Estimating intentions to speak in social settings

Speaker intention estimation using accelerometer data and non-verbal vocal behaviour

Waded Oudhuis

Supervisor: 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: Waded Oudhuis
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Amira Elnouty
Daily supervisors: Litian Li, Jord Molhoek, Stephanie Tan

Abstract

Computers having the ability to estimate intentions to speak can improve human-computer interaction. While plenty of research has been done on next-speaker prediction, they differ from intentions to speak since these rely only on the person themselves. Previous research was done on inferring intentions to speak using accelerometer data with some useful results. This paper expands on that research by adding non-verbal vocal behaviour as an additional modality, making the model multimodal. The model is trained on successful intentions to speak, and tested on successful and unsuccessful intentions to speak. Part of the dataset was annotated for unsuccessful intentions to speak and the signals in these annotations were analyzed. In conclusion, using non-verbal vocal behaviour is a much more reliable indicator of successful intentions to speak than accelerometer data. Using a combination of both improves the score slightly, but not significantly. Training on unsuccessful intentions to speak is likely needed to estimate these reliably. Additional modalities could be investigated to possibly improve the model further.

1 Introduction

As the usage of technology continues to grow, the significance of human-computer interaction is becoming more and more prevalent. While computers excel at performing actions based on human instructions and programming, a gap exists when the action of the human does not properly represent their intention. Artificial intelligence lacks the ability to interpret social cues and nuances of human behaviour. However, by improving AI's ability to interpret a person's intentions, human-computer interaction would benefit greatly.

Possible applications of the detection of intentions to speak can include giving notifications in virtual meetings to let other participants know someone wanted to speak but was not given the turn. Additionally, for direct human-computer interactions, it could be possible for the AI to detect when a person did not perform their intended action and prompt them to still do so.

In research done by Li et al. [1], they attempted to detect intentions to speak using a body-worn accelerometer. In their work, they are able to estimate these intentions better than random guessing.

While they have built a model using accelerometer data as a singular modality, more research can be done in using multiple modalities to estimate intentions to speak. Specifically, non-verbal vocal behaviour (also referred to as audio for simplicity) is a compelling modality to consider. Previous research shows that the combination of movement and speech is a reliable indicator of intentions to speak [2]. The combination of accelerometer data that can capture movement and audio features that can capture vocal behaviour can thus yield reliable results. Therefore, this research aims to combine these two modalities to estimate intentions to speak in social settings.

1.1 Research questions

The main question this paper answers is:

To what extent can we estimate intentions to speak by combining non-verbal vocal behaviour and accelerometer data?

Sub-questions that will also be investigated are:

- How do estimations based on non-verbal vocal behaviour alone compare to those based on combining non-verbal vocal behaviour and accelerometer data?
- How do estimations based on accelerometer data alone compare to those based on combining non-verbal vocal behaviour and accelerometer data?
- What cues indicate unrealized intentions to speak for different people?

1.2 Related work

In research done by Petukhova and Bunt [2] on speaker selection it was found that certain turn-taking behaviour signals can be reliably observed. Four people classified 2,396 segments that contained cues and observing these results, the combination of repetitive head movements and the use of filler words was seen as a good indication of an intention to take the turn. Posture shifts on their own were also seen as a strong indicator of turn-initiation. Furthermore, they found that showing more than one cue to take the turn resulted in a higher chance of getting the turn. From this, it shows that the combination of movement and vocal behaviour can be good modalities to detect intentions to speak.

Hadar et al.[3] found that head shifts have a role in regulating turns and showing the boundaries of turns. Posture shifts mainly occurred at the beginning of taking the turn, and usually began right before starting to speak and continued until after speaking had started. Ishii et al. [4] also found that head movement is a reliable indicator in predicting the next speaker.

In research done on next-speaker prediction, Ishii et al. [5] found that mouth-opening transition patterns (MOTPs) are an effective predictor for the next speaker in multi-party conversations. While the dataset used for this research does not have video footage containing all the participants' faces, it is likely audible MOTPs are included in the non-verbal vocal behaviour.

