To What Extent Does Speech Behavior Signal Social Contingency? Senior Thesis

Maciej Zielonka, Rebecca Ramnauth,
Brian Scassellati

May 2021

Abstract

The goal of this paper is to investigate whether low-level features in an individual's speech, as well as the structural composition of an interaction between two social agents can signal the presence or lack of Social Contingency in speech. In this paper, **Social Contingency** is defined as one's sensitivity and responsiveness to another social agent's speech and behavior. Past research has found that features extracted from individual spoken utterances can be used to create models of engagement with up to 90% accuracy. This paper builds off of this past work in two ways: first, we investigate which low-level features play the largest roles in engagement modeling; next, we use an engagement model to measure how often individuals appear to be engaged in a conversation, and whether coupling this measurement with silence between utterances can give any insight towards the concept of social contingency. In this paper, we present a Random Forest Classifier that can classify engagement from a spoken utterance with 83% accuracy by extracting a combination of low-level audio features. Using an Analysis of Variance (ANOVA) measure, we determined that of these features, the top ten highly-contributing features fell under the categories of Mel-frequency Cepstral Coefficients (MFCCs), Chroma values, Root Mean Squared (RMS) energy, and tonal centroid features. The Random Forest Classifier was then used to measure how often individuals were engaged or disengaged during conversations. These measurements were visualized against the average duration of each silent period, the increase or decrease of silence over the course of a conversation, and the confidence of the model. The results suggest that, although further investigation is needed, these simple structural features together can signal whether an individual is socially contingent or not.

1 Introduction

1.1 Social Contingency

Imagine reading a book on a park bench. You hear two people walking by, engaged in conversation. Maybe you do not even understand the language they are speaking in, but based on the intonation and the frequency of either individual's responses, you can tell that they are both actively participating in the interaction. Soon after, a second pair of individuals walks by. This time however, one of the participants speaks slower, or more quietly, or perhaps they do not even respond at all. It quickly becomes clear to you that this individual is not fully invested in the interaction. You look up, and sure enough, they are looking at their phone, not fully attentive to the person next to them.

The difference between these two interactions is that in the first, both participants are socially contingent, while in the second, the texting individual is not. **Social Contingency** is defined as one's sensitivity and responsiveness to another social agent's speech and behavior. Of course, visual cues provide the bulk of information when determining an individual's responsiveness. Looking up from your book confirmed your suspicions that the texter was not paying attention. But clearly, thanks to past experiences with social situations, most humans can use the auditory signals themselves to make the distinction between interactions that are socially contingent and those that are not.

The question then becomes, what are those auditory signals, to what extent does speech behavior signal social contingency? And might we be able to extract high and low-level features of a conversation to create a model that would aid automated systems make their own classifications of the social contingency in human-robot interactions?

1.2 Purpose

The recognition of social contingency is a social skill that is one that we take for granted, but one that is also incredibly valuable. If while trying to communicate something to another individual, we see that they are not attentive and receiving the information, we might call to them to bring their attention back, or we might wait until they bring their attention back to the conversation themselves. An understanding of social contingency is especially important in settings involving children. In a classroom for example, the instructor must be able to recognize when a student is not actively participating in class or groupwork, or the student might not absorb the necessary knowledge or skills to succeed in the class. Having recognized this lack of engagement, the instructor will adjust their approach accordingly; perhaps they will call on the student more often, or recall a past strategy that succeeded in maintaining the student's attention.

The ability to automate the recognition of social contingency could prove very useful in the field of socially assistive robotics (SAR). SAR is a field of robotics which aims to support individuals in the development of social, lan-

guage, or academic skills. The Social Robotics Lab at Yale focuses on the study of SAR by building models of human social behavior, especially the development of early social skills [1]. A lot of their work involves developing robots that help children, particularly those with Autism Spectrum Disorder (ASD), improve their social skills such as eye contact, social and emotional understanding, and perspective-taking [6]. Many of their robots benefit from making the same adjustments as the teacher described above; when the child interacting with them is clearly distracted, they attempt to recall the participant's attention.

Many robots can already use visual cues to create automated classifications of engagement and disengagement [5]. Using computer vision, robots can distinguish between the presence and lack of eye contact, facial expressions, and make accurate predictions of a participant's social contingency. However, automating the extraction of audio features and using those to aid in the classification of social contingency could have useful applications. Audio tools could be factored into the robot's decision-making when working to maintain engagement over the course of an interaction. An audio classifier could also improve the robot's functionality, as it could make classifications even when a participant is out of the robot's view. This would prove especially helpful in extending the robot's utility to group settings such as classrooms. A robot would not need to have a camera pointed at every individual, but could instead rely on the audio classifier to keep track of which participants are socially contingent, and which require more attention from the robot.

