Estimating Intentions to Speak Using Body Poses in Social Interactions

Luning Tang
L.Tang-2@student.tudelft..nl

Supervisor: Hayley Hung
h.hung@tudelft.nl

TU Delft

1 Background Information

- Sometimes, people do not get a chance to express their thoughts in a social context.
- This intention can sometimes be shown by unintentional body movements. [1]
- Enabling social agents to use body poses to estimate the intentions to speak can increase the efficiency of conversations.

Premise:
There is an existing model trained based on **accelerometer data** to estimate the intention of speaking. [1]

Problem:
- **More social cues** to show the intention
- Accelerometer data does not capture the posture shifts accurately and can be influenced by **irrelevant vibrations.**
- Accelerometer is **not scalable**

Solution:
Look directly into the **body postures** extracted from the cameras as body behaviour

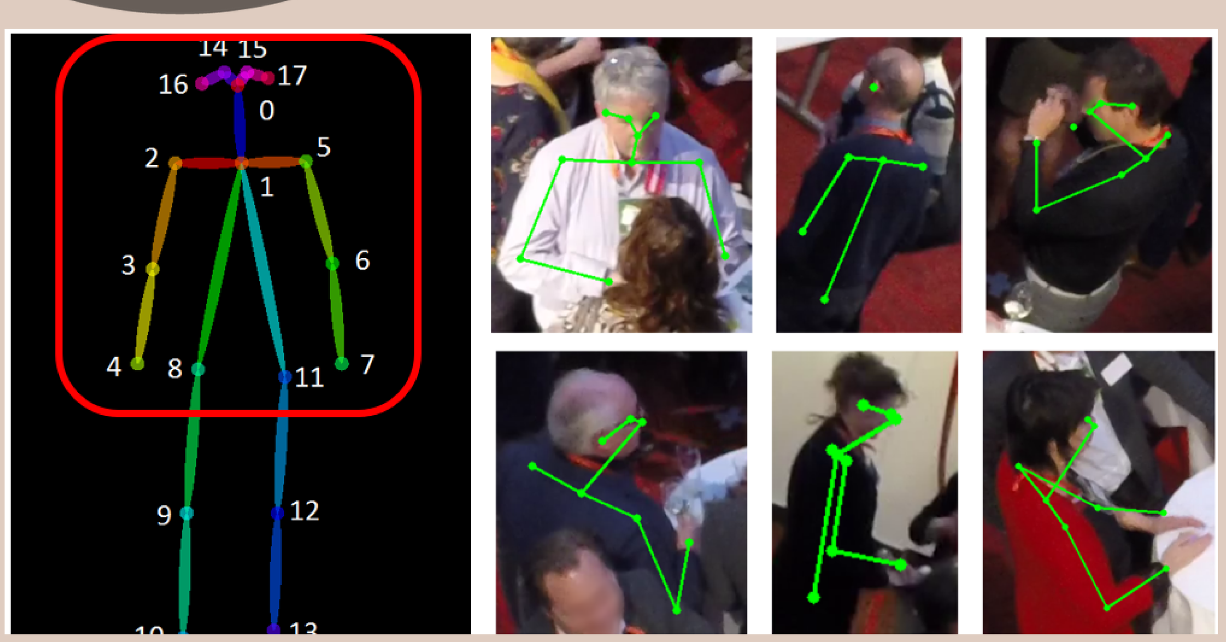


Figure 1: Selected key points (left) and detected skeletons from the Rewind dataset (right).
Left Image is adapted from OpenPose (https://github.com/ArtificialShane/OpenPose/blob/master/doc/media/keypoints_pose.png), right image is from Vargas Quiros et al. [2]

2 Research question

Can a model be trained by the body postures in-the-wild to estimate people's intention of speaking with similar or higher performance than
- **a random guessing model**
- **the existing model trained with accelerometer data?**

