Understanding the Perception of Courteous and Humorous Behavior using Prosodic and Lexical Features

Sidd Jagadish, Ranjay Krishna‡, Gabriele Carotti-Sha∓*

siddj@stanford.edu, rak248@stanford.edu, gcarotti@stanford.edu
*Department of Statistics, Stanford University
‡Department of Computer Science, Stanford University
∓Symbolic Systems, Stanford University

Abstract

Computationally understanding the perception of a speaker's characteristics is an important task speech processing, behavioral outcomes and dialogue systems. Our work focuses on the classification and understanding the perception of two such characteristics: courtesy and humor. We analyze ratings surveyed from the SpeedDate corpus to build models using prosodic and lexical features of the speakers. We find that lexical features like *Hedge-words* and prosodic features like *Intensity Min SD* are important for detecting courteous while *Exclamations* and listener's *Intensity* are important for funny. Using an Adaboost Classifier, we detect the funny with an accuracy of 0.74 and courteous at 0.76.

Index Terms: Prosody, Speed Date, AdaBoost, Gender differences in prosody, factor analysis, courteous, funny

1. Introduction

Efforts in linguistic understanding encounter inevitable difficulties whether the interpretative entity is a human or a machine. For example, any communicative exchange across agents runs the risk of misrepresentation of one's self or of their interlocutor. It has been shown that given a dialogue between just two agents, human judgments about each other can be subject to great discrepancies (R. Ranganath et al. 2009), and this is potentially due to a failure in recognizing each other's affective cues. To face this issue, it is useful to invoke the notion of *interpersonal stances*, as described by Schaerer (2000, 2003). An interpersonal stance is the way in which interlocutors pose themselves with respect to other agents within a given exchange. In particular, this notion incorporates affective stances, such as flirtatiousness, awkwardness, or courteousness, whose expression can be detected via subtle cues from a speaker's voice or words. This observation leads to the definition of two sets of communicative features: prosodic, pertaining to the physical characteristics of the vocal signal, and lexical, pertaining to the semantic content of the words pronounced.

Examples of prosodic features are the fundamental frequency contour, or pitch, of a given speaker's voice, or the mean intensity they use to convey a particular emotive state (anger, excitement, humor). On the other hand, the number of times a given word is repeated, either throughout an entire conversation or only in response to a preceding speech segment, are considered lexical features. Thanks to the detailed and insightful research conducted by R. Ranganath et al. (2013), it has been possible to extract both sets of features from a series of speed date encounters, which were recorded and transcribed. Despite the relative artificiality of conversations within this particular context, this very artificiality can be favorable in an empirical study: it poses constraints on how much time interlocutors have to establish an exchange, and also generates very specific interpersonal stances (listed in the following sections), all of which are relevant to the date being either a success of a failure. By applying NLP classification methods to the extracted feature sets, it is possible to generate a fairly accurate predictor of whether, given a pair of interlocutors, one of the two found the other to be flirtatious, intelligent, etc. and vice versa.

We build our own models upon the previously conducted research on the same dataset, and apply the same techniques in order to detect stances such as courteousness, sincerity, and humor.

2. Related Works

There is a large literature on detecting social meaning. Ang et al. (2002) [8] investigate the detection of annoyance and frustration. Lee & Narayanan (2002) [7] discuss the classification of positive and negative emotion. Ranganath et al. (2013) use the same data set to predict the labels of awkwardness, friendliness, flirtatiousness, and assertiveness, finding that prosodic and linguistic features can indeed predict these labels quite well [1]. In addition, they find that the most relevant predictors vary significantly across labels and genders. Ranganath et al. (2009) find that using stacked autoencoders to find low-dimension representations of the lexical data significantly improves prediction accuracy [6]. There is a wide literature on the study of humor, but surprisingly little of this research has studied the acoustic nature of humor – instead choosing to focus largely on semantics. Menninghaus et al. (2014) find that clear regard for rhyme and meter enhances the perceived humor of couplets [9]. Purandare & Litman (2006) study differences in prosody between laughter-preceding and non-laughter-preceding statements in the situational comedy *FRIENDS*[10]. Of course, these are jokes told on a scripted telivision show, with clear breaks in dialogue for "canned laughter." Here, instead of rehearsed jokes, we predict the perception of funniness based on acoustic features of conversational speech. While there is an abundance of research on humor, very little research has been conducted on the perception of courteousness.

