Inferring Segments of Speaking Intention Using a Body-worn Accelerometer

Enhancing social interaction with AI-powered systems

Nils Achy

Supervisor: Hayley Hung

[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: Nils Achy
Final project course: CSE3000 Research Project
Daily supervisors: Litian Li, Jord Mohoek, Stephanie Tan
Thesis committee: Hayley Hung, Amira Elnouty

Abstract

This research paper proposes a deep learning model to infer segments of speaking intentions using body language captured by a body-worn accelerometer. The objective of the study is to detect instances where individuals exhibit a desire to speak based on their body language cues. The labeling scheme employed is a binary string, with "0" indicating no intention to speak and "1" indicating the presence of an intention to speak on a defined window size of 40 (corresponding to a 2 seconds segment recorded at a frequency of 100 Hz scaled down to 20 binary points per second). In this experiment, a real-life social event dataset was employed, and intentions to speak were manually annotated. A 10-minute segment from the dataset was selected and annotated using the ELAN software. The annotations included two categories: realized intentions, where individuals intended to speak and actually did so, and unrealized intentions, where individuals displayed intentions to speak but did not take their turn. The dataset consisted of 255 segments with realized intentions and 31 segments with unrealized intentions. Additionally, 255 negative samples were included, representing instances where no intention to speak was observed throughout the entire segment. To address the class imbalance inherent in the dataset, the model was evaluated using the Area Under the ROC Curve (AUC) metric using 5-fold cross validation. The model was tested on realized intentions, unrealized intentions, and a combination of both. Its performance was compared against a baseline model, which is assumed to always predict the start of the intention in the middle of the segment. Also, a new model was built that enables classification using varying window sizes as input. The classification task is performed to provide a comparison to another study [1] in which the training segments were predetermined window sizes instead of precisely annotated segments. The results of the study indicate that the deep learning models perform consistently better than the baseline on the segmentation task, and surpasses the performance of the model trained exclusively on window sizes on the classification task. This not only demonstrates the potential of body language as an informative cue for inferring speaking intentions, but also suggests that a supervised learning where intentions are identified with greater precision can lead to a superior outcome.

1 Introduction

The pace of development for AI-powered systems is in continuous growth and as a result, interacting with such systems in various forms, including chatbots, robots, and home assistants, will become increasingly popular. Although such voice interactive systems have achieved impressive levels of performance, one of their current limitations is the lack of "human-touch". Carrying a conversation with a home assistant or a robot does not yet provide the same social interaction experience as with fellow humans.

Several factors, such as the absence of voice nuances, strict and uniform language use, and limited adaptation to the person they communicate with, contribute to this disparity. This includes their inability to recognize when one or more of its interlocutors want to contribute in the group discussion. If they could detect such behavior, robots and other conversational agents could create a more engaging environment for conversation. For instance, they could turn towards the person and welcome them to speak up politely. Enabling conversational agents to recognize when someone wants to speak up would not only enhance the "human-touch" experience of interacting with non-human agents but also create a more equitable environment. Robots could serve as mediators to ensure that everyone has the opportunity to express their opinions and thoughts. In order to make significant advancements in the development of our interactive systems, it is needed to dedicate research efforts to the study of speaking intentions. This area of investigation holds key importance, as it can pave the way for further progress.

Previous work by L. Litian et al. [1] attempts to develop an estimator by recording and annotating data from a body-worn accelerometer in a social gathering of 13 individuals that meet the required criteria (i.e., every participant is in the frame of one of the four cameras during the extract under study, and wears a voice audio recorder and a body-worn accelerometer). They divide intentions into realized (how an individual behaves before starting to speak) and unrealized (when an individual exhibits observable behaviors, either visual or auditory, that are interpreted as indicative of their desire to speak but they do not actually take their turn in the conversation).

Despite obtaining some promising insights on how body posture changes with the intention to speak, their model is a classifier that predicts whether an individual has the intention to speak within 4 given time windows (1, 2, 3, and 4 seconds before the individual starts speaking). The model does not identify precise start and endpoints of speaking intention from the discussion, and thereby does not provide a comprehensive understanding of the underlying structure of these intentions. This limitation raises questions about when the intentions truly commence and how they may vary among different individuals and across different types of intentions.

