Estimating Intentions to Speak Using Body Postures in Social Interactions

Leveraging Different Machine Learning Techniques for Accurate Estimation of Intentions to Speak In-the-Wild

Luning Tang[1], **L.Tang-2@student.tudelft.nl**

Supervisor(s): Hayley Hung[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Luning Tang
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Amira Elnouty, Litian Li, Jord Molhoek, Stephanie Tan

Abstract

Everyone has the intention to speak sometimes. Allowing agents to estimate people's intention of speaking can increase conversation efficiency and engagement. The intention of speaking can be expressed by multiple modalities as social cues. In order to add values to existing accelerometer-based research [1], this research aims to build a model on body postures and explore how it performs on both successful and unsuccessful intention cases. The time segments of successful intentions are automatically generated and the segments of unsuccessful intentions are annotated in a small time period. The model uses poses extracted from the successful intention segments and evaluated on both successful and unsuccessful cases. It is shown that body posture is an effective modality to predict the intention while there are problems like visibility based on camera angles and lack of context while combining data from multiple angles. More modalities are to be added to enhance the model's generalisability and reliability.

1 Introduction

Subpar situations are not rare in social scenarios, as people have different personalities and backgrounds that may result in them not being able to express their opinions when they want to. These types of situations usually happen when the person is too shy or there are some dominant people who keep talking [1]. Petukhova and Bunt indicated that various modalities as social cues can happen when people have the intention, such as leaning forward or a half-open mouth [2]. Meanwhile, with the rapid development of human-computer interaction, AI can also be used in social science fields, such as leading a conversation. Enabling them automatically to detect participants' social cues and invite people to talk by inferring the intention of speaking can increase the efficiency of conversation, and have participants to be more engaged and active. The agent will also be considered more approachable as everyone wants to be heard.

Currently, there is only limited research on different modalities and their relations with the intention to speak. Although there are models built for predicting the next speakers [2] [3] [4], most researches on intention to speak are within the field of social science and psychology [5]. Remarkably, Li et al.'s work connects those fields with computer science by building a residual neural network (RNN) model to predict the intention to speak based using data captured by body-worn accelerometers. The experiment is done in the wild, which means they are not conducted in a controlled lab setting, but rather in a real-life context to make the model more robust and generalisable. However, most modalities like body movement expressed by accelerometer data and vocal behaviour by audio data have a scalability problem as it is hard to collect audio and accelerator data from many participants. According to Vargas Quiros et al., [6], collecting the audio of the participants can also more easily raise privacy concerns, and considerable noise can be introduced by a large number of sound sources.

Making the intention to speak estimation model support body postures as a powerful modality can mitigate the scalability and privacy problem, and enhance the existing accelerometer-based approaches to increase the robustness. Therefore, the main research question of this project is: **Can a model be trained by the body postures in-the-wild and estimate people's intention of speaking with similar or higher performance than the random guessing model and the existing models trained with the accelerometer?** The hypothesis is also made: A model trained with only body poses data can predict intentions of speaking with similar or higher performance than models trained with the accelerometer.

This project is inspired and built from the existing work to give more insights into both the field of social sciences and computer sciences. At the end of this project, it is expected that by using data from an in-the-wild dataset, a model that is trained by the body poses can estimate the intention of speaking with similar or higher performance than the model trained by the accelerometer data in Li et al.'s work [5] and it can be integrated with other modalities-based models.

The report is structured as follows: Section 2 provides the relevant literature in the field, as a foundation for this research study. Section 3 discussed the selected methods used to approach the research question. Section 4 details the annotation process undertaken before the experiments and the observations found. Section 5 outlines the experimental setup and the corresponding results obtained from these experiments. Those results are discussed in Section 7. Section 8 makes a conclusion of this research and areas that can be further investigated.

2 Background

This research is inspired and built upon some existing works, in the following three related fields: turning-taking in Section 2.1, the next speaker estimation2.2, intention to speak estimation in Section 2.3.

2.1 Turn-taking

Turn-talking happens between two conversation turns, and it is usually for the speaker to give the speaking right to another person. It is an essential aspect of any interactive conversation and involves highly complex mechanisms and phenomena, according to Petukhova et al. [2]. In order to capture the moment when people start to have the intention to speak, knowing the mechanism of how turn-taking works is necessary. Sacks et al. [7] have shown that there are three rules when turn-taking is happening:

1. The current speaker C selects the next speaker N, then C must stop and N should continue.
2. C does not select anyone, and whoever self-selects first gains the right to speak for the next conversation round.
3. If no one self-selects, C may continue.

Sometimes it is hard for people to grasp the correct timing, so there are also pauses, gaps, and overlaps during the transition [8]. These moments are usually the time when people

might show their intention to speak as they want to take the turn. Petukhova and Bunt also show there are different social cues influencing the next turn in the turn-taking, such as the body and verbal behaviour signals [2].

2.2 Next Speaker Estimation

Even though people who have the intention to speak might not become the next speaker, the field of next speaker estimation still adds great value to this research. Ishii et al. have investigated predicting the next speaker with various modalities related to body pose, such as eye gazing, head movement, Mouth-Opening Transition Patterns (MOTP), and so on [3] [4]. Those researches show that cues like month open degree and the amplitude of head rotation can be used effectively in turn-taking to predict whether the person will hold the turn, which indicates those cues can also be used to predict whether the current speaking person has the intention to continue speaking or other people have intentions to start speaking.