3. The SpeedDate Dataset

We utilize the SpeedDate dataset introduced by [1] to build our model. The dataset contains approximately 1100 heterosexual 4-minute speed dates. Each date is stored as a wav file recording

	Male					Female				
Variable	Intens.	PitchMax	PitchMin	TurnDur	Intens.Var	PitchMax	Intens.Var	TurnDur	PitchMin	Intens.Mean
tndur.mean	0.03	0.26	-0.22	**0.94**	-0.05	0.17	-0.03	**0.97**	-0.15	0.04
pmin.mean	0.16	-0.23	**0.85**	-0.24	0.11	-0.21	-0.07	-0.21	**0.89**	0.07
pmax.mean	0.13	**0.94**	0.01	0.23	0.0	**0.92**	0.03	0.23	-0.1	0.05
pmean.mean	0.18	0.53	**0.69**	-0.13	0.07	0.72	-0.02	0.02	0.48	0.19
psd.mean	-0.12	**0.71**	-0.38	-0.06	0.01	**0.88**	0.09	-0.11	-0.28	-0.05
imin.mean	0.54	0.09	0.13	-0.27	-0.21	-0.02	-0.75	-0.06	0.07	0.46
imax.mean	**0.93**	0.13	0.06	0.19	0.0	0.1	-0.16	0.06	0.03	**0.9**
imean.mean	**0.99**	0.05	0.12	0.06	-0.02	0.07	-0.48	-0.01	0.1	**0.87**
tndur.sd	0.00	0.12	-0.1	**0.87**	-0.04	0.05	0.04	**0.87**	-0.04	0.03
pmin.sd	0.01	-0.03	**0.83**	-0.07	0.16	-0.06	0.02	-0.09	**0.87**	0.02
pmax.sd	-0.12	**-0.78**	-0.09	-0.17	0.08	**-0.67**	0.01	-0.1	0	-0.05
pmean.sd	-0.09	-0.04	0.13	-0.16	0.27	0.11	0.11	-0.3	0.3	0.03
psd.sd	-0.20	-0.42	-0.01	-0.28	0.21	-0.27	0.06	-0.33	0.12	-0.01
imin.sd	0.07	0.00	0.02	0.02	0.36	0.0	0.36	0.03	0.14	0.02
imax.sd	**-0.65**	-0.08	0.07	-0.06	**0.66**	0.0	**0.87**	-0.12	-0.07	-0.35
imean.sd	-0.27	-0.09	0.12	-0.01	**0.95**	0.04	**0.97**	-0.14	-0.11	-0.12
Var Explained	0.18	0.16	0.13	0.13	0.10	0.17	0.17	0.13	0.13	0.12

Table 1: *Factor Analysis Loadings Matrices*

of the speed date along with text files of the dates annotations. On average, each date contains 812 words, with an average of 406 words per speaker. These worded are divided into an average of 93 turns per date where the speaker changes to the other.

3.1. Feature Extraction

We use OpenSMILE to extract Prosodic features from audio files of the speed date. Lexical features are extracted from the transcripts with the help of the LIWC dictionary.

3.1.1. Prosodic Features

Prosodic features describe the speech of the speaker. In our model, we include features such as *F0, intensity, RMS Amplitute, turn duration* and *pitch*. For each of these different features, we calculate the *min, max, standard deviation (SD), range, sd sd* and *mean*.

