Detecting Personality Traits in Conversational Speech

Liam Kinney∗
Computer Science
Stanford University
liamk@stanford.edu

Anna Wang
Computer Science
Stanford University
annaxw@stanford.edu

Jessica Zhao
Computer Science
Stanford University
jesszhao@stanford.edu

Abstract

Detecting the personality implications of speech is integral to building human-level dialogue systems. Such systems should know not only what constitutes meaningful natural language, but also what makes for lexically and prosaically rich speech. By deriving the social perception of speech in addition to the meaning of the natural language, dialogue systems can perform better in conversation. We investigate the salience of different speech-derived features on the perception of three attributes (funny, awkward, and flirty). The experiments include tests with several different learning models across various compositions of five feature sets. We reveal which combinations are most attuned to identifying each personality attribute.

1 Introduction

The goal of our project is to explore the performance of different models and feature sets on classifying perceived traits about a speaker. The models are trained on human-human conversational data from the Stanford Speed Dating Corpus. Participants were asked, after a short conversation, to rank their conversational partners in various traits. We focus on the perceptions for awkward, funny, and flirty.

2 Background Work

Though not a strictly academic realm, the world of self-help books provides insights into and may benefit from this project. While presenting this work, our team was approached by Phil, a man who helps international students be more personable to Americans; he, as many others at varying levels of English ability, was curious as to what makes someone come off as funny, flirty, or awkward.

Dewitte (2017) recently posited that what separates a person with a sense of humor (i.e. who is funny) from a person without is a scheme of response generation and selection: those who generate a lot of jokes and are sensitive to negative appraisal will be considered humorous by their peers. Our research corroborates this theory by providing evidence for self-deprecating behavior as a marker of funniness.

Our results suggest that a flirty person utilizes lively speech – namely speech that is louder and at a higher pitch range. This is consistent with advice for flirting given by Rabin and Lagowski (1993). "You may not know it," she starts her book, "but there’s a flirt in you". Flirting is all about confidence, and lively speech is confident speech.

Speech is perceived as awkward when a speaker avoids personal reference, such as when s/he dives deeply into a specific, often seemingly irrelevant, topic. Nitkin (2000) cites this as cornerstone behavior of children with socially developmental deficiencies, like ADHD and Asperger’s syndrome.

∗Code for speeddate_classifier available on github: https://github.com/AnnaXWang/speeddate_classifier
The results of our experiments suggest that awkward speech is indeed quieter and less personal (i.e. fewer uses of “I”).

In order to build a model that could effectively identify the three traits in speech, we needed a robust labeled dataset and a curated set of features. We took both from [Ranganath (2009)], who also provides initial insight into the task of predicting a speaker’s observed personality features from the acoustics of speech. The paper mounts experiments to explore the limitations of humans and machines in determining social cues. For example, their model performed much better than humans in correctly evaluating intent of flirtation (71.5% to 56.2%). The experiments below constitute a continuation of these socially-observable trait predictive modeling efforts.

3 Approach

3.1 Features

Lexical Features: [Ranganath (2009)] We pulled 20 lexical features per speaker per conversation that quantified the number of backchannel utterances, appreciations, questions, next turn repair indicator questions, laughs, turns, dispreferred responses, filled pauses, total words, past tense words, metadata words, you-related words, we-related words, I-related words, words expressing assent, swear words, words implying insight, words expressing a poor sentiment, words of a sexual nature, and words related to eating/drinking. For more details about our lexical features, refer to Ranganath’s paper.

Aggregated Acoustic Features: [Ranganath (2009)] The rationale behind the use of acoustic features that rely on pitch, rate and energy, as made clear in [Mohammadi and Vinciarelli (2012)], is that they are the most important aspects of prosody and also the most salient markers when deriving personality perception from speech. Modeled from Jurafsky’s paper, we take 19 acoustic features per speaker per conversation that quantified F0, pitch range, RMS, rate of speech, and turn duration. For F0, pitch range, and RMS, we also extracted the values for their min, max, standard deviation, standard deviation of the standard deviation, min standard deviation, and max standard deviation. For more details about our aggregated acoustic features, refer to page 3 of Ranganath’s paper.