3 Methodology

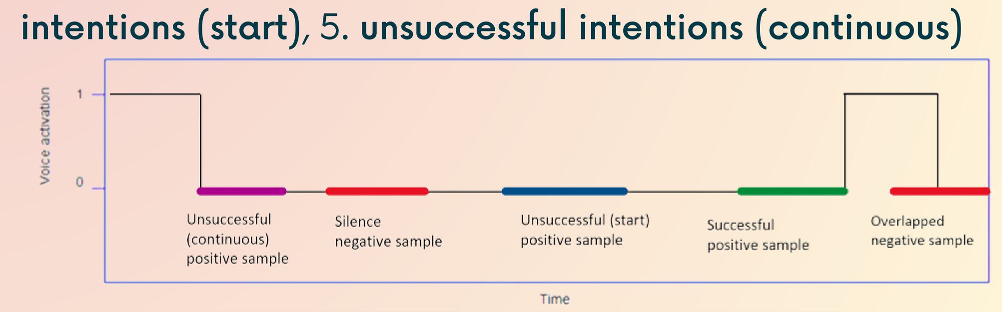
Dataset
- Rewind dataset [2]
- Collected from Dutch professional networking event
- Audios collected from a subset of people
- Videos from 4 cameras

Successful Intentions
- Successfully gets the turn after having the intention [1].
- Speaking status from Voice Activity Detection (VAD)
- 1- 4 seconds segments are generated automatically before each speaking.

Unsuccessful Intentions
- Fail to get/hold the turn after the intention [1].
- Unsuccessful intentions to start and continue
- Manual annotations from the research group within 1:00:00 - 1:10:00

Training Time Segments
- outside of 1:00:00 to 1:10:00
- Successful intentions

Testing Time Segments
- Inside of 1:00:00 to 1:10:00
- Five experiments 1. **all intentions** (successful + unsuccessful intention), 2. **Successful intentions**, 3. **unsuccessful intentions**, 4. **unsuccessful intentions (start)**, 5. **unsuccessful intentions (continuous)**



Pose tracks
- A set of tracks (skeletons across frames) from cameras 2 & 3.
- A pose skeleton: 13 key points from the upper body in each time frame (20 frames/s) [2].
- A key point: a coordinate (x, y) and a confidence score.

Training Samples
- Poses extracted from the tracks with the given training time segments.

Testing Samples
- Poses extracted from the tracks with the given testing time segments.

Model
- Adapted from existing model [1][2]
- Residual neural network (RNN)
- Three convolution layers
- Kernel sizes 3, 5, and 7

3-fold cross validation

Trained Model
- Input: the pose skeletons extracted from a person in a time window
- Output: a binary classification of whether the person has the speaking intention

Prediction & Evaluation
- Test on five experiments
- Evaluation metric: AUC
- Each experiment is conducted 100 times to obtain the mean and the standard deviation of the AUC.

4 Annotation Findings

53 annotations

All annotations:
- 77% head movements
- 57% posture shifts
- 51% arm/hand movements

Annotations (start):
- 87.5% of head movements
- 62.5% of posture shifts
- 50% arm/hand movements