3.1.2. Lexical Features

Using LIWC, introduced by [2], we extract *hedge words* like *sort of* and *I guess* that indicate uncertainty. We also gather *meta* words that represent common words in the given speed date scenario and *academic* words like *PhD* and *research*. Finally, we record occurrences of discourse markers.

3.1.3. LIWC Features

We use the LIWC software to classify words into specific topic groups like *love, hate, food, negate* and *drink*. We count the word occurrences within each speed date for each of these word topics.

3.1.4. Accomodation Features

[3], [4] and [5] demonstrate through their work the convergence of vocal intensities between interlocutors. To capture this convergence, we extract features from a speaker that accommodate the previous speaker's speech. These features include the *Rate of Speech* of the two speakers over time, the number of *functional* and *content* words also used in the other's previous turn and *laughter* that immediately proceeds the other's laugh.

3.2. Feature Normalization

Before we use the features in our model, we normalize the relevant lexical feature with the total word count in the conversation. We then normalize all the features such that they all have zero mean and unit variance.

4. Exploratory Analysis

Before attempting to classify speakers as funny or not, we examine the underlying dimensions along which the data varies. To do, so we use exploratory factor analysis. To determine the number of factors, we use a scree test and various non-graphical measures, including parallel analysis and an optimal coordinates test, as described in [14]. We find an optimal number of factors $k = 5$ for both males and females, conducted separately. Figure 1 shows the two scree plots.

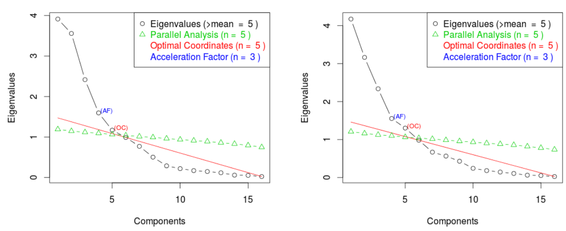

Figure 1: *Determining Number of Factors*

Interestingly, although we find that the various factors for male and female speech are similar, they explain different amounts of variation in the data. For males, the first factor reflects high intensity values and low intensity variation, explaining 18% of the variation in the male prosodic data. For females,

Classification Type			Funny				Courteous			
Gender	Participant	Predictors	SVM RBF	SVM Lin	AdaBoost	L1SVM	SVM RBF	SVM Lin	AdaBoost	L1SVM
Male	Self	Prosody	0.71	0.65	0.67	0.67	0.75	0.71	0.71	0.68
Male	Other	Prosody	0.64	0.69	0.65	0.67	0.63	0.66	0.64	0.63
Male	Both	Prosody	0.70	0.69	0.66	0.65	0.71	0.68	0.70	0.69
Male	Self	Lexical	0.64	0.59	0.59	0.61	0.63	0.61	0.64	0.63
Male	Other	Lexical	0.57	0.55	0.6	0.58	0.63	0.55	0.60	0.64
Male	Both	Lexical	0.63	0.57	0.61	0.55	0.65	0.61	0.65	0.65
Male	Both	Both	0.74	0.73	0.68	0.67	0.76	0.65	0.70	0.69
Female	Self	Prosody	0.63	0.53	0.59	0.57	0.62	0.60	0.61	0.59
Female	Other	Prosody	0.60	0.56	0.62	0.60	0.66	0.65	0.62	0.61
Female	Both	Prosody	0.64	0.59	0.61	0.57	0.68	0.63	0.66	0.62
Female	Self	Lexical	0.55	0.53	0.55	0.54	0.63	0.56	0.58	0.61
Female	Other	Lexical	0.63	0.63	0.63	0.65	0.58	0.55	0.55	0.57
Female	Both	Lexical	0.63	0.63	0.62	0.62	0.57	0.55	0.56	0.59
Female	Both	Both	0.64	0.59	0.63	0.58	0.60	0.60	0.63	0.60

Table 2: *Classification Results.*

the fifth and final factor had very similar loadings, only explaining 12% of the variation in the data. For females, variation in intensity was one of the most important factors, accounting for 17% of variation in the data, whereas for males, variation in intensity only accounted for 10% of the variation in the data. It may be interesting in the future to investigate how sources of variation in prosody differ across genders and contexts.