Segmented Acoustic Features: These features are similar to the Aggregated Acoustic Features, but the 19 features are collected per turn in the conversation rather than for the conversation as a whole.

Word Embeddings: We used 100 dimension pretrained GLoVe word embeddings from the Twitter corpus to attain a vector representation for every word in a conversation [Pennington and Manning (2014)].

Dialogue Embeddings: We calculated a 600 dimension vector representation for the dialogue comprising a speaker’s turn in a conversation by using sent2vec by Pagliardini (2017), which is trained on English Wikipedia unigrams.

3.2 Baseline Linear Classification Model

Our baseline predictive model is a logistic regression classifier. We used the model to (1) establish a baseline for deeper learning models and (2) explore the salience of different features within the sets to the model’s predictions. The model implements regularized (L2) logistic regression. The training algorithm uses the one-vs-rest (OvR) scheme. We experimented with inputting Lexical, Aggregated Acoustic, Segmented Acoustic, and Dialogue Embeddings separately and in various combinations.

3.3 Neural Network Models

Model 1: Single Input Feedforward This neural network takes a single feature set as input and sends it through one hidden layer before softmaxing.
Model 2: Dual Input Feedforward This neural network takes two concatenated feature sets as input and sends them through a single hidden layer and then softmax.

Model 3: Single Input Unidirectional RNN This neural network runs a RNN with LSTM over a single input and sends the output through a single hidden layer and then softmax.

Model 4: Single Input Bidirectional RNN This neural network runs a RNN with LSTM bidirectionally over a single input and sends the output through a single hidden layer and then softmax.

Model 5: Bidirectional RNN Hybrid This neural network runs a RNN with LSTM bidirectionally over one input and concatenates the output with another feature set before sending the complete outputs through a single hidden layer and then softmax.

Model 6: Dual Input Bidirectional RNN This neural network runs two RNNs over separate inputs, concatenates the outputs, and sends the complete outputs through a hidden layer and then softmax.

3.3.1 Model Equations

For every one of our models, the hidden layer fed into a softmax layer utilizes the same equations. Let us represent the output of the first stage of input processing as $y$ and the final probabilities over the classes as $p$. Let $W_1$ and $b_1$ be the trainable weights and biases respectively.

$$o = W_1 y + b_1$$
$$p = \text{softmax}(o)$$
Now, for each of our models, we only have to define \( y \). For Model 1, the input to the neural network is simply \( y \). For Model 2, let us denote the two inputs as \( x_1 \) and \( x_2 \) respectively, so \( y = \text{concat}(x_1, x_2) \). For Model 3, let us denote the output of the forward RNN as \( z_1 \), so \( z_1 = y \). For Model 4, let us additionally denote the output of the backwards encoding as \( z_2 \), so \( y = \text{concat}(z_1, z_2) \). For Model 5, \( y = \text{concat}(\text{concat}(z_1, z_2), x_1) \). For Model 6, let us denote the forwards and backwards encoding of the second input as \( v_1 \) and \( v_2 \). \( y = \text{concat}(\text{concat}(z_1, z_2), \text{concat}(v_1, v_2)) \).

After experiencing a large amount of overfitting for these models on our small dataset, we decided to conclude our experimentation and not to try more complex models.

4 Experiments

4.1 Dataset Preparation

The conversations in our dataset followed a turn-taking format. We separated the data by speaker and by conversation. Each speaker was ranked on a scale of 1-10 for each attribute by their conversationalist. Following Ranganath (2009), we used the top 10% of labels as positive classifications and the bottom 10% of labels as negative classifications. We discarded conversations in which a rank was not provided. The preprocessing of the transcript files involved some data sanitation that included replacing numbers with hashtags and padding punctuation with spaces. The preprocessing of the audio files involved a split based on pauses between speech and the frequency of the speech. The dataset comes with a Perl script that performs this split, courtesy of Ranganath et al.