1.1 Contribution

This paper seeks to expand on the findings from [1] and make further progress by attempting to estimate segments of intention-to-speak, from the moment an individual is assumed to want to speak to the moment when the individual starts speaking if the intention is realized and to the moment he/she is assumed to have started speaking if the intention is not realized, using body language data gathered by a body-worn accelerometer. In typical segmentation tasks, the goal is to identify all instances of intention-to-speak by examining the entire timeline. However, this research takes a different approach by focusing on already extracted segments and aiming to pinpoint the exact start and end points of intention-to-speak within those segments. This approach allows for performing segmentation analysis as well as performing classification to compare with the work conducted by [1]. In this study, the specific segments of interest can be studied in isolation, providing a valuable opportunity to investigate specific patterns, trends, or anomalies within those segments. It is interesting to know if this supervised process can yield better result than using pre-defined window sizes. Despite the advantages of using additional modalities to enhance the model's performance, this paper focuses specifically on demonstrating the potential of body language in inferring speaking intentions as it aims to provide a fair comparison of the results with the model presented in the study conducted by [1]. In alignment with their research, which solely utilized accelerometer data, this study will also exclusively employ accelerometer data. If successful, a potential outcome of this research could be a deeper understanding of the structure of conversational speech, from the perspective of body language and its relation to the progression from intention to articulation.

Research question: 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?

2 Related work

Turn-taking

A significant aspect to consider is the influence of cultural differences on turn-taking behaviors. T. Stivers et al. [2] conducted an analysis of 10 languages with diverse characteristics. Their research confirmed the presence of strong universals in turn-taking mechanisms, despite these cultural variations. These universals include a positive time offset in response and, on average, no more than half a second of overlap with the previous speaker. Additionally, the factors indicating response time appeared to be consistent across all languages.

However, the authors acknowledge that ethnographic reports on certain languages, such as Danish, suggest a subjective impression of longer response times. They argue that this perception can be explained by humans' hypersensitivity to even minor changes in response time, often within the range of less than 100 ms, where in reality this delay can simply be attributed to the time required to articulate the first syllable; which may differ among languages.

In another work, by H. Sacks et al. [3], the authors establish a set of systematic rules that appear to govern the organization of turn-taking. They propose that turn allocation in conversation is a collaborative process, guided by shared understandings among participants. Rather than following a predetermined order, turn allocation is negotiated in real-time, contingent upon various cues and signals present in the interaction. The authors emphasize the concept of ”transition relevance places” (TRPs), identifying these as points in the conversation where a current speaker can appropriately yield the floor to the next speaker. TRPs are marked by syntactic completion points, prosodic cues, or other indicators that indicate a suitable moment for transition.

Their analysis also acknowledges the occurrence of overlaps and interruptions in conversation, which challenge the notion of a clean turn-taking process. However, they demonstrate that these phenomena are not random but rather systematic and rule-governed. By examining recorded conversational data, they provide insights into how participants manage and resolve overlaps and interruptions within the turn-taking system.

Verbal cues

G. Skantze [4] has made significant contributions to the understanding of verbal cues and demonstrated that they play a vital role in analyzing conversational speech, providing valuable insights into the intentions and dynamics of communication. According to his research, lip smacks, filler words, and inhaling have emerged as significant indicators. These cues, when observed and understood, can offer valuable information about the speaker's cognitive and emotional states, the organization of the conversation, and social interactions.

Lip smacks, characterized by a brief clicking or smacking sound produced by the lips, often occur during pauses or transitions in speech, indicating an intention to continue speaking or marking a momentary hesitation. Filler words, such as ”um,” ”uh,” or ”like,” are another set of verbal cues that can provide valuable insights. These words often appear during pauses or when speakers are searching for words or formulating their thoughts. Inhaling, although primarily a physiological process, can also act as a relevant verbal cue. The sound of inhalation can signify the speaker's intention to take a breath before continuing their speech. It may indicate pauses, turn-taking, or shifts in emphasis.

In her paper, Schaffer D. [5] explores how intonation serves as a cue to indicate intentions in conversation. She highlights that variations in pitch, stress, and rhythm can convey signals of a speaker's intention to continue or yield their turn in a conversation. Intonation patterns, such as rising or falling pitch contours, can indicate whether a speaker is making a statement, asking a question, or signaling their desire to maintain or relinquish their speaking turn. Schaffer emphasizes the significance of intonation as a powerful cue that contributes to the smooth flow and coordination of conversation.

Non-verbal cues

The research conducted by P. Bull (1983) and later by F. Poyatos (2002), focus on the significance of nonverbal communication in interpersonal interactions. Bull's study examines the relationship between body movement and communication, emphasizing the role of nonverbal cues, such as gestures, posture, and facial expressions, in interpreting social interactions. Poyatos' research explores the multifaceted nature of nonverbal communication across different disciplines, investigating cultural influences, sensory interactions, speech, and conversation in shaping nonverbal communication patterns.

Common nonverbal cues discussed in the research papers include facial expressions, gestures, eye contact, proximity, touch, and body movements. Facial expressions convey emotions and intentions, while gestures and body postures provide additional meaning during communication. Eye contact signals interest or disengagement, and proximity reflects personal space and cultural norms. Touch communicates warmth, trust, and familiarity. Body movements, such as orientation and posture, contribute to nonverbal communication, indicating interest, disinterest, or discomfort.

