Estimating Intention To Speak Using Non-Verbal Vocal Behavior

Julie van Marken[1]

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: Julie van Marken
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Amira Elnouty
Daily supervisors: Litian Li, Jord Molhoek, Stephanie Tan

Abstract

This research aims to answer the question whether non-verbal vocal behavior can be used to estimate intention to speak. To answer this question data from a dutch social networking event is used to gather intentions to speak. The intentions to speak are split up in two categories: successful and unsuccessful intentions. The unsuccessful intentions are further split up into two categories: unsuccessful intentions to start speaking and unsuccessful intentions to continue speaking. The perceived unsuccessful intentions to speak are gathered by manually annotating a 10-minute segment of the networking event and successful intentions to speak are automatically extracted using Voice Activity Detection. From the audio, non-verbal vocal features are extracted to train a machine learning model to predict if there is an intention to speak. The model is trained on successful intentions to speak and evaluated on both successful and unsuccessful intentions to speak. From the experiment results it was concluded that the model predicted intention to speak better than random guessing.

1 Introduction

With the rise of ChatGPT Artificial Intelligence is a hot topic. AI can serve many purposes. In this project AI will be put in the context of a chairperson to lead a conversation. One of the most important tasks when one leads a conversation is ensuring everyone who wishes to speak up gets the opportunity to do so. This will allow for a better conversation, as more opinions get to be heard. Additionally, it can create a greater sense of belonging for the participants, as they get the feeling their voice is being heard. This of course raises the question 'How do you know when somebody wants to say something?'. Humans give subtle cues to show they want to speak up such as a slight head movement, opening of the mouth, saying certain words, and many more [1]. Humans have a natural instinct to pick up on these cues. In this project it be researched if AI can be trained to detect the intention to speak as well. This project is done as part of a research group where five students of the TU Delft look into using different modalities to estimate intention to speak.

Although different researches have been done into the prediction of the next speaker [2] [3], limited research has been done into the intention to speak. Next speaker predictions only take successful intentions to speak, meaning a participant wanted to say something and was successfully able to do so, into consideration. Intention to speak not only concerns itself with the successful cases, but also the unsuccessful cases. These unsuccessful cases indicate when a participant wanted to say something but for some reason was unable to do so. If AI were to be able to identify these unsuccessful intentions to speak, it could interject and allow the participant to speak up, allowing for a better conversation.

Li et al. (2023) [4] did research on using accelerometer data, which captures information about the general body movement, to estimate the intention to speak and were able to train a model to estimate the intention to speak better than random guessing. This is a promising start of the research into estimating intention to speak and can be built upon with research into different modalities.

One of those modalities is non-verbal vocal behavior. Non-verbal vocal behavior refers to all vocal cues, except for the meaning of words. Examples of non-verbal vocal cues are lip smacks, audible breathing, pitch and intonation. Non-verbal behavior is of great importance in conversations [5] and can greatly impact how the meaning of an utterance is perceived. As non-verbal vocal cues can be observed before turn taking [1], this could be a useful modality for estimating the intention to speak.

1.1 Research Question

This research aims to answer the question: **can non-verbal vocal behavior be used to estimate intention to speak?** To measure how well non-verbal vocal behavior can be used to estimate intention to speak, two sub questions will have to be answered:

- Can non-verbal vocal behavior be used to estimate the intention to speak better than random guessing?
- How does the performance of non-verbal vocal behavior compare to accelerometer data?

To answer these questions a similar experimental set up will be followed as Li et al. [4]. Unsuccessful intentions to speak will be annotated from a data set, after which non-verbal vocal features will be extracted to train a model to estimate the intention to speak using machine learning.

1.2 Related Work

This research will build upon previously conducted research in three related fields of study: turn taking, next speaker prediction and estimation of intention to speak. To understand how the intention of speech can be estimated, it is important to know what cues participants use when they take the turn to occupy the speaker role and whether these cues can be used to predict the next speaker.

Turn Taking

Turn management is an essential part of a multi party conversation. In order to take the turn and occupy the speaker role, the participant must signal their intent to have the next turn. Research into the cues that are observed before turn taking can help identify cues that could be used to show intention to speak. Petukhova and Bunt [1] did research into how participants signal their intention to have the next turn. They discovered that, beside gaze redirection and posture shifts, a range of audible expressions may be used to signal the intention to have the next turn, including filler words, repetitive sounds and other vocal sounds. Additionally, they found that mouth and lip movements correlate to turn initiating segments. Although these cues are not non-verbal vocal behavior by nature, certain mouth movement can manifest in audible

cues such as lip smacks caused by the opening of the mouth. Petukhova and Bunt also found that even though some modalities can be used by themselves, combining different modalities results in a better success rate in obtaining the next turn.

Predicting Next Speaker

The cues discussed in the 'Turn Taking' section to show intent to obtain the next turn, can be used to predict the next speaker. As mentioned previously, prediction of the next speaker is closely related to estimating the intention to speak. When a modality can be used to successfully predict the next speaker, it is likely it can also be used to estimate the intention to speak. Ishii et al. (2014) did research into the next speaker prediction using several different modalities [2] [3] [6]. Although they did not conduct research into using non-verbal vocal behavior specifically, they did do research into using mouth opening patterns [2] and respiration [6]. Both of these cues can be picked up by audio when they are loud enough. With the research into respiration it was discovered that a speaker inhales more rapidly and quickly after an utterance if they intend to keep the turn. It was also discovered that the next speaker tends to take a bigger breath in than the other participants. Using these findings they were able to construct a model that was effective for predicting the next speaker on average 900 ms before the next utterance. With the research into mouth opening patterns it was found that the next speaker often narrowly opens their mouth directly after closing it, an action which can occasionally result in an audible lip smack. Although mouth opening patterns were able to be used to successfully predict the next speaker, combining this modality with eye-gaze behavior allowed for a better prediction.