2 Background

The problem of classifying human speech in conversation is not a new one. Researchers have already built models to classify engagement of utterances of individuals. In [4], the F0 parameter, and 12 Mel-frequency cepstral coefficients (MFCCs), and voice quality parameters to achieve a random forest classifier that is 88% accurate. In [8], pitch, energy, and formant frequencies were passed into a model that combined a support vector machine (SVM) and a hidden Markov model (HMM) to create a classifier that is 63% accurate.

A similar problem to recognition of engagement is that of audio emotion recognition. The combination of MFCCs, pitch, and energy has proven to yield a model that can classify emotions with 95% accuracy [7]. Similarly, [3] discusses the use of MFCCs, pitch, energy, as well as other speech parameters when classifying emotions. The classifications of engagement and emotions are unsurprisingly related.

Classification of engagement by itself has limitations, however. Certain one-word responses may not have enough variability to be classified as engaged or disengaged. Similarly, moments of disengagement or distraction may arise when an individual is thinking of what to say, or how to respond. An individual might be entirely engrossed in the conversation, and classifying a single utterance might not paint a big enough picture to recognize that. This is how social contingency differs from engagement. The purpose of this paper is to see

whether the structural features of a conversation surrounding individual utterances paint that picture by asking if we can take the problem of engagement one step further to better our understanding of social contingency.

3 Data

3.1 Audio Data

The data used for this project were taken from the study done for [5]. In this study, children were brought into a room (either individually or in groups of three), with two socially assistive robots. The robots would then speak to each other, enacting a scene, pausing at pre-designated points to allow the participants to choose the robots' next course of action. The audio and visual data was then used to classify models of engagement and disengagement for the participants. This dataset was initially chosen because it focused on modeling human engagement in individual and group settings. The models focused primarily on visual cues, but incorporated speech as well. The goal of this paper was to see how well audio signals on their own supported the results of the study. However, upon scouring the dataset, these human-robot interactions contained almost no speech from the participants. The children sat in place, pressing buttons in order to choose the next step in the story, and only the robots spoke. Therefore, this dataset from the study did not contain enough participant speech to use for the purposes of this paper.

Instead, the study's followup interviews were used. These were discussions done in a separate room, away from the robots, between an adult member of the lab and each of the individual participants. The adults followed one of three sets of questions, depending on which of the stories the robots were enacting. They followed the same format, asking the participants basic questions about the story, such as *"What happened in the story?"*, *"How did the robot feel at this time?"*, *"Why do you think they felt this way?"*, and *"What option for the story did you choose next?"* among others. The participants would respond to each question as they wished. If a participant appeared to be stuck on a question, after a certain period of time, the interviewer would ask if the participant wanted more time or to move on to the next question.

This dataset proved to be very useful, because it offered a wide range of responses; some participants gave very short answers, some gave very long ones; some responses were extremely enthusiastic, and others were entirely disinterested; some participants hardly answered any questions, while some did more talking than the interviewer; some participants would start with answers that appeared to be disengaged, but would appear to be more engaged as the interview went on. This was therefore a rich dataset that would offer a lot of material for any classifier.

The format of the data did contain its own limitations, however. Although the interview format provided an element of control in that all the questions were the same, and the participants all dealt with the same scenarios, it also

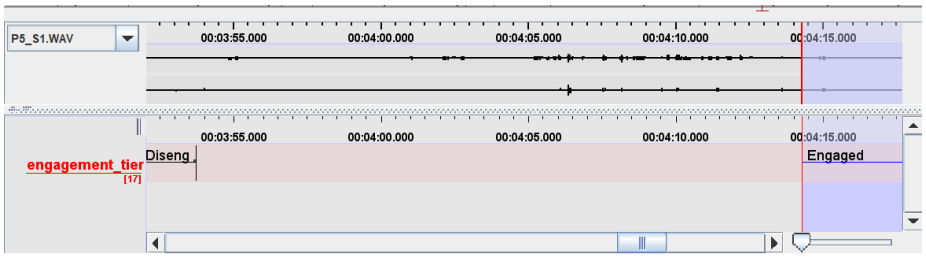

Figure 1: Screenshot of an interview annotation. Participants' responses were labeled as Engaged or Disengaged

meant that one of the participants was unusable for a classifier. The purpose of the interviewer in the dataset is to ask questions and keep the participant as engaged as possible. Every question posed by the interviewers was asked in an engaged manner, almost immediately after each of the participants' responses. Where the responses in the interviews provided a rich dataset, the questions of the interviewers did not. It became necessary, then, to only extract the features from the participants' responses.