3 Methodology

3.1 Information about the dataset

To conduct the experiment, data from a real-life social event was used, which recorded audio, video, and gathered data from a body-worn accelerometer for some of the participants. The dataset was explored collaboratively with 4 other students researching on a different subtopic and modality. The focus was on the REWIND dataset [6], which consists of a 1.5-hour business networking event. The first half of the event consisted of workshops where participants were assigned specific topics for discussion. During the second half, participants were encouraged to engage in free conversations, providing a more natural conversational setting. Four cameras were strategically positioned to capture the event from different angles. The focus was on a 10 minute extract of the event, precisely from 1:00:00 to 1:10:00. This particular extract was chosen as it corresponds to the one utilized by [6], ensuring a fair comparison, and offers the advantage of capturing participants engaged in free conversations, thus reflecting a practical setting outside of controlled environments. Like in the case of the study conducted by [1], among the 58 participants who attended the event, only 13 could be successfully identified during the extract on at least one of the four cameras, and were also wearing a body-worn accelerometer device and a voice audio recorder.

3.2 Generating samples

Segments of intentions were manually extracted from the dataset using the ELAN software [7] and categorized into realized and unrealized intentions, similar to the approach undergone by [1]. The unrealized intentions were used to evaluate the performance of the model. To generate samples suitable for a segmentation task, the labels were represented as binary strings with a maximum defined length. In this encoding scheme, a value of “0” indicated that no intention to speak had been detected, while a value of “1” represented the presence of an intention to speak. The end of the string corresponded to the precise moment when the person began to speak. For realized intentions-to-speak, this moment was identified by detecting a spike in activity on the voice audio recorder, indicating the start of speech, or in this case the end of the intention. This approach excluded filler words, which were considered part of that intention. Since the dataset was in the Dutch language, a list of filler words specific to that language was defined during the annotation process. The research team, including three members who were native Dutch speakers, collaborated to establish this list. For unrealized intentions, this precise moment was less straight forward and had to be approximated instead. The annotation of unrealized intentions for each participant followed the annotation of realized intentions, which proved helpful in identifying patterns in individual behavior. This process contributed to establishing a ground truth for each participant. For instance, it was observed that some participants initiated speech immediately after a posture shift, while others did so after a certain time delay following a tongue click. The same cues were utilized to annotate the endpoint of the intention to speak. Appendix A.1 illustrates a list of all the defined cues along with their presence or absence in each unrealized intention found.

A visual representation of a label is given in Figure 1. By employing this binary string representation, the model was effectively identifying and segmenting periods of speech intention within the given input data.

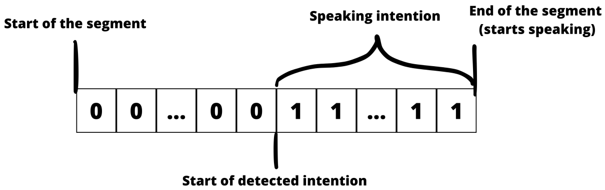


Figure 1: Binary string representing the model’s output

To generate the training dataset, data collected by the body-worn accelerometer was used. Specifically, the readings corresponded to the data collected during the timeframe encompassing the moment a person began to speak and a fixed interval before it. The length of this timeframe was predetermined and consistent with the data from the labels.

For the classification task, the labels were defined using the standard binary encoding (1 indicating the detection of an intention, 0 otherwise). The model was trained and tested using the same dataset associated with the detected intentions, but following two distinct approaches. In the first approach, the model was trained and tested using a uniform window size for all samples, similar to the method employed by [1]. In the second approach, the model was trained and tested using manually extracted segments of speaking intentions. The idea was to test the benefits of using the segments extracted during the segmentation task to build a classification model.

3.3 Building and evaluating the model

To accurately predict the segments, a predictive model was constructed using a Recurrent Neural Network (RNN), known for its effectiveness in sequence modeling tasks. To assess the model’s performance, its predictions were compared to a baseline model, for which the exact behaviour will be explained in the Experimental Set Up section.

To ensure a comprehensive evaluation, this process was repeated using three different configurations for the test set. This provided a robust assessment of the model’s performance across 3 different criteria:

- Only test on **unrealized** intentions to speak
- Only test on **realized** intentions to speak
- Test on a combination of both (**realized and unrealized**)

For the classification task, a modified version of the Recurrent Neural Network (RNN) was developed, enabling it to handle variable input sizes. This modification allowed the RNN model to be trained and tested on the extracted segments that have varying lengths.

4 Experimental Setup and Results

4.1 Experimental Setup

Extracting realized intentions

In order to obtain a substantial amount of data, an automated approach using a voice activity detector (VAD) was considered for extracting segments corresponding to the intention to speak in the training dataset. However, this approach proved unsuitable for achieving the research objective. The aim was to precisely define segments of speech intentions, necessitating the identification of exact start and end points in time. When provided with an audio file, the VAD has the capability to detect voice activity, thereby indicating the beginning of speech. However, it is limited to detecting only the onset of someone speaking and does not provide information about the intention that precedes the speech. This limitation is what forced the researchers from [1] to employ fixed window sizes, which assumed uniform duration for all intentions. This contradicted the purpose of the current research, which sought to explore a different approach.