In other research by Ishii et al. [6] on respiration patterns in next speaker prediction, they found that "The next speaker takes a bigger breath toward speaking in turn-changing than listeners who will not become the next speaker". Additionally, for predicting the next speaker, the inspiration and amplitude of the inhalation were found to be an effective indicators. With the accelerometer sensor on the torso, it could pick up on respiration data [7].

Filler words such as 'ok', 'so', 'well' and 'yeah' frequently appear in conversation. They usually fulfil one of two functions: back-channelling, which shows engagement with the other speaker, or an intention to take the turn. Research done on the pitch contours associated with these cue phrases by Hockey [8] shows that different pitch contours can indicate different intentions, including turn-taking.

1.3 Background

In a social setting, the different people participating alternate turns. A turn of a person happens when they are speaking and the others are listening. There are four interactions happening in the turn exchange during a conversation: turn-taking, turn-grabbing, turn-giving and turn-keeping [9]. Turn-taking occurs when someone seizes the turn that is available. Turn-grabbing or interrupting happens when someone seizes a turn that is not available. Then there is turn-giving, where someone has the turn but signals that they are done and someone else can take the turn. They can also take it again themselves if no one else does. Turn-keeping happens when someone has the turn and gives cues that they are not done speaking and want to keep it. This research will be looking at intentions to take or grab the turn. These intentions are considered successful or realized if they get it and unsuccessful or unrealized otherwise.

In the context of this research and in accordance with the research by Li et al. a distinction is made between 'start' and 'continue' when referring to turn-taking. A turn-taking intention is considered 'start' if the person did not have the turn before, and 'continue' if they did. [1] Intentions to speak will be considered 'unrealized' if the person does not get the turn after showing cues and 'realized' if they do.

Section 2 will explain the methodology used in this research and annotations done on unrealized intentions to speak. The observations made from the annotations will be discussed in section 3. The results of the experiments are found in section 4. In section 5, the ethical aspects of the research are explained. The conclusion, limitations and future work can be found in section 6.

2 Methodology

This section outlines the various design choices made for this research. Section 2.1 describes in detail the dataset chosen. In section 2.2 the extraction of the realized intentions to speak will be explained. Section 2.3 details the procedure of the annotations of unrealized intentions to speak. How the features for the non-verbal vocal behaviour are extracted is discussed in section 2.4. Lastly, section 2.5 explains the evaluation measurements used for testing the model.

2.1 REWIND dataset

This research uses the REWIND dataset [10], a Dutch dataset that contains video, audio and accelerometer data from a social networking event. During this event, the people walk around freely and have conversations with each other. The video data consists of four overhead cameras in different corners of the room. Most participants are in view of at least one camera most of the time. The audio is taken from a select group of participants wearing a microphone and a group of participants wearing a body-worn accelerometer. Some participants have all of these three modalities, which is the group this research will look at, and consists of 13 participants. This will result in an easier comparison of the estimations for research done on different modalities. Using only the footage of in-the-wild social networking, around 1 hour and 50 minutes of data is left. Since this research aims to compare to previous research [1], using the same dataset leads to the most meaningful comparison.

2.2 Realized intentions to speak

To extract realized intentions to speak, first, Voice Activity Detection (VAD) and speaker diarization were used to find when a participant was speaking as done by Vargas-Quiros [10]. VAD creates an array of equal length to the audio segment, with a 1 if there was speech detected and a 0 if not. Speaker diarization distinguishes different speakers from the same audio segment, making it possible to see when the participant is speaking and when the microphone is picking up other people's voices. If the segment of speaking is longer than 1.5 seconds, it is considered a 'turn'. This is to remove backchannels from the VAD, as no turn-taking or turn-grabbing is happening then. As shown in figure 1, moments in between speaking that were less than 1.5 seconds were smoothed out, to prevent moments between sentences or brief pauses in the turn to be detected as intentions to speak. Then, as visualized in figure 1, the segment length x before the start of the turn was taken and labelled as a realized intention to speak. [1] The value of x depends on the window size, which, for this research was either 1, 2, 3 or 4 seconds. The model will be trained and tested on all four window sizes to explore the differences in performance among them.