3.2 Annotations

With the data in hand, the next step was to create annotations that could be used by a classifier. The data was hand-annotated using ELAN software. As only the participants' responses were to be used, only their responses were annotated. Ten interviews were annotated, with each response receiving a label of *Engaged* or *Disengaged*, depending on the manner of the response. A total of 335 annotations were marked from these ten interviews.

Once these annotations were saved, each of these ten interviews was loaded using the Python library librosa, with a sampling rate of 22050 frames per second. The array slices corresponding to each annotation were saved as comma-separated value (.csv) files, along with their binary annotation of 0 or 1 for disengaged and engaged, respectively.

There are limitations to the way this data was collected. First of all, ideally more interviews would have been annotated. However, due to time constraints, only ten were. In addition, hand-annotating engagement relies heavily on the annotators subjective opinion. This project would have benefited from having more than one annotator, and doing a majority-vote style annotation. Finally, using ELAN, or any hand-annotating software, there is possibility of imprecision when beginning and finishing annotation segments.

4 Building the Engagement Classification Model

4.1 Feature Extraction

Similarly to past research, this paper focused on supervised learning techniques to create an engagement classification model. Building off of past research, the features extracted from each utterance included F0, MFCCs, and Root Mean Square (RMS) Energy. The final list of features used in the model is:

- Mean and Standard Deviation of 13 MFCCs
- The Mean F0 value
- The Means, Standard Deviations, and Skews of
  - RMS
  - Zero Crossing Rate (ZCR)
  - a 12-bin Chromagram
  - 12-bin Chroma Energy Normalized
  - 12-bin Constant-Q Chromagram
  - the Spectral Centroid
  - the Spectral Flatness
  - the Spectral Contrast
  - the Spectral Roloff
  - the Spectral Bandwidth
  - the Mel Spectrogram
- The Mean and Standard Deviation for 6 tonal centroid features

For a total of 131 features. The clips from which these features were extracted were of variable length, but the features were extracted using a hop length of 512 and a window size of 2048.

4.2 Model Analysis

The classifiers considered for this paper were: Logistic Regression, Random Forest Classifier, a Support Vector Machine, Decision Tree, Linear Discriminant Analysis, and Naive Bayes. A train/test split of 70%/30% was used. The accuracy, as well as the f1-score was used to determine the best model. The accuracy of a model is the measure of the number of correct predictions made over the total predictions made. However, for the sake of this project, the goal is not only to have the model be as accurate as possible, but to minimize the number of false positives. In any system that aims to detect social contingency in order to strategically adjust its behavior, a false positive would result in a

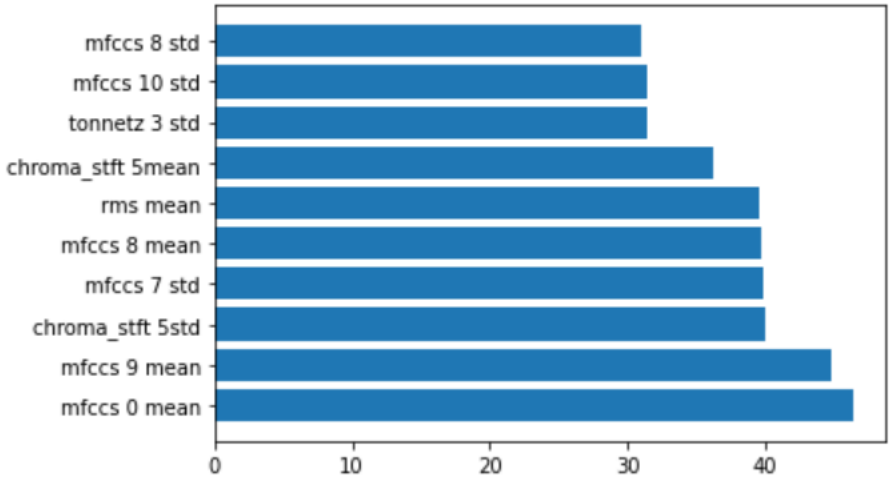

Figure 2: The ten features which contribute most to the classification of engagement. The feature labels are on the y-axis, and the normalized scores from the ANOVA F-Test are on the x-axis

failure to recognize a disengaged individual, which would prevent any attention recall from occurring. Therefore, the F1 score is also taken into account, as it factors false positives into its measurement. The results for these measurements were:

Classifier	Accuracy	F1 Score
Random Forest	83%	84%
SVM	80%	81%
Decision Tree	77%	79%
Naive Bayes	76%	78%
Linear Discriminant Analysis	65%	70%

A K-Fold Cross Validation of 10 splits yielded a Random Forest Classifier accuracy of 77.6%. With these results, and factoring the relative speed of classification, the Random Forest Classifier was chosen for this project.