An exploration of the dataset was conducted, searching for patterns that could serve as simplified assumptions for a rule-based approach to segment extraction using the VAD. The results revealed that, although patterns were identifiable among the 13 individuals being studied, they were too diverse to establish a universal rule. Some individuals manifested an intention to speak by leaning backward, while others did so by noding their heads, and still others by remaining motionless, occasionally signaling their intention to speak with a tongue click or inhaling. These differences are documented more in details in A.1. This indicated that extracting data using a rule-based approach would have needed a complex multimodal extractor that falls beyond the scope of the current study. Consequently, under the current experimental conditions, segments had to be manually annotated to generate samples.

To accurately identify the start and end points of an intention-to-speak, it was crucial to understand which indicators to consider during the annotation process and as already mentioned and illustrated in A.1, these indicators vary significantly. The literature helped in trying to define a ground truth in the annotation process, especially the cues found in the turn-taking mechanism mentioned by [8] and [3], and the non-verbal cues mentioned by [9] and [10]. A.2 provides detailed descriptions of two instances during the annotation process, along with an explanation of the cues employed in each case.

During the 10-minute extract and for the 13 participants under study, 255 segments of speaking intentions were identified and annotated. During the data exploration of the annotated segments, it was discovered that the average duration of the segments was approximately 1.32 seconds, with the longest segment spanning 1.91 seconds. Based on this analysis, the decision was made to set the segment length to 2 seconds for the segmentation task.

The first intriguing discovery from the study was the notable variation in the time gap between the appearen of indicators signaling an intention to speak and the actual commencement of speech, even within the same individuals. This finding emphasizes the likely significance of adopting supervised learning techniques to infer instances of speaking intentions, rather than relying on pre-defined time windows.

Extracting negative samples

To ensure the model's ability to distinguish between cases where participants did not intend to speak, negative samples were included. These negative samples were carefully selected to avoid any overlap with the positive samples, although they could overlap with each other. To generate these negative samples, the minimum start time and maximum end time of any detected intention to speak for a specific participant were used as the lower and upper bounds, respectively. A new start time was randomly generated within this range, and the segment length was added to determine the end time of the new segment. Subsequently, it was verified whether this new segment overlapped with any realized, or unrealized segments. If an overlap occurred, the method was recursively called, and if no overlap was found, the segment was considered a valid negative sample. The label associated with this new segment was assigned zeros and added to the dataset.

For each participant, an equal number of negative segments were generated as the number of realized intentions found for them. This approach ensured consistency in the training process and provided each participant with an equal representation during training. This summed up to 510 samples for training and testing the model. In the classification task, negative samples were also included during the testing phase of unrealized intentions following a similar approach.

Generating the samples

To extract the accelerometer data from a specific segment, the AccelExtractor class provided by [1] was used. By providing the person's ID, start time, and end time, this class returns a matrix consisting of three arrays that represent the three parameters recorded by the accelerometer device during the corresponding interval for that particular person. The accelerometer data is recorded at a frequency of 100 Hz but scaled down to 20 binary points per second.

Each mask in the segmentation task had a duration of 2 seconds and consequently each feature array from a sample had a length of 40. The corresponding label array also had a size of 40, where the values were set to zero from the beginning until the start of the intention, and then set to one from that point until the end of the segment. For negative sample points, this array was filled with only zeros. For the classification task, the size of the feature array was found by multiplying the window size (in seconds) by 20.

Building and evaluating the model

For the segmentation task, the model used for prediction is a Recurrent Neural Network built using pytorch [11] with a first simple layer of 64 units and relu as activation function, and a Dense layer of 40 units (the output array length) and a sigmoid activation function. The model's performance was evaluated using a 5-fold cross validation technique. To ensure fairness in evaluation, the unrealized intentions were tested using the same trained models that were generated at every fold. To evaluate both the realized and unrealized intentions together, the annotated segments of the unrealized intentions were concatenated with the validation set before performing the evaluation.