Intention to Speak

Where prediction of next speaker only concerns itself with successful intentions to speak, research into intentions to speak also takes the unsuccessful cases into consideration. Although no research in this field has been done into using non-verbal vocal behavior, there has been research done using a different modality. The main research this paper will build upon is research conducted by Li et al. [4] into estimating intention to speak using accelerometer data . In their research a machine-learning model was trained on successful intentions to speak. The model was then evaluated on both the successful intentions and unsuccessful intentions, which were manually annotated from a 10-minute segment of the REWIND data set [7]. It was concluded that, although accelerometer data captured useful information, it was not enough to reliably capture intention to speak. Li et al. [4] also noted that in the REWIND data set occasional mouth opening patterns could be heard in the form of tongue clicks and lip smacks. As these cues were observed at least once for 7 of the 13 annotated participants of the data set, they are likely not too person-specific and could be used to infer intention to speak.

Wlodarczak and Heldner [8] investigated whether breathing cues could be used to identify hidden turn taking events, an unrealised intention to take the turn. They discovered that breath holds produced towards the beginning of an exhalation potentially indicates an unsuccessful intention to take the turn. Although these breath holds are not captured in audio, they do strengthen the belief that breathing patterns can be an important indicator of turn changing and intention to speak, and therefor should not be overlooked.

Heldner et al. [9] did research into the pitch difference between back channels and turn changes. It was discovered that back channels, which are brief responses or acknowledgements such as "yeah" or "uh-huh", are similar in pitch to the end of the previous utterance of another participant, while pitch distances in turn exchanges tend to be larger between the utterances. This finding could be useful in the field of estimation to speak. When a large pitch distance is noticed between a filler word and the previous utterance, it could be an indication that there was an intention to speak instead of a back channel.

2 Methodology

To answer the research question a model was trained to predict the intention to speak. After an initial exploration of the data set annotations were made of unsuccessful intentions to speak and Voice Activity Detection (VAD) processing was done to allow for the extraction of successful intentions to speak. This was then used to create the intervals of positive successful intentions. These intervals, combined with extracted features from the audio, were then used to create samples. These samples were then used to train and test a model. In this chapter the steps taken will be discussed in more detail. A visualization of the pipeline can be seen in figure 1.

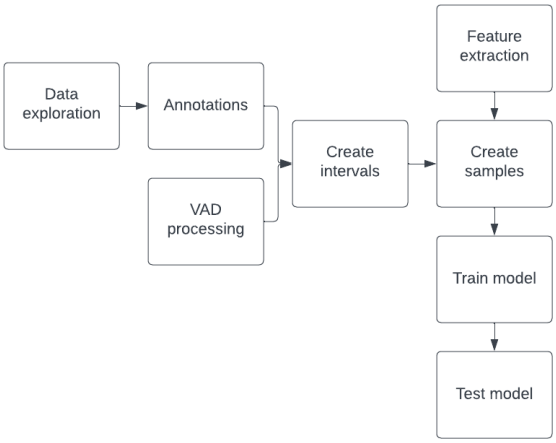

Figure 1: Methodology pipeline

2.1 REWIND Data Set

Before gathering data, a data set must be decided upon. For this research it was decided to use the REWIND data set [7]. The REWIND data set is a two hour recording of a dutch social networking event. During these two hours

the participants are able to walk around freely and talk to other participants. The event is recorded by several overhead cameras. A subset of these participants is wearing an accelerometer or a microphone. In total, there are 24 participants equipped with both an accelerometer and a microphone. The reason for choosing this data set is that the presence of audio-, video- and accelerometer-data allows for many different modalities to be extracted. This will allow for comparisons between the different modalities, which could be valuable to see how well the different modalities perform in the estimation of intention to speak. Additionally, in the context of the research group, choosing a data set that encapsulates multiple modalities allows for all group members, who do similar research into different modalities, to annotate the same data set which will result in more data to train and test the model.

From this data set a ten-minute segment where participants were able to walk around freely was annotated for unsuccessful intentions to speak, as used by Li et al. [4]. The use of the same segment allowed for a comparison to the annotations made by Li et al. To ensure that all projects within the research group could use the same date, only participants that were clearly visible on at least one of the cameras, wearing a microphone and wearing an accelerometer were annotated. These requirements left 13 participants to be annotated.

2.2 Data Set Exploration

Before the start of the annotation, the data set was explored to observe what cues people tend to give to show their intention to speak. It was noted that for a subset of participants an audible lip smack could be heard before the start of their turn, as can be seen in figure 2. This finding was also reported by Li et al. [4]. These findings align with the previously discussed finding of Otsuka et al. [2] that the next speaker often narrowly opens their mouth directly after closing it, resulting in an audible lip smack.

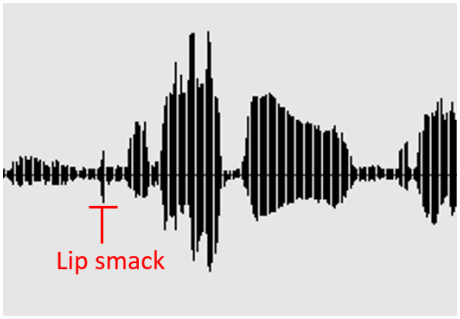

Figure 2: Lip smack before speech

Additional to the lip smacks, deep breaths were also observed before the start of a speaking turn, as can be seen in figure 3. This observation aligns with the findings of Ishii et al. [6] that the next speaker tends to take a bigger breath before the turn change. These findings indicate that the deep breaths observed in the audio could indicate intention to speak.

There also seemed to be a noticeable pitch difference between back channels and an (attempted) start of a sentence.

