Relationships Between Prosodic-Linguistic Features and High-Level Descriptors of Speed Dates

Sidd Jagadish, Ranjay Krishna and **Gabriele Carotti-Sha**
Department of Statistics, Stanford University
Department of Computer Science, Stanford University
Symbolic Systems, Stanford University

Abstract

We extracted lexical and prosodic features of speech from 1493 speed dates at Stanford University. Each speed date was accompanied by information about what each participant thought of the other, including 1-10 assessments of intelligence, funniness, and sincerity. Here, we investigate prediction of these assessments from linguistic and prosodic features. We find that for many labels, lexical and prosodic features can powerfully predict our high-level descriptors.

1 Data

We were given, courtesy of Jurafsky (2013), a dataset consisting of (FILL IN NUMBER OF SPEED-DATES) heterosexual speed dates. For each date, we have two .wav files corresponding to the microphones attached to each participant. We also have high-level descriptors of each participant, including their he. Beyond this, we have 1-10 self-assessments of flirtatiousness, friendliness, ...

2 Feature Extraction

Prosodic features were extracted using openSMILE. Lexical features were extracted in Python, with the help of the LIWC dictionary.

3 Prosodic Features

4 Exploratory Analysis

4.1 Sparse PCA

Before building any classifier, we must first unveil any underlying structure in our predictor matrix. We begin with (FILL-IN:# of inital prosodic features) prosodic features. Considering that we only have 1493 full data points, we clearly want to reduce the dimesion somehow. We begin by running a Sparse Principal Component Analysis, as outined in Zou(2006). This method places an L1-Norm penalty on the loadings vectors in Principal Component Analysis, arriving at a sparse reduction. As seen in the chart below, most of our non-zero coefficients correspond to MFCC features. This is worrisome, as this means that most of the variation in our prediction data is attributable to MFCC features, which largely indicate the vowels that our speaker is making. This is understandable, but ultimately not relevant to our pursuits. For this reason, we choose to discard the MFCC features for the remainder of our anaysis.

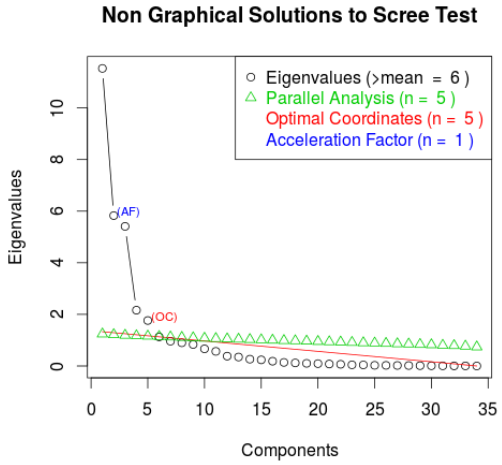
4.2 Factor Analysis

Even after removing MFCC features, we have FILL IN NUMBER OF PROSODIC FEATURES per speaker per date. This is more than we would like to have in order to have easily interpretable results. As such, we wanted to run a factor analysis on our prosodic predictors. It is important to consider that in this data set, much of the variation in the data may be attributable to variation in gender. As we did not want gender information to be our most important latent factor, we first ran factor analysis separately for the two groups. We found very similar results for the two separate factor analyses. The results are shown below.

We see that for both sexes, the first factor largely corresponds to maximum pitch third to min pitch, and fourth to turn duration. We see differences in the second and fifth factors – the second factor for males, which largely corresponds to loudness, is the fifth factor for females. The fifth factor for males, which largely corresponds to variation in loudness, is the second factor for females. Thus, our latent factors are very similar for the two sexes. When we pool the two sexes, we again get very similar factor coefficients (shown below). We will use these factors for prosodic features for the remainder of this paper.

We ran a factor analysis on our prosodic pre-

dictors, uncovering the following useful latent themes. We see our first factor corresponds to pitch, the second roughly to intensity, the third to turn duration, the fourth to variation in intensity, and the fifth to variation in pitch. To determine the number of factors to use, we used parallel analysis (Hayton 2004) and a scree plot (shown below) uncovering 5 factors.