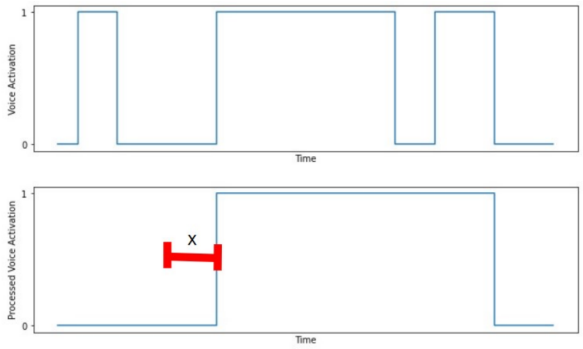

Figure 1: Extracting realized intentions to speak using VAD (figure taken from Li et al. [1])

2.3 Unrealized intentions to speak

For unrealized intentions to speak, footage of the dataset was annotated when there was an intention to speak that was not realized. Naturally, this was only done on perceivable intentions to speak. For these annotations, a group of five people, of which three were native Dutch speakers and two with a slight understanding of Dutch, annotated a 10-minute clip from the REWIND dataset. The participants annotated were the 13 participants that were on video and had a microphone and accelerometer. For the first few annotations, all group members annotated the same participant and compared their segments to agree on what would be categorized as an unrealized intention to speak. From this came the following rules:

- An intention to speak will be considered unrealized if the participant does not get the next turn after the intention,

or if it is followed by another intention.

- An intention will be categorized as 'continue' if the participant had the turn before having the intention and 'start' if they did not.
- If the participant gets the turn briefly after the intention, but is quickly interrupted, this will not be considered an unrealized intention to speak.
- An unrealized intention will last from the first perceivable cue until the last perceivable cue.

The cues chosen to pay attention to when annotating were decided on by observing the dataset and previous literature. The size of the annotation was not fixed, since the size of the segment is dependent on the window size taken from x seconds before the final cue until the final cue. The observations from the annotations will be discussed in section 3.

2.4 Feature extraction

After segmenting the realized intentions to speak, feature extraction for the audio was done using openSMILE (open-source Speech and Music Interpretation by Large-space Extraction) [11]. The features extracted are those in the Geneva Minimalistic Acoustic Parameter set (GeMAPS) [12], of which the latest version is included in openSMILE. GeMAPS is a minimalistic voice parameter set including various paralinguistic features purposed for automatic voice analysis. The feature set aims to be a common baseline of a standard acoustic parameter set. The 25 features extracted from the audio files were reduced to 10 using Principle Component Analysis (PCA) [13] as a dimensionality reduction method. PCA transforms the original features by computing the covariance matrix. The 10 eigenvectors of the covariance matrix (principal components) with the largest eigenvalues are selected for the reduced feature vector. This way, the components with the largest variance will be included. This vector is what was used for the model. From calculations done on the covariance matrix, it was found that using 10 features would retain 95% of the information of the data, which is why a reduction to 10 features was chosen (figure 2).

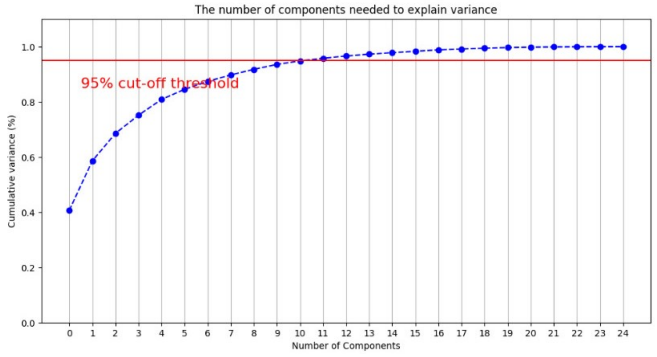

Figure 2: Cumulative variance of number of features for PCA visualized by J. van Marken using format of Mikulski [14]

For posture shifts and head movements, accelerometer data was used to compute the movement of a participant. The data is taken from a body-worn accelerometer that measures the rate of change velocity of the wearer on the x, y and z-axis. This was chosen over the body pose extraction from the video for simplicity and easier comparison with previous research. Additionally, using an accelerometer also allows for possibly picking up on respiration patterns, which is not possible from body pose [7].