2.3 Intention to Speak Estimation

In Li et al.'s research of speaking intention estimation, accelerometer data extracted with a certain time window size before actual speaking are labelled as successful intention cases and are used to train a neural network model [1]. This research shows that the accelerometer modality is effective to predict the intention to speak and the selection of different window sizes can affect the results. However, there are small noise movements captured by the accelerometer which might also affect the prediction and including extracted pose as a modality can potentially mitigate that. Li et al. also categorized the speaking intention into the following two cases:

- **Successful intention**: The person successfully gets the turn after having the speaking intention.
- **Unsuccessful intention**: The person does not get the turn after a certain amount of time after having the speaking intention.
  - Unsuccessful starting intention: The person wants to start his/her conversation but fails.
  - Unsuccessful continuous intention: The person wants to continue his/her conversation but fails.

3 Methodology

To approach the research question and achieve the objective, some methods were used. This section talks about the critical decision of the project: the chosen dataset in 3.1, how the successful intentions were obtained in 3.2 (unsuccessful intention extraction is in Section 4), how body pose was extracted and represented in 3.3, the models in 3.4, and the selected evaluation method 3.5. Figure 1 shows the entire workflow of data processing for the model training and testing.

3.1 Dataset

There was two datasets, Rewind (LaRed) and MULAI taken into consideration as they both provided accelerometer data, high-quality audio and videos. Access to these two datasets had been approved.

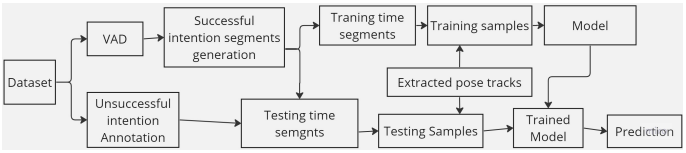


Figure 1: Pipeline of the entire workflow of data processing for training and testing the model

Rewind was collected from a professional networking event held in the Netherlands with Dutch speakers where the participants can walk around and talk with random people [6]. It contains:

- Videos: 12 overhead cameras and 4 side-elevated cameras with a 2-hour length. In this project, only the side-elevated cameras are used.
- Audios: collected from wearable microphones from a subset of participants.
- Accelerometer data: Collected from wearable accelerometer sensors from a subset of participants.

This dataset holds priority for several reasons. Firstly, it is more of an in-the-wild meeting, namely, people have enough freedom to choose who to talk to, what to talk to, and how to talk. Therefore, the model trained by this dataset is closer to real-life settings. Secondly, people are standing so there is more room and possibility for them to make different body poses, which makes the data diverse and a more adaptable and generalizable model will be trained. Thirdly, there is some existing work about speaking intention estimation with accelerometer data from Li et al. [1] done based on this dataset, and the analysis of the videos and implementation of the body pose extraction done by Vargas Quiros et al. [6]. It is efficient for this project to get inspiration and reuse some codes from it.

This project only considered time frames from 99800 to the end. Before that, the participants were either listening to music/watching a presentation, or talking with assigned partners about a specific topic, which was not considered in the wild enough.

MULAI was collected in a lab setting when two sitting people are interacting with each other [9]. It contains 357-minute recorded video, audio, physiological data streams, and hundreds of annotations about laughter and respirations. It has the advantage that the people talk in English. However, because of the three reasons mentioned above.

3.2 Successful intention extraction

The successful intention happens right before people get the turn to speak and its time segments were extracted automatically. First, it was necessary to know when people were speaking. This was done in the work from Li et al. based on diarized binary voice activation detection (VAD) on the audio that was collected by the microphone [1]. In their work, it was mentioned that there were three problems: activation disturbed by (1) noises from other people, (2) short backchannels, and (3) short pauses. This project also followed the same rules to deal with this problem. To handle the first problem,

diarized VAD minimized the effect of other people's speaking. The second and third problems could be handled by considering speaking shorter than 1.5 seconds as "not speaking" (i.e., set to 0) or pauses shorter than 1.5 as "speaking" (i.e., set to 0). Lastly, after getting the VADs, x seconds before speaking (i.e., when speaking status went from 0 to 1) were extracted as the segments of successful intention, as Figure 2 showed [1]. If a generated time segment overlapped with the previous speaking, the segment will be abandoned. In this research, the size x is chosen to be 1, 2, 3 and 4 seconds in the research to evaluate the influence on the prediction with different amounts of poses.

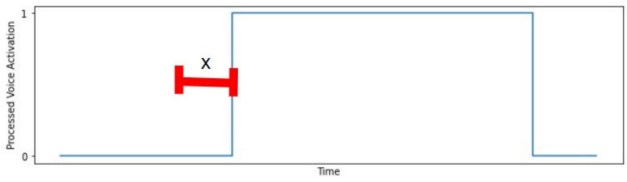


Figure 2: x seconds before the actual speaking are extracted as the successful intention speaking samples. (Figure from Li et al. [1])

3.3 Data extraction and representation

There are three types of representation for the body poses: Skeleton-based [10], Contour-based, and Volume-based [11]. Skeleton-based model is to capture the pose structure with connected key points or joints[10]. Contour-based uses the shape of contour or silhouette information as body pose. Volume-based represents the pose in three-dimensional and is useful when the captured pose needs to be personalized [11]. According to Rhodin et al.[11]'s work, skeleton-based representation is the simplest and most fundamental. The subsequent order is contour-based and then volume-based representations. Each representation builds upon the one preceding it. This research chose the **skeleton-based representation** because it is the most straightforward and there is no need for personalized poses.