5 Classifiers

5.1 Prosody Only

First, we attempted to classify our labels, in particular funniness and intelligence, by our prosodic factors alone. Here, we are faced with the following important decision. To classify one participant's (the listener's) perceptions of the other participant (the speaker), whose prosodic features should we use. Clearly both the listener's and the speaker's prosodic features should be informative – here, we first attempted to use only the speaker's prosodic features, as we would like to answer the question of 'How much information about funniness/intelligence/sincerity is contained in a speaker's prosody?'

To answer this question, we fit three models for each of 12 labels, seen in the table below. The first model is a classification SVM with a linear kernel, where our model is described by (TODO: Put in SVM model). Our second model is a classification SVM with a radial basis fuction kernel. Our third model is a AdaBoost model with decision trees as weak classifiers. We note now that the latter two models allow for non-linear decision boundaries, as we would expect for the task at hand. As such, we expect these classifiers to outperform the linear kernel SVM.

Below, we see the average results of 5-fold cross-validation for our three different binary classification methods, for each of 10 different labels. We see that, at a baseline of 50% (each of our training examples consisted of half positive cases and half negative cases), many of these classifiers do not perform extremely well. However, we note that for the labels studied in Jurafsky & Ranganath (2013), performance of our AdaBoost classifier is comparable to their L1-penalized SVM. It seems that the accuracy lost from feature inclusion (their model included prosodic features from *both* speakers, as well as lexical and "accomodation" features), we have gained from model choice. In general, our AdaBoost classifier and our SVM with a RBF kernel outperform our SVM with a linear kernel, indicating that we may have non-linear classification boundaries here. In any case, we have shown that, especially for labels such as intelligence and funniness, the speaker's prosodic features include useful information for classification.

5.2 Explanatory Power of Self Features vs. Other Features

Another interesting question here is comparing the explanatory power of one's own speech to that of their partner's speech. Do the listener's prosodic features in fact reveal more about what they think of the speaker than the speaker's prosodic features do, and are the same features important in making these classifications. Here, I run the same 3 classifiers as above on the same labels, achieving the following results.

	Self Features			Other Features		
	RBF	Linear	AdaBoost	RBF	Linear	AdaBoost
s_fndly	0.600	0.536	0.592	0.568	0.554	0.544
s_flirt	0.590	0.546	0.590	0.553	0.531	0.542
s_awk	0.540	0.486	0.554	0.579	0.560	0.550
s_assert	0.558	0.532	0.557	0.536	0.519	0.501
o_fndly	0.583	0.530	0.560	0.552	0.541	0.528
o_flirt	0.568	0.548	0.575	0.546	0.529	0.542
o_awk	0.540	0.505	0.543	0.563	0.574	0.546
o_assert	0.571	0.554	0.561	0.581	0.561	0.555
o_attrct	0.576	0.581	0.602	0.599	0.576	0.593
o_sincre	0.552	0.538	0.538	0.531	0.550	0.536
o_intell	0.639	0.641	0.607	0.561	0.545	0.546
o_funny	0.623	0.591	0.619	0.618	0.604	0.608
o_ambits	0.578	0.536	0.554	0.541	0.547	0.528
o_crteos	0.589	0.559	0.562	0.549	0.561	0.553

We see that in general, participant 1's features are better predictors of labels assigned to particant 1 than participant 2's features are. For intelligence, this difference is particularly stark (p

¡ 0.01). We interpret these results as the following: Some information about the speaker's apparent intelligence is revealed through this speaker's prosody. Some information about the listener's perception of the speaker's intelligence is revealed through the listener's speech. In the case of intelligence, we find that the speaker's prosody is much more informative than the listener's prosody. In humor and awkwardness, however, this is not the case. As found in our lexical analysis (NEED TO INCLUDE LAUGHTER AS A FEATURE IN LEXICAL ANALYSIS), both of these labels are closely associated with laughter. We now investigate whether the laughter in funny conversations differs from the laughter in awkward conversations.

References

Ronan Collobert and Jason Weston. 2008. A unied architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*.