But which features contribute most to the classification of engagement? An Analysis of Variance (ANOVA) F-Test was performed 100 times, averaging the results to determine the features that most influence the classification of a segment of audio as engaged or disengaged.

The results of this ANOVA F-Test are presented in Figure 2. As expected from their relevance in previous research, MFCCs are large contributors to the classification of engagement, making up six of the top ten features. Similarly unsurprisingly, the RMS values find themselves in the top ten as well. Past research did not include the chromagram (chroma_stft) and tonal centroid (tonnetz) values, so this analysis presents a novel finding in their role in audio

classification.

In fact, a model using these top ten features alone performs comparably, with 78% accuracy and an F1 Score of 80%. For a list sorting features by contribution to classification, as well as other forms of model analysis, please consult the code in model_evaluation.ipynb, which goes with this paper.

5 Extracting Structural Features

As mentioned previously, however, the classification of engagement of individual utterances might not necessarily paint the whole picture of social contingency. This section describes the process of extracting the structural features of a conversation.

5.1 Diarization

Thinking back to the example of the park bench, one of the main indicators that the individual texting was not paying attention might have been that their responses were infrequent, or that perhaps they did not respond at all. Therefore, in order to recognize social contingency in an interaction, it is necessary to be able to distinguish between the speech of different agents, and the silence between those periods of speech. In addition, since the model was built only for the child participants of the interviews, and not the adult interviewers, the first step for the structural feature extraction was that of diarization.

Diarization is the process of partitioning audio input into sections based on the speaker. Research on how to improve diarization, particularly unsupervised diarization, where no previous data about the speakers is known, is still ongoing. In this paper, two diarization techniques are proposed, both of which proved to be moderately successful.

5.1.1 Diarization: Silence as a Cluster

In the first method the Python library Resemblyzer is used. Resemblyzer's VoiceEncoder creates continuous embeddings of overlapping slices of preprocessed audio files. These embeddings are then assigned one of three labels using a spectral clustering algorithm where the number of clusters is 3. The concept behind this technique is that in the interviews, there are two speakers, and separate noise that is mostly just silence. Figure 3 shows the projections of the embeddings, and indeed, it is possible to make out three clusters from these. The cluster with the lowest amplitude is then determined to be the silent cluster.

5.1.2 Diarization: Filtering Out Silence First

In this method, the silence is filtered out initially from the audio input, and the clustering only occurs on genuinely spoken audio. Segments of silence are first detected using the Python library Pydub. All other segments of audio are interpreted as speech. These other segments are passed into Resemblyzer's

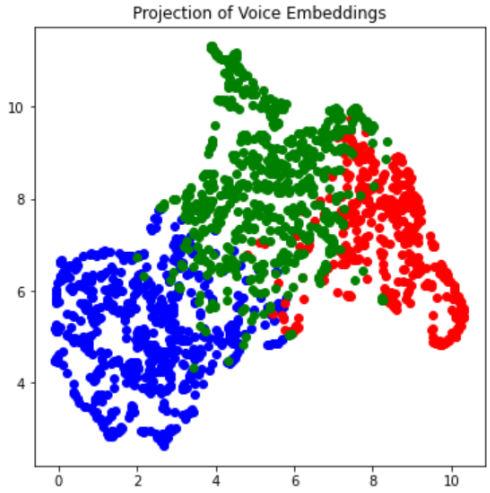

Figure 3: A visualization of the voice embeddings described in 5.1.1 with their cluster labels. Three different clusters become apparent

VoiceEncoder, and spectral clustering is performed on the resulting embeddings, with the number of clusters set to 2. With the silence cluster already set, the embeddings are labelled. The first person to speak is given the "Adult" label, and the second label is the "Child" label, as the interviewers are the first speakers in almost all of the interviews.

Both of these methods generate a breakdown of the structure of the interviews, with segments labeled by start time, end time, and whether the speaker is "Adult", "Child", or "Silence". Overall, listening through most of the interviews, the diarization is mostly accurate, with occasional mislabelled segments.

