ESTIMATING INTENTIONS TO SPEAK USING MULTIMODAL DATA

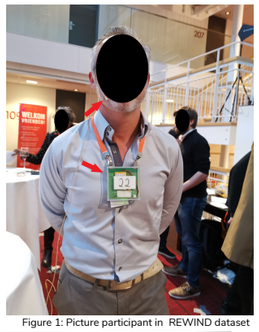

Figure 1: Picture participant in REWIND dataset

Estimating intentions to speak can help human-computer interaction, detecting moments when a person had an intention that was not realized. Using multimodal data, we can look at combining different modalities to gain a better estimation than from one modality alone.

1 BACKGROUND

- In research done by Li et al. [1], they attempted to estimate intentions to speak using accelerometer data. In their work, they were able to estimate these intentions better than random guessing.
- Research shows that head movements and posture shifts can be reliable indicators of the start of a turn [2]
- Additionally, non-verbal vocal behaviour is an interesting modality to consider, as pitch contours can indicate turn-taking [3].
- Using both movement and speech is compelling since research shows this combination is a reliable indicator of intentions to speak [4].

2 RESEARCH QUESTION

Can we estimate intentions to speak by combining non-verbal vocal behaviour and accelerometer data better than using accelerometer alone?

3 METHODOLOGY

- REWIND dataset [5], a database that contains video, audio and accelerometer data from a Dutch social networking event was used.
- To extract realized intentions to speak, VAD was used to find when a participant spoke for longer than 1.5s. The segment x (1, 2, 3 or 4 seconds) before the speaking was extracted (figure 2).
- For detecting unrealized intentions to speak, the research group annotated the data of 13 participants and segmented moments that could indicate an unsuccessful intention to speak. These intentions are split into start and continue.

Figure 2: VAD pre-processing [1]

- Audio features are extracted from the segments using openSMILE. The 25 features of the eGeMAPS parameter set are extracted and reduced to 10 using PCA.
- For movement, data is taken from a body-worn accelerometer (seen in figure 1) that measures change of velocity on the x, y and z-axis.
- For evaluation of the model, the AUC score is used.

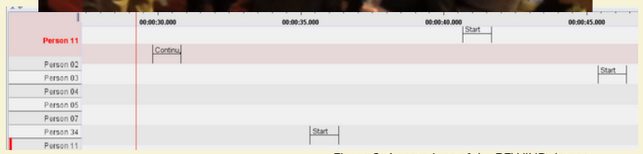

Figure 3: Annotations of the REWIND dataset

4 ANNOTATIONS

A group of 5 people, of which 3 native Dutch speakers, annotated a 10-minute clip of the REWIND dataset. The participants annotated were the 13 participants that were on video, had a microphone and an accelerometer sensor. The frequency of the signals annotated can be found in figure 4.

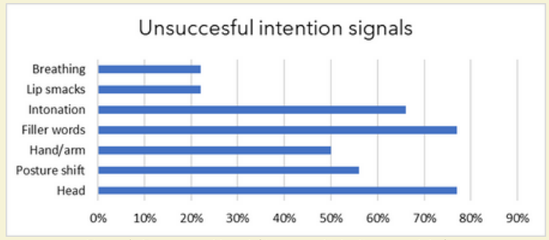

Figure 4: Frequency of turn-taking cues of intentions annotated

5 EXPERIMENTAL SETUP

- 5 different tests sets were used to validate the model: All, successful, unsuccessful, unsuccessful (start), and unsuccessful (continue).
- The tests were run on accelerometer data only, non-verbal vocal behaviour only and a combination of both
- For every modality, experiments were run for each combination of window size and test set. All experiments ran 100 times, after which the mean AUC score and standard deviation were computed