5. Evaluation Approach

5.1. Methodology

We divide the dataset into two separate groups based on gender. For each gender, we train a classifier for detecting the perception of funny and courteous characteristics.

We create a training set for a given characteristic for a given gender by splitting the dataset into deciles. We use the top decile as positive and the bottom decile as negative examples.

5.2. Classifiers

We start by using an SVM with a linear kernel. To improve our classifications, we next enforce an L1 norm on the SVM to compensate for the large number of lexical and prosodic features we extract. In case our decision boundaries are non-linear, we also train a RBF kernelized SVM. However, we lose interpretability of the features by using an RBF. So, we train a fourth Adaboost Classifier with decision stumps as weak classifiers to maintain interpretability while also allowing for a non-linear decision boundary.

6. Classification Results

First, we run our four classifiers on our prosodic features only. Despite only looking at acoustic features of these conversations, we achieve promising results – an accuracy of 75% for courteousness and 71% for funniness. These results are shown in Table 2. Interestingly, when attempting prediction from prosody alone, we find that, for male evaluators, prediction based only on the evaluator's speech features outperforms prediction based only on the interlocutor's prosodic features ($p < 0.01$, paired t-test). This may be because our evaluator *reveals* perceptions of the other person through their prosody. We do find, however, that the interlocutor's speech is also informative. In addition, we find that it is far easier to use prosody and lexical features to predict males' evaluation of women than it is to predict women's evaluation of males, as for each classification task, our prediction accuracies are significantly greater than those of the female evaluator counterpart.

Beyond these two results (that self features are more informative than other features for males and that male evaluators are easier to predict than female evaluators), we may consider that male prosodic features are more informative about our labels than female prosodic features. We see that with male evaluators, self features are always more informative than other evaluators, but with female evaluators, other features are sometimes more informative. For example, when we predict females' evaluation of males' funniness using the females' lexical features, we achieve an accuracy of only 0.55; however, when we use males' lexical features, we get an accuracy of 0.65. Here, the males' features are more informative ($p < 0.01$, paired t-test).

Furthermore, we see that in general prosodic features outperform lexical features, and there is not a huge improvement from independently using prosodic features or lexical features to using both. This may be due to overfitting in the number of features included – we note that the L1-penalized SVM improves for the inclusion of more features, even if it is still outperformed by the nonlinear classifiers. Data reduction, such as that achieved by stacked autoencoders, may be desirable for improved performance from lexical features.

7. Analysis

7.1. Adaboost Weights

By limiting the depth of the trees in Adaboost to 1, we are able to unravel the top features that were used by each tree in Adaboost. We aggregate these top features by summing across the weights associated with each tree split to capture the features with the highest weights.

Since Adaboost learns different classifier each time, we perform a 10 fold feature aggregation and averaging the sum of weights for each feature split. We perform this aggregation by training on lexical and prosodic features independently. Table 4 shows the top features that adaboost splits on. In the table,

#	Funny		Courteous	
	Lexical	Prosodic	Lexical	Prosodic
1	Other hear	Other imin SD	syllables per turn	imax mean
2	Other youKnow	Other pslope mean	probably	Other prange mean
3	Other see	tndur SD	like	Other pmin SD
4	Other achieve	Other tndur SD	relativ	ptmax SD
5	Other LaughAcc	Other pslnjmp mean	Other you know	itmin SD
6	syllables per turn	tndur mean	humans	Other pslnjmp mean
7	program	pquan mean	cogmech	Other.imin.SD
8	Other past	imax mean	sad	pquan mean
9	program	Other tndur mean	Other anxious	ptmin SD
10	Other pronoun	Other pmax mean	i think	iquan mean

Table 3: *Feature Splits with Highest Weights in Adaboost.*

Other refers to the listener's features.