2.5 Evaluation

The model used for this research is adapted from the models created by Li et al. [1] for detecting intentions to speak using accelerometer data and Vargas-Quiros [15] for speaker status detection using video, pose and accelerometer data. The model used is a residual neural network, aiming to classify a window of 1000-4000ms as either an intention to speak or not.

The AUC-ROC curve (or AUROC) is the performance measurement that was used to evaluate the model. The ROC is a probability curve, plotting the true positive rate against the false positive rate. The AUC is the area underneath the ROC curve. The worst performance for AUC is 0.5, which means there is no distinction between the classifications, a random classifier. The best AUC score is 1 when the model perfectly classifies 0 to 0 and 1 to 1.

3 Annotations unrealized intentions to speak

Using the procedure described in section 2.3, the group had a total of 53 annotations. All these annotations were compared and agreed upon by at least two group members. Of these annotations, 32 were labelled 'start' and 21 were labelled 'continue'.

3.1 Observations

From the annotations, multiple turn-taking cues were annotated. These included: posture shift, head movement, arm/hand movement, use of filler words, intonation, lip smacks, throat clearing and audible breathing. It was observed that 77% of the intentions included head movement, 56% showed posture shifts and 50% hand/arm movement. Also notably, 77% of the intentions included a filler word, and of these 41 intentions, 35 also had intonation that could indicate the intention. Conversely to the observations paper of Li et al. [1] only 22% of the annotations contained lip smacks or audible breathing. It must be stated, however, that especially intentions to speak that contain filler words are relatively easy to spot, compared to more subtle intentions. This could skew the results to show a higher percentage for the use of filler words than that reflects the real world. Nevertheless, these findings further support the research question focusing on movement and audio cues. The accelerometer sensor should also be able to measure the head movement in the total body acceleration.

3.2 Comparison

Comparing our annotations to those of Li et al. [1] there are a lot of differences. While 22 of the annotations are in total agreement (time stamp and label), for 43 annotations there is disagreement as seen in table 1. These disagreements are

counted from annotations from both Li et al. and ours and are the annotations that did not have an agreement pair. For 3 segments the same time stamps were annotated, but they did not have the same 'start' or 'continue' label. From investigating these disagreements, one of the things that stood out is Li et al. classified someone starting to speak and getting interrupted after a few words of their turn as an unsuccessful intention to speak. As stated in 2.3, we counted this as being interrupted and thus not an unsuccessful intention to speak, as the participant did briefly have the turn. Furthermore, while Li et al. had one person doing the annotations, this group always had at least two people looking at the same annotation segment. This could also explain the differences.

	Start	Continue
Total agreement	15	7
Disagreement	21	22
Mislabeled	3	3

Table 1: Comparison annotations.

The observations concerning posture shifts and head movement being correlated with intentions to speak are in line with the findings of Petukhova and Bunt[2]. Although the data in that research is English, their observations concerning 'filler words' indicating intentions to speak, while English, are also in accordance with our annotations. Since their research has a video containing all the participants' faces, their results about lip-opening patterns yield better results than the annotations from REWIND, which are solely based on audible mouth-opening patterns.

4 Experimental Setup and Results

The experimental setup used for this research will be described in section 4.1. Section 4.2 will explain the results of the experiments in detail.

