# Extracting Social Meaning: Identifying Interactional Style in Spoken Conversation

**Dan Jurafsky**
Linguistics Department
Stanford University
jurafsky@stanford.edu

**Rajesh Ranganath**
Computer Science Department
Stanford University
rajeshr@cs.stanford.edu

**Dan McFarland**
School of Education
Stanford University
dmcfarla@stanford.edu

## Abstract

Automatically extracting social meaning and intention from spoken dialogue is an important task for dialogue systems and social computing. We describe a system for detecting elements of **interactional style**: whether a speaker is *awkward*, *friendly*, or *flirtatious*. We create and use a new spoken corpus of 991 4-minute speed-dates. Participants rated their interlocutors for these elements of style. Using rich dialogue, lexical, and prosodic features, we are able to detect flirtatious, awkward, and friendly styles in noisy natural conversational data with up to 75% accuracy, compared to a 50% baseline. We describe simple ways to extract relatively rich dialogue features, and analyze which features performed similarly for men and women and which were gender-specific.

## 1 Introduction

How can we extract social meaning from speech, deciding if a speaker is particularly engaged in the conversation, is uncomfortable or awkward, or is particularly friendly and flirtatious? Understanding these meanings and how they are signaled in language is an important sociolinguistic task in itself. Extracting them automatically from dialogue speech and text is crucial for developing socially aware computing systems for tasks such as detection of interactional problems or matching conversational style, and will play an important role in creating more natural dialogue agents (Pentland, 2005; Nass and Brave, 2005; Brave et al., 2005).

Cues for social meaning permeate speech at every level of linguistic structure. Acoustic cues such as low and high F0 or energy and spectral tilt are important in detecting emotions such as *annoyance*, *anger*, *sadness*, or *boredom* (Ang et al., 2002; Lee and Narayanan, 2002; Liscombe et al., 2003), speaker characteristics such as *charisma* (Rosenberg and Hirschberg, 2005), or *personality* features like extroversion (Mairesse et al., 2007; Mairesse and Walker, 2008). Lexical cues to social meaning abound. Speakers with links to depression or speakers who are under stress use more first person singular pronouns (Rude et al., 2004; Pennebaker and Lay, 2002; Cohn et al., 2004), positive emotion words are cues to agreeableness (Mairesse et al., 2007), and negative emotion words are useful cues to deceptive speech (Newman et al., 2003). The number of words in a sentence can be a useful feature for extroverted personality (Mairesse et al., 2007). Finally, dialog features such as the presence of disfluencies can inform listeners about speakers' problems in utterance planning or about confidence (Brennan and Williams, 1995; Brennan and Schober, 2001).

Our goal is to see whether cues of this sort are useful in detecting particular elements of conversational style and social intention; whether a speaker in a speed-dating conversation is judged by the interlocutor as *friendly*, *awkward*, or *flirtatious*.

## 2 The Corpus

Our experiments make use of a new corpus we have collected, the SpeedDate Corpus. The corpus is based on three speed-dating sessions run at an elite

private American university in 2005 and inspired by prior speed-dating research (Madan et al., 2005; Pentland, 2005). The graduate student participants volunteered to be in the study and were promised emails of persons with whom they reported mutual liking. Each date was conducted in an open setting where there was substantial background noise. All participants wore audio recorders on a shoulder sash, thus resulting in two audio recordings of the approximately 1100 4-minute dates. In addition to the audio, we collected pre-test surveys, event scorecards, and post-test surveys. This is the largest sample we know of where audio data and detailed survey information were combined in a natural experiment.

The rich survey information included date perceptions and follow-up interest, as well as general attitudes, preferences, and demographic information. Participants were also asked about the conversational style and intention of the interlocutor. Each speaker was asked to report how often their date's speech reflected different conversational styles (awkward, friendly, flirtatious, funny, assertive) on a scale of 1-10 (1=never, 10=constantly): "How often did the other person behave in the following ways on this 'date'?". We chose three of these five to focus on in this paper.