6 RESULTS

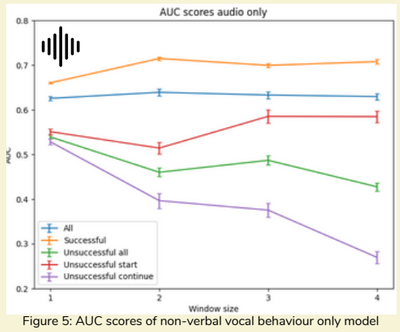

Figure 5: AUC scores of non-verbal vocal behaviour only model

- The audio only model performs best on successful intentions to speak
- The highest AUC score for this is 0.7147 at the 2s window
- Unsuccessful intentions to 'start' perform better than 'continue'
- The performance of the model spreads out more the bigger the window size

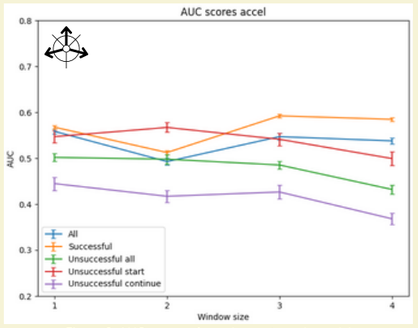

Figure 6: AUC scores of accelerometer only model

- Accelerometer data performs worse than audio, peaking at 0.5922 at 3s
- The model performs worst at the 2s window for successful intentions with a score of 0.5124, while this was best for the non-verbal vocal behaviour
- The unsuccessful intentions are closer to the successful, and the drop-off is less, compared to audio.

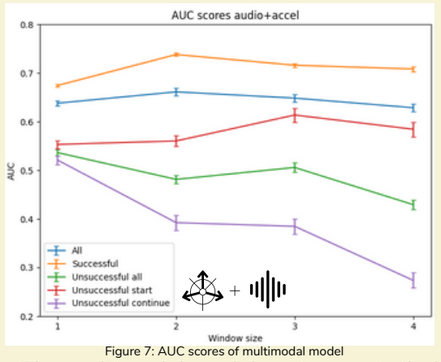

Figure 7: AUC scores of multimodal model

- The multimodal model performs similarly to the audio only model, but slightly better.
- The successful intentions perform best with scores of 0.7379 and 0.7159 for the 2 and 3 second window
- The non-verbal vocal behaviour seems to have a bigger influence on the model than accelerometer data

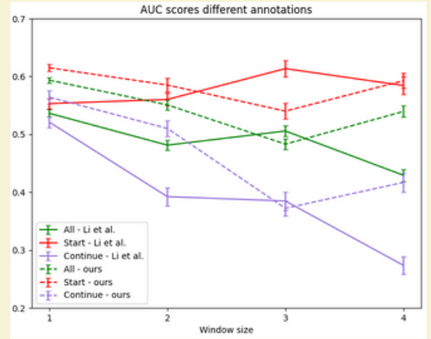

Figure 8: AUC scores of different annotations of multimodal model

- For both annotations, the intentions labelled 'start' perform a lot better than those labelled 'continue'.
- The performance difference in the 'continue' intentions in the 2s and 4s windows is relatively big
- Generally, our annotations perform better than those of Li et al.

7 CONCLUSIONS

- From the annotations, it was found that head movement, filler words and intonation could be reliable indicators for unsuccessful intentions to speak. Posture shifts and arm/hand movement were also common indicators.
- 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 estimating 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.

8 LIMITATIONS

- The model is only trained on successful intentions, not on unsuccessful ones.
- There is a chance unsuccessful intentions to speak are included in the negative samples.
- Because the annotations were done by the research group with knowledge of the different topics, there could be possible bias in annotations done.

9 FUTURE WORK

- Adding additional modalities to the model, such as lexical information, body pose or video could improve the model.
- Annotating a larger part of the dataset would allow for training on unsuccessful intentions to speak

RELATED LITERATURE

[1] Litian Li et al. Inferring intentions to speak using accelerometer data in-the-wild. TU Delft, 2023. Unpublished
[2] 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):237245, Sep 1984
[3] 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. Cite-seer, 1993
[4] 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 DialogueFull Papers (June 2009).
[5] 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.

AUTHOR Waded Oudhuis
CONTACT W.J.Oudhuis@student.tudelft.nl
SUPERVISOR Hayley Hung
ADVISORS Litian Li, Jord Molhoek, Stephanie Tan

TU Delft