Our final datasets contained 543 examples in the Funny dataset, 506 examples in the Flirty dataset, and 656 examples in the Awkward dataset. We randomly split the data into training, dev, and test datasets for each attribute, adhering to the 70/20/10 split.

4.2 Results and Evaluation

The main evaluative measure for the classification task was percent accuracy on the test set. The tables below show the comparative performance of different feature sets and of different learning models on the binary classification task. The highest performance on the test set for each attribute is highlighted.

For our neural network models, we tuned hyperparameters for hidden size (HS), learning rate (LR), batch size (BS), dropout keep percentage (D), and optimizer (Opt). We tried various combinations of features that showed promising improvement as model complexity increased.

Overall, we are pleased that our neural networks models outperformed our baseline logistic regression models in classifying every attribute. In fact, our neural network model for flirty achieved scores that are on par with the accuracy score reported for identifying flirtation in Ranganath (2009). As we experimented with inputting combinations of feature sets and tuning hyperparameters, we noticed that Dialogue Embeddings performed better than Word GloVe Embeddings. This makes sense, because the overall information conveyed in a single turn in a conversation is more important than the meaning of words that are spoken across different turns. Our Bidirectional RNN Hybrid (Model 5) with Dialogue Embeddings and Aggregated Acoustic features achieved the highest scores for Awkward and Flirty, while the Dual Input Bidirectional RNN (Model 6) with Dialogue Embeddings and Segmented Acoustic features achieved the highest score for Funny. This is expected, as Model 5 and 6 are our most complex models and processed the most textual and acoustic information for the classification task. It is interesting that funny was the only attribute that was classified better with the addition of Segmented Acoustic features. We posit that identifying humor requires more nuanced acoustic information because humor tends to appear in flashes during a conversation while awkwardness and flirtation persist throughout a conversation. Therefore, valuable acoustic information for classifying Funny may have been lost in the Aggregated Acoustic Features.
<table>
<thead>
<tr>
<th>Lexical</th>
<th>Aggregated Acoustic</th>
<th>Segmented Acoustic</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev: 63.89%</td>
<td>test: 58.18%</td>
<td>dev: 62.04%</td>
</tr>
<tr>
<td>Dialogue Embeddings</td>
<td>Aggregated Acoustic+Lexical</td>
<td>Lexical+Dialogue Embeddings</td>
</tr>
<tr>
<td>dev: 68.52%</td>
<td>test: 58.18%</td>
<td>dev: 70.37%</td>
</tr>
<tr>
<td>dev: 64.82%</td>
<td>test: 65.46%</td>
<td>dev: 65.74%</td>
</tr>
</tbody>
</table>

Table 1: Performance of Logistic Regression for Funny

<table>
<thead>
<tr>
<th>Lexical</th>
<th>Aggregated Acoustic</th>
<th>Segmented Acoustic</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev: 58.02%</td>
<td>test: 57.58%</td>
<td>dev: 53.44%</td>
</tr>
<tr>
<td>Dialogue Embeddings</td>
<td>Aggregated Acoustic+Lexical</td>
<td>Lexical+Dialogue Embeddings</td>
</tr>
<tr>
<td>dev: 58.02%</td>
<td>test: 54.55%</td>
<td>dev: 64.12%</td>
</tr>
<tr>
<td>dev: 61.07%</td>
<td>test: 57.58%</td>
<td>dev: 64.89%</td>
</tr>
</tbody>
</table>

Table 2: Performance of Logistic Regression for Awkward

5 Observations

We explore some of the implications that the results above have on predicting the personality traits in question. The tables below inform us which features within each set were most salient to the task. (Note that there was no such analysis done for the sentence embeddings feature set, since the features are esoteric values from the embedding matrix that capture prosody, and cannot be interpreted individually.) This data yields insights into the aspects of speech to which humans attend when making personality judgments about their conversational partners.

The experiments with different models allow us to make claims about the underlying technology that should be used in dialogue systems that attempt to serve up socially conscious dialogue.