We acquired acoustic information by taking the acoustic wave file from each recorder and manually segmenting it into a sequence of wavefiles, each corresponding to one 4-minute date. Since both speakers wore microphones, most dates had two recordings, one from the male recorder and one from the female recorder. Because of mechanical, operator, and experimenter errors, some recordings were lost, and thus some dates had only one recording. Transcribers at a professional transcription service used the two recordings to create a transcript for each date, and time-stamped the start and end time of each speaker turn. Transcribers were instructed to mark various disfluencies as well as some nonverbal elements of the conversation such as laughter.

Because of noise, participants who accidentally turned off their mikes, and some segmentation and transcription errors, a number of dates were not possible to analyze. 19 dates were lost completely, and for an additional 130 we lost one of the two audio tracks and had to use the remaining track to extract features for both interlocutors. The current study focuses on the 991 remaining clean dates for which we had usable audio, transcripts, and survey information.

## 3 The Experiments

Our goal is to detect three of the style variables, in particular awkward, friendly, or flirtatious speakers, via a machine learning classifier. Recall that each speaker in a date (each conversation side) was labeled by his or her interlocutor with a rating from 1-10 for awkward, friendly, or flirtatious behavior. For the experiments, the 1-10 Likert scale ratings were first mean-centered within each respondent so that the average was 0. Then the top ten percent of the respondent-centered meaned Likert ratings were marked as positive for the trait, and the bottom ten percent were marked as negative for a trait. Thus each respondent labels the other speaker as either positive, negative, or NA for each of the three traits.

We run our binary classification experiments to predict this output variable.

For each speaker side of each 4-minute conversation, we extracted features from the wavefiles and the transcript, as described in the next section. We then trained six separate binary classifiers (for each gender for the 3 tasks), as described in Section 5.

## 4 Feature Extraction

In selecting features we drew on previous research on the use of relatively simple surface features that cue social meaning, described in the next sections.

Each date was represented by the two 4-minute wavefiles, one from the recorder worn by each speaker, and a single transcription. Because of the very high level of noise, the speaker wearing the recorder was much clearer on his/her own recording, and so we extracted the acoustic features for each speaker from their own microphone (except for the 130 dates for which we only had one audio file). All lexical and discourse features were extracted from the transcripts.

All features describe the speaker of the conversation side being labeled for style. The features for a conversation side thus indicate whether a speaker who talks a lot, laughs, is more disfluent, has higher F0, etc., is more or less likely to be considered flirtatious, friendly, or awkward by the interlocutor. We

also computed the same features for the *alter* interlocutor. *Alter* features thus indicate the conversational behavior of the speaker talking with an interlocutor they considered to be flirtatious, friendly, or awkward.

### 4.1 Prosodic Features

F0 and RMS amplitude features were extracted using Praat scripts (Boersma and Weenink, 2005). Since the start and end of each turn were time-marked by hand, each feature was easily extracted over a turn, and then averages and standard deviations were taken over the turns in an entire conversation side. Thus the feature F0 MIN for a conversation side was computed by taking the F0 min of each turn in that conversation side (not counting zero values of F0), and then averaging these values over all turns in the side. F0 MIN SD is the standard deviation across turns of this same measure.

Note that we coded four measures of f0 variation, not knowing in advance which one was likely to be the most useful: F0 MEAN SD is the deviation across turns from the global F0 mean for the conversation side, measuring how variable the speakers mean f0 is across turns. F0 SD is the standard deviation within a turn for the f0 mean, and then averaged over turns, hence measures how variable the speakers f0 is within a turn. F0 SD SD measures how much the within-turn f0 variance varies from turn to turn, and hence is another measure of cross-turn f0 variation. PITCH RANGE SD measures how much the speakers pitch range varies from turn to turn, and hence is another measure of cross-turn f0 variation.

