# Estimating intentions to speak using Lexical information

## Leveraging Lexical Information to Facilitate Social Interactions with Artificial Agents

**Ferhan Yildiz**[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: Ferhan Yildiz
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Amira Elnouty, Litian Li, Jord Mohoek, Stephanie Tan

## Abstract

This research paper implements, evaluates, and compares two approaches, a machine learning (ML) approach and a rule-based approach, aimed to estimate intentions to speak. The ML approach trains lexical information extracted from time windows surrounding speech events. The rule-based approach looks for specific keywords or utterances to identify intentions to speak. The results show that the ML approach is a more favourable solution to the problem due to its adaptability and potential for improvement. Sample generation and parameter tweaking showed to be vital to the performance of the model, with its best performance being when it predicted unsuccessful intentions to continue speaking. This study concludes that a machine learning approach can be a viable solution for estimating intentions to continue speaking, with there being future use cases in conversational systems and human-computer interactions.

# 1 Introduction

Artificial intelligence has seen a great surge in interest and development since the start of the twenty-first century. With this increase in development, new use cases for this technology have surfaced, some of them placing artificially intelligent systems in social settings [1]. These social artificial agents owe their usefulness to their capability of interacting with humans. To interact with humans in a social sense, the artificial agent will need a measure of when the human intends to talk so that the agent knows to give the human the room to speak. Interruptions or overlaps can cause frustration and confusion in social situations [2], which makes this an important task for the agent. Giving the agents this ability would allow for them to be used in social settings, and it can also allow for more specialized use cases such as a system that leads meetings.

To this end, Li et al. worked on a paper where they estimate intentions to speak using accelerometer data in-the-wild [3]. Here they attempt to use existing methods for next speaker prediction to infer intentions to speak-in-the-wild. However, in their report they do not extensively research whether certain audio cues, mainly lexical information and back channels, have a strong correlation to this estimate. This report will mainly be focused on this concept. In particular, by seeing if it would be more beneficial to look at the lexical information of someone speaking to estimate their intention to continue speaking, or if it would be better to strictly look at a more rule-based approach such as trying to pay attention to any audio cues someone might give that intends to talk. For example, Petukhova and Brunt [4] mention cues such as someone saying ”Hmm”, ”Umm” or making a tongue clicking sound before they start talking, in their paper about turn-taking.

To fill the knowledge gap in this area of research, this paper focuses on the topic of estimating intentions to speak using lexical information. Starting with an explanation of the methodology used in this paper in Section 2, followed by a recounting of the experimental setups and results in Section 3. After that, the ethical topics surrounding the research are discussed in Section 6. Then the paper is finalized with a discussion based on the results and the conclusions and future work in Sections 5 and 4.

## 1.1 Related Work

Although the field of research surrounding the estimation of speaking is a relatively young one, a lot of research has already been performed in fields that are related to this research paper. For example, a related topic called turn-taking, which is the process by which speakers in a conversation alternate and manage who speaks, has already had a lot of research put into it, with some turn-taking papers even focusing on lexical content [5] [4].

Lia et al. proposed a Recurrent Neural Network (RNN) based model which takes the joint embedding of lexical and prosodic contents as its input to classify utterances into turn-taking related classes [6]. It even goes so far as to try to time the turn-taking. To estimate if an utterance is an attempt to take the turn to speak, Lia et al. trained their RNN on word embeddings of words spoken during conversations in their dataset. They used these embeddings alongside prosodic features to train their model, which gave them an accurate estimate of if an utterance was an attempt to start speaking. Although their research is extremely related to what is being attempted in this research paper, they do not focus explicitly on lexical contents. Alongside that, they use a database of Japanese conversations, which might return different results than if a model were to be trained on Dutch conversations, as will be done in this research.

Li et al. investigate the viability of a model trained only on accelerometer data that attempts to estimate both successful and unsuccessful intentions to speak better than random guessing [3]. To approach this problem, they simplify it into a classification problem, where someone either has the intention of speaking or they don’t. This might not perfectly represent how it works in reality, since intention to speak might work more like a scale, where for example someone slightly has the intention to speak. Nonetheless, it is a vital simplification as it allowed for the problem of estimating intentions to speak to be turned into a classification problem. This report will also make use of the same simplification.

Petukhova and Brun [4] have performed research in a very similar area to this report. In their paper they examine verbal and nonverbal cues that participants use to signal their intention to take the turn to speak in multi party dialogue. During the examination they observed that lexical information, such as words or specific phrases, was a valid identifier to spot turn-taking behavior. They mention that lexical information could reliably signal if someone wants to take the turn speaking. They do also mention that other cues such as gaze direction and posture shifts can also be reliable signals for intentions to take the turn.