The second method is slightly more accurate. However, neither of these techniques are perfect. Creating a better distinction between the two speakers, and generating a more accurate distinction between speaking and non-speaking segments would greatly improve the results of this project.

5.2 Feature Extraction

As mentioned, the data all follow the standard interview format of question and answer. Therefore, the main indicators of social contingency occur in between the questions. Each segment labelled as "Child" in the diarization labelling was run through the engagement classifier, and a tally of engagement and disengagement was kept. The tally of the number of times the child did not respond, i.e., the "Adult" label appeared twice in a row without a "Child" label in between, was also recorded. The duration of silent segments and the confidence levels of the engagement model were also extracted as structural features. In the end, the features extracted from the interviews used to determine

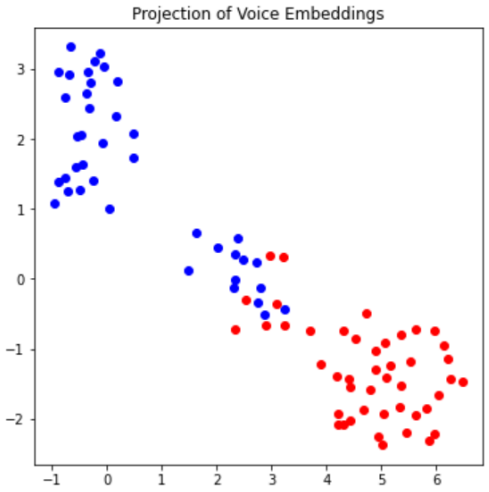

Figure 4: A visualization of the voice embeddings described in 5.1.2 with their cluster labels. The two clusters here are more clear than in Figure 3

social contingency were:

- **Number of Engaged Responses** – How often are the child's responses classified as engaged?
- **Number of Disengaged Responses** – How often are the child's responses classified as disengaged?
- **Number of No Answers** – How often does the child not respond to the adult's question?
- **Average Duration of Silence** – What is the average length, in seconds, of each silent segment?
- **Std of Duration of Silence** – What is the standard deviation, in seconds, the silent segments in an interview?
- **Silence Slope** – After running linear regression on the durations of silence, do the periods of silence increase or decrease in length, and how steep are those changes?
- **Silence Before Engaged Responses** – What is the average duration of silence before an engaged response?
- **Silence Before Disengaged Responses** – What is the average duration of silence before a disengaged response?
- **Silence Before No Answer** – What is the average duration of silence that the adult waits before speaking again in the event of a no-answer?

- **Confidence of Engaged Classification** – How confident is the model, on average, in its classification of engaged responses in the interview?
- **Confidence of Disengaged Classification** – How confident is the model, on average, in its classification of disengaged responses in the interview?

6 Results and Analysis

The results of building a model that can classify the engagement of an utterance with 83% accuracy and the extraction of other structural features of a conversation suggest that social contingency can be discerned from the relationships between these features. A few of these relationships are presented in this section.

Figure 5 depicts the relationship between the number of engaged responses in an interview and the model's confidence in its classification of engagement. This

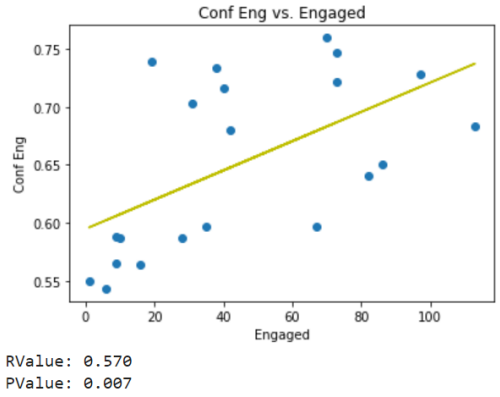

Figure 5: *A graph of the confidence in the model's predictions of engagement against the total number of engaged responses in an interview. The model is more confident the more engaged responses there are*

result might suggest that if an individual remains engaged throughout most of a conversation, their engagement is reflected in each individual response. Their responses themselves more regularly contain distinctive features that allow the model to more certainly predict engagement. This relationship unsurprisingly suggests that the low-level features of each utterance in a conversation have a direct relationship with the social contingency of an interaction.