4.1 Model

To validate the model 3-fold cross-validation is used with batch size 32. The test set of the model is the 10-minute segment that was annotated (01:00:00 - 01:10:00), to allow for testing on unsuccessful intentions to speak. The model is trained on all other data. The model was trained and tested with different window sizes: 1, 2, 3 and 4 seconds. As explained in 2.2, positive samples are automatically extracted with the same length as the window size. The negative samples were taken from the rest of the data, but excluded data labelled as an intention to speak, both extracted and annotated. The ratio of positive to negative samples is 1:20.

The multiple modalities are combined using late fusion, where the mean of the output masks was taken. The mask length is 20 per second, meaning it will be 20 for a 1s window, 40 for a 2s window etc. This is the same length as the model used for the accelerometer only by Li et al. [1].

There are five different test sets used for experiments. Experiment 1 contains all cases, including all the annotated unrealized intentions to speak and the extracted successful intentions to speak. Experiment 2 contains only the extracted successful intentions. In experiment 3 there are only the annotated unrealized intentions. Experiment 4 contains only the unsuccessful cases labelled 'start' and experiment 5 only those labelled 'continue'. The tests, with the exception of the annotation comparison, were all done with the annotations of Li et al. to have a more meaningful comparison. Separate experiments were run for audio alone, accelerometer alone and the combination of accelerometer and audio for easy comparison between the modalities. For each test set and window size combination, the model was tested 100 times to account for the randomness in choosing the test samples. From these results, the mean AUC score was computed as well as the standard deviation.

4.2 Results

Estimating intentions to speak using non-verbal vocal behaviour only.

Figure 3 shows the AUC values for the different test sets for all window sizes. From this, notably, experiment 2 using only successful intentions to speak in the test set gives the best result. The model gives the best results for a window size of 2s testing only on successful intentions to speak with an AUC score of 0.7147. Window sizes 3 and 4 also give similar, although slightly lower, AUC scores with 0.6993 and 0.7077 respectively. The standard deviation of experiment 2 is also very low, showing a lot of consistency over the 100 experiments. The lowest is 0.0019 for window size 1 and the highest is 0.0050 for 4s. Unsurprisingly, the model performs best on successful intentions since it was trained on this. Testing on the unsuccessful intentions, the model performs worse. For the 1s window, the AUC scores are very close together, but they spread out more the bigger the window size gets. Of the unrealized intentions to speak, the highest AUC scores are in the 'start' category, peaking at 3s with an AUC of 0.5852. The intentions labelled 'continue' performed especially poorly, 0.2697 for 4s. Since the 'continue' intentions happen after the participant had the turn before, part of this turn could be included in the segments, especially for the larger window sizes. This could explain the low performance compared to the 'start' intentions.

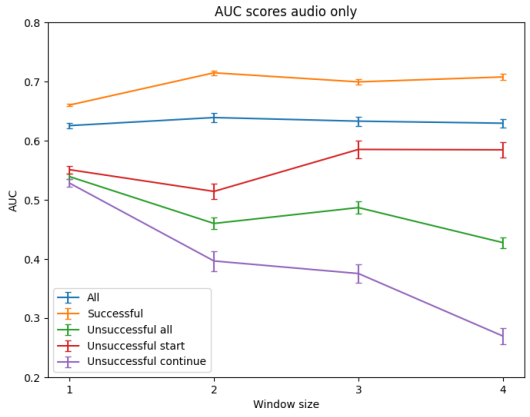

Figure 3: AUC score audio

Estimating intentions using accelerometer only.
Figure 4 shows the AUC scores of the different window sizes and test sets for accelerometer data only, abbreviated to accel. This research aimed to replicate the model from Li et al. [16] as well as possible, and the successful intentions to speak are very similar to the revised scores from their paper. The performance for unsuccessful intentions to speak is slightly different but still follows around the same patterns. This difference can be explained by randomness in sampling.

Overall, accelerometer data performs worse than non-verbal vocal behaviour for successful intentions to speak, peaking at 0.5922 for the 3s window. Remarkably, accelerometer data alone scores worst at the 2s window with a score of 0.5124 while this window size was best for audio.