5 Experiments & Results

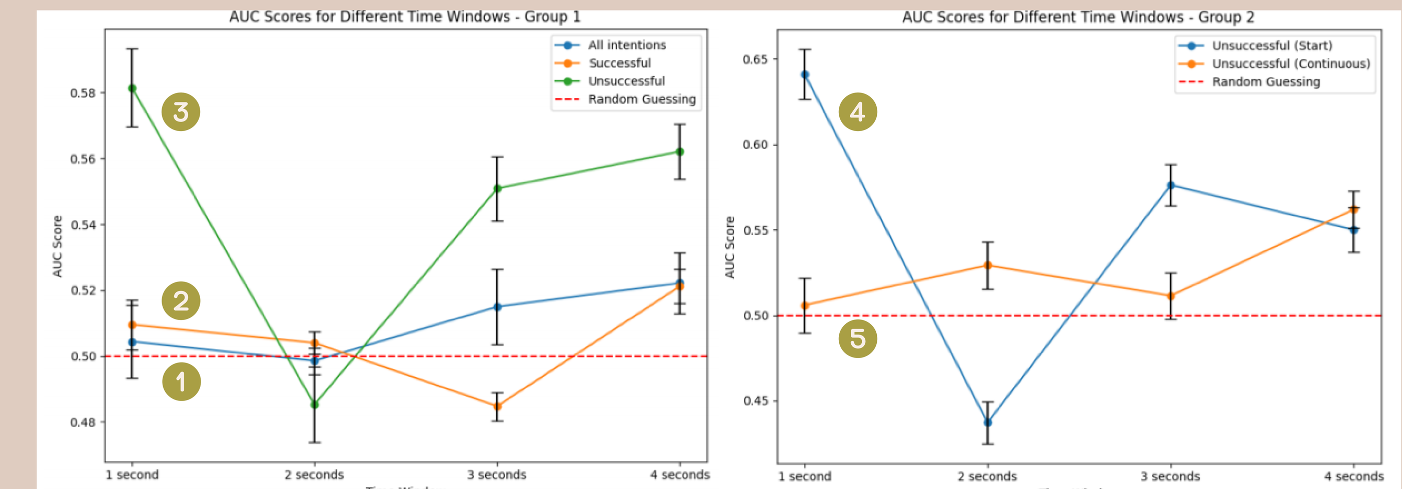


Figure 2: Means and STDs of the AUC of the model with 4 window sizes (5 experiments)

AUC scores	1 sec	2 secs	3 secs	4 secs
All intentions	0.00013	3.04e-22	1.73e-23	1.3e-42
Successful	0.99928	1.59e-20	0.0057	7.83e-64
Unsuccessful	2.42e-85	1.00000	2.67e-73	5.90e-89
Unsuccessful (Start)	1.82e-99	1.00000	2.34e-81	9.81e-61
Unsuccessful (Continuous)	0.00039	1.06e-38	3.05e-13	1.37e-77

AUC scores	1 sec	2 secs	3 secs	4 secs
All intentions	1.00000	7.94e-26	1.00000	1.00000
Successful	0.99928	1.00000	1.00000	1.00000
Unsuccessful	1.17e-06	1.00000	4.25e-76	1.02e-109
Unsuccessful (Start)	2.49e-32	1.00000	2.47e-62	1.23e-78
Unsuccessful (Continuous)	1.00000	6.19e-76	1.31e-60	4.83e-103

Figure 3: p-values of t-test with random guessing (left) and accelerometer model (right)

- Chosen variables: **training batch 131**, **pose without confidence scores** (26 features), **combined pose data** (cameras 2 and 3)
- Window 4 is the best while Window 2 is the worst (Figure 2).
- Performs better with unsuccessful intentions as annotations have less noise than the automatically generated ones.
- Overperform **random guessing** as the window size increases as there are more contexts (Figure 3 left).
- Overperform **model with accelerometer data** in unsuccessful intention prediction, as the potential interference from the speaking activity captured by the accelerometer (Figure 3 right).

6 Conclusions & Future Work

- **RNN model with body postures (13 key points) for speaking intention estimation.**
- **5 experiments + 4 window sizes**
- **Speaking intention: successful + unsuccessful (better performance)**
- **Overperform random guessing as window size increases**
- **Overperform accelerometer with unsuccessful intentions**

- **Explore better combination + longer window sizes**
- **More annotations** on both successful and unsuccessful intentions
- **Combined modalities** (accelerometer data, non-verbal vocal behaviours, lexical information)

Reference:

[1] Litian Li, Jord Molhoek, and Jing Zhou. Inferring Intentions to Speak Using Accelerometer Data In-the-Wild. Intelligent Systems Department MSc group project, page 20, 1 2023.
[2] Jose Vargas-Quiros, Stephanie Tan, Chirag Raman, Ekin Gedik, Laura Cabrera Quiros, and Hayley Hung. Rewind dataset: Speaking status detection from multimodal body movement signals in the wild. IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.