## 1.2 Research Question

In order to research the validity and potential effectiveness of the task of using lexical information to estimate intention to speak, the following research questions were procured:

- Is a machine learning (ML) based approach viable to estimate intention to speak?
- Is a rule-based approach viable to estimate intention to speak?
- When estimating intention to speak, is it be better to take an ML approach based on lexical information, or a rule based approach which looks for specific utterances?

# 2 Methodology

To answer the question of the viability of each method mentioned in the research question, an implementation has to be created for each of them. Namely, a program that, given a time frame of a conversational participant, can classify if this time frame has an intention to speak or does not.

Such a program can be created by either training a model given training data, or using a rule-based approach. In either case, these implementations have to make use of a data set, which will also have to be annotated and processed for the trained model implementation. This will be discussed in 2.1. Furthermore, the actual implementation of the model and rule based approach will be discussed in 2.2 and 2.3 respectively. Finally, this section discusses the evaluation metric that will be used in section 2.4. A flow chart of the machine learning model can be found in figure 1.

## 2.1 Data

The chosen data, as well as the annotation and processing of it, is a crucial part of this research, as it can dictate what the results look like. In the following subsections, the specifics of these elements are discussed as well as the motivations behind the choices.

**Dataset**

For the dataset of this research, mainly two datasets were discussed, namely the REWIND dataset[1] and the MULAI dataset [7]. The REWIND dataset consists of the data collected from a social networking event, where participants were invited into an indoor space to freely roam and talk to other participants while standing, with the language spoken being Dutch. Some participants were given microphones which recorded their speech. This speech can be used to extract lexical information, which can be used to train the model. Therefore this seemed like a good candidate for a dataset. The Multimodal Database of Laughter, MULAI, consists of data captured by dyadic human-human interactions which include both spontaneous and task-induced laughter. Here all of the audio is recorded and the laughter is annotated, and the language spoken was English. For the purposes of this research the REWIND dataset seemed like a better fit since it has a better emphasis on natural in-the-wild conversational flow, which is ideally what the model should be trained on.

**Annotation**

To eventually test the data in an accurate way, a subset of the data was taken and annotated based on unrealized intentions to speak. There was the option to take the annotations made for the master student's project [3] which were annotated by one person, or assign 5 student researchers to annotate the same data as well as a larger piece of the data. The chosen route was to assign 5 student researchers. Mainly because this would allow more data to be annotated, and the fact that there being multiple annotators would be a good way to account for biases during annotating, which is discussed more in Section 6.3.

Similar to the annotation process in the paper created by Li et al. [3], the unrealized intentions to speak were subdivided into two categories, intentions to start speaking, and intentions to continue speaking. Unrealized intentions to start speaking were annotated when the annotator noticed that a participant had the intention to start speaking but was unsuccessful in speaking. Intentions to continue speaking were annotated similarly, with the difference being that this label was used when a participant already had the word while such an intention was picked up.

The unrealized intentions were spotted by annotators based on a specific rule-set:

- A cue is considered an unrealized intention to speak when it happens more than 2 seconds before the person is able to get the word.
- A cue is considered an unrealized intention to speak when there is no other intention to speak before actual speaking.
- A cue is considered an intention to continue speaking if it happens within 2 seconds of the person stopping their last sentence.
- A cue is considered an intention to start speaking if it happens more than 2 seconds after the person has stopped talking.

However, a lot of it came down to instinctively picking up these unrealized intentions based on cues from the participants, such as shifts in body pose, or mutters. To ensure the validity of the annotations, all of the annotated data has been annotated by at least two student researchers who, after annotating, discussed their annotations with each other to ensure that they had annotated similarly. The annotation was done using the ELAN software [8]. In total, the annotation process was applied for 13 participants for a 10 minute time window, and a total of 52 unrealized intentions to speak were recovered of which 32 were intentions to start speaking and 20 were intentions to continue speaking, with 77.3 percent of all annotated intentions being preceded by a filler word.

**Processing Data**

To use the data in a meaningful way such as to train and test the model, it must first be processed by extracting and labeling relevant parts of the data (the Data Extraction step in figure 1). To train the model, instances where there are realized intentions to speak and instances where there is no intention to speak are required. These instances are acquired through the generation of samples. This process is expanded upon

[1] Unpublished dataset by Jose Vargas-Quiros, Hayley Hung, and Laura Cabrera-Quiros