In the segmentation task, the model does not adhere to a binary classification framework. While metrics like Intersection-Over-Union (IoU) are more commonly used for segmentation problems, the preferred choice was the AUC score. This metric is a popular choice when dealing with class imbalance which was likely to be the case in this experiment [12]. It is used to evaluate the performance of binary classification models by measuring the overall quality of the model's predictions across different classification thresholds. A curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) as the classification threshold varies. The area under this curve, called the Receiver Operating Characteristic (ROC) curve, represents the AUC score, which ranges from 0 to 1. An AUC score of 1 indicates a perfect classifier that achieves perfect discrimination between the positive and negative classes. A classifier with an AUC score of 0.5 performs no better than random guessing, indicating that its predictions are essentially random. Although it might seem logical to evaluate the model's performance against a random guesser, such a comparison would not be fair due to the underlying structure of the labels, which can lead to highly promising results. In this case, the labels consist of binary strings that begin with zeros and end with ones. The only variation among the labels is the position of the first occurrence of "1." Consequently, the labels exhibit significant similarity to one another, and there is a high likelihood of accurately identifying segments containing true positives (as the start and end positions of the labels are almost always the same). To ensure a fair comparison, the model should be evaluated against a model that takes this underlying structure into account. One approach is to assume that a comparable model always predicts the first occurrence of "1" to be located in the middle of the segment, specifically at index 20. All the entries before that index contain only zeros and all the entries after that index only contain ones. The model can then be tested against this baseline model using the same respective test sets (realized/unrealized/combination).

To facilitate the use of the AUC score for the segmentation task, each entry in the prediction window array was treated as an individual prediction. The AUC score was subsequently computed based on the transformed labels and averaged. A similar approach was used by [13] in detecting and inferring segments of laughter on the same dataset. The refactored code used in the experiment is hosted on Github and can be publicly accessed here [14].

The results derived from the trained model and the baseline model are shown on a Box plot for realized, unrealized, and a combination of both. A box plot is a graphical representation of a dataset's summary statistics [15]. It is constructed as follows:

- **The box:** The box in the plot represents the interquartile range (IQR), which encompasses the middle 50% of the data. The lower edge of the box represents the first quartile (Q1), and the upper edge represents the third quartile (Q3). The length of the box therefore depicts the spread of the central data.
- **The line within the box:** This line represents the median, which is the middle value of the dataset when arranged in ascending order.
- **The whiskers:** The whiskers extend from the box and indicate the range of the data. By default, the whiskers extend to 1.5 times the IQR beyond the first and third quartiles. Data points beyond the whiskers are considered outliers and are typically represented as individual points.
- **Outliers:** Data points that fall outside the whiskers are considered outliers and are often shown as individual points beyond the whiskers.

The model used by [1] for classification does not allow for varying window sizes, making it unsuitable for the intended purpose. As this limitation hinders a truly realistic comparison, this research used an alternative model, a modified version of a Recurrent Neural Network (RNN) that allows for the utilization of different window sizes as input, to perform the classification evaluation on both window sizes and extracted segments. The model comprises a first layer of Long Short-Term Memory (LSTM) with a size of 64. Then, a Dense layer of size 1 is added for binary classification, using a sigmoid activation function. Just like the metric utilized for evaluating the classifier developed by [1], the AUC score was also employed in this context as an evaluation metric. The results are also obtained using 5-fold cross validation on the 3 different criteria (realized, unrealized, and both) and visualized using a Box plot for the 4 different window sizes (1, 2, 3, and 4 seconds) and for the supervised segments.

4.2 Results

For the segmentation task, the performance achieved on each of the 5 folds for the 3 test sets are shown on the respective Box and Whisker plots in Figure 2a, 2b, and 2c. Table 1 and 2 display the average scores and standard deviation respectively obtained by the trained and baseline model on the 3 criteria.

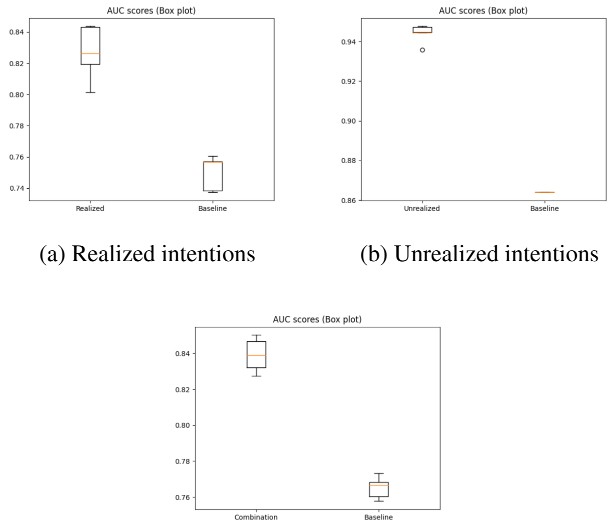
Table 1: Average AUC score and S.D for the trained model

	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

For the classification task, the performance achieved on each of the 5 folds for each respective window size and for the supervised case on the 3 criteria are resepctively shown in Figure 3a, 3b, and 3c. Tables 3, 4, 5, 6, and 7 also display the average score and standard deviation.