The graph shown in Figure 6 represents the total number of disengaged and engaged responses in an interview. This graph shows that generally, interactions with high levels of engagement will have low levels of disengagement. These results confirm the belief that engagement factors heavily into the definition of social contingency. The graph can almost be split in half by the line of best fit; above the line are interactions that are not socially contingent, as they contain

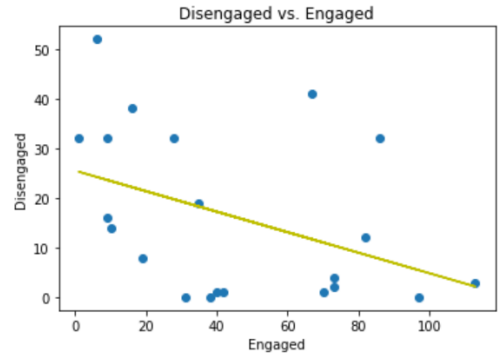

Figure 6: *A graph of the disengaged responses against the engaged responses. A steep decrease which confirms suspicions that individuals in a conversation who are engaged will stay that way throughout most of the conversation*

high levels of disengagement, and therefore lack of sensitivity towards the other social agent, and on or below the line are relatively high levels of engagement and therefore socially contingent interactions.

Figure 7 shows a direct relationship between the number of no responses and the number of disengaged responses. This contributes to the understanding of social contingency, as individuals that respond less often are also more likely to not be engaged in the conversation when they do respond.

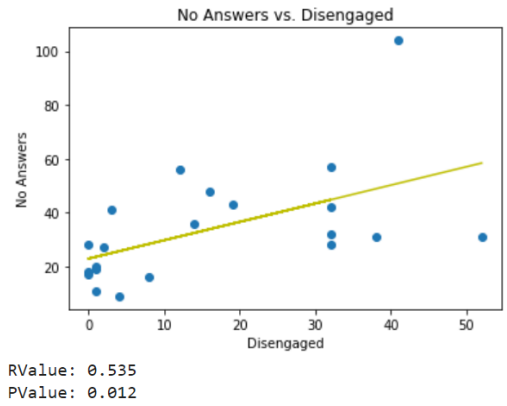

Figure 7: *How many times a child did not respond to a question vs how many of their responses were disengaged. We see a direct relationship and two separated clusters*

Figure 8 depicts how the number of engaged responses relates to the aver-

age duration of silence The relationship between the two affirms a certain belief

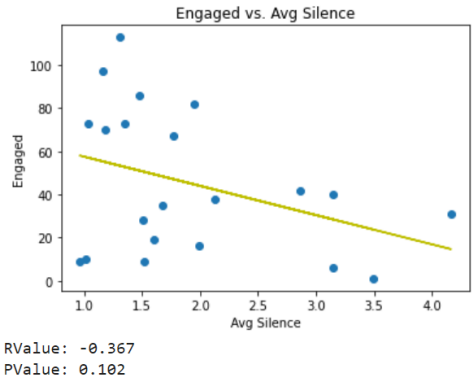

Figure 8: *The number of engaged responses vs. the average duration of each period of silence. Unsurprisingly, we see a negative relationship. Children with engaged responses answered questions more quickly. We see the formation of potentially three clusters*

regarding social contingency; individuals with higher levels of engagement also take less time to respond. This suggests that the duration of silence is a structural feature that can contribute to our understanding of social contingency. The visualiztion also shows the formation of two or three clusters: high engagement/low silence, low engagement/low silence, and low engagement/high silence. Many of the datapoints in the low engagement/low silence cluster also have low disengagement as shown in Figure 9, and low numbers of no answers as shown in Figure 7. Therefore, these datapoints are ambiguous with respect to social contingency. These individuals might have few long, engaged responses instead of many short ones. It might be useful to record the total duration of speech, rather than just a tally.

The visualization of the total number of disengaged responses versus the average duration of each period of silence (Figure 9) is not quite as exciting, and the results were not what one might have hoped. Although there appears to be a slight increase in the average duration of silence as the number of disengaged responses increases, the results are less significant than expected for the definition of social contingency. One might expect the formation of two clusters, where the socially contingent interactions would have short periods of silence and low numbers of disengaged responses. The relationship between these two features is not quite as clear. This could be a result of error in diarization, or simply not having enough interviews with high numbers of disengaged responses.