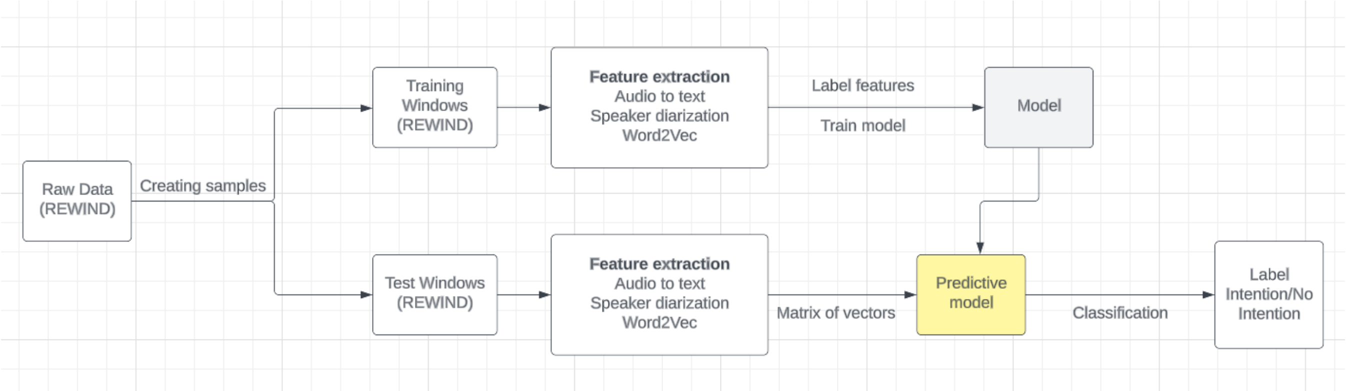

Figure 1: Flowchart of the machine learning process

in Section 3.1. The realized intentions to speak are extracted from the audio file by taking the frame of time before a participant starts speaking, since a participant is bound to have the intention to speak before they start speaking. The instances where there are no intentions to speak are extracted by looking at when a participant's voice activates for a small amount of time. Here lies a simplifying assumption that these utterances of small amounts of time are back channels and not a participant attempting to have the word. Throughout the annotating process, the annotators agreed that these short utterances seemed to be backchannels in the vast majority of their use cases. Because of this, the simplification most likely has a negligible effect on the output of the model. The exact rules for the sample generation are expanded on in section 3.1.

Furthermore once the time frames are collected, the lexical contents must be extracted. This is done by first transcribing the full audio files using Automatic Speech Recognition model Whisper [9]. After the audio is transcribed, the time frames are matched to the transcription to see which words were said in a given time frame. After that, the words are embedded using the library Word2Vec, which gives a vector for each word, with words with similar meanings being given similar vectors [10] [11]. Since the dataset for this research paper is in Dutch, this project uses Word2Vec vectors trained on Dutch corpora [12]. All of these vectors are then stored in a large file which is used to train the model. Word2Vec offers a representation of the similarity between words with regard to their semantic and syntactic relationship to each other. With the nature of this project in mind, it is important to consider the relationship between words to estimate intentions to speak, which is why Word2Vec was chosen as opposed to other techniques such as Bag-Of-Words or TF-IDF [13], which don't fit as well due to their lack of contextual understanding.

## 2.2 Machine Learning Approach

To investigate the research question regarding the ML model, and the validity of this approach, a model must first be created. The master student project mentioned earlier in the paper refactored a previously designed model [14]. The model is a residual neural network model which consists of three convolution layers with respective kernel sizes 3, 5, and 7. This model was intended to predict the speaking status of a given person, which is not exactly what the master students were intending to predict, so they refactored it [15]. The main difference here is that the refactored model aims to predict intentions to speak rather than speaking status. Also, the refactored model only uses data that is X seconds before the act of speaking, with X being a parameter to be changed to find the optimal value for performance. The other large difference being that the previously designed model uses three modalities fused into one input: audio, video, and accelerometer data, whereas the master students' work only uses the accelerometer data.

To use lexical information in training rather than accelerometer data, another refactored version of the code was created [16]. This refactored version uses the vectors extracted by the use of Word2Vec on the transcribed audio files from the data set as explained in 2.1. The details of how these elements all fit together are discussed in Section 3, and more details of how the model work can be found in Section 3.1.

## 2.3 Rule-Based Approach

The rule based approach consists of gathering a dictionary of words that are highly likely to be used by a conversational participant that has an intention to speak. By gathering these words, a program can be written to specifically search for these words in given time windows. When the program finds one of these words in the time window, it labels the window as an intention to speak.

## 2.4 Evaluation Metric

To assess the performance of the implementations, the AUC is used, or Area Under the Receiver Operating Characteristic(ROC) curve. This is a widely used metric for the evaluation of classification models. The ROC curve plots the true-positive rate on the y-axis and the false-positive rate on the x-axis, so the curve encapsulates how the two rates fare against each other. Since both axes range from 0 to 1, the AUC will also range from 0 to 1. Furthermore, a perfect classifier would show in the form of a vertical line going from (0,0) to (0,1), and then a horizontal line from (0,1) to (1,1). A classifier that randomly guesses would then have an AUC score of 0.5 with the line being diagonal from (0,0) to (1,1). So the performance of our implementations can be measured based on the AUC.

# 3 Experimental Setup and Results

This section aims to provide details of the implementations of the machine learning model and the rule-based approach as well as provide the results achieved by them. First, the model implementation is discussed in Section 3.1, followed by the implementation of the rule-based approach in Section 3.2, finished with a look at the results in Section 3.3.