The scores for testing on unsuccessful intentions to speak are closer to those of the successful ones compared to audio. Also, the drop off in performance seen in figure 3 for the 4s window is a lot smaller for accelerometer data. This can be explained by the accelerometer data being less affected by overlapping with a previous turn than non-verbal vocal behaviour.

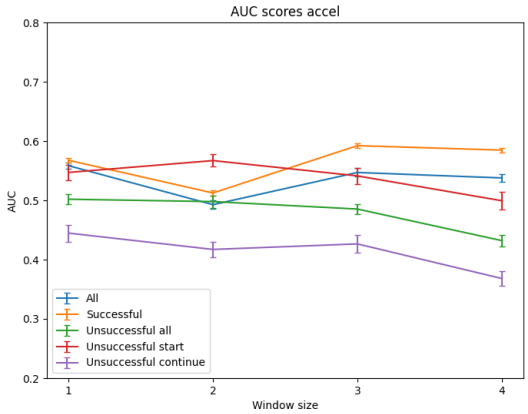

Figure 4: AUC score accelerometer

Estimating using non-verbal vocal behaviour and accelerometer data.
Looking at figure 5 and 6 it shows that the multimodal model performs similarly to the audio-only model, albeit slightly better. For this model, the successful cases perform best with AUC scores of 0.6742, 0.7379, 0.7159 and 0.7082 for the different window sizes. The non-verbal vocal behaviour seems to have a bigger influence on the performance of the model than the accelerometer data. This is logical since the non-verbal vocal behaviour performs a lot better than the accelerometer data. While the combination of the modalities shows improvement in the model, these improvements are small. The same holds for the unrealized intentions to speak.

One-tailed t-tests were done to compare the multimodal performance to accelerometer alone. Then, the p-values of these tests were computed. The p-value is the probability that the null hypothesis is true. A value smaller than 0.0001 shows that the difference is significant and the null hypothesis can

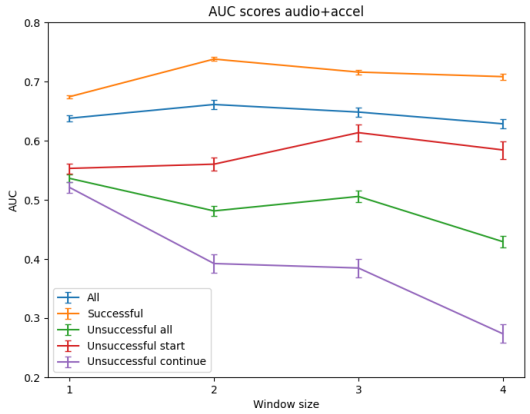

Figure 5: AUC score multimodal audio+accel

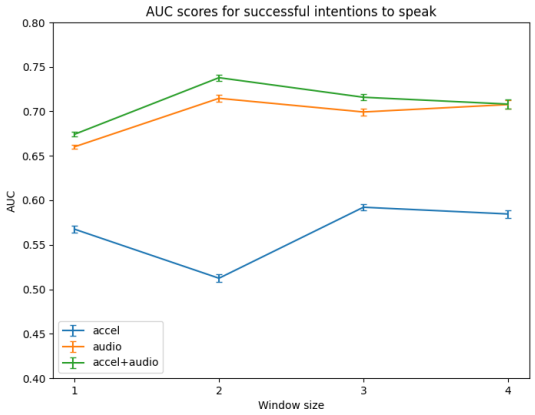

Figure 6: AUC scores comparing modalities for successful intentions to speak

be rejected in favour of the alternative hypothesis. A value of 0.9999 or higher shows that the null hypothesis is likely true. In this case, the null hypothesis H_0 is: The multimodal model performs similarly or worse than the accelerometer model. The alternative hypothesis, H_1, is: The multimodal model performs better than the accelerometer model. These values, shown in table 2, when green mean that the null hypothesis can be rejected and the multimodal model performs significantly better. The values coloured red mean that the multimodal model performs worse than the accelerometer model.

p-value	1s	2s	3s	4s
all	<0.0001	<0.0001	<0.0001	<0.0001
realized	<0.0001	<0.0001	<0.0001	<0.0001
unrealized	<0.0001	>0.9999	<0.0001	0.9740
start	0.0001	>0.9999	<0.0001	<0.0001
continue	<0.0001	>0.9999	>0.9999	>0.9999

Table 2: Table showing p-values of audio+accel compared to accel only

Different annotations.