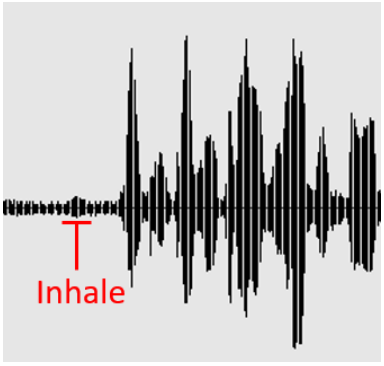

Figure 3: Inhale before speech

Although it seemed promising, when context was removed it proved to be extremely difficult to distinguish between the two. This aligns with the previously mentioned findings of Heldner et al. [9] that to distinguish between back channels and an attempt to take the turn, the pitch can be compared to the previous utterance of another participant. The larger the pitch difference between the two participants, the higher the chance of it being an intention to speak. Unfortunately, in the chosen data set, REWIND, not all participants have been equipped with microphones. This results in many conversations being between a participant with a microphone and a participant without a microphone. Due to not all participants wearing microphones, it is not possible to compare pitches to the pitch of the previous utterance of another participant. Therefor it will be out of scope for this research, although this could be interesting to explore in future research.

2.3 Annotation of Unsuccessful Intentions

To extract the unsuccessful intentions to speak from the data set, annotation was done using the software Elan [10]. To annotate a participant their audio, captured by the microphone, was isolated and the camera where the participant was best visible was looked at. When a participant displayed a cue that they has an intention to speak, the time frame where the cue occurred was annotated. Similar to Li et al. [4], unsuccessful intentions to speak were split into two different labels: intention to start speaking and intention to continue speaking. The intention to start speaking was annotated when there was a perceived unsuccessful attempt to take the turn. The intention to continue speaking was annotated when the participant already had the turn and attempted to keep the turn.

As mentioned previously, this data set was chosen so multiple group members could use the same data set which resulted in the annotating of more data, in this case more unsuccessful intentions to speak. When all the data is annotated by a single annotator, the data could potentially be biased towards what that one individual considers to be an intention to speak, which could result in skewed data.

A downside of having multiple annotators is that it can be challenging to have consistency across the different anno-

tators. In an effort to create consistency for the annotation process, all five annotators started out by annotating the same participant. After the initial annotation was completed all the results were compared to check for consistency and agree on what would be considered an unsuccessful intention to speak and how it would be annotated. Additional to the label, for each unrealized intention, the perceived cue that led to the assumption of it being an unrealized intention was annotated. The annotated cues were categorized into ”posture shift”, ”head movement”, ”arm/hand movement”, ”filler words”, ”intonation”, ”lip smack” and ”inhaling”. One annotation could involve multiple cues.

After this initial annotation, the other participants were annotated by at least two group members who compared their annotations to ensure there was consistency between the annotations. After the annotation of the 10-minute segment of the 13 participants was finished, the results were compared to the annotations of Li et al., which will be discussed in section 3.2. In total 53 unsuccessful intentions were annotated, consisting of 32 unsuccessful intentions to start speaking and 21 unsuccessful intentions to continue speaking. The results of the annotations will be discussed in detail in section 3.1.

2.4 Sampling Strategy

Li et al. [4] used a sampling strategy where positive samples were automatically extracted using microphone activition to extract the segment right before a participant started speaking. In the positive samples no overlap with speech was allowed. The negative samples were randomly picked intervals that did not overlap with the positive samples. As the negative samples are randomly picked, they are likely to include fragments of the participant speaking. If the same sampling strategy as Li et al. [4] were to be applied here, where positive samples had no overlap with speech and negative samples were randomly sampled, the model would likely be trained on the absence of speech rather than the presence of an intention to speak. To avoid this it is important to get positive samples that overlap with speech. Therefor it has been decided to allow for overlap with speech in the positive samples. An additional benefit of allowing overlap with speech is that there were more positive samples to train the model on.

An alternative approach that was considered was following the same sampling of positive samples as Li et al. [4] and only choosing random negative samples that did not overlap with speech. Although this approach would have solved the problem of the model associating speech with negative samples, it would have been difficult to find enough negative samples that fit the given criteria.

2.5 VAD Processing for Successful Intention Extraction

To extract the successful intentions to speak, the same method was used as Li et al. [4]. Microphone activitiy was used to automatically extract successful intentions to speak. Using microphone activity the start of a turn can be found and a time window before the microphone activity can be extracted, as we assume that before the start of a turn some intention to speak must have been shown to get the turn. Using this method three problems were identified by Li et al. [4] ”(1) microphone activity due to noise/other people speaking, (2) microphone activity because of short backchannels and (3) microphone deactivity when a speaker has a short pause but keeps the turn”. Fortunately, the REWIND data set [7] provides diarized binary VAD (Voice activity Detection) files for all participants. This removes microphone activity due to background noise or other people speaking and thus ensures there is only microphone activity when the participant wearing the microphone talks. The latter two problems were solved by using preprocessing. A microphone activition shorter than 1.5 seconds is set to 0, meaning ’not speaking’, as these are likely to be back channels and not an actual turn. Li et al. also set pauses shorter than 1.5 seconds to 1, meaning ’speaking’, but in this research only pauses shorter then 0.5 seconds were set to 1. This decision has been made to allow for more overlap with speech in the positive samples. Therefor the simplifying assumption has been made that even during a short pause the participant must have shown some intention to keep the turn and should be treated as an successful intention to speak. This assumption has the additional benefit of creating more positive samples to train the model with as the VAD will extract more successful intentions. A visualization of the preprocessing and time window extraction is shown in figure 4.