## 3.1 Model implementation

As mentioned in 2.2, for the purposes of this research a refactored version of a previously implemented model is be used [15]. The model is a residual neural network model. It consists of three convolution layers with respective kernel sizes: 3, 5, and 7. This section aims to expand on the workings of the model, mainly focusing on the preprocessing, training, and testing.

### Generating Samples

The main function of the preprocessing step is to generate samples for the model to train and test on. For the model to be trained properly, both positive and negative samples must be generated. Therefore the sample generation is done by creating time windows for instances of successful intentions to speak and instances where there are no intentions to speak. As discussed in 2.1, Li et al. [3] make use of the simplifying assumptions that there is an intention to speak X seconds before a person starts speaking, and that utterances that last less than Y seconds are not intentions to speak. These assumptions are used to generate samples through code rather than having annotators go through all of the data, which would be an incredibly expensive process.

Li et al. [3] used these assumptions to create positive successful intention samples by taking X amount of seconds before someone starts speaking. Speaking detection is done mainly by using processed Voice Activation Detection (VAD) files. Negative samples were generated by taking windows of time where there was no speaking according to the VAD files.

In the context of training the model using lexical information, the before mentioned approach for sample generation is suboptimal. This is because the time window before someone starts speaking in the general case does not have any lexical information to train on. To leverage lexical information fully, a different sample generation rule set was created. Namely, instead of taking the window of time X seconds before a participant starts talking, with the new approach the window is taken beginning from the start of speech and lasting until X seconds after the start of speech. There is also the condition that this speaking turn lasts more than Y seconds. The Y seconds condition is there to only have valid speaking turns as successful samples, and backchannels as negative samples. Negative samples in this case are utterances that last less than Y seconds. Examples of both the positive and negative sample generation can be seen in Figure 2. It's important to note that this kind of approach would train a model to predict intentions to continue speaking rather than intentions to start speaking.

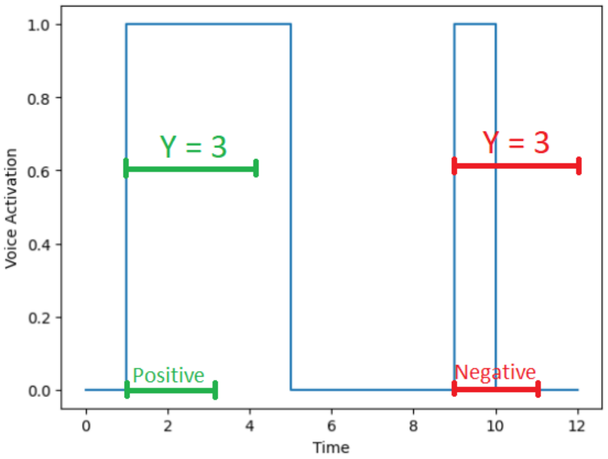

Figure 2: Sample generation illustration. The left utterance shows a positive sample, for which the utterance lasted more than Y = 3 seconds. The right side shows a negative sample.

For this research project, the time window generation rules of Li et al. are tested as well as the rules specifically crafted with lexical analysis in mind. After the time windows are generated, examples are created by attaching the relevant information to these time windows. In the case of this research paper, the attached information is the participant id, start-time, end-time, and the feature vector of the window.

### Training

After the preprocessing phase has finished the generated examples are used in the training phase. During this phase, a model is generated which aims to identify patterns that indicate intentions to speak. Training consists of this model using the examples to generate an output, which is then compared with the actual labels to compute the loss. This loss is generated using binary cross-entropy loss, which is a suitable function due to the binary nature of this problem.

The model parameters are then tweaked automatically with the aim of minimizing this loss. This loss calculation and parameter tweaking is then repeated over multiple iterations until the performance reaches a plateau and stops improving.

### Testing

During the testing phase, the predictive model's labeling capabilities are evaluated. The model is presented with training data which was separated from the dataset before any of the training happened. This was done so that the model is not trained and tested on the same data, since doing so would cause a bias in the results. The examples are however generated using the same method as is explained in the preprocessing subsection.

During the evaluation, the model tries to label the time windows correctly given the parameters that it has after the training phase is finished. The output labels are compared to the actual labels to evaluate the performance. The result is outputted using the AUC metric referenced in section 2.4. This ensures an evaluation which indicates how well the model would perform on new in-the-wild data.

## 3.2 Rule-Based Approach Implementation

The rule-based approach, as an alternative to the machine learning-based model, is a more simplistic method of tackling the task of estimating intentions to speak. This approach consists of using predefined rules to estimate if someone wishes to have the word. The goal is to make such estimations based on the usage of certain keywords.

### Word selection

The first step in this approach is to gather a dictionary which includes only the words that are believed to be indicative of an intention to speak. Some examples can be found in section 1. This dictionary is crucial, as the entire performance of this approach is dependent on the selection of these words.