OpenPose is a popular framework that can efficiently and accurately extract skeleton-based body poses in a multi-person setting [10]. Cao et al. propose an architecture that does part detection and association jointly and use a greedy algorithm to parse poses with high quality regardless of the numbers and spatial interference people. This is very useful in the Rewind setting as there are many overlapped individuals in each camera view.

In order to connect the detected poses from OpenPose across frames, Vargas Quiros et al. implement a step-wise method to obtain high-quality tracks by associating chest keypoint in each individual frame [6]. For each frame, there is a set of poses gained from OpenPose and each of them is compared to the head (the latest pose belonging to the track) of existing tracks for assignment. Tracks with heads older than R_th frames are excluded from the comparison. The assignment problem is solved by calculating the Euclidean distance between the chest key points of poses and using the Hungarian algorithm (If the distance is larger than a threshold, the pose cannot be assigned to the head). After comparing with all heads, the unassigned tracks are assigned to new tracks. Linear interpolation is used to maintain track continuity when the newly assigned poses from non-consecutive frames are not the immediately preceding frame. Figure 3 (right) shows the detected skeletons from people with different poses.

For each pose skeleton, there are 17 joints, and only the 13 key points from the upper body were used because some people have their lower bodies blocked by others, as shown in Figure 3 (left). Each key point has three feature values: x and y coordinate and a confidence score, and therefore each skeleton is represented by an array of 39 entries.

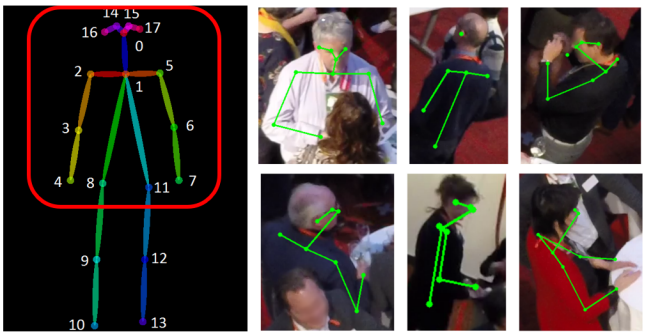


Figure 3: The left figure shows the selected key points from a pose skeleton ((Figure adapted from OpenPose [12]. The right figure group shows the detected body poses from people in the Rewind dataset with algorithms from Vargas Quiros et al.. (Figure from Vargas Quiros et al. [6])

3.4 Models

There were two models considered currently: the neural network model and catboost. The input feature of the model is a Pose skeleton extracted from a person with a time frame and the output label is a binary classification of whether the person has the intention to speak in this frame.

The neuron network model performs well when there is a large amount of data and when the data is consistent and homogeneous. It is suitable in this situation because there are a lot of successful cases automatically generated in the program, and the data is homogeneous as it only contains the coordinates of the joints. Among the different neuron network models, the convolutional neural network was used for this project because it handles well on large images. The architecture and the hyperparameters were tuned based on cross-validation.

Catboost is an algorithm for gradient boosting on the decision tree model. It performs well when the data is diverse and there is not much training data. There is also no need to tune the parameters because the default ones already perform very well. Unfortunately, this model was not tried due to time constraints and too few annotations (about 50)

3.5 Evaluation method

The evaluation criterion was the AUC (Area Under the Curve) score. This metric is commonly used to evaluate the performance of a binary classification model. It is the area under the Receiver Operator Characteristic (ROC) curve, which is

a probability curve of true positive rates (TPR) against their false positive rates (FPR). It also handles the case of an imbalanced class well, which happens when there are many more positive cases than negative cases.

AUC measures the ability of a model to distinguish classes. When AUC is 1, the model can distinguish all the cases correctly and the model predicts all cases with the opposite label when AUC is 0. When AUC is higher than 0.5, the model is better than predicting randomly and can classify negative cases from positive cases because the model gets more true positives and true negatives than false positives and false negatives. The higher AUC the model has, the better it can classify between the positive and negative cases, which means a better classifier the model is.

4 Unsuccessful Intention Annotations

To address unsuccessful speaking intentions, annotation was done by five members of the entire research group, guided by specific conventions. Unsuccessful intentions cannot automatically be generated as successful intentions in Section 3.2 because the person does not get the turn after the intention. Section 4.1 talks about what to annotate and Section 4.2 is about how to do it. Statistic results of the annotations can be found in Section 4.3 and Section 4.4 discussed some interesting observations found during the process.

4.1 Contents

While it is possible to reuse the existing annotation samples done by Li et al., the research group annotated independently. Annotating the data was not only about gaining the results but also about observing people's social cues during the annotating process (as in Section 4.4). In addition, there were more people annotating (compared to only one person who did the annotation in Li et al.'s work). The plan was to do more annotations within other time slices. However, it was not done due to issues of lack of space for the dataset and the time limit.