Figure 10 corresponds to the results in Figure 8, in that the duration of silence before an engaged response decreases as the number of engaged responses decrease. Once again, we see the relationship that is expected of social contingency, in that highly engaged individuals will take less time to respond during

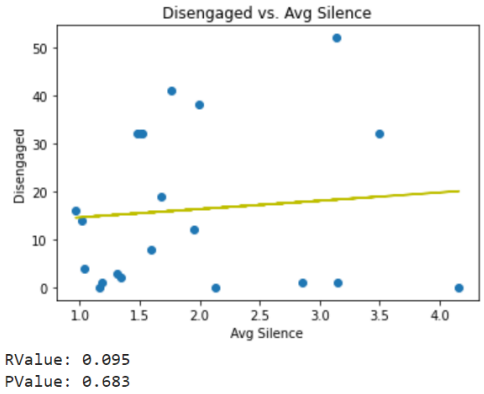

Figure 9: *A graph of the number of disengaged responses against the average duration of silence between responses. A slight increase of silence in disengaged responses, though maybe not as steep as one would have expected*

a conversation. We see, however, another factor that might contribute to our

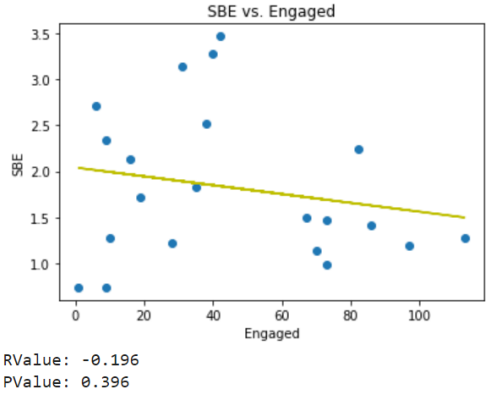

Figure 10: *Silence Before Engaged Responses vs. the number of engaged responses. We obtain similar results to Figure 8*

understanding of social contingency. Even non-socially contingent individuals may have a few engaged responses. Longer periods of silence before those responses could indicate that despite periods of engagement, they are not socially contingent, as suggested by the split in the graph.

Figures 11 and 12 present the relationships between the change in the duration of silence and disengaged and engaged responses. One might expect a socially contingent individual to become more engaged in a conversation as it progresses, and therefore take less time to respond over the course of the conversation, as suggested in the previous graphs. One might also expect the opposite

to be true for a non-socially contingent individual; namely, that periods of silence will decrease in length as an individual becomes less engaged in the conversation. Figures 11 and 12 suggest that this may be partially true. Periods of silence do appear to increase more steeply as disengagement increases, but this result is not necessarily significant. High-engagement conversations do not necessarily decrease the duration of each silence period over the course of the conversation, but the periods of silence remain relatively steady throughout the interaction. These relationships, therefore, could contribute to our understanding of social contingency, but maybe not to the extent that we expected.

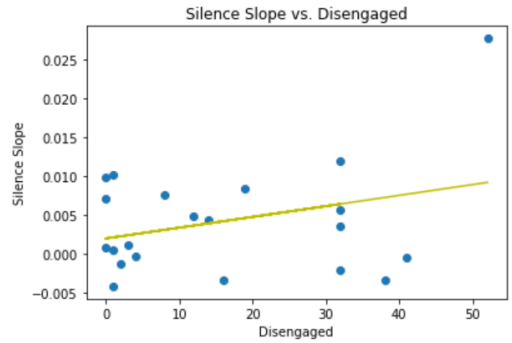

Figure 11: *How the durations of silence change vs the number of disengaged responses. We see the formation of binary split, and a direct relationship, where periods of silence get longer as the number of disengaged responses increases*

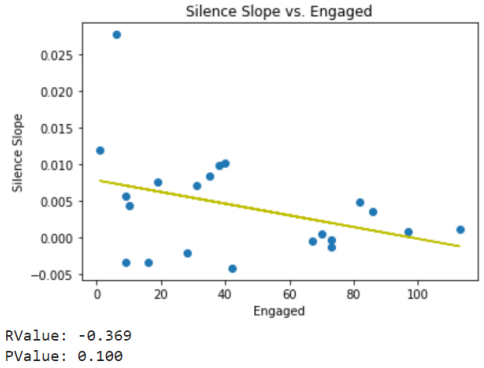

Figure 12: *Change in durations of silence vs the number of engaged responses. We also see a split between the two, but the periods of silence become shorter when there are more engaged responses*

These are only some of the visualizations from the extracted structural features. For the rest, please consult the code in Social_Contingency_Notebook.ipynb associated with this paper.