The word dictionary was procured by two means. The majority of the words were obtained by including all of the words which appeared to proceed intentions to speak during the annotation process. The exact process is described in Section 2.1, what is important to note here is that when an annotator found that an intention to speak was signaled by the use of a word, this word was noted down and eventually included in the dictionary. In addition to this, a group of peers was also asked for their input on which words they thought should be included in this dictionary. The received suggestions coupled with the words retrieved during the annotation process amounted to the final dictionary. The finalized list of words can be found in appendix section A.2.

### Logic

Once the dictionary has been formed, the logic of this approach is quite straightforward. Namely, the logic consists of, given a time window much like the one in subsection 3.1, seeing if any of the words in the dictionary are used in this window. If there are such words said within the window, the window is classified as an intention to speak, else the window is labeled as having no intention to speak.

### Testing

Testing of the rule-based approach is done by taking the training data and generating time windows of this data. The program then applies the logic to these windows and labels them accordingly. After the labels are created, they are compared to the actual labels and evaluated using the AUC score. The AUC metric was picked so that it would be easy to compare the accuracy of this approach compared to the machine learning approach.

## 3.3 Results

After running the tests for the different approaches, the results were analyzed. The results as well as the discussion regarding them, will be discussed further in this section.

### Machine Learning Approach

Using the window-creation logic that was used in the paper of Li et al. [3], the following AUC scores were obtained. The 1 and 2-second windows have an AUC score very close to 0.5, which would mean it's on par with a random guessing classifier. The 3 and 4-second time windows have a slightly worse performance of around 0.49. This poor performance across all time windows is most likely caused because of the window-creation logic used here. Taking the window of time before someone starts speaking inherently clashes with the idea of analyzing lexical information, since the window of time before speech will generally not have speech, because if it did it would be included in the speech itself, and the window of time before that would be taken as a sample.

| Time Window (s) | AUC-ROC |
|---|---|
| 1 | 0.503 (0.006) |
| 2 | 0.508 (0.003) |
| 3 | 0.489 (0.003) |
| 4 | 0.487 (0.004) |

Table 1: AUC-ROC scores for different time window sizes performed using three-fold cross-validation for 5 repetitions.

As for the negative samples it might still be the case that a window includes a backchannel that was not identified as a speaking turn, which could explain the deviations that the 3 and 4-second time windows have from the 0.5 AUC score baseline, since a larger time window has a larger probability to capture one of these backchannels and train using it.

Comparing these results to the results of Li et al. [3] shows that their model trained on accelerometer data outperforms the model trained on lexical information when estimating all intentions to speak.

The results obtained by using the window-creation logic that was proposed in section 3.1 with Y = 2 seconds can be found in Table 2. As expected, the AUC score for the predictions regarding intentions to continue speaking scored relatively well, with the exception being when the time window has a length of 1 second. This could be because a 1-second time frame does not offer enough lexical information to draw valid conclusions from regarding intentions to continue speaking.

The results obtained by taking Y = 1 can be found in Table 3. Generally, the scores for the predictions on all intentions to speak are better. One notable change is that at the 2-second time window the unsuccessful intention to continue speaking performs worse. Another notable change is that the 4-second time window suddenly has an incredible 0.7315 AUC when predicting unsuccessful intentions to continue speaking. The full results for all of the other experiments can be found in appendix A.1.

Table 2: AUC-ROC Scores for testing all intentions to speak, and testing unsuccessful intentions to continue speaking with Y = 2 using three-fold cross-validation for 100 repetitions.

| Window | All intentions to speak | Unsuccessful (Continue) |
|---|---|---|
| 1 Second | 0.5083 (0.005) | 0.4981 (0.010) |
| 2 Seconds | 0.5026 (0.007) | 0.5731 (0.013) |
| 3 Seconds | 0.4993 (0.006) | 0.5310 (0.011) |
| 4 Seconds | 0.4997 (0.006) | 0.5274 (0.014) |

The changes made to the sample generation logic seem to have caused a favourable effect for the AUC scores when estimating unsuccessful intentions to continue speaking. In both cases for Y, the 1-second windows seem to be performing close to a random baseline which would suggest that this window size is too brief to capture meaningful information

Table 3: AUC-ROC Scores for testing all intentions to speak, and testing unsuccessful intentions to continue speaking with Y = 1 using three-fold cross-validation for 100 repetitions.

| Window | All intentions to speak | Unsuccessful (Continue) |
|---|---|---|
| 1 second | 0.5056 (0.004) | 0.4965 (0.007) |
| 2 seconds | 0.5120 (0.003) | 0.4584 (0.009) |
| 3 seconds | 0.5079 (0.006) | 0.5435 (0.008) |
| 4 seconds | 0.5312 (0.006) | 0.7315 (0.007) |

regarding intentions to speak using lexical information.