Though the annotation was conducted independently, the annotation decisions are made similarly to Li et al's annotation process. The group chose the 10-minute slice (1:00:00 to 1:10:00) from the Rewind dataset same as Li et al., because it was made available earlier than the entire dataset.

The participants who were visible from the cameras, with a wearable microphone and accelerometer, were annotated using the software Elan [1] [1]. Though this research does not require accelerometer data, people without accelerometers are excluded to keep consistent with research on other modalities. In the end, there were 13 participants annotated. Annotators labelled two unsuccessful intention cases: start and continue, as Li et al. categorized in Section 2.3

4.2 Rules and Procedures

The annotation has to do with human intuition, which might vary between individuals and cause inconsistency, so the group strives to minimize these variations. All five annotators (three native Dutch speakers and two that have a slight understanding of Dutch) started by annotating the same participant and comparing results to gain insights and learn from others. After a few rounds, annotations were done pairwise to speed up the procedure while keeping consistent. The group was planning to use Inter-Annotator Agreement [13] as the guideline to validate the annotations, but due to time issues Moreover, the research group discussed and made the following conventions together:

- Unsuccessful speaking only happens 4 seconds before the actual speaking (The maximum window size of successful intention is 4 seconds as mentioned in 3.2)
- A segment is considered an unsuccessful intention to speak only when there is no other intention to speak before actually speaking.
- Continuous unsuccessful speaking only happens within 4 seconds after the speaking
- Start unsuccessful speaking happens after 4 seconds of the end of the actual speaking
- The segments start at the first perceivable cue of intending to speak and end after the last perceivable cue.

4.3 Statistic Results

In total, 50 samples were annotated across the 13 participants. The total numbers, mean values and standard deviations of the annotation samples are shown in Table 1. The statistic shows there are more unsuccessful starting intentions than continuous ones, and most samples have durations ranging from 1 to 3 seconds. The person with the most annotations marked has 12 intentions and the one with the least annotations has only 1 intention marked.

	Counts	Means	STDs
All unsuccessful	50	2.398	1.022
Unsuccessful starting	30	2.496	1.141
Unsuccessful continuous	20	2.250	0.787

Table 1: Statistic data for different annotation categories

4.4 Observations

Unsuccessful intentions are usually caused by interruptions. An unsuccessful starting intention happens mostly when the person fails to interrupt someone, while an unsuccessful continuous intention happens mostly when the person is successfully interrupted by someone else.

The degree people engage in the conversation influences the extent of the social cues to convey speaking intentions. There are fewer unsuccessful intentions of an active person when he/she is talking to an inactive person. This person keeps talking and sometimes just interrupts when there is something to say. The occurrences of having intentions but have not obtained the opportunity are scarce (Only 2 intentions are found in the 10-minute slice). When active people are talking, it is more likely to predict unsuccessful intentions as they all want to get the turns.

People tend to have some unrealised movements and behaviours when they are in different speaking statuses. Given the Rewind dataset scenarios, when people are not talking, they are usually drinking or eating. People also tend to have

[1] https://archive.mpi.nl/tla/elan

more movements while talking than having the intention to speak, and the reason can be body language is used more to support the speaking content.

Multiple social cues are also marked during the annotation process. Body movements are very common to be used to indicate unsuccessful speaking intentions. Around 77% of intentions include head movements (e.g. glazing away, from looking down to looking at the other person, nodding), 57% contain posture shift (e.g. changing weights) and 51% show arm or hand movement. Meanwhile, vocal cues are also frequent with 77% filler words (e.g. 'ja', 'en', 'eh', 'ik') and 66% changing intonation. Moreover, there are 62.5% unsuccessful start annotations including posture shift and 87.5% including head movement.

5 Experimental Setup and Results

This section presents the setup of experiments and the corresponding results and observations. The Settings and workflows of how the model was built and tested are described in Section 5.1. Section 5.2 shows the evaluation results of different experiments run with the model and their significance are analysed in Section 5.3.

5.1 Model Settings and Workflow

This model is built upon the codes from Vargas Quiros [14] and Li [15] by incorporating the pose context. The refractor code of this research is shown in [16]. It is a residual neural network (RNN) which successfully performs the speaking status and speaking intention prediction with accelerometer data [6] [1]. There are three convolution layers with kernel sizes 3, 5, and 7 respectively. The pipeline of data processing for training and testing can be found in Figure 1, and the procedure shows as follows.

Firstly, the raw audios are converted into lists of zeros and ones (VAD) to classify the speaking status for successful intention extraction as discussed in Section 3.2. Secondly, positive samples and negative samples of both training and testing data and their corresponding labels are generated. For different

Only successful intention samples are used as training data because there are not enough unsuccessful intention annotations to train. Testing data are only collected in the annotation time frame 1:00:00 to 1:10:00 and are divided into five categories for different experiments. The following list illustrates the detailed information of data used for testing and training and Figure 4 shows the visualisation of how testing data is gained.