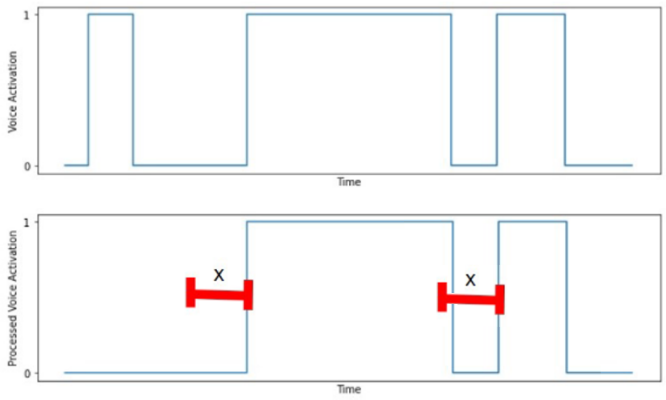

Figure 4: Extracting successful intentions. Figure adapted from Li et al. [4]

2.6 Feature Extraction

To train and test a model on non-verbal vocal behavior, features needed to be extracted from the audio. To extract the non-verbal vocal features from the audio files, OpenSmile [11] was used. OpenSmile is a commonly used software for research into prediction using audio. [12] [13]. OpenSmile has default feature sets, one of them being The eGeMAPS set [14], an abbreviation for the extended Geneva Minimalistic Acoustic Parameter Set. This feature set is commonly used for voice research and allows for the extraction of 25

low-level audio descriptors intended for use in para-linguistic speech applications. Para-linguistics are the aspects of spoken communication that do not involve words, such as tone changes and throat clearing. The usage of this feature set allowed for the extraction of vocal features, such as pitch and volume. Additionally, as it is used for para-linguistics, it was expected that the lip smacks and deep breathing mentioned in the 'Data set exploration' section could also be captured in this feature set.

After the 25 low-level descriptors were extracted, dimensionality reduction was performed. Dimensionality reduction is a commonly used pre-processing step to remove noise and reduce training time [15]. To speed up the training time, PCA (Principal Component Analysis) was performed. PCA is particularly useful when the dimensions of the input features are high and there exist multi-colinearity between the features. As can be seen in figure 5, the features that were extracted fit both these criteria. The color of the cell represents how related the two features are to each other. The lighter the color, the higher the correlation.

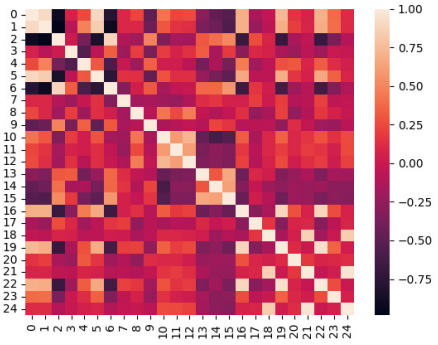

Figure 5: heat map of the set of extracted features

To choose the right number of components, Skicit-learn [16] was used to plot the cumulative variance. The cumulative variance shows the accumulation of variance for each number of principal components. The visualisation of the cumulative variance can be seen in figure 12.

To retain most of the information extracted, the explained variance threshold was set to 95%. The lowest number of components that surpassed this threshold was 10, hence the amount of principal components was set to 10. Reducing the amount of features from 25 to 10 allowed for a significant decrease in the training time while retaining most of the variance.

2.7 Machine Learning Model

For the model, the model used by Li et al. [18] was refactored to use audio features as input instead of accelerometer data. The model is a residual neural network trained by 3-fold cross-validation, aiming to classify a sample as either an intention to speak or not an intention to speak. The

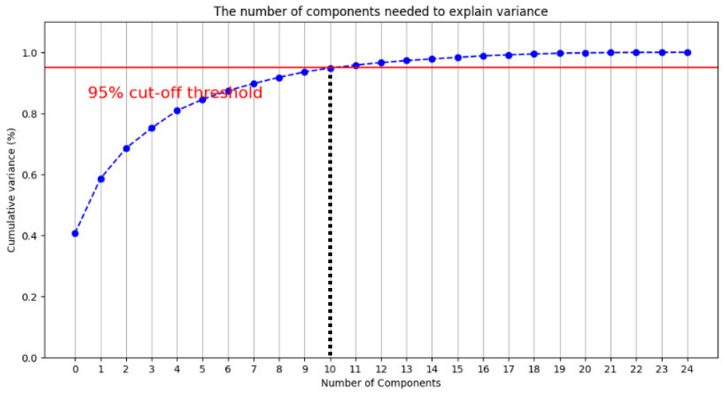

Figure 6: cumulative variance, visualized using format of B. Mikulski [17].

model outputs a binary classification of whether the sample contained an intention to speak or not.

Due to the limited amount of unsuccessful intentions that were annotated, the model was only trained on successful intentions to speak, extracted using the method described in section 2.5. The training samples are taken from the 2-hour recording, excluding the 10-minute time frame that was used for the annotations (1:00:00-1:10:00). The annotated unsuccessful intentions and the extracted successful intentions of the 10-minute segment were used as the test set.

To evaluate the performance of the model, the AUC (Area Under Curve) score was used. The AUC score is calculated as the area underneath the ROC (Receiver Operating Characteristic Curve), which maps the relation between the true positive rate and the false positive rate of the model. The AUC score ranges from 0 to 1. The larger the AUC, the better the model performs. An AUC score of 0.5 is on par with random guessing, hence the AUC score is usually expected to be in the range of 0.5 to 1.