A noteworthy result is the AUC for predicting unsuccessful intentions to continue speaking at a 4-second window with Y = 1. This result might be pointing at the fact that larger windows are more likely to capture important contextual aspects, which would explain the high AUC score. This increase could also be caused by the fact that there is more overlapping speech in a 4 second time window. If a participant is interrupted, it is a given that they had an unrealized intention to continue speaking. This could possibly be captured significantly more in 4 second time windows.

When comparing the AUC scores for unsuccessful intentions to continue speaking with the results of Li et al. [3], it is apparent that the results achieved in this report are much more favourable with all of the time windows except for the 1-second time window heavily outscoring the results of Li et al.

### Rule-Based Approach

The rule-based approach was ran on the entire set of test examples for realized and unrealized cases. This was only done for one iteration because of the deterministic nature of the approach. The AUC score was obtained by comparing the labels predicated by the Rule-Based Approach to the actual labels with the logic explained in Section 2.4. Furthermore, this procedure was only ran for the 1-second time window, since the emphasis of this approach is to focus on the first few words said. Any words said after the first second of someone speaking can therefore be disregarded in this approach.

The AUC score that was obtained by using this approach was 0.4898, which is slightly worse than a random classifier. This poor performance could be attributed to the flawed nature of the proposed approach. It could for example be the case that, although the word ”ja” is classified as an intention to speak due to it being seen during annotation, it also is used when there is no intention to speak. It could be the case that a word is used as an intention to speak in 20 percent of the cases, and that there is no intention to speak in the remaining 80 percent of its uses. Something along this line could explain why the rule-based approach is returning a poor AUC score.

Another reason is the binary nature of the rule-based system. During the annotation it was determined that around 77.36 percent of the annotated intentions to speak included the use of a filler word. This leaves 22.64 percent of intentions to speak with no such filler words. The binary nature of the rule-based system forces the cases in this 22.64 percent to be labeled as having no intentions to speak, which could again drive down the AUC.

## 4 Conclusions

This research aimed to answer the questions introduced at the start of the report, namely:

- Is a machine learning (ML) based approach viable to estimate intention to speak?
- Is a rule based approach viable to estimate intention to speak?
- When estimating intention to speak, is it better to take an ML approach based on lexical information, or a rule based approach which looks for specific utterances?

Running the experiments with the machine learning-based approach halted different results which mainly were influenced by different sample generation logic and different window sizes. With the sampling logic of Li et al., the scores all hovered very closely to an AUC score of 0.5 which seemed to have been caused by the contradicting nature of the window generation and the attempt to analyze lexical information. This led to the model getting AUC scores close to 0.5, which means that it is performing on par with a random baseline.

With the customized sample generation logic specifically designed for the estimation of intentions to continue speaking using lexical analysis, the model performed relatively better. As expected it achieved its highest AUC score when predicting unsuccessful intentions to continue speaking. Particularly with a 4-second time window and taking Y = 1.

As for the rule-based approach, which offered a deterministic method of estimating intentions to speak by solely relying on the usage of words, the performance was poor with the AUC resembling that of a random baseline. For this approach however, it is important to note that the performance is heavily dependent on the selection of the keywords.

In conclusion, the machine learning approach using lexical information has the possibility to be viable in the estimation of intentions to speak, particularly when estimating intentions to continue speaking. It is important to note that the success of this approach is highly dependent on the logic behind the sample-taking as well as the parameters chosen, such as window size.

A well-thought-out rule-based approach has the possibility to be a simple solution to the problem, however, the findings from this report do not give any solid backing to the viability of the approach as-is. The approach is heavily reliant on the selection of keywords. A more extensive and thought-out research should be conducted to find these words and gauge the effectiveness of the approach properly.

## 5 Discussion and Future work

During this research, two approaches were constructed and evaluated based on their ability to estimate intentions to speak. The results provide valuable insight and directions for possible future work. The insights also could lead to developments in real-world applications in human-computer interactions.

### 5.1 Machine Learning Model Discussion

Despite the lacking performance of the machine learning approach with the initial window generation, when custom sample generation logic was applied the adaptability of the model and approach came to light. The model had particularly promising results when predicting unsuccessful intentions to continue speaking within a 4-second time window. This opens up a use case in situations like online meetings, group discussions, and even public debates where it is important to gauge a speaker's intent on continuing speaking. With further studies that perform a more in-depth analysis to identify optimal parameters and sample-taking logic, this approach has the possibility to be used for systems that intend to manage conversations, possibly reducing interruptions and annoyances in conversations.

As for another future direction, to train the model mentioned in this report, the sample generation made use of smoothed-over voice activity detection (VAD) files. It would be extremely valuable to research the optimal type of VAD file to use of this type of research (most likely raw VAD files) since the type of VAD file used for the window generation might have had a tremendous effect on the results for this research. It could for example be the case that some backchannels were filtered out.