- Training samples (outside of 1:00:00 to 1:10:00)
  - Positive samples: Successful intention samples that are automatically generated
  - Negative samples: Samples that do not overlap with the positive training samples
- Testing samples (inside of 1:00:00 to 1:10:00)
  - Positive samples:
    1. All intentions: Both successful and unsuccessful intention testing samples
    2. Successful intentions: Successful intention samples that are automatically generated
    3. Unsuccessful intentions: All annotations
    4. Unsuccessful intentions (start): Annotations that are labelled as "start"
    5. Unsuccessful intentions (continuous): Annotations that are labelled as "continuous"
  - Negative samples: Samples that do not overlap with successful and unsuccessful intention test samples (positive testing samples)

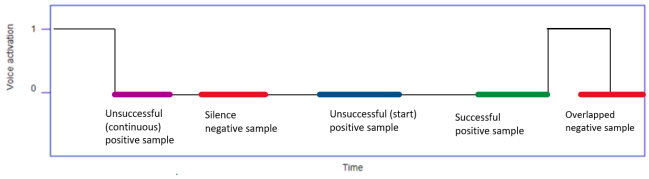


Figure 4: Visualization of both positive and negative testing samples (Figure adapted from Li et al. [1]

Thirdly, body poses are extracted from the pose tracks with corresponding time segments. The pose tracks are from Vargas Quiros's pose prepossessing[14], as discussed in Section 3.3, and is noticeable that only tracks from cameras 2 and 3 are available. Maybe it is because these two cameras were positioned to capture videos from a large common area but with different angles. Last, those body poses and their labels are put in batches to train and test the model.

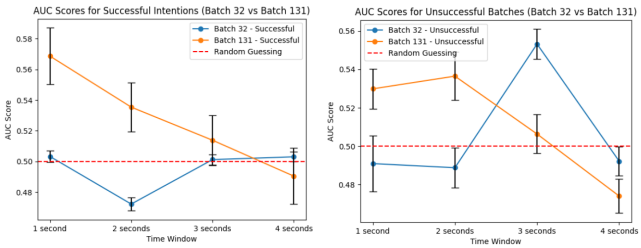
Each time there are four different models trained based on the four different window sizes (as mentioned in 3.2) and the four models will be tested based on the five testing cases. For the unsuccessful intention cases, even if the time segment overlapped with the previous speaking with a certain window size, the segment is still kept dues to the lack of samples. With the evaluation metric AUC (Section 3.5), each testing procedure was conducted 100 times to get the mean and the standard deviation of the AUC ROC scores to obtain a reliable evaluation of the variability in the results.

5.2 Evaluation Result

Various variables are explored in this research to evaluate their effect on the performance of models. For each variable, a control variable experiment is conducted.

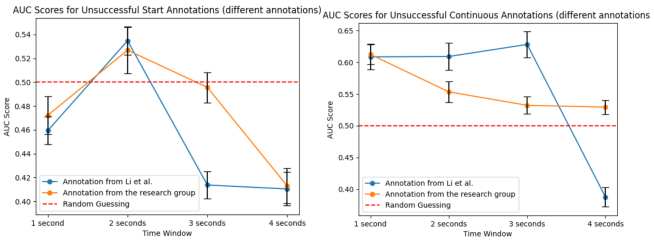
- The **batch size** of the model: 32, 131
- Different **annotations** for the model evaluation: from Li et al., from the research group
- **Pose features**: with confidence scored, without confidence scores
- **Source cameras** of the poses: camera 2, camera 3, the combination of camera 2 and camera 3

Batch size The batch size refers to the number of training samples put in a single forward pass and backpropagation and there is a trade-off. Smaller batch size helps the model generalize better while larger batch sizes can lead to more efficient training with fewer parameter updates. In this research, batch



(a) Successful Intentions (b) Unsuccessful Intentions

Figure 5: Comparison of AUC ROC between models trained with batch sizes 32 and 131



(a) Unsuccessful start (b) Unsuccessful continuous

Figure 7: AUC scores of unsuccessful intention evaluated using testing data from annotations provided by Li et al. and the research group

sizes 32 and 131 are tested and their AUC scores for the successful and unsuccessful intention cases are shown in Figure 5. The models use the research group's annotation as unsuccessful intention testing samples, and poses with confidence scored collected camera 2.

As Figure 5 indicates, the model performs well with batch 131 when the window size is 1 or 2 seconds in both successful and unsuccessful cases, then its performance decreases as the window size increases. The model with batch 32 performs better with unsuccessful test data (especially with window size 3) and there is no obvious trend with the window size. Even though generally the standard deviation of AUC is higher for the model with batch 131, it has a relatively better performance in both successful and unsuccessful, it is chosen for the fixed batch size to continue the rest of experiments.

Pose features The pose skeleton per frame is represented as an array of 39 entries as mentioned in Section 3.3, and a model is initially trained with this. Another model is trained based on pose skeletons without the confidence scores (array with 26 entries). This experiment is to check whether the confidence scores add valuable information to the intention estimation. Other variables are controlled as both models use batch size 131, annotations from the research group and pose collected from camera 2.