3 Results

3.1 Annotation Results

As mentioned in section 2.3, a total of 53 unsuccessful intentions to speak were annotated, 32 of which were intentions to start speaking and 21 intentions to continue speaking. The average length of an annotated time window was slightly lower than 2,4 seconds. 90,6% of the annotations included some body movement as a cue: 56,6% included posture changes, 77% included head movements and 50,9% included some hand or arm movement. Although 90,6% of the annotations containing body movement as a cue is significant, it is important to note that in many cases body movement was not the main cue observed. For example, an intention would be annotated based on the observation of a specific filler word and only after it was already classified as an intention to speak, the body movement would be identified as an additional cue. 77,4% of the intentions included filler words

and in 35 of those 41 intentions intonation was observed as a cue as well. Intentions containing filler words were easier to spot than other more subtle cues such as lip smacks and body movement. Words such as ”maar”, ”en” or a start of a sentence were seen as a clear indication of an intention to speak. As these cues were the easiest to spot, there is a possibility that a disproportionate percentage of annotated intentions contain filler words. In 22,6% of the intentions lip smacks were observed and in 22,6% of the intentions deep breathing was observed, which aligns with the findings of Li et al. [4]. The cues related to non-vocal verbal behavior were ”intonation”, ”lip smack”, ”throat clearing” and ”inhaling”. In 83.1% of the annotated unsuccessful intentions at least one of these cues was observed, supporting the idea that non-verbal vocal behavior can be used to estimate intention to speak. A full list of the annotations can be found in Appendix B.

3.2 Comparing Annotations

Of the 53 unsuccessful intentions annotated, only 22 were also annotated with the same label by Li et al. [4]. As intentions can be shown by subtle cues, difference between the two sets of annotations was expected. However, an agreement of only 41,4% was significantly lower than expected. To find out where these disagreements stemmed from, both annotations were analysed to find the most common agreements and disagreements.

Agreements

When unsuccessful intentions to speak were annotated the same by Li et al. [4], it is assumed that there must have been a stronger cue shown than intentions that were disagreed upon. To test this assumption, the model was tested on a window size of 1 second on the prediction of unsuccessful intentions. This experiment was run for four different sets of annotations: the annotations done by Li et al., the annotations done by this research group, only the annotations that were agreed on and all the annotations combined. Each experiment was repeated 100 times to get reliable results. As can be seen in figure 7, the model performed better on the agreed upon intentions than the other sets of annotations . Although these are promising results that support the assumption made, these results are based on only 22 samples. To get more reliable results more samples are needed.

Disagreements

In 4 cases the same time was annotated, but a different label was assigned. When a participant had finished their first sentence but then lost the turn again Li et al. [4] would annotate this as an unsuccessful intention to start speaking whereas in these annotations such a case was assigned with the ’continue’ label. This disagreement comes down to a different interpretation of when the turn has been successfully taken. Another noticeable difference was the annotation of sudden interruptions. Li et al. annotated a sudden interruption as an unsuccessful intention to continue speaking whereas they were left out of scope in these annotations. The reason for leaving these cases out is that it was decided that when an interruption happened in the middle of a sentence there had been no chance to show intention to continue speaking.

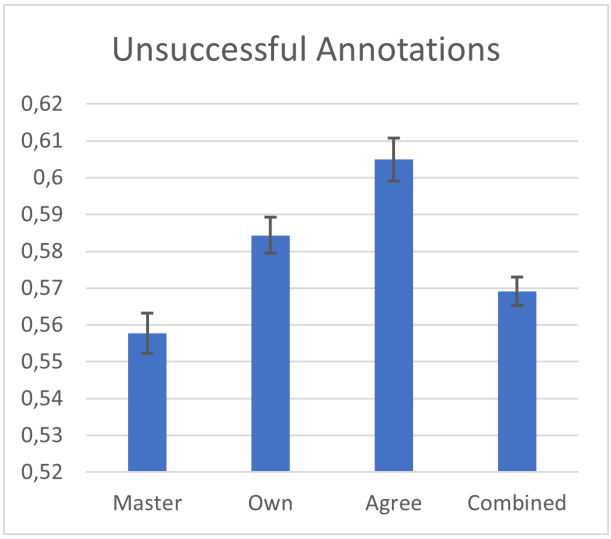

Figure 7: Visualisation of AUC scores for the different annotations

3.3 Performance Evaluation

The same experimental approach was used as by Li et al. [4]. The AUC scores were calculated for four different time windows: 1, 2, 3 and 4 seconds. The successful intentions for the different windows were created by extracting the corresponding amount of seconds before the voice activation, as described in section 2.5. For the annotated unsuccessful intentions the end time of the annotation was taken and the start time was the corresponding amount of seconds before the end time.

For each window five different experiments were run. Each experiment was used to test a different set of intentions to speak: prediction of all intentions to speak (both successful and unsuccessful), prediction of successful intentions to speak, prediction of unsuccessful intentions to speak, prediction of unsuccessful intentions to start speaking and prediction of unsuccessful intentions to continue speaking. To get reliable results, each experiment was repeated 100 times. The table containing the results in full can be found in Appendix A. The results of the first three experiments are visualized in figure 8 and the results of the last 2 experiments, where the unsuccessful intentions are further split into ’start’ and ’continue’, can be seen in figure 9.

As can be seen in figure 8, for every time window the model performs better on the successful intentions to speak than the unsuccessful intentions to speak. This was to be expected as, due to the limited amount of unsuccessful intentions to speak, this is what the model was trained to predict. An interesting difference occurs when the window size is increased to 2 seconds. Whereas the AUC score of unsuccessful intentions decreased, the score of the successful intentions to speak significantly increases from 0.655 to 0.730. An explanation for this observed difference could be that to successfully obtain a turn participants have to show their intention for a longer period of time. After the