### 5.2 Rule-Based Approach Discussion

The rule-based approach ran into its limitations due to the selection of keywords and the essence of the solution. A more extensive look should be given into what would be a more effective set of keywords. This could be done by manually annotating a larger set of data, and noting down all of the words that seem to be used primarily in cases where a person has the intention to speak. This would however require a larger and more diverse dataset, to ensure that the set of words is not honed toward one group of people or one type of social situation.

### 5.3 General Discussion and Conclusion

Moreover, exploring the effect of combining different modalities could be of tremendous use. For example, a model that integrates both lexical as well as acoustic information to predict speaking intentions. Leveraging the strength of the different modalities would improve the performance, robustness, and accuracy of the model. Especially because of the fact that during the annotation process it was determined that all types of cues were noticed for each unrealized intention to speak, with filler words and head movements being noticed 77.3 percent of the time, and intonation and posture shifts being noticed 66.0 and 56.6 percent of the time respectively.

Finally, it is also important to consider the impact that linguistic and cultural differences can have on speaking intentions. Different cultures and languages have different ways of communicating, so it follows that the intentions to speak might manifest themselves in different way across different cultures and languages.

To summarize, while there are definitely improvements to be made and further research to be done, this report provides steps towards understanding and predicting speaking intentions, specifically using lexical information. The insights derived from this report can aid in the development of future conversational systems and ultimately lead to better and smoother human-computer interactions.

## 6 Responsible Research

The integrity and reliability of this research project hinge on the fact that its methodologies are reproducible and ethical. While such a section is often omitted from scientific papers, it is still important to recognize the potential ethical implications and shortcomings of this project, as well as the possible issues with the reproducibility of the research. This section aims to reflect on these matters critically in relation to the research, addressing reproducibility concerns as well as ethical concerns. The intention is to provide an account of the measures taken to uphold responsible research and some of the shortcomings in this regard.

### 6.1 Unpublished Dataset

The dataset that is used in this research project, REWIND, is at the time of this research project, a yet to be published dataset. What follows is that the experiment done in this project can not be verified by just anyone, rather only the people that are able to get access to this dataset, since if you do not have access to it, there is no way to verify that the data and annotations that are made are valid. However, the dataset paper is attached to reputable names in the field, and it can be accessed by contacting any of the individuals involved.

With that being said, every researcher that had access to the dataset had to sign an End User License Agreement, agreeing that they would make their best efforts to keep the data private, and not use anything they come across in the dataset to jeopardize the privacy of the participants.

### 6.2 Dataset Biases

The REWIND dataset, as explained before in Section 2.1, is data extracted from the conversations between people participating in a social networking event in the Netherlands. This means that the main spoken language is Dutch. Furthermore, even though the participants varied in gender and ethnicity, there was a lack of diversity in age. The majority of the participants appeared to be middle aged with close to no participants being young adults (less than 30 years old). This lack of age diversity could lead to a bias in our machine learning model because of the fact that some age groups tend to use different words and sentence structures than other age groups. Other than that it might also be the case that older age groups give off different lexical cues when they wish to speak, due to their more seasoned way of speaking. This bias can unfortunately not be compensated for since it would require adding a younger group of participants to the networking event retroactively, which is not possible. It is however assumed that the model will perform with similar accuracy across all age groups, since language in its essence is used very similarly between all ages, despite some slight differences as mentioned before in this paragraph.

### 6.3 Fair Annotating

The annotation process is a process that is very susceptible to bias, especially in the case of our research project. Deciding if someone has an unrealized intention to speak based on any subliminal or non-subliminal cues that they give off is something that can become extremely subjective. In some instances a person saying ”Yes...” can be an unrealized intention to speak, and in some instances it can just be a back channel. It ends up being a judgement call of the annotator which can lead to inconsistencies and wrong annotations. To combat this subjectivity, any data that was annotated for the purposes of this project was done in pairs. Every time a part of the data had to be annotated, two student researchers would annotate the same piece of data and meet with each other after the individual annotation to confirm that they reached the same conclusions. In case of differences in the members' annotations, they discussed to reach an agreement.

This process was intended to combat the subjectiveness of the annotation, but of course did not completely remove it as there were still some cases where two different annotators wouldn't agree, mainly in cases where one annotator thought that a case was a valid intention to start/continue speaking, and the other one did not. These cases were primarily caused due to ambiguity in the speaker's social cues. For the sake of time management in these cases they would settle on one either way without being fully convinced. With more time and resources at hand, it would have been more useful to bring in another annotator to decide.

## Acknowledgements

This project was created with the extremely helpful supervision of Hayley Hung, Jord Molhoek, Litian Li, and Stephanie Tan, who offered their guidance and valuable feedback for this project. Also, the members of Group 22 who despite focusing on different research questions and modalities were of extreme help because of their willingness to exchange information and tips as well as their cooperation in annotation.