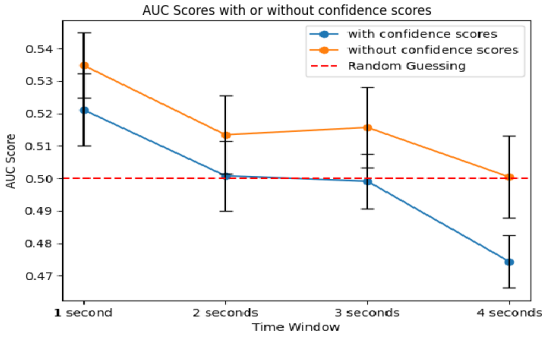


Figure 6: AUC scores of all intentions as testing data to models trained with confidence scores (39 features) and without confidence scores (26 features)

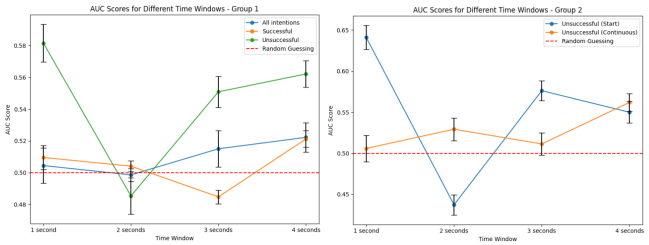
As figure 6 shows, it is interesting to notice that model trained with 26 features is better than the one with 39 with all four window sizes while using all intentions as test data. However, the variation of the AUC scores of the model trained without confidence scores seems to be slightly bigger than the one trained with confidence scores, and the AUC of the model with confidence is higher than the one without confidence in some evaluations with successful and unsuccessful testing data, which indicates that the confidence score indeed influences the model training, by sometimes providing valuable information and sometimes interfering with the prediction.

Annotation comparison A trained model is tested based on Li et al.'s annotation compared to annotations made by the five members of the research group (Section 4). This is to explore if the annotation from the research group adds value. A graph of the AUC of unsuccessful intention experiments is made for comparison. As Figure 7 shows, the evaluation with the research group's annotations has a smoother trend than the one with Li et al.'s annotation. Although the model performs quite well in the first 3 seconds with unsuccessful continuous data from Li et al.'s annotations, the overall model performance is higher on the research group's annotation.

Source cameras As mentioned in Section 5.1, there are only pose tracks in cameras 2 and 3, but is feasible to compare and combine the data due to the similar regions of the camera. Based on this, three sets of experiments were conducted: one using data solely from camera 2, another using data solely from camera 3, and a third using combined data from both camera 2 and camera 3. The three models are trained with 131 batches, pose with 26 features and evaluated on annotations from the research group.

In the experiment that uses the combined data, if both cameras have a corresponding track, the means and stds of the confidence scores in the tracks are compared. There is a composite score $0.6 * mean - 0.4 * std$ calculated the means and stds of the confidence scores in the tracks, where 0.6 is the weight for mean and 0.4 is the weight for standard deviation. The track with the higher score will be selected.

Table 2 shows the AUC results of the model from the combined cameras and the comparison with the models trained solely from cameras 2 and 3 is marked by different colours. Red indicates the result is higher than the ones in both cameras. Blue indicates the result is lower than the ones in both cameras. Black indicates the result is between the ones from the cameras. It seems that there are some performance degradation happens when the time windows are in 2 and 3 sec-



(a) Group 1 (b) Group 2

Figure 8: Visual plots of results in Table 2. Group 1 shows the AUC with all intentions, successful, and unsuccessful intention test data (First three rows in the table). Group 2 shows the results with unsuccessful start and continuous intention test data. (Last two rows in the table

onds, while all the improvements happen when the time windows are in 1 and 4 seconds. Because this combined model has the highest performance of all other models among its results and the rest results are relatively stable, it serves as the final model to perform the rest statistical analysis to answer the research question.

AUC scores	1 sec	2 secs	3 secs	4 secs
All intentions	0.5044 (0.011)	0.5095 (0.008)	0.5150 (0.011)	0.5222 (0.009)
Successful	0.4986 (0.004)	0.5040 (0.003)	0.4848 (0.004)	0.5212 (0.005)
Unsuccessful	0.5815 (0.012)	0.4373 (0.012)	0.5509 (0.010)	0.5621 (0.008)
Unsuccessful (Start)	0.6408 (0.015)	0.4853 (0.011)	0.5761 (0.012)	0.5500 (0.013)
Unsuccessful (Continuous)	0.5059 (0.016)	0.5292 (0.014)	0.5113 (0.013)	0.5619 (0.011)

Table 2: Mean and Standard deviation of AUC ROC results for all five experiments with 4 window sizes with combined pose data from cameras 2 and 3. Red: Higher than both cameras. Blue: Lower than both cameras. Black: Between the cameras' results

5.3 Statistic Analysis

In order to answer the research question, two statistic tests are performed to analyse the significant difference between results from the final model in Section 5 with the random model and the model trained with accelerometer data from Li et al.[1], and the one-tailed p-values are shown on Table 3 and 4 respectively. The green colour indicates the null hypothesis is rejected and the red colour indicates that there is not enough evidence to reject.

Comparison to random guessing model With the assumption that the model is independent of random guessing, a one-sample t-test is performed. The null hypothesis H_0 is "The model has a lower or equal performance than the random guessing" and the alternative hypothesis H_1 is "The model has a higher performance than the random guessing". The AUC results of the model are compared to the AUC scores of the mean of the random guessing model, which is 0.5 as mentioned in Section 3.5, with a conservative threshold of 0.001. Table 3 shows the one-tailed p-values calculated from the one-sample t-test.