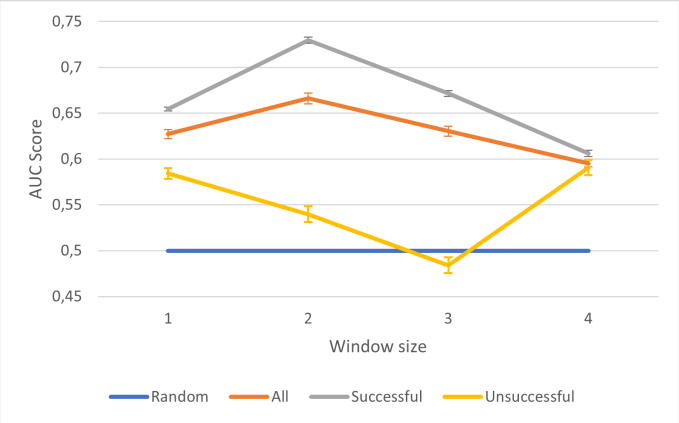

Figure 8: Visualisation of AUC scores for the first three experiments

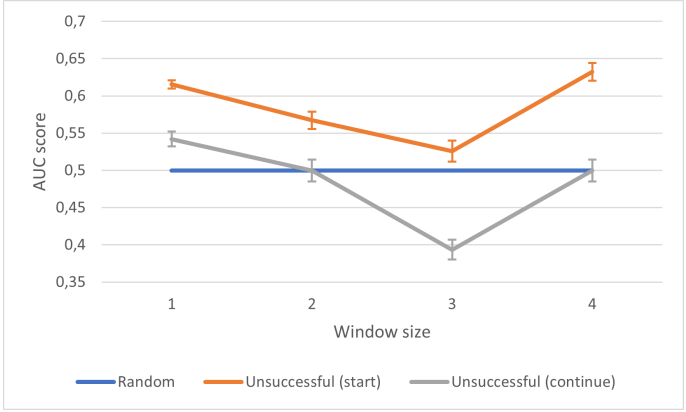

Figure 9: Visualisation of AUC scores for start and continue

initial increase for the 2 second window size, the AUC score decreases as the window size increases. This finding could indicate that most of the intentions to speak are shown closer to the start of speech. Increasing the window size adds more data that is less informative, making it more difficult for the model to pick up on patterns related to intention to speak. This finding was also reported by Li et al. [4].

The unsuccessful intentions also show a decrease in the AUC score as the window size increases, excluding the window size of 4 seconds. Annotated cues such as "lip smacks" and "breathing" tend to happen within one second of the start of speech in successful cases, which aligns with the assumption that most of the useful cues happen close to speech. Hence the shorter the time window, the better the AUC score. The high AUC score for the window size of 4 seconds was an unexpected result, given this assumption. As mentioned above, when more data that is less informative is added, it becomes more difficult to pick up on patterns related to intentions to speak. It is possible that in the 4 second time window, there is so much data unrelated to intention to speak that the model picked up other patterns, unrelated to intention to speak. This assumption is speculative as it is based on limited data samples, more data samples are needed in future research to draw reliable conclusions.

To get a more detailed insight into the unsuccessful intentions, the results of experiment 4 and 5, where the unsuccessful intention to start speaking and the unsuccessful intentions to continue speaking are separated, can be seen in figure 9. The AUC score of the intention to start speaking is significantly higher than the score of the intention to continue speaking, similar to the results found by Li et al. [4]. Li et al. suggested that the intention to continue speaking is not as long-lasting as the intention to start speaking as they already have the turn and therefor do not need to show social clues. These results seem to support that suggestion.

3.4 Comparison to Previous Work

To answer the research question of whether non-verbal vocal behavior can be used to estimate intention to speak the results will be compared to random guessing and the results found by Li et al. [4], where accelerometer data was used to estimate intention to speak. To compare whether the results are better t-tests were performed. The p-values are compared to the threshold of 0.001.

Comparison to Random Guessing

The null hypothesis of the t-test is "The model performs worse or the same as random guessing". As mentioned in section 2.7, the mean AUC score of random guessing is 0.5. The p-values of the t-test can be found in table 10. The green cells contain values that indicate the result is significant, the red cells contain values that indicate that the null hypothesis can not be rejected. On the window size of 1 second the model outperforms random guessing for every experiment. As the window size increases the model still outperforms random guessing, except for on the unsuccessful intentions to continue speaking.

p-value	Window 1	Window 2	Window 3	Window 4
All	<0.0001	<0.0001	<0.0001	<0.0001
Successful	<0.0001	<0.0001	<0.0001	<0.0001
Unsuccessful	<0.0001	<0.0001	1.000	<0.0001
Start	<0.0001	<0.0001	<0.0001	<0.0001
Continue	<0.0001	0.5517	1.000	0.5517

Figure 10: P-value for t-test comparing the model to random guessing

Comparison to Accelerometer Data

To compare the results to those of Li et Al. [4], the null hypothesis of the t-test is "The model performs worse or the same as the model for Accelerometer Data". As can be seen in figure 11, the model consistently outperforms the model of Li et al. [4] on successful intentions to speak, whereas their model, with the exception of the window size of 1 seconds, performs better or similar on the unsuccessful intentions.

4 Responsible Research

The research is currently not entirely reproducible. The experiment has been made more reproducible by making

p-value	Window 1	Window 2	Window 3	Window 4
All	<0.0001	<0.0001	<0.0001	<0.0001
Successful	<0.0001	<0.0001	<0.0001	<0.0001
Unsuccessful	<0.0001	1.000	0.0157	<0.0001
Start	<0.0001	1.000	0.4848	<0.0001
Continue	1.000	1.000	1.000	<0.0001

Figure 11: P-value for t-test comparing the model to the model for accelerometer data

the code open source on GitHub. All the steps taken to train and test the model have been documented in this paper. However, the data set used in this research is a unpublished data set approved by the ethics board of Delft University. This limits the reproducibility to those who have access to the data set. Although the research is currently not entirely reproducible, the paper related to this data set is in preparation for submission to a journal. After the publication of this paper the data set can be released publicly, making the research reproducible.

As the data set used, REWIND, is currently not publicly available, an EULA (End User License Agreement) had to be signed to gain access to the data set. By signing this EULA it was agreed that the data set would not be further distributed by the user and sufficient security measures for protecting the personal data should be taken by the user. This agreement was upheld by keeping the data locally and not sharing it with anyone. Additionally, before publishing the code to Github, all files containing data from the original data set were removed from the code.