(a) Realized intentions

(b) Unrealized intentions

(c) Combination

Figure 2: Box plots displaying the AUC scores of the trained model and the baseline model for the segmentation task

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

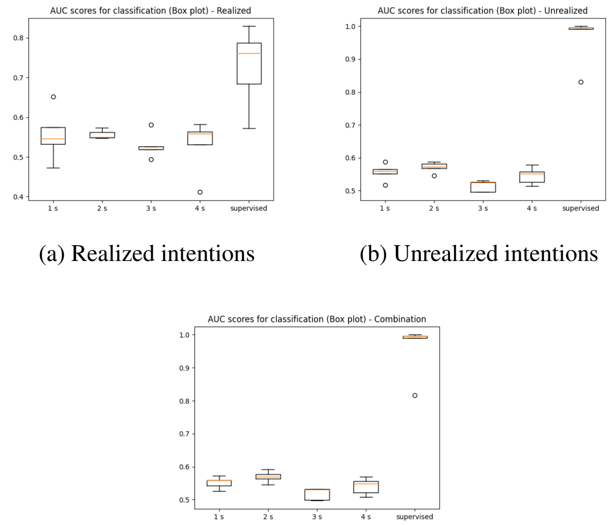
	Realized	Unrealized	Combination
Average	0.528	0.514	0.518
S.D	0.032	0.017	0.018

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



(a) Realized intentions

(b) Unrealized intentions

(c) Combination

Figure 3: Box plots displaying the AUC scores of the model trained and tested on different window size and on a supervised case for the classification task

On the segmentation task, the trained model demonstrates better performance compared to the baseline model across the 3 evaluation criteria, including realized intentions, unrealized intentions, and their combined evaluation. All the results were also depicting a relatively small standard deviation. The observed lower performance in the realized intentions with respect to the unrealized ones could be attributed, at least in part, to the inclusion of negative samples in the test set of the realized intentions. It is plausible that the model displays reduced accuracy when dealing with negative samples, leading to a higher rate of false positives. In such cases, the model mistakenly predicts a speaking intention (assigning a value of 1) when the true label is actually 0. Consequently, this discrepancy contributes to a decrease in the overall AUC score, despite still performing better than the baseline. This finding suggests that the model has a tendency to overestimate the occurrence of speaking intentions. Overall, the results demonstrate the potential of the proposed deep learning model in inferring speaking intentions from body language captured by a body-worn accelerometer, although further analysis and evaluation may be necessary to assess the robustness and generalizability of the model across different datasets and contexts. For example, the model was tested against a single "baseline" model, relying on a simplistic assumption. However, there is room for employing diverse evaluation approaches that can be tested against and potentially surpass the performance of the trained model.

On the classification task, the performance of the supervised model surpassed that of the models trained and evaluated with fixed window sizes in all three evaluation criteria. The results demonstrated a notable consistency with a relatively low standard deviation. These findings highlight the significant benefits of employing a supervised learning process to infer cases of intentions to speak in a dataset. Furthermore, an intriguing observation is that the results obtained for the window sizes align with the findings of [1] regarding unrealized intentions (referred to as “unsuccessful” in their research). This alignment suggests that the 2-second segment demonstrates the most promising outcomes among the four fixed window sizes. However, the alignment does not hold for realized intentions (referred to as “successful” in [1]), where a window size of 1 second yields better results according to their findings, whereas the present research indicates improved results in the 2-second segment.

It is worth mentioning that, both on the classification and the segmentation task, the models were trained on a limited dataset due to the manual annotation process. Also, only one metric (AUC) was used to evaluate the model’s performance, and some different approaches could be more representative for the task at hand, especially for the segmentation task. Lastly, it is crucial to note that the models used in the experiment were not fine-tuned and were not subjected to extensive testing with various sets of hyperparameters and/or layers.

5 Conclusions and Future Work

In summary, the experiment highlighted the significance of body language and nonverbal cues in understanding speaking intentions. It is worth noting that a model solely trained on accelerometer data showed great segmentation capability by effectively identifying both positive and negative instances within the given window size. Although it was inferred that the model performs worse on the negative cases, the AUC scores achieved by the trained model surpassed the ones obtained with the baseline model for all 3 criteria with minimal standard deviation. Furthermore, the supervised learning appears to bring great improvement to the classification task, suggesting that focusing on highly qualitative data to infer instances of speaking intention can be the next crucial step in building more realistic estimators that can be used in a practical setting to improve the quality of human-computer interactions.

Several potential ideas for improvement were identified, including training and testing the model on a larger dataset and incorporating data from diverse cultures and languages. For example, a rule-based approach could be employed to extract the data directly from the Voice Activity Detector (VAD). By developing a multi-modal extractor, it would be possible to extract segments of speaking intentions in significant quantities, without relying on manual annotation. This not only saves time but also overcomes the limitations imposed by the availability of human resources for manual annotation. Furthermore, the annotations on intentions to speak from the dataset were carried out manually and individually, relying on assumptions derived from existing literature on turn-taking and conversational speech. It is important to acknowledge that this process is subjective, and to ensure the accuracy and reliability of the annotations, it is recommended to have multiple individuals performing the annotations. Lastly, fine-tuning the model and assessing its performance using different metrics are also factors worth considering. By exploring these ideas, further advancements can be made in leveraging body language for inferring speaking intentions.