## References

[1] Alessio Antonini and Lucia Lupi. Social AI for engaging UbiComp. In *Proceedings of the Halfway to the Future Symposium 2019*. ACM, November 2019.

[2] E. A. Schegloff. Accounts of conduct in interaction: Interruption, overlap, and turn-taking. In J. H. Turner, editor, *Handbook of Sociological Theory*, pages 287–321. Springer US, 2001.

[3] L. Li, J. Molbroek, and J Zhou. Inferring intentions to speak using accelerometer data in-the-wild. TU Delft, 2021.

[4] V. Petukhova and H. Bunt. Who's next? speaker-selection mechanisms in multiparty dialogue. In *Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers, SEMDIAL*, 2009.

[5] Stephen C. Levinson. Turn-taking in human communication – origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14, January 2016.

[6] Chaoran Liu, Carlos Ishi, and Hiroshi Ishiguro. Turn-taking estimation model based on joint embedding of lexical and prosodic contents. In *Interspeech 2017*. ISCA, August 2017.

[7] D. Nazareth D. Heylen M.-P. Jansen, K. Truong. Introduction mulai: A multimodal database of laughter during dyadic interactions. *European Language Resources Association*, 2020.

[8] Max Planck Institute for Psycholinguistics. ELAN (Version 5.8), 2023. Accessed: 2023-05-26.

[9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[10] KENNETH WARD CHURCH. Word2vec. *Natural Language Engineering*, 23(1):155–162, December 2016.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[12] Stephan Tulkens, Chris Emmery, and Walter Daelemans. Evaluating unsupervised dutch word embeddings as a linguistic resource. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

[13] Wojciech Zamojski, Jacek Mazurkiewicz, Jarosław Sugier, Tomasz Walkowiak, and Janusz Kacprzyk, editors. *Contemporary Complex Systems and Their Dependability*. Springer International Publishing, 2019.

[14] josedvq. Lared dataset. https://github.com/josedvq/lared_dataset, 2023.

[15] Litian Li et al. Refactored model for inferring intentions to speak. https://github.com/llt-warlock/testProject, 2023.

[16] ferhany2002. Lexical information speaking estimation model. https://github.com/ferhany2002/UniversityRepo, 2023.

# A Appendix

## A.1 AUC-Scores

Table 4: AUC-ROC Scores for Experiments with speaking turns classified as larger than 1 second

| Window | All intentions to speak | Successful | Unsuccessful | Unsuccessful (Start) | Unsuccessful (Continue) |
|---|---|---|---|---|---|
| 1 second | 0.5056 (0.004) | 0.4988 (0.003) | 0.4787 (0.004) | 0.4704 (0.007) | 0.4965 (0.007) |
| 2 seconds | 0.5120 (0.003) | 0.5232 (0.005) | 0.4753 (0.005) | 0.4896 (0.007) | 0.4584 (0.009) |
| 3 seconds | 0.5079 (0.006) | 0.5256 (0.005) | 0.5079 (0.006) | 0.4898 (0.008) | 0.5435 (0.008) |
| 4 seconds | 0.5312 (0.006) | 0.5153 (0.008) | 0.6543 (0.006) | 0.6103 (0.006) | 0.7315 (0.007) |

Table 5: AUC-ROC Scores for Experiments with speaking turns classified as larger than 2 seconds

| Window | All intentions to speak | Successful | Unsuccessful | Unsuccessful (Start) | Unsuccessful (Continue) |
|---|---|---|---|---|---|
| 1 Second | 0.5083 (0.005) | 0.5204 (0.005) | 0.4868 (0.007) | 0.4785 (0.008) | 0.4981 (0.010) |
| 2 Seconds | 0.5026 (0.007) | 0.4941 (0.007) | 0.5318 (0.008) | 0.4971 (0.012) | 0.5731 (0.013) |
| 3 Seconds | 0.4993 (0.006) | 0.5372 (0.007) | 0.4822 (0.005) | 0.4430 (0.010) | 0.5310 (0.011) |
| 4 Seconds | 0.4997 (0.006) | 0.5133 (0.007) | 0.4786 (0.008) | 0.4345 (0.012) | 0.5274 (0.014) |

## A.2 Selected Words for Rule-based Approach

| Word | Obtained by |
|---|---|
| maar | Annotation |
| van | Annotation |
| nou | Annotation |
| ja | Annotation |
| en | Annotation |
| uh | Annotation |
| eh | Annotation |
| ook | Annotation |
| ik | Annotation |
| uhm | Peers and Annotation |
| nee | Peers and Annotation |
| juist | Peers and Annotation |
| ehm | Peers |
| hmm | Peers |
| dus | Peers |
| alhoewel | Peers |
| kijk | Peers |
| toch | Peers |

Table 6: Selected words for rule-based approach.