| | |
|---|---|
| F0 MIN | minimum (non-zero) F0 per turn, averaged over turns |
| F0 MIN SD | standard deviation from F0 min |
| F0 MAX | maximum F0 per turn, averaged over turns |
| F0 MAX SD | standard deviation from F0 max |
| F0 MEAN | mean F0 per turn, averaged over turns |
| F0 MEAN SD | standard deviation (across turns) from F0 mean |
| F0 SD | standard deviation (within a turn) from F0 mean, averaged over turns |
| F0 SD SD | standard deviation from the f0 sd |
| PITCH RANGE | f0 max - f0 min per turn, averaged over turns |
| PITCH RANGE SD | standard deviation from mean pitch range |
| RMS MIN | minimum amplitude per turn, averaged over turns |
| RMS MIN SD | standard deviation from RMS min |
| RMS MAX | maximum amplitude per turn, averaged over turns |
| RMS MAX SD | standard deviation from RMS max |
| RMS MEAN | mean amplitude per turn, averaged over turns |
| RMS MEAN SD | standard deviation from RMS mean |
| TURN DUR | duration of turn in seconds, averaged over turns |
| TIME | total time for a speaker for a conversation side, in seconds |
| RATE OF SPEECH | number of words in turn divided by duration of turn in seconds, averaged over turns |

Table 1: Prosodic features for each conversation side, extracted using Praat from the hand-segmented turns of each side.

### 4.2 Lexical Features

Lexical features have been widely explored in the psychological and computational literature. For these features we drew mainly on the LIWC lexicons of Pennebaker et al. (2007), the standard for social psychological analysis of lexical features. From the large variety of lexical categories in LIWC we selected ten that the previous work of Mairesse et al. (2007) had found to be very significant in detecting personality-related features. The 10 LIWC features we used were *Anger, Assent, Ingest, Insight, Negemotion, Sexual, Swear, I, We,* and *You*. We also added two new lexical features, "past tense auxiliary", a heuristic for automatically detecting narrative or story-telling behavior, and *Metadate*, for discussion about the speed-date itself. The features are summarized in Table 2.

### 4.3 Dialogue Act and Adjacency Pair Features

A number of discourse features were extracted, drawing from the conversation analysis, disfluency and dialog act literature (Sacks et al., 1974; Jurafsky et al., 1998; Jurafsky, 2001). While discourse features are clearly important for extracting social meaning, previous work on social meaning has met with less success in use of such features (with the exception of the 'critical segments' work of (Enos et al., 2007)), presumably because discourse fea-

| | |
|---|---|
| TOTAL WORDS | total number of words |
| PAST TENSE | uses of past tense auxiliaries *was, were, had* |
| METADATE | *horn, date, bell, survey, speed, form, questionnaire, rushed, study, research* |
| YOU | *you, you'd, you'll, your, you're, yours, you've* (not counting *you know*) |
| WE | *lets, let's, our, ours, ourselves, us, we, we'd, we'll, we're, we've* |
| I | *I'd, I'll, I'm, I've, me, mine, my, myself* (not counting *I mean*) |
| ASSENT | *yeah, okay, cool, yes, awesome, absolutely, agree* |
| SWEAR | *hell, sucks, damn, crap, shit, screw, heck, fuck** |
| INSIGHT | *think*/thought, feel*/felt, find/found, understand*, figure*, idea*, imagine, wonder* |
| ANGER | *hate/hated, hell, ridiculous*, stupid, kill*, screwed, blame, sucks, mad, bother, shit* |
| NEGEMOTION | *bad, weird, hate, crazy, problem*, difficult, tough, awkward, boring, wrong, sad, worry,* |
| SEXUAL | *love*, passion*, loves, virgin, sex, screw* |
| INGEST | *food, eat*, water, bar/bars, drink*, cook*, dinner, coffee, wine, beer, restaurant, lunch, dish* |

Table 2: Lexical features. Each feature value is a total count of the words in that class for each conversation side; asterisks indicate that suffixed forms were included (e.g., *love, loves, loving*). All except the first three are from LIWC (Pennebaker et al., 2007) (modified slightly, for example by removing *you know* and *I mean*). The last five classes include more words in addition to those shown.

tures are expensive to hand-label and hard to automatically extract. We chose a suggestive discourse features that we felt might still be automatically extracted.