6 Responsible Research

The research utilized data obtained from real individuals, enabling the identification of their faces through camera footage, as well as the ability to listen to their discussions with other participants during the recorded 1.5-hour session. From the audio recordings, it was possible to deduce personal information such as their names, ages, residential addresses, workplace details, and other relevant data. Given the sensitive nature of this information, an End-User License Agreement (EULA) was signed to ensure confidentiality. To maintain data ownership, the collected data was not shared on platforms that do not guarantee ownership rights, opting instead for the use of Surfdrive instead of platforms like WeTransfer. Additionally, while the experiment’s code was hosted on GitHub, the data files remained exclusively on local computers and were never committed to any remote repository. No individual other than the research team that signed the EULA has had direct or indirect access to the data. Visual illustrations of the event are displayed in the appendix. Special effort was made to blur the faces ensuring that no individual can be personally identified in these illustrations. In terms of the practical use of the model, the mapping is from accelerometer data to an estimated speaking intention. The model does not receive any audio, video, or extracted body pose information as input. Consequently, the absence of such trivial information ensures that personal identification of the participants cannot be derived from it. This characteristic makes the model suitable for practical use, as it alleviates any privacy concerns related to personal identification.

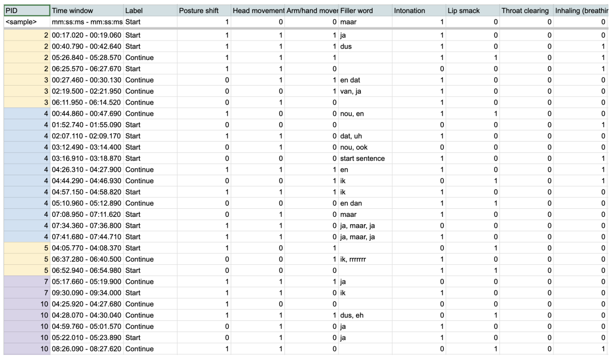
Acknowledgements

I would like to thank Hayley Hung, Jord Molhoek, Litian Li, and Stephanie Tan, for supervising this project and for providing valuable feedback. Next, I would like to thank the fellow students of Group 22, Ferhan Yildiz, Julie van Marken, Waded Oudhuis, and Luning Tang, that worked on a different research question and subtopic for their help during the annotation of the dataset and with whom I exchanged ideas throughout the research.

A Appendix

A.1 Cues Employed in the Annotation Process

A list of predefined cues, along with the absence or presence of these cues for each participant in the annotated unrealized segments, was constructed in collaboration with four other teammates. The matrix illustrating this annotation of cues is presented in Figure 4. The same set of cues was employed individually during the annotation process for the realized intentions. By utilizing this collaborative approach, a comprehensive analysis of the cues exhibited in both realized and unrealized intentions was facilitated, enabling a better understanding of the dynamics of speaking intentions. Table 8 displays the percentage of occurrence of each cue in the annotated intentions:



PID	Time window	Label	Posture shift	Head movement	Arm/hand mover	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

Figure 4: Screenshot of the annotated spreadsheet matrix showcasing cues and their presence or absence for each unrealized intention

Table 8: Occurrence of each cue in the annotations (as a percentage)

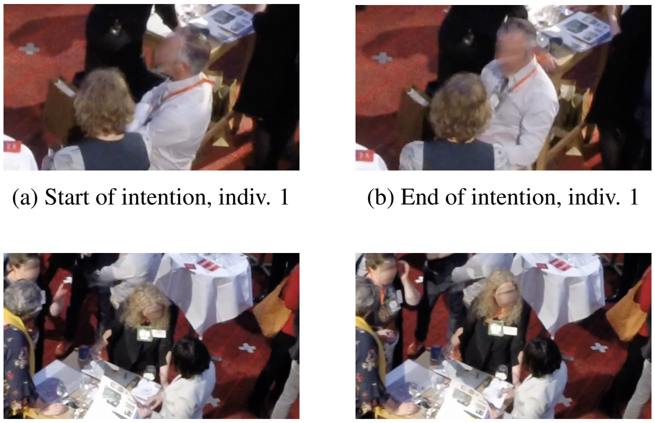
	%
Posture shift	57
Head movement	77
Arm/hand movement	51
Filler word(s)	77
Intonation	66
Lip smack	22
Throat clearing	1.8
Inhaling	2.3

Among the identified cues, the most frequently occurring ones, in descending order, are head movements, the use of filler words, specific intonations, posture shifts, and arm and/or hand movements. On the other hand, throat clearing and inhaling were observed less frequently. It is crucial to acknowledge that these findings are derived from the analysis of a limited sample size of only 52 segments. To ensure greater realism, it is essential to incorporate a larger volume of data in future analyses.

