Using non-verbal vocal behaviour to estimate intention to speak

1: Background

- In conversations, people who have the intention to speak up do not always get the chance to do so.
- Can AI be used to pick up on these intentions to speak?
- Master students of the TU Delft were able to successfully use accelerometer data to estimate intention to speak better than random guessing, although they noted that more modalities were needed to get a reliable estimation. [1]

2: Research Question

- **Can non-verbal vocal behaviour be used to estimate intention to speak?**
- Compare the results to random guessing and the results of the accelerometer model.

3: Related Work

- Mouth opening patterns, resulting in lip smacks, can be used to predict the next speaker.
- respiratory patterns can be used to predict the next speaker.
- Pitch difference can be used to distinguish between back channels and the start of a sentence.

4: Method

Machine Learning Model

- The existing model [2] was refactored to take non-verbal vocal features as input.
- The model was trained on successful intentions and evaluated on both unsuccessful and successful intentions to speak extracted from the REWIND data set [3].

Annotation of Unsuccessful Intentions

- Unsuccessful intentions (separated into intention to start speaking and intention to continue speaking) were annotated in a 10-minute segment.
- In total 53 unsuccessful intentions were identified: 32 unsuccessful intentions to start speaking and 21 intentions to continue speaking.
- Of the annotations 22,6% contained lip smacks, 22,6% contained breath and 66,0% contained intonation.

Successful Intentions

- Successful intentions to speak were automatically extracted using VAD (Voice Activation Detection) by extracting the segment before the start of speech.
- Pre-processing was done to set activations shorter than 1,5 seconds to 0 and pauses shorter than 0,5 seconds to 1.

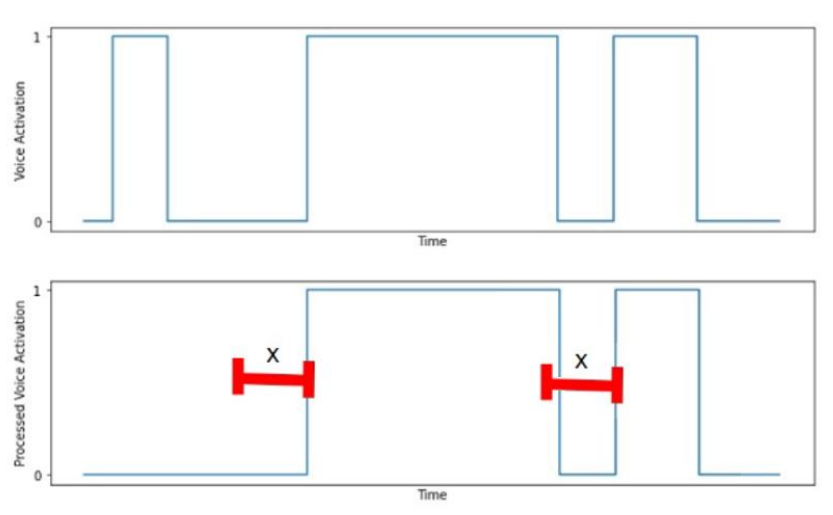

Figure 1: VAD pre-processing

Extracting non-verbal vocal features

- The eGeMAPS default feature set of OpenSmile [4] was used to extract 25 low-level audio descriptors related to paralinguistics.
- To speed up the training time, PCA was used to reduce the number of features from 25 to 10.

5: Evaluation

- To evaluate the performance of the model the AUC score was used.
- The model was evaluated for 4 different window sizes (1, 2, 3 and 4 seconds) and 5 experiments were run for different sets of intentions: all, successful, unsuccessful, unsuccessful (start) and unsuccessful (continue).