Figure 7 shows the comparison of performance for the model for unsuccessful intentions between the annotations done by Li et al. [1] and ours. For both annotations, the intentions labelled 'start' perform a lot better than those labelled 'continue'. Notably, the performance difference in the 'continue' intentions in the 2s and 4s windows is quite big. Since the segment chosen for the unrealized intention to speak is based on the noted end time of the segment, a slight consistent difference in end time could also explain the differences. Mainly, given the high number of disagreements discussed in 3.2, it is unsurprising there are big differences in performance. It can be seen, however, that the annotations of this project generally perform better than those of Li et al..

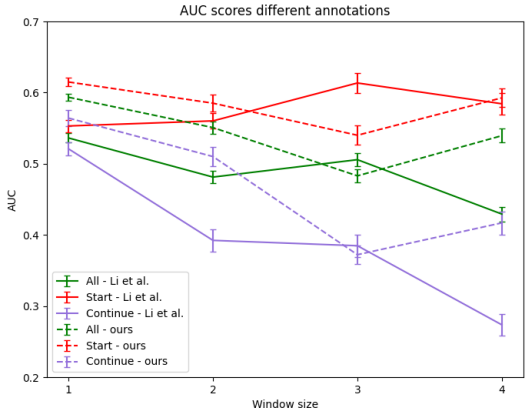

Figure 7: AUC score multimodal unsuccessful annotations Li et al. vs ours

5 Responsible Research

The research will be reproducible by anyone with access to the dataset and the repository used to write the code for the project [17]. The time stamps of the unrealized intentions to speak are also part of this repository.

To ensure the privacy of all participants, an End User License Agreement (EULA) was signed by all research group members. In this EULA, it was specified that the data may not be shared with third parties, and may not be used to identify the people in the dataset. Therefore, all work done with the dataset where video or audio footage was visible or audible was done in a private environment, and no data was shared without being fully anonymized.

Given that the annotations are done based on human intuition, it is safe to assume some bias might be present. To account for this bias, the research group made sure no annotations were done by one person alone, and all annotations were at least agreed upon between two people. If there was further uncertainty, the specific annotation was discussed with all five group members.

Another important aspect to look out for is confirmation bias. The knowledge of the different research being done by members of the research project group (body pose, lexical information, non-verbal vocal features), could make the annotator hyper-aware of these features when looking for unrealized intentions to speak. The group was aware of this possible bias and made an effort to approach the task objectively.

6 Conclusions and Future Work

This research estimates intentions to speak using multimodal data. It builds on previous research done on this same topic using accelerometer data as a sole modality. For this research, non-verbal vocal behaviour was added as a modality. Intentions to speak are split into successful intentions, which are extracted from the data, and unsuccessful intentions, which were manually annotated. From these annotations, it was found that head movement, filler words and intonation could be reliable indicators for unsuccessful intentions to speak. Additionally, posture shifts and arm/hand movement were common indicators.

Testing the model shows that non-verbal vocal behaviour is a good modality to infer intentions to speak, performing best on successful intentions to speak. Accelerometer data alone, while performing better than random guessing in successful intentions to speak, performed much worse than non-verbal vocal behaviour. Combining the modalities showed a slight increase in performance compared to non-verbal vocal behaviour.

Limitations

It is unsurprising the model performs a lot better on successful intentions to speak than on unsuccessful ones since the model is only trained on successful intentions. While it shows unsuccessful intentions can be estimated to some degree, a model trained on unsuccessful intentions to speak would likely give better results or a more accurate result.