A.2 Annotations: Understanding with Examples

2 participants were selected at random to illustrate the procedure followed during the annotation phase. Figures 5a, 5b and 5c, 5d, respectively show for these 2 participants the change in body posture and head movement between the start of the detected intention to the moment the individuals start speaking. The change in body posture and head movement are evident behaviours occurring during an intention to speak. Participants intending to speak shift their position, redirecting their gaze towards their interlocutor just before initiating speech. These behaviors are consistent with the existing literature on non-verbal cues in conversational speech, with the notable contributions from Bull P. [9] and Poyatos F. [10] as described in the section on related works.

Important cues can also be derived using vocal information. The verbal cues considered when annotating intentions to speak were derived from the literature on verbal cues [5]. These include the presence of lip smacks, illustrated in Figures 6a and 6b, as well as the presence of inhaling, illustrated in Figures 6c and 6d extracted from the voice audio recorder (VAR) of the same two participants. Both are illustrated as a screenshot of the VAR from their respective segments. These captured verbal cues provided insightful information during the annotation process as these cues frequently serve as the initial indicators and thus define the beginning of the intention.



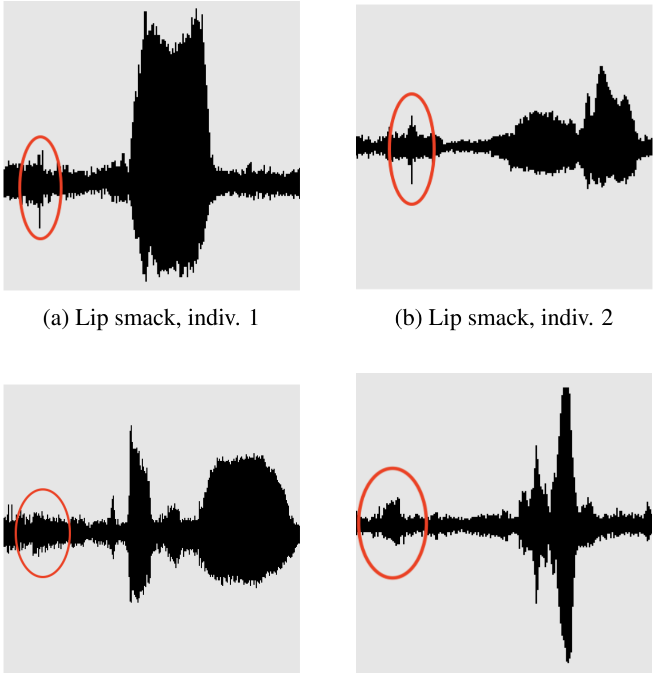
(a) Start of intention, indiv. 1

(b) End of intention, indiv. 1

(c) Start of intention, indiv. 2

(d) End of intention, indiv. 2

Figure 5: Body language at the start and end of an intention to speak for 2 participants



(a) Lip smack, indiv. 1

(b) Lip smack, indiv. 2

(c) Inhaling, indiv. 1

(d) Inhaling, indiv. 2

Figure 6: Lip smacking and inhaling for 2 participants

References

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

[2] Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009.

[3] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In JIM SCHENKEIN, editor, *Studies in the Organization of Conversational Interaction*, pages 7–55. Academic Press, 1978.

[4] Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech Language*, 67:101178, 2021.

[5] Deborah Schaffer. The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics*, 11(3):243–257, 1983.

[6] josedvq. Lared dataset. https://github.com/josedvq/lared_dataset.

[7] Max Planck Institute for Psycholinguistics, The Language Archive. ELAN (Version 6.5). Computer software, 2023. Retrieved from https://archive.mpi.nl/tla/elan.

[8] Lie Lu, Hao Jiang, and HongJiang Zhang. A robust audio classification and segmentation method. In *Proceedings of the Ninth ACM International Conference on Multimedia*, MULTIMEDIA '01, page 203–211, New York, NY, USA, 2001. Association for Computing Machinery.

[9] P. Bull. *Body Movement and Interpersonal Communication*. Wiley, 1983.

[10] F. Poyatos. *Nonverbal Communication Across Disciplines: Culture, sensory interaction, speech, conversation*. Nonverbal Communication Across Disciplines. J. Benjamins Publishing Company, 2002.

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[12] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[13] Jose Vargas-Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung. Impact of annotation modality on label quality and model performance in the automatic assessment of laughter in-the-wild, 2023.

[14] Refactored model for inferring segments of intentions to speak. Accessible onhttps://github.com/Nilsachy/ResearchProject.

[15] Ronald Boddy and Graham L. Smith. *Statistical Methods in Practice: For Scientists and Technologists*. John Wiley & Sons, 2009.