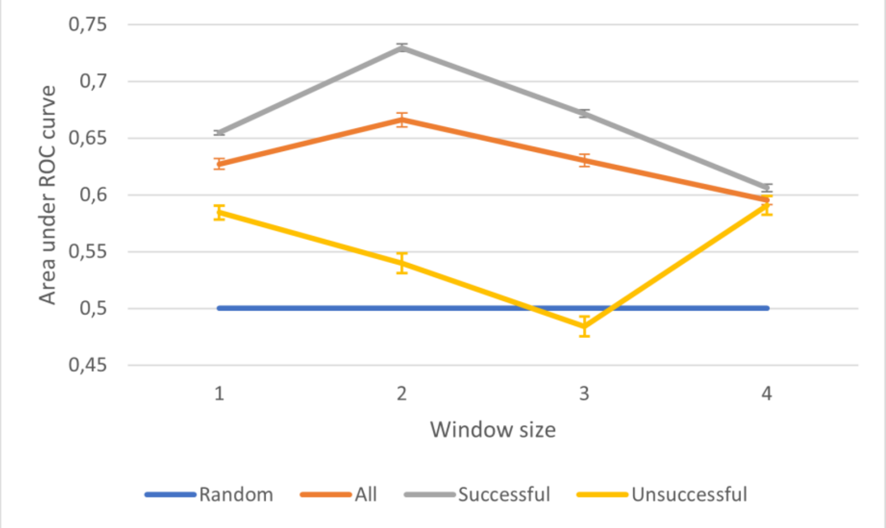

Figure 2: Graph of the first 3 experiments

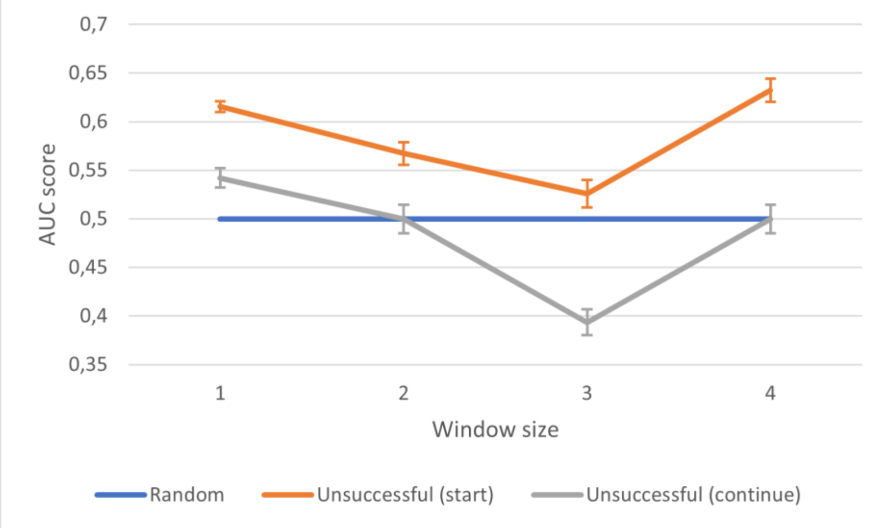

Figure 3: Graph of unsuccessful intentions

- To model performs best on successful intentions to speak and smaller window sizes.
- The good results at the 4 s window size were unexpected, the model could have picked up on a different pattern.

6: Conclusions

- The model consistently outperformed random guessing on all intentions except for the unsuccessful intentions to continue speaking.
- The model performed better than the accelerometer model on successful intentions to speak, except for window size 1, and performed slightly better on the unsuccessful intentions.

7: Future Work

- Annotate more data so the model can also be trained on unsuccessful intentions.
- Add different modalities to improve the performance.
- Use a data set where all participants wear microphones so the pitch difference can be researched further.

8: References

[1] Li et al. Inferring intentions to speak using accelerometer data in-the-wild. TU Delft, 2023.
[2] https://github.com/llt-warlock/unrealizedIntention
[3] Unpublished dataset by Jose Vargas-Quiros, Hayley Hung and Laura Cabrera-Quiros
[4] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 25.

Author: Julie van Marken **Contact: J.A.vanmarken@student.tudelft.nl** **Supervisor: Hayley Hung**