7.1.1. Funny

Based on the results for funny, words related the other person using words in the LIWC categories *hear, youKnow, see* and *achieve* are good indicators that the other person finds the speaker funny. The *achieve* category correspond to words like ẃint́hat has been expressed by the listener after the speaker delivers a joke. Words corresponding to *hear* and *see* demonstrate that the listener it paying attention and acknowledging the speaker's speech. The 5^{th} feature is laughter accommodation where the listener laughs in reaction to the speaker's laugh.

The prosodic features for funny also correspond well with the lexical features as shorter *turn duration*, which is the 4^{th} and 5^{th} on the prosodic features list, between the two people usually results in more words in the LIWC *see* and *hear* categories of acknowledgement. These features also show that a high intensity min SD from the listener and a high pitch slope mean are good indicators of funny.

Interestingly, we find that the relevant features vary by gender, as shown in Table 5. We find that females respond to males varying pitch and intensity, whereas males respond to females with high *average* pitch and high *average* intensity. In addition, males' own pitch is very important in predicting their evaluations of funniness. Hedge words are important lexical predictors for both genders. When males utter many exclamations, they find the female funny. Females respond to accommodating laughter from their male partners.

7.1.2. Courteous

Discourse words like *like, probably and I think* are good lexical indicators for detecting courteous. The honesty weakening of there sentences by using such words makes them appear more courteous. Also, talking about *humans* and *cognitive processes* are important. The use of anxious and *fillers* like *you know* are the only features of the listener that have a high weight.

Low intensity max mean with small pitch range SD and pitch min SD are important prosodic features. This implies that large changes in speech are perceived as non being courteous.

Again, the relevant features vary by gender, as shown in Table 5. We see that females respond strongly to male accommodation features – clearly it makes intuitive sense that males with high values for accommodation features will be judged to be more courteous. Males respond strongly to variation in intensity and pitch from the female, just as they did for funniness.

7.1.3. Comparing Funny and Courteous

It is interesting to note that the speaker's features are the most important when determining whether the speaker will be perceived as courteous. On the other hand, the listeners features are more important for detecting funny.

8. Future Work

The model we build has a lot of features and reducing dimensionality of this data by learning hidden variables from an sparse auto-encoder (particularly for the lexical features, as in [6]) might give us more insight into how we can create a shared representation of the lexical with the prosodic features. In addition, applying simultaneous dimension reduction techniques, such as Canonical Correlations Analysis, across audio, visual and text modes may provide us with interesting insights as to how these media interact.

We have found here that in general, the evaluator's speech is more informative than the other participant's. In addition, we have found that, in general, male speech is more informative about males' evaluations of females than female speech is about females' evaluations of men. In the future, we would like to be able to determine the magnitude of these two effects – it is difficult to tease apart these effects using the methodologies employed in this paper.

The SpeedDate dataset we use also contains ratings of people's intelligence, sincerity and other characteristics. An interesting area to explore would find correlations between these characteristics and their corresponding prosodic and lexical features.

Adding another source of features, such as video from video speed dates is another extension to this project. It would be insightful to see how the contribution of the current features will change if visual features are also added.

9. Conclusion

In this paper, we evaluate the perceptions of two separate characteristics, funny and courteous, in short meetings like speed dates. Our initial factor analysis of the features show clear distinctions between variations captured amongst males and females. We use different classifiers and obtain a 74% and 64% accuracy when predicting funny for males and females respectively. Furthermore, we attain 76% and 63% accuracy when predicting courteous for the same genders. Again, while analyzing the features that correspond to the highest weights in