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

Table 3: One-sided p-values of one-sample t-tests comparing AUC scores to random guessing model. Red: the model has a lower or equal performance; Green: the model has a better performance

Comparison to Model trained accelerometer data A paired t-test is conducted as accelerometer and pose data are dependent due to their reliance on body movements, leading to a dependency in the AUC scores of models trained using these two data types. The null hypothesis H_0 is "The model has a lower or equal performance than the model trained with accelerometer data" and the alternative hypothesis H_1 is "The model has a higher performance than the model trained with accelerometer data". Given the limited access to Li et al.'s full results, a normalized assumption is made. Using the provided mean and standard deviation for each experiment, a list of AUC values is generated with a normal distribution, and the AUCs from this research are compared to them in a paired t-test. The one-paired p-values are shown in Table 4.

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

Table 4: One-sided p-values of paired t-tests comparing AUC scores to model with accelerometer data. Red: the model has a lower or equal performance; Green: the model has a better performance

6 Responsible Research

It is essential for the research to be responsible and ethical. This section focuses on the ethical considerations that arise during the research process. Four main fields are discussed: the legitimacy of the dataset, the justification of the data trimming, the potential bias, and the reproducibility of the research.

Dataset: This project uses the dataset Rewind, which is approved by the ethical broad from Delft University of Technology and will be released to the general research community soon. All the participants are well informed and give their consent before their data is recorded and collected. An agreement form is signed up to protect the privacy of data. They are not stored or shared by third-party clouds (e.g. Google Drive) and no sensitive (e.g. profile picture of a participant) information is put in the report or the respiratory.

Bias: The modal trained in this research might have a bias because its training data is only collected from Dutch people. It is possible that Dutch people use different body poses during the conversation to express their intention. When people make these body poses, it is more likely for the model to predict that they have the intention to speak. In a conversation with people from different cultural backgrounds, a social agent who uses this trained model might be easier to identify the speaking intention of people who are from Dutch backgrounds compared to other backgrounds. This can be a bias against people from other cultures.

Reproducibility: The research cannot be fully reproduced as the dataset might limit the reproducibility of the search. All the experiments are done based on the information from this dataset and people can reuse the methods only if they require access to the dataset in Delft University of Technology. Other methods can be reproduced. Some of the existing public codes are reused as mentioned in Section **??**, and the repository of the original code is also put in the reference list. To fully reproduce the annotation results might be challenging as it has to do with human intuition. Still, there are conventions mentioned in Section 4 so there should be no significant differences.

7 Discussion

One notable trend is that the model performs well when the window size is 4 seconds for almost all cases. As Table 2 shows, AUC scores with 4 seconds are improved and relatively high for all five experiments after combining the pose data from cameras 2 and 3. Different from some other social cues, it is observed that people show their intention to speak through body postures not only in the last moment before they actually speak. As mentioned in 4.4, some people tend to do some other things (e.g. eating and drinking) while listening to others, and stop the current work (e.g. putting down the glass) in a few seconds before they actually want to speak. A larger window size is more likely to capture those contexts and the combination of the pose data migrates the potential noise brought by it.

However, the combination also brings performance degradation for almost all five experiments with a window size of 2 seconds, it is possible that the information that a 2-second window contained is not as pure as from a 1-second window but also not as complete as a 4-second window. Moreover, the combination only compares the mean and standard deviation of the confidence scores but a particular view might result in better/worse capturing of the important key points that matter. This might be the reason that with the experiment of all intention testing data, the combination only improves the result of 4 seconds, due to the lack of context from the other three window sizes.

Results of unsuccessful intentions show higher variability compared to successful intentions. This can be due to the potential bias and inconsistency of the annotation, as mentioned in Section 4.2. However, the annotations have a higher quality compared to the automatically generated time segments, as they are generated when there are obvious social cues to indicate the intention, which of a considerable percentage contains body movements (Section 4.4), while the successful intention samples have more noise. Another noteworthy finding is that unsuccessful start intention has the highest AUC of 0.6408 among all the results. When people have unsuccessful start intentions, it is observed that this is often because of interruption and most of them have body movements like posture shifts, and head or hand movements (Section 4.4). It is inferred that the cues are usually not as long-lasting as in other situations because an unsuccessful start intention happens before people fail to get the turn and the cure is to show the underlying unexpected.

According to the t-tests, it is shown that the model trained with body pose indeed can predict the intention of speaking better than the prediction from random guessing using a window size of 3 and 4 seconds. The possible reasons are discussed above as longer window frames provide more contexts which makes the decision-making during the combination to be more reasonable. While comparing to the model trained with accelerometer data from Li et al., it is noticeable that the model shows more advantages while predicting the unsuccessful intention situations. This might be due to the more complete annotations from the five student researchers but can also be because people's intentions are more likely to be shown based on the body postures than accelerometer data, which means people might move their body to a greater extent while they cannot get the speaking turn than just subtle movements.