Another part of responsible research to take into consideration is the selection bias. As the unsuccessful intentions are annotated by the researchers themselves, there is a chance that, subconsciously, the intentions that do not fit the hypothesis get overlooked. To minimize this bias, multiple project members, who all do research into different modalities, have annotated the data.

5 Conclusions and Future Work

5.1 Conclusions

This project aims to answer the question is intentions to speak can be predicted used non-verbal vocal behavior. The intentions to speak are split up in two categories: successful and unsuccessful intentions. The unsuccessful intentions are further split up into two categories: unsuccessful intentions to start speaking and unsuccessful intentions to continue speaking. The perceived unsuccessful intentions to speak are gathered by manually annotating a 10-minute segment of the REWIND dataset and successful intentions to speak are automatically extracted using Voice Activity Detection. After the experiment results it was concluded that the model predicted intention to speak better than random guessing, especially on the smaller window sizes of 1 and 2 seconds. On the window size of 1 second the model also predicts the intention to speak better than the model using Accelerometer data on all intentions to speak, except for the unsuccessful attempts to continue speaking. On the larger window sizes the model outperforms the Accelerometer model on successful intentions to speak, whereas the Accelerometer performs better on the unsuccessful intentions. Although the model already performs better than random guessing, more research can be done to improve the prediction of intention to speak.

5.2 Future Work

Although a subset of cues that show intentions to speak can be captured in non-verbal vocal behavior, there are many cues, such as movement and lexical information, that can not be captured. Future work could be done into combining different modalities, which will allow for more cues to be captured, to improve the estimation of intention to speak. A modality that could be especially interesting to combine with non-verbal vocal behavior would be accelerometer data. As mentioned by Wlodarczak and Heldner [8], breathing can be a strong indication of intention to speak and it can be captured by both audio and accelerometer data. It would be interesting to see if combining two modalities that are capable of capturing the same cue could aid in improving the accuracy of the estimation.

As mentioned in the Data set exploration chapter, the lack of every participant wearing a microphone made it impossible to explore the findings of Heldner et al. (2012) [9] in relation to estimating intention to speak. Their findings in distinguishing between back channels and other utterances using the vocal features of the previous utterance of another participant could be a valuable tool in using non-verbal vocal behavior to estimate intention to speak and is worth exploring more in the future.

As shown in the results, there is a significant difference between the AUC score of unsuccessful intentions to start speaking and unsuccessful intentions to continue speaking. Annotating more data in the future would allow for more research in how these two unsuccessful intention differ from each other to explain this difference in scores.

Acknowledgements

I would like to thank Hayley Hung, Stephanie Tan, Jord Molhoek, and Litian Li for supervising this project and providing valuable feedback through out. I would also like the thank my group members who I collaborated with for the annotation of the data set and provided their insights. Lastly, I would like to thank my group member Waded Oudhuis for the collaboration on the refactoring of the model.

References

[1] V. Petukhova and H. Bunt, "Who's next? speaker-selection mechanisms in multiparty dialogue."

[2] 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, p. 70, 2019.

[3] 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.

[4] L. Li, J. Molbroek, and J. Zhou, "Inferring intentions to speak using accelerometer data in-the-wild," 2023, tU Delft.

[5] M. Argyle, "Non-verbal communications in human social interaction," in *Non-Verbal Communication*, R. Hinde, Ed. Cambridge University Press, 1972, pp. 243–268.

[6] R. Ishii, K. Otsuka, S. Kumano, and J. 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*, ser. ICMI '14. Association for Computing Machinery, 2014, p. 18–25. [Online]. Available: https://doi.org/10.1145/2663204.2663271

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

[8] M. Wlodarczak and M. Heldner, "Breathing in conversation," *Frontiers in Psychology*, vol. 11, 2020. [Online]. Available: 10.3389/fpsyg.2020.575566

[9] M. Heldner, J. Edlund, and J. Hirschberg, "Pitch similarity in the vicinity of backchannels," in *Proc. Interspeech 2010*, 2010, pp. 3054–3057.

[10] "Elan [computer software] (version 6.4)," 2022, nijmegen: Max Planck Institute for Psycholinguistics. The Language Archive. [Online]. Available: https://archive.mpi.nl/tla/elan

[11] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," pp. 1459–1462.

[12] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6705–6709.

[13] D. Chakraborty, Z. Yang, Y. Tahir, T. Maszczyk, J. Dauwels, N. Thalmann, J. Zheng, Y. Maniam, N. Amirah, B. L. Tan, and J. Lee, "Prediction of negative symptoms of schizophrenia from emotion related low-level speech signals," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6024–6028.

[14] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing." TRANSACTIONS ON AFFECTIVE COMPUTIN.

[15] S. Velliangiri, S. Alagumuthukrishnan, and S. Iwin, "A review of dimensionality reduction techniques for efficient computation," *Procedia Computer Science*, vol. 165, pp. 104–111, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920300879

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] B. 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).