Additionally, because of the way the negative samples are chosen, there is a chance unsuccessful intentions to speak

are included in these. This could skew the results for unsuccessful intentions to speak, and possibly even successful intentions to speak if the cues are very similar. This could be solved by having a bigger set of annotated data and only choosing negative samples so unsuccessful intentions are excluded.

Future work

Firstly, adding additional modalities to the model would be interesting to research and see what could improve the estimation. These modalities include video, body pose and lexical information [2-8], [10]. Secondly, as mentioned in the limitations, annotating a larger part of the data will allow for training on unsuccessful intentions to speak and thus more accurate estimations. While some parameters were optimized, looking more into parameter tuning and the choices made for VAD processing could also improve the performance of the model.

7 Acknowledgements

I want to thank Hayley Hung for supervising this project and giving us guidance and feedback throughout. I thank Stephanie Tan for joining in on our meetings and providing helpful insights. I would also like to give special thanks to Litian Li and Jord Molhoek for helping us and supporting us throughout this project by explaining their work and giving their insights. I would like to thank my group members for collaborating on the annotations. I would also like to thank my group member Julie van Marken for collaborating on the refactoring of the model. Lastly, I would like to thank Ruud de Jong for providing us access to the REWIND dataset.

References

[1] L. Li, J. Molbroek, and J Zhou. Inferring intentions to speak using accelerometer data in-the-wild. TU Delft, 2021.

[2] Volha Petukhova and Harry 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*, Stockholm, Sweden, Jun 2009. SEMDIAL.

[3] Uri Hadar, Thorsten Steiner, E.C. Grant, and F. Clifford Rose. The timing of shifts of head postures during conservation. *Human Movement Science*, 3(3):237–245, Sep 1984.

[4] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. Predicting next speaker based on head movement in multi-party meetings. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2319–2323, 2015.

[5] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. Prediction of who will be next speaker and when using mouth-opening pattern in multi-party conversation. *Multimodal Technologies and Interaction*, 3:70, 10 2019.

[6] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. Analysis of respiration for prediction of "who will be next speaker and when?" in multi-party meetings. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 18–25, New York, NY, USA, 2014. Association for Computing Machinery.

[7] Guanzheng Liu, Yanwei Guo, Qingsong Zhu, Bang-Yu Huang, and Lei Wang. Estimation of respiration rate from three-dimensional acceleration data based on body sensor network. *Telemedicine Journal and E-health*, 17(9):705–711, Nov 2011.

[8] Beth Ann Hockey. Prosody and the role of okay and uh-huh in discourse. In *Proceedings of the Eastern States Conference on Linguistics*, pages 128–136. Citeseer, 1993.

[9] Mathieu Jégou and Pierre Chevaillier. A computational model for the emergence of turn-taking behaviors in user-agent interactions. *Journal on Multimodal User Interfaces*, 12(3):199–223, May 2018.

[10] Jose Vargas-Quiros, Stephanie Tan, Laura Cabrera-Quiros, Chirag Raman, Ekin Gedik, and Hayley Hung. Rewind dataset: Speaking status detection from multimodal body movement signals in the wild. *Unpublished*.

[11] Florian Eyben, Martin Wöllmer, and Björn Schuller. opensmile - the munich versatile and fast open-source audio feature extractor. In *Proc. ACM Multimedia (MM)*, pages 1459–1462, Florence, Italy, October 2010. ACM.

[12] Florian Eyben, Klaus R. Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, Apr 2016.

[13] Ian T. Jolliffe and Jorge Cadima. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202, Apr 2016.

[14] Bartosz Mikulski. Pca-how to choose the number of components? Available at https://www.mikulskibartosz.name/pca-how-to-choose-the-number-of-components/ (2019/06).

[15] Jose Vargas-Quiros. lared_dataset. https://github.com/josedvq/lared_dataset, 2020. Accessed: June 5, 2023.

[16] Litian Li. testProject. https://github.com/llt-warlock/testProject, 2022. Accessed: June 5, 2023.

[17] Waded Oudhuis. researchProject. https://github.com/WadedOudhuis/researchProject, 2023. Accessed: June 25, 2023.