Overall, the splits and the clusters visible in these figures suggest that structural features aside from engagement and disengagement can signal the presence or lack of social contingency in an interaction. Although engagement and disengagement do make the clearest distinctions, the number of times an individual does not respond and the relationship between responses and silence can give a better understanding of how socially contingent a person actually is throughout a conversation. On their own, the features may not be enough, but being able to see that an individual has been giving mostly disengaged responses, that the silence before their engaged responses is relatively long, and that the duration of silence in between responses is increasing, could paint a better picture of that individual's social contingency. The results suggest that the structural features extracted could signal a binary classification of socially contingent or not.

7 Future Work

Although the results of this paper are promising, as mentioned, there have been certain limitations. The study of social contingency through audio features would benefit greatly from some more work and improvements.

To start, data from real-life conversations, not interviews, might improve certain aspects. Conversations are a two-way street, and it could be interesting to see how analysis of social contingency changes when classifying the social contingency not just of one person, but both people.

If sticking with the interview format, the engagement classifier would be improved by annotating more interviews and giving the classifier more training data.

On the topic of annotations, it would be useful to have more than just one annotator. Since distinguishing between engagement and disengagement could be a subjective matter, having multiple people annotate sections of the conversation could improve the accuracy of the model. Using a majority-voting system, sections could be annotated by several people as engaged or disengaged. Furthermore, since the focus is purely on the low-level features, it could even be useful to have annotators that do not speak the language marking down the annotations, so they cannot be influenced by the words spoken, but rather only the manner of speaking. With an improved dataset, we can be more certain of the model's accuracy.

There are many more low-level features that can be extracted to improve audio classification. [2] describes many of them. Including other spectral, voice quality, or Teager Energy Operator (TEO) features not included in this paper might contribute to the improvement of the engagement classification of the model. Furthermore, many other structural features of the conversation could be extracted. Since the data was in an interview format, the ratio of how often either person spoke was forgone. However, including frequency of speach or

duration of speech, among other structural features, might also improve the understanding of social contingency.

Finally, speaker diarization, especially unsupervised diarization, is still a growing field. Improving the manner of diarization will ensure more accurate breakdowns of the structural features of a conversation. This may involve eliminating noise through filtering, improving the embeddings of utterances into feature space, or improving the spectral clustering to assign labels.

The sources of error in this paper do not detract from the fact that there is evidence that structural features of a conversation can signal the presence or lack of social contingency. A better understanding of social contingency can improve the way teachers interact with students, particularly through virtual school. Socially assistive robots can adjust their strategies by factoring these features into their decision logic. Social contingency has substantial implications, and I am excited to see which direction audio classification goes next.

8 Acknowledgements

A huge thanks to Rebecca Ramnauth, my mentor for this project, for meeting with me weekly, always wishing me a happy Friday, and making sure this project got done.

Thank you to Scaz for being my advisor, and his Social Robotics Lab for giving me the opportunity to work with them.

I'd like to also thank Iolanda Leite and her team from [5] for providing the data and the past work to build off of.

Thanks to the rest of CS @ Yale for all the stress, but also all the fun and excitement that made me the programmer I am today.

A special thanks to Stanley Eisenstat who taught me that Computer Science can be very scary and challenging at times, but it is always worth the trouble. Rest in peace, Professor.

Thank you to my friends and family who supported me throughout my life and time at Yale. You've seen me at my lowest, so thanks for sticking through it all.

And finally, a thank you to the man who got me started on this journey, Stephen Young. Thank you for teaching me the mantra that has stuck with me throughout the years, and thank you for being one of the main reasons I did Computer Science at Yale.

References

[1] Yale university social robotics lab, Nov 2020.

[2] Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020.

[3] Starlet Ben Alex, Leena Mary, and Ben P. Babu. Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features. *Circuits, Systems, and Signal Processing*, 39(11):5681–5709, November 2020.

[4] Christy Elias. Analysis of speech parameters as indicators of engagement in conversation. 2020.

[5] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scassellati. Comparing models of disengagement in individual and group interactions. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, page 99–105, New York, NY, USA, 2015. Association for Computing Machinery.

[6] Brian Scassellati, Laura Boccanfuso, Chien-Ming Huang, Marilena Mademtzi, Meiying Qin, Nicole Salomons, Pamela Ventola, and Frederick Shic. Improving social skills in children with asd using a long-term, in-home social robot. *Science Robotics*, 3(21), 2018.

[7] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema, and Shiny Rajan. Emotion recognition from audio signals using support vector machine. In *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 139–144, 2015.

[8] Chen Yu, Paul M. Aoki, and Allison Woodruff. Detecting user engagement in everyday conversations. *CoRR*, cs.SD/0410027, 2004.