#	Funny				Courteous			
	Lexical		Prosodic		Lexical		Prosodic	
	Male	Female	Male	Female	Male	Female	Male	Female
1	excl	Other LaughAcc	pquan mean	Other tndur SD	syll turn	like	pquan mean	Other prange mean
2	pronoun	Other hear	Other pmean mean	Other imin SD	probably	syll turn	iquan SD	Other pslnjmp mean
3	Other achieve	Other youKnow	Other pquan mean	Other iquan mean	Other anx	Other LaughAcc	Other imin SD	imean SD
4	Other see	kindOf	iquan mean	pquan SD	Other like	Other FuncAcc	imax mean	Other pslope mean
5	achieve	Other program	Other pslope mean	Other pmean SD	aLittle	sad	Other pmin SD	ptmin SD
6	Other youKnow	like	Other imin SD	pmean mean	pronoun	Other youKnow	Other imean SD	ptmax SD
7	Other like	Other iThink	Other prange SD	tndur SD	they	Other friend	imean mean	pquan mean
8	Other you	syll turn	Other pslnjmp mean	pmean SD	humans	relativ	itmin SD	Other imin mean
9	shehe	Other see	tndur mean	Other tndur mean	work	posemo	imin mean	imax SD
10	cogmech	Other affect	imin SD	Other ptmax mean	shehe	Other future	imin SD	imax mean

Table 4: *Gender Based Feature Splits with Highest Weights in Adaboost.*

our classifiers, we see a distinct different between males and females. Models that train using one gender perform better than a model that uses features from both genders. Finally, we see that male prosody tend to be more revealing about their perceptions than female prosody.

10. Acknowledgements

We thank Dan Jurafsky for giving us access to the Speed Dating Dataset and for his continued guidance during this project.

11. References

[1] Ranganatha, R., Jurafsky, D., McFarland, D.A., “Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates“, in Computer Speech & Language Volume 27, Issue 1, Pages 89-115, 2013

[2] Pennebaker, J. W., Booth, R., Francis, M., “Linguistic inquiry and word count: LIWC2007“ operators manual. Tech. rep., University of Texas, 2007

[3] Natale, M., “Convergence of mean vocal intensity in dyadic communication as a function of social desirability“, in Journal of Personality and Social Psychology 32 (5), 790-804, 1975

[4] Ireland, M.E., Slatcher, R.B., Eastwick, P.W., Scissors, L.E., Finkel, E.J., Pennebaker, J.W., “Language style matching predicts relationship initiation and stability“, in Psychological Science 22 (1), 39, 2011

[5] Levitan, R., Hirschberg, J., “Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions“, in ACL 2011

[6] Ranganath, R., Jurafsky, D., McFarland, D. “It’s Not You, It’s Me”, in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.

[7] Lee, C., Narayanan, S. and Pieraccini, R. “Combining acoustic and language information for emotion recognition.” INTERSPEECH. 2002.

[8] Ang, J. , et al. ”Prosody-based automatic detection of annoyance and frustration in human-computer dialog.” INTERSPEECH. 2002.

[9] Menninghaus, W., et al. ”Sounds funny? Humor effects of phonological and prosodic figures of speech.” Psychology of Aesthetics, Creativity, and the Arts 8.1 (2014): 71.

[10] Purandare, A., and Litman, D. ”Humor: Prosody analysis and automatic recognition for f* r* i* e* n* d* s*.” Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2006.

[11] Feinberg, David R., Jones, B. Little, A., D. Michael Burt, and Perrett, D. 2005. ”Manipulations of Fundamental and Formant Frequencies Inuence the Attractiveness of Human Male Voices.” Animal Behaviour 69 pp: 561-568.

[12] Mairesse, F. et al. ”Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text.” J. Artif. Intell. Res.(JAIR) 30 (2007): 457-500.

[13] Gravano, A., et al. ”Acoustic and prosodic correlates of social behavior.” (2011).

[14] Hayton, J., Allen, D., and Scarpello, V.. ”Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis.” Organizational research methods 7.2 (2004): 191-205.