5.1 Funny

The results from Table 1 indicate that features derived from prosody and sentence embeddings are the most relevant when producing funny speech. Table 7 reveals that the amount of fluctuation in pitch/frequency and the rate of speech are the acoustic components that make for funny speech. Knowing that acoustic features outweigh lexical, we can argue that pitch patterns and speech rate are greater deciders of humor than generic lexical patterns. However, the salience of sentence embeddings suggests that what is being said is far from irrelevant. We can posit that varied-pitch speech at a certain rate, using novel sentences, is funny speech.

This is consistent with the observations made by Menninghaus (2014) about humor in poetry. Meter, Menninghaus found, has an integral effect on how funny poetry/song is perceived to be. Meter in a poetry context parallels rate of speech in a conversational context. The results in Table 7 show that it is a high variance in pitch that makes for funny speech.

In terms of Lexical features, shown in Table 7, the amount of laughing was an obvious determiner. Less predictable was the effect of the number of self-references and of words expressing dispreference. After examining at some example inputs, it is possible that the combination of these two lexical trends indicate self-deprecating behavior, which Dewitte (2017) posits is a cornerstone of any humorist’s emotional toolkit.

5.2 Awkward

Acoustically derived and lexically derived features have a nearly equal effect on how awkward speech is perceived. Taken together, the two feature sets are very good – more predictive than any experiment on the other two traits – at the task. As shown in Table 2, the lexical feature set performs slightly better than the acoustic.
<table>
<thead>
<tr>
<th>Lexical Aggregated Acoustic Segmented Acoustic</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev: 64.36%</td>
</tr>
<tr>
<td>Dialogue Embeddings Aggregated Acoustic+Lexical Lexical+Dialogue Embeddings</td>
</tr>
<tr>
<td>dev: 56.44%</td>
</tr>
<tr>
<td>dev: 62.38%</td>
</tr>
</tbody>
</table>

Table 3: Performance of Logistic Regression for Flirty

<table>
<thead>
<tr>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS:100</td>
</tr>
<tr>
<td>HS:150</td>
</tr>
<tr>
<td>HS:50</td>
</tr>
<tr>
<td>HS:100</td>
</tr>
<tr>
<td>HS:100</td>
</tr>
<tr>
<td>HS:50</td>
</tr>
<tr>
<td>HS:50</td>
</tr>
<tr>
<td>HS:100</td>
</tr>
</tbody>
</table>

Table 4: Performance of Neural Network models for Funny

The salient lexical features allow us to mount some novel analysis into what makes a conversation awkward. Table 8 reveals that speaking about the context for the conversation is the tendency most closely tied to perceived awkwardness. This kind of speech belies the speaker’s either disinterest in or inability to engage with their partner; the awkward speaker acknowledges the inherently contrived context of the speed date rather than embracing it. The number of words in the conversation (namely, when there are few, as shown in Table 8) are clearly salient because neither party would be easily able to nor want to draw out an awkward conversation. Finally, the amount of self-reference is notable because it appears as an indicator of funny speech as well. As is clear from the difference in mean values between the negative and positive cases in Table 8, a speaker is heard as awkward when they refer to themselves fewer times.

On the acoustic front, the strength of aggregated acoustic features is consistent when examining awkward’s most salient features in Table 8. Aggregated acoustic values, including the standard deviation of a standard deviation and the mean of a standard deviation (for F0) are the best predictors. This corresponds with conversational consistency, as one can naturally have awkward moments apparent in pitch and intensity when on a speed-date (for example, a strained chuckle), yet the classification of awkward as a personality trait becomes more justifiable if there are more frequent occurrences as determined across an aggregated conversation.

5.3 Flirty

As shown in Table 9, the segmented acoustic feature set is the most predictive set by far for the perceived flirtiness of speech. Looking at Table 9, we see that RMS, a measure of amplitude per turn, is the key determinant of the acoustic suite. Given that the RMS Mean and RMS Min are higher in positive labelings than in negative labelings, we can safely say that a louder person is perceived as more flirty. The third feature in the table implies that a higher pitch range, too, is indicative of flirty speech.

