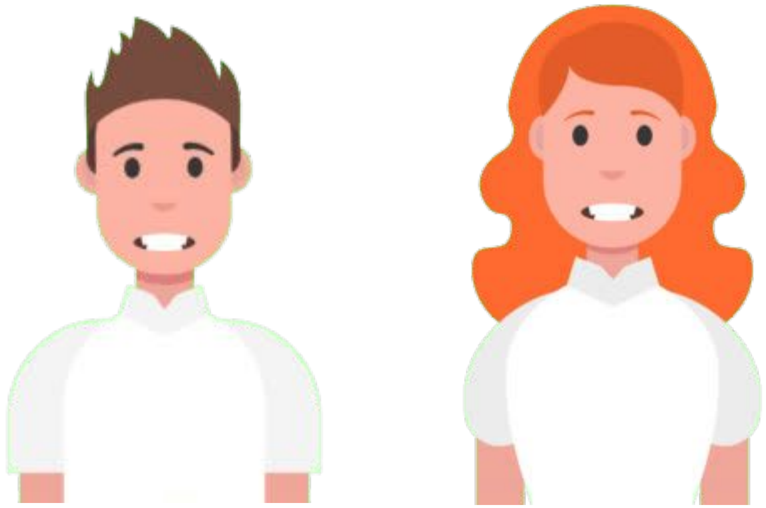
- Total Yaw angle=Head angle + Torso Angle
- Yaw angle and rotational speed of Head is more than torso [4]
- Head and torso should move synchronously to have a human-like motion [5]
- ROS controller manager (JTC) used in simulation, given PID values for head and torso
- Time required for trajectories proportional to next measured angle difference



Parameters	Head	Torso
Kp	26	26
Ki	0.0065	0.0065
Kd	1.0	1.0

Two Person Case

- Design augmentations to full OOP design
- Each detected face computes and stores its own features, classification results, and output variables
- Adaptive to unique characteristics of each face, can be easily expanded to any # of participants



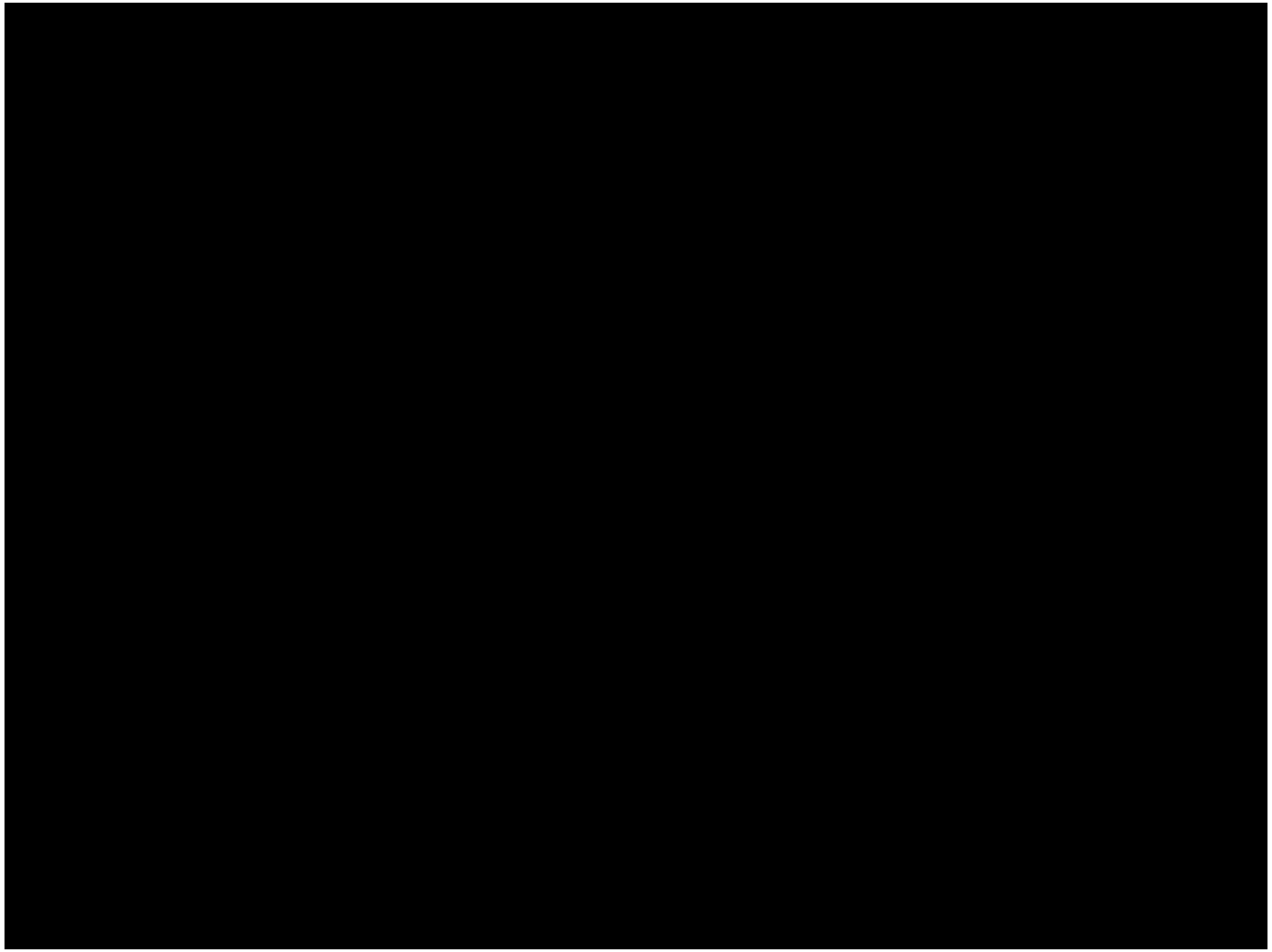
SCENARIO 2: TWO PERSONS

Two Person Case

- Actively able to switch between single and double speaker scenarios
- Tracking: Robot uses midpoint to track; attempts to keep both speakers in frame
- Speech identification: Both arms used to signal speech presence independently

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

Two Person Demo



Future work

- Tracking speakers from 0 through 360 degrees, including foot actuation along with head and torso
- Gaze/head pose information to estimate who each speaker is communicating with
- More intelligent speaker identification (better discrimination between mouth behaviours)
- Multi-modal implementation with audio (e.g, identify speakers outside frame, actuate to include them)

References

[1] Spyridon Siatras, Nikos Nikolaidis, Michail Krinidis, Ioannis Pitas. Visual Lip Activity Detection and Speaker Detection Using Mouth Region Intensities. IEEE, 2008

[2] Liu Peng, Wang Zuo-Ling, (2006). Audio-visual voice activity detection

[3] Rodríguez, C. (2019). Measuring Shoulder Abduction in a Healthy and Young Population: A Feasibility Study

[4] Horn, Marina, Manish Sreenivasa, and Katja Mombaur. "Optimization model of the predictive head orientation for humanoid robots." 2014 IEEE-RAS International Conference on Humanoid Robots. IEEE, 2014.

[5] Courtine, G., & Schieppati, M. (2003). Human walking along a curved path. I. Body trajectory, segment orientation and the effect of vision. European Journal of Neuroscience, 18(1), 177-190.

Thank you for your attention!