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

A Full Results

AUC scores	Window 1	Window 2	Window 3	Window 4
All intentions to speak	0.6272 (0.0049)	0.6661 (0.0061)	0.6304 (0.0055)	0.59534 (0.0041)
Successful	0.546 (0.0020)	0.7296 (0.0034)	0.6715 (0.0033)	0.6061 (0.0032)
Unsuccessful	0.5843 (0.0059)	0.5397 (0.0087)	0.4841 (0.0086)	0.5910 (0.0084)
Unsuccessful (start)	0.6155 (0.0055)	0.5671 (0.0116)	0.5260 (0.0141)	0.6323 (0.0117)
Unsuccessful (continue)	0.5422 (0.0099)	0.4998 (0.0149)	0.3935 (0.0133)	0.4998 (0.0149)

Figure 12: Mean and standard deviation of the AUC scores

B Annotation full

PID	Time window	Label	Posture shift	Head movement	Arm/hand moven	Filler word	Intonation	Lip smack	Throat clearing	Inhaling (breathir
<sample>	mm:ss:ms - mm:ss:ms	Start	1	0	0	maar	1	0	0	0
2	00:17.020 - 00:19.060	Start	1	1	1	ja	1	0	0	0
2	00:40.790 - 00:42.640	Start	1	1	1	dus	1	0	0	1
2	05:26.840 - 05:28.570	Continue	1	1	1		1	1	0	1
2	06:25.570 - 06:27.670	Start	1	1	0		0	0	0	1
3	00:27.460 - 00:30.130	Continue	0	1	1	en dat	1	0	0	0
3	02:19.500 - 02:21.950	Continue	0	0	1	van, ja	1	0	0	0
3	06:11.950 - 06:14.520	Continue	0	1	0		1	0	0	0
4	00:44.860 - 00:47.690	Continue	1	0	0	nou, en	1	1	0	0
4	01:52.740 - 01:55.090	Start	0	0	0		0	0	0	1
4	02:07.110 - 02:09.170	Start	1	1	0	dat, uh	1	0	0	0
4	03:12.490 - 03:14.400	Start	0	1	0	nou, ook	0	0	0	0
4	03:16.910 - 03:18.870	Start	0	0	0	start sentence	1	0	0	1
4	04:26.310 - 04:27.900	Continue	1	1	1	en	1	0	0	1
4	04:44.290 - 04:46.930	Continue	0	0	1	ik	0	1	0	1
4	04:57.150 - 04:58.820	Start	1	1	1	ik	1	0	0	0
4	05:10.960 - 05:12.890	Continue	0	0	0	en dan	1	1	0	0
4	07:08.950 - 07:11.620	Start	0	1	0	maar	1	0	0	0
4	07:34.360 - 07:36.800	Start	1	1	0	ja, maar, ja	0	0	0	0
4	07:41.680 - 07:44.710	Start	1	1	0	ja, maar, ja	1	0	0	0
5	04:05.770 - 04:08.370	Start	1	0	1		0	1	0	0
5	06:37.280 - 06:40.500	Continue	0	0	1	ik, rrrrrrr	1	0	0	0
5	06:52.940 - 06:54.980	Start	0	0	0		1	1	0	0
7	05:17.660 - 05:19.900	Continue	1	1	1	ja	0	0	0	0
7	09:30.090 - 09:34.000	Start	1	1	0	ik	1	0	0	0
10	04:25.920 - 04:27.680	Continue	1	0	0		0	0	0	0
10	04:28.070 - 04:30.040	Continue	1	1	1	dus, eh	0	1	0	0
10	04:59.760 - 05:01.570	Continue	0	1	0	ja	1	0	0	0
10	05:22.010 - 05:23.890	Start	0	1	0	ja	1	0	0	0
10	08:26.090 - 08:27.620	Continue	1	1	0		0	1	0	1
11	00:35.322 - 00:36.322	Start	1	1	1		0	1	0	0

PID	Time window	Label	Posture shift	Head movement	Arm/hand moven	Filler word	Intonation	Lip smack	Throat clearing	Inhaling (breathir
<sample>	mm:ss:ms - mm:ss:ms	Start	1	0	0	maar	1	0	0	0
11	00:50.231 - 00:51.231	Start	1	1	1	nou	0	0	0	0
11	00:52.129 - 00:53.129	Start	1	1	1	ik	1	0	0	0
11	04:44.610 - 04:45.610	Continue	1	1	0		1	0	0	0
11	08:24.946 - 08:25.946	Start	0	1	0		1	0	0	0
17	00:04.640 - 00:07.170	Start	1	1	1	ja	1	0	0	0
17	02:07.110 - 02:08.590	Continue	0	1	0	dus, eh, en	1	1	0	1
17	03:54.790 - 03:56.770	Continue	1	1	1	dus, eh	1	1	0	1
17	09:46.720 - 09:48.630	Start	1	1	0	ja, nou	1	0	0	0
22	04:05.900 - 04:08.380	Start	1	1	1	ja	1	0	0	0
22	05:19.730 - 05:22.010	Start	1	1	1	start sentence	1	0	0	0
22	05:55.920 - 05:59.270	Start	1	1	1	start sentence	1	0	0	0
22	08:51.540 - 08:56.200	Continue	1	1	1	ja, eh	1	0	0	0
22	09:15.410 - 09:17.510	Start	1	1	1	start sentence	1	0	0	0
23	09:03.100 - 09:05.700	Continue	0	1	0	nee	1	1	0	0
27	00:35.363 - 00:38.454	Continue	0	0	0	en, eh	0	0	0	0
27	00:44.909 - 00:48.909	Start	0	1	1	ja, dus, eh	0	0	0	0
27	01:26.181 - 01:28.090	Continue	0	0	1	dus	0	0	0	0
27	07:52.000 - 07:56.363	Start	1	1	1		0	0	1	0
34	07:06.545 - 07:12.090	Start	0	1	1	ja, juist, eh	1	0	0	1
34	07:19.545 - 07:20.727	Start	0	1	0	ja	0	0	0	0
34	07:53.363 - 07:58.545	Start	1	1	0	ja, eh	1	0	0	1
34	08:35.636 - 08:38.636	Start	0	1	0	dus, eh	0	0	0	0
35	03:34.080 - 03:36.580	Start	0	1	1	ja, zo dat	1	0	0	0
CUE %			0,5660377358	0,7735849057	0,5094339623	0,7735849057	0,6603773585	0,2264150943	0,01886792453	0,2264150943
CUE COUNT			30	41	27	41	35	12	1	12