The lexical features are similarly helpful in picking out determinants for flirty speech. As could be expected, flirty people laugh more often. It also makes sense that those who ask more questions are perceived as flirty, because questions show that the person is actively interested in getting to know his/her conversationalist better. Finally, people who use fewer verbal crutches like "uh" and "um" are also perceived as flirty, probably because they appear more confident.
### Table 5: Performance of Neural Network models for Awkward

<table>
<thead>
<tr>
<th>Model</th>
<th>Scores (dev</th>
<th>test)</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1: Lexical</td>
<td>60.77%</td>
<td>63.33%</td>
<td>HS:100</td>
</tr>
<tr>
<td>Model 1: Aggregated Acoustic</td>
<td>64.28%</td>
<td>64.06%</td>
<td>HS:200</td>
</tr>
<tr>
<td>Model 1: Dialogue Embeddings</td>
<td>58.46%</td>
<td>60.00%</td>
<td>HS:100</td>
</tr>
<tr>
<td>Model 2: Lexical + Aggregated Acoustic</td>
<td>66.92%</td>
<td>66.66%</td>
<td>HS:200</td>
</tr>
<tr>
<td>Model 3: Word GloVe Embeddings</td>
<td>60.77%</td>
<td>63.08%</td>
<td>HS:100</td>
</tr>
<tr>
<td>Model 3: Dialogue Embeddings</td>
<td>58.46%</td>
<td>65.00%</td>
<td>HS:100</td>
</tr>
<tr>
<td>Model 4: Dialogue Embeddings</td>
<td>54.62%</td>
<td>66.67%</td>
<td>HS:100</td>
</tr>
<tr>
<td>Model 4: Segmented Acoustic</td>
<td>61.54%</td>
<td>61.67%</td>
<td>HS:50</td>
</tr>
<tr>
<td>Model 5: Dialogue E. + Aggreg. Acoustic</td>
<td>59.23%</td>
<td>71.67%</td>
<td>HS:100</td>
</tr>
<tr>
<td>Model 6: Dialogue E. + Seg. Acoustic</td>
<td>57.69%</td>
<td>63.33%</td>
<td>HS:50</td>
</tr>
</tbody>
</table>

### Table 6: Performance of Neural Network models for Flirty

<table>
<thead>
<tr>
<th>Model</th>
<th>Scores (dev</th>
<th>test)</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1: Lexical</td>
<td>65.57%</td>
<td>64.00%</td>
<td>HS:200</td>
</tr>
<tr>
<td>Model 1: Aggregated Acoustic</td>
<td>66.00%</td>
<td>55.43%</td>
<td>HS:200</td>
</tr>
<tr>
<td>Model 1: Dialogue Embeddings</td>
<td>64.44%</td>
<td>60.00%</td>
<td>HS:50</td>
</tr>
<tr>
<td>Model 2: Lexical + Aggregated Acoustic</td>
<td>67.00%</td>
<td>66.00%</td>
<td>HS:100</td>
</tr>
<tr>
<td>Model 3: Word GloVe Embeddings</td>
<td>57.14%</td>
<td>66.67%</td>
<td>HS:100</td>
</tr>
<tr>
<td>Model 3: Dialogue Embeddings</td>
<td>60.00%</td>
<td>68.00%</td>
<td>HS:50</td>
</tr>
<tr>
<td>Model 4: Dialogue Embeddings</td>
<td>59.00%</td>
<td>70.00%</td>
<td>HS:100</td>
</tr>
<tr>
<td>Model 4: Segmented Acoustic</td>
<td>59.00%</td>
<td>66.00%</td>
<td>HS:50</td>
</tr>
<tr>
<td>Model 5: Dialogue E. + Aggreg. Acoustic</td>
<td>55.00%</td>
<td>73.00%</td>
<td>HS:100</td>
</tr>
<tr>
<td>Model 6: Dialogue E. + Seg. Acoustic</td>
<td>38.00%</td>
<td>70.00%</td>
<td>HS:50</td>
</tr>
</tbody>
</table>

### 6 Error Analysis

We hypothesize that the biggest cause for error in our neural net models is the small size of our datasets. We were only able to train on about 400 samples per epoch, which caused severe overfitting on the train set and a large variance between our dev and test scores. Further, we do not control for physical attractiveness and body language during the speeddate, which no doubt affects perceptions for awkward, flirty, and funny. Finally, we did not control for male vs. female speakers. If we had done so, the model may have performed better because certain features may have been more salient for one gender. For example, females probably have a larger pitch range in general so a large pitch range for a male may be a strong indicator for humor.