There are a few limitations in the research: The number of unsuccessful intention annotations is limited. Only pose data from cameras 2 and 3 are used and people who are not in the two cameras are excluded. Only the key points from the upper body are used so intentions through poses on the lower body cannot be predicted. The approach of combining pose data from cameras 2 and 3 is also not optimal. Only the mean and standard deviation of the confidence scores are compared but there can be situations where one key point or one skeleton is more important than the other The negative training samples might overlap with the unsuccessful intention cases, as there is no annotation done for training.

8 Conclusions and Future Work

In this research, people's intentions of speaking are estimated by using body posture as the modality. The intention of speaking is spitted into successful and unsuccessful intentions, and unsuccessful intentions can be spitted into the start and continuous intentions. The successful intentions are generated automatically while the unsuccessful intentions are annotated by five student researchers in a research group. A model is built based on poses extracted from combined cam-

eras by the successful intention samples with four different window sizes, and five experiments are conducted to evaluate the model: all intentions, all successful intentions, all unsuccessful intentions, unsuccessful start and continuous intentions.

The model outperforms the random guessing model along with the increased window sizes, as larger window size provides more context and thus increases the success rate of pose combination from different cameras. While compared to the model that is trained by the accelerometer data, the model shows more advantages in estimating the speaking intentions in unsuccessful intention situations, as people might perform more body posture shifts than subtle movements.

8.1 Future work

Combined modalities: In this project, only body poses are considered, while the model trained by accelerometer overperforms in half on the experiments as Table 4 shows. This indicates that the accelerometer and pose data can sometimes complement each other. In addition, in Jose et al.'s work of predicting speaking status [6], it is shown that the model trained with raw video frames as modality has a higher performance. Moreover, in the research group, there are other student researchers who investigated other modalities like non-verbal vocal behaviour and lexical information. As observations from Section 4.4, those modalities collected from audios are also included in many speaking intention cases. Those modalities can provide additional value to the intention of speaking estimation and a multi-modalities model can be more generalised, robust and with better performance.

More annotations: Currently only around 50 annotations of unsuccessful intention cases are done across 10-minute time periods. More annotations of unsuccessful intentions should be done across the longer time period so that unsuccessful intention data sets are big enough to train a model. A model trained with both successful and unsuccessful intentions will be more generalised as both are considered intentions of speaking. It is also worth looking into annotations of successful intentions, which have higher quality than the automatically generated ones. It is expected that the model trained with that will have higher performance.

Acknowledgements

The author would like to extend sincere appreciation to Hayley Hung, Litian Li, Jord Molhoek, and Stephanie Tan for their supervision of this project, invaluable guidance and feedback. Gratitude is also to Jos´e Vargas Quiros for generously offering assistance and answering questions. Moreover, the author would like to acknowledge the fellow students who attended the midterm presentation, provided valuable feedback, and asked insightful questions. Finally, the author would like to thank Ruud de Jong for granting access to the Rewind and MULAI datasets.

References

[1] L. Li, J. Molhoek, and J. Zhou, "Inferring Intentions to Speak Using Accelerometer Data In-the-Wild," *Intelligent Systems Department MSc group project*, p. 20, 1 2023.

[2] V. Petukhova and H. Bunt, "Who's next? speaker-selection mechanisms in multiparty dialogue," in *Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers, SEMDIAL*, 2009. [Online]. Available: http://semdial.org/anthology/Z09-Petukhova_semdial_0009.pdf

[3] R. Ishii, K. Otsuka, S. Kumano, R. Higashinaka, and J. Tomita, "Prediction of who will be next speaker and when using mouth-opening pattern in multi-party conversation," *Multimodal Technologies and Interaction*, vol. 3, no. 4, 2019. [Online]. Available: https://www.mdpi.com/2414-4088/3/4/70

[4] R. Ishii, S. Kumano, and K. Otsuka, "Predicting next speaker based on head movement in multi-party meetings," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2319–2323.

[5] W. J. Levelt, *Speaking: From intention to articulation*. MIT press, 1993.

[6] J. Vargas-Quiros, S. Tan, C. Raman, E. Gedik, L. Cabrera-Quiros, and H. Hung, "Rewind dataset: Speaking status detection from multimodal body movement signals in the wild," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*.

[7] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, p. 696–735, 1974.

[8] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, p. 555–568, 2010.

[9] M.-P. Jansen, K. P. Truong, D. K. Heylen, and D. S. Nazareth, "Introducing MULAI: A multimodal database of laughter during dyadic interactions," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4333–4342. [Online]. Available: https://aclanthology.org/2020.lrec-1.534

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[11] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt, "General automatic human shape and motion capture using volumetric contour cues," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 2016, pp. 509–526.

[12] ArtificialShane, "keypoints of pose," https://github.com/ArtificialShane/OpenPose/blob/master/doc/media/keypoints_pose.png.

[13] R. Artstein, "Inter-annotator agreement," *Handbook of linguistic annotation*, pp. 297–313, 2017.

[14] josedvq, "Lared dataset," https://github.com/josedvq/lared_dataset.

[15] llt warlock, "Refactored model for inferring intentions to speak," https://github.com/llt-warlock/testProject.

[16] LibbyTang, "Refactor code of intention to speak estimation based on body pose," https://github.com/LibbyTang/researchProject.