TU Delft

ENHANCING SOCIAL INTERACTIONS WITH AI-POWERED SYSTEMS

Nils Achy

1. INTRODUCTION

- AI-powered systems like chatbots, robots, and home assistants becoming increasingly popular.
- Conversations with these systems do not provide the same social interaction experience.
- Disparity caused by the absence of voice nuances, strict language use, and limited adaptation.
- **Inability to recognize when someone in a group discussion wants to contribute.**
- If robots and conversational agents could detect this behavior, they could create a more engaging environment for discussion. Ex: serve as mediators to ensure that everyone has the opportunity to express their thoughts (information gain).

2. RELATED WORK

- Previous work by Litian Li et al. [1] aims to infer instances of speaking intentions by training a model on accelerometer data using fixed window sizes.
- Does not provide a comprehensive understanding of the underlying structure of these intentions.
- When does the intention truly start?

[1] Jing Zhou Litian Li, Jord Molhoek. Inferring intentions to speak using accelerometer data in-the-wild, 2023. Unpublished.

3. CONTRIBUTION

- Infer segments (finding the start and end time) of speaking intentions **(segmentation)** on the same dataset using data captured by a body-worn accelerometer as input data to the model.
- Infer instances of speaking intentions **(classification)** using varying segment sizes of accelerometer data instead of fixed window sizes and compare with [1].

How can body language, captured by a body-worn accelerometer, be utilized to estimate segments of speaking intentions in time, and does a supervised learning process improve the performance of detecting such cases?

4. EXPERIMENT

- REWIND dataset (business networking event) of in-the-wild data of conversational speech for 1.5h in Dutch language (limitation)
- First half of the event participants engage on assigned topics, the second half in free conversation
- 13 people visible on camera during the 10-minute extract (1:00:00 to 1:10:00) wearing an audio recorder and accelerometer device

5. METHODOLOGY

1 REALIZED / UNREALIZED

2 USE VAD?
- Multi-modal extractor required to find the start time of intention

NOT feasible

instead

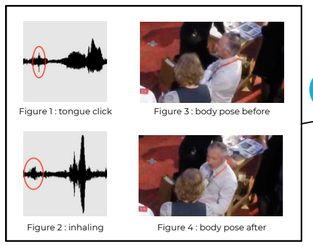

Figure 1 : tongue click
Figure 3 : body pose before
Figure 2 : inhaling
Figure 4 : body pose after

3 SUPERVISED
- Manual annotation
- 255 realized (positive), 31 unrealized & 255 negative
- Duration of Intention below 1.91 seconds

4

5 CLASSIFICATION
- Modified RNN allowing for varying input sizes
- Standard binary encoding for the label

5 SEGMENTATION
- 2-second segments
- RNN using accelerometer data as input arrays (size 40)
- Target binary vector (see Figure 1) of length 40

Against 4 models trained on fixed window sizes: (1, 2, 3, and 4 sec.)

6 EVALUATION
- AUC ROC metric
- 5-fold cross validation
- On realized, unrealized, and combination

Against baseline model assuming intentions always start in the middle

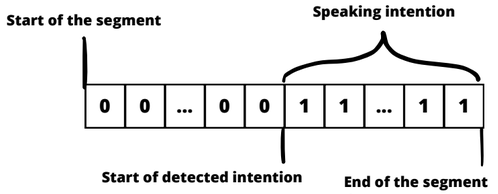

FIGURE 1: Target vector example

6. SEGMENTATION RESULTS

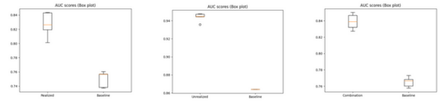

Table 1: Average AUC score and S.D for the trained model

	Realized	Unrealized	Combination
Average	0.827	0.944	0.840
S.D	0.018	0.005	0.010

Table 2: Average AUC score and S.D for the baseline model

	Realized	Unrealized	Combination
Average	0.750	0.864	0.765
S.D	0.011	0.0	0.006

- Trained model outperforms the baseline model (label structure explains high scores)
- Results show a small standard deviation across the 5 folds.
- Lower performance was observed in realized intentions compared to unrealized intentions.
- Possible reason for lower performance is the inclusion of negative samples in the test set of realized intentions.
- Model may have reduced accuracy when dealing with negative samples, leading to a higher rate of false positives.
- Model tends to overestimate the occurrence of speaking intentions.
- Despite overestimation, the model still performs better than the baseline.

7. CLASSIFICATION RESULTS

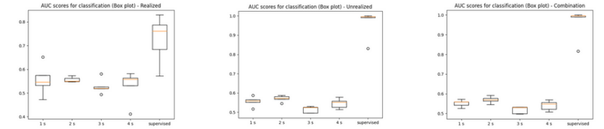

Table 3: Average AUC score and S.D for window size = 1

	Realized	Unrealized	Combination
Average	0.556	0.556	0.552
S.D	0.065	0.026	0.018

Table 4: Average AUC score and S.D for window size = 2

	Realized	Unrealized	Combination
Average	0.556	0.571	0.569
S.D	0.011	0.016	0.017

Table 5: Average AUC score and S.D for window size = 3

	Realized	Unrealized	Combination
Average	0.528	0.514	0.518
S.D	0.032	0.017	0.018

Table 6: Average AUC score and S.D for window size = 4

	Realized	Unrealized	Combination
Average	0.529	0.545	0.541
S.D	0.068	0.026	0.025

Table 7: Average AUC score and S.D for the supervised case

	Realized	Unrealized	Combination
Average	0.727	0.963	0.960
S.D	0.10	0.074	0.080

- Supervised model outperforms models with fixed window sizes in all three evaluation criteria and in all 4 window sizes.
- Results show consistency and low standard deviation.
- Supervised learning provides benefits in intention inference.
- Alignment with findings from [1] for unrealized intentions (2-second window size yields the most promising outcomes).
- Misalignment with findings from [1] for realized intentions (2-second window size in the present research, 1-second for [1]).

8. CONCLUSION & LIMITATIONS

- Model trained on accelerometer data demonstrates effective segmentation capability within the 2-second segment. Performs better on positive instances but achieves higher AUC scores on all 3 criteria compared to the baseline model with low standard deviation.
- Supervised learning brings significant improvement to the classification task; suggesting that qualitative data can play a role in building more realistic estimators for inferring speaking intentions.
- Potential future work includes the use of a larger dataset, incorporating data from more languages, and employing a rule-based approach to build a multi-modal extractor (to use a VAD).
- Manual annotations of speaking intentions are subjective, recommending multiple individuals for accuracy and reliability
- Fine-tuning the model and assessing performance using different metrics are important considerations.
- Exploring these ideas can lead to advancements in our ability to infer speaking intentions and improve Human-Computer interactions.