#### 6.1 Funny - Model 6: Dual Input Bidirectional RNN

**Dialogue Excerpt:** "[laughter] ah-huh [laughter]. This is the letter E. I think I’m letter E."

**Salient Acoustic Features:** Pitch Range SD: 80.96, F0 SD: 50.86, Rate of Speech: 3.49

This speaker was incorrectly classified as funny, but the facetious nature of the dialogue and the occurrences of laughter makes it difficult for even humans to classify correctly. One key consideration is that the model is blind towards the data of the opposite conversationalist. For example, the speaker’s laughter is not an accurate indicator of humor if the other person does not also laugh. In the future, we can experiment with feeding language features from the other speaker. The pitch range standard deviation, F0 standard deviation, and rate of speech are all on the higher end of values. We can see from our feature salience analysis that funny speakers on average have higher values for these 3 features. Thus, we can understand why our model incorrectly classified this example.

#### 6.2 Flirty - Model 5: Bidirectional RNN Hybrid

**Dialogue Excerpt:** "Yes I know, it turns me on, turns me on... But how are you doing?"

**Salient Acoustic Features:** RMS Mean: 74.41, RMS Min: 60.63, Pitch Range: 195.53
This example was incorrectly classified as not flirty. Clearly, the model did not pick up on the sexual innuendo contained in the dialogue excerpt. We hypothesize that the model failed to correctly classify because of shortcomings of the dialogue embeddings. The sent2vec pretrained vectors are generalized and not tailored for this task. Therefore, “turn me on” in the input text may not have been accurately weighted as important in the embedding values. The salient acoustic features for Flirt also had very similar values for negative and positive classifications, which probably made it difficult for the model to predict correctly based on acoustic information.

### 7 Conclusion

This project attempted two different but related tasks: to analyze the relative performance of acoustic and lexical features in classifying perceptible social traits, and to design the best system for classifying those traits. The former task is a sociological one, the first chapter in a data-driven self-help book; the latter task seeks to further the field of spoken language processing.

Future work could stem from either set of results. Our finished system could play a part in a speech generation module that seeks to produce more socially appropriate, human-like speech. A more robust sociological study could try to suss out whether males find flirty features in conversation with females where the female parties do not. Given more data, researchers could extend our model to other traits, potentially even to the five factor model of personality (McCrae and Costa (1987)).

Society trusts and relies on speech systems more than ever as practical every-day human-computer interfaces. Two of the big hurdles left for the field of spoken language processing to surmount are understanding human conversational speech and generating believable human speech. Our research is a step forward in both of these tasks.

### Acknowledgments

We would like to thank Andrew Maas and Jiwei Li, our instructor and head TA for CS 224S. Particular thanks go out to Dan Jurafsky, who was prompt and generous during our email correspondence. Final thanks to Google for providing a GPU subscription through their platform.
### Acoustic Feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMS Mean</td>
<td>Min: 63.83</td>
<td>Mean: 77.81</td>
</tr>
<tr>
<td>RMS Min</td>
<td>Min: 51.08</td>
<td>Mean: 63.68</td>
</tr>
<tr>
<td>Pitch Range</td>
<td>Min: 137.47</td>
<td>Mean: 208.00</td>
</tr>
</tbody>
</table>

### Lexical Feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td># Laughs</td>
<td>Min: 0.00</td>
<td>Mean: 3.51</td>
</tr>
<tr>
<td># Uh/Ums</td>
<td>Min: 62.00</td>
<td>Mean: 83.40</td>
</tr>
<tr>
<td># Questions</td>
<td>0.00</td>
<td>Mean: 12.91</td>
</tr>
</tbody>
</table>

Table 9: Most Salient Features Analysis for Flirty

### References