Four particular dialog acts were chosen as shown in Table 3. *Backchannels* (or continuers) and *appreciations* (a continuer expressing positive affect) were coded by hand-built regular expressions. The regular expressions were based on analysis of the backchannels and appreciations in the hand-labeled Switchboard corpus of dialog acts (Jurafsky et al., 1997). Questions were coded simply by the presence of question marks.

Finally, *repair questions* (also called NTRIs; next turn repair indicators) are turns in which a speaker signals lack of hearing or understanding (Schegloff et al., 1977). To detect these, we used a simple heuristic: the presence of 'Excuse me' or 'Wait', as in the following example:

FEMALE: Okay. Are you excited about that?
MALE: Excuse me?

A *collaborative completion* is a turn where a speaker completes the utterance begun by the alter (Lerner, 1991; Lerner, 1996). Our heuristic for identifying collaborative completions was to select sentences for which the first word of the speaker was extremely predictable from the last two words of the previous speaker. We trained a word trigram model[1] and used it to compute the probability $p$ of the first word of a speaker's turn given the last two words of the interlocutor's turn. We arbitrarily chose the threshold .01, labeling all turns for which $p > .01$ as collaborative completions and used the total number of collaborative completions in a conversation side as our variable. This simple heuristic was errorful, but did tend to find completions beginning with *and* or *or* (1 below) and wh-questions followed by an NP or PP phrase that is grammatically coherent with the end of the question (2 and 3):

(1) FEMALE: The driving range.
(1) MALE: And the tennis court, too.

(2) MALE: What year did you graduate?
(2) FEMALE: From high school?

(3) FEMALE: What department are you in?
(3) MALE: The business school.

We also marked aspects of the *preference structure* of language. A *dispreferred* action is one in which a speaker avoids the face-threat to the interlocutor that would be caused by, e.g., refusing a request or not answering a question, by using specific strategies such as the use of *well*, hesitations, or restarts (Schegloff et al., 1977; Pomerantz, 1984).

Finally, we included the number of instances of laughter for the side, as well as the total number of turns a speaker took.

### 4.4 Disfluency Features

A second group of discourse features relating to repair, disfluency, and speaker overlap are summarized

[1] interpolated, with Good Turing smoothing, trained on the Treebank 3 Switchboard transcripts after stripping punctuation.

| | |
|---|---|
| BACKCHANNELS | number of backchannel utterances in side (*Uh-huh.*, *Yeah.*, *Right.*, *Oh, okay.*) |
| APPRECIATIONS | number of appreciations in side (*Wow*, *That's true*, *Oh, great*) |
| QUESTIONS | number of questions in side |
| NTRI | repair question (Next Turn Repair Indicator) (*Wait*, *Excuse me*) |
| COMPLETION | (an approximation to) utterances that were 'collaborative completions' |
| DISPREFERRED | (an approximation to) dispreferred responses, beginning with discourse marker *well* |
| LAUGH | number of instances of laughter in side |
| TURNS | total number of turns in side |

Table 3: Dialog act/adjacency pair features.

in Table 4. Filled pauses (*um*, *uh*) were coded by

| | |
|---|---|
| UH/UM | total number of filled pauses (*uh* or *um*) in conversation side |
| RESTART | total number of disfluent restarts in conversation side |
| OVERLAP | number of turns in side which the two speakers overlapped |

Table 4: Disfluency features

regular expressions (the transcribers had been instructed to transcribe all filled pauses). Restarts are a type of repair in which speakers begin a phrase, break off, and then restart the syntactic phrase. The following example shows a restart; the speaker starts a sentence *Uh, I* and then restarts, *There's a group...*:

Uh, I–there's a group of us that came in–

Overlaps are cases in which both speakers were talking at the same time, and were marked by the transcribers in the transcripts:

| | |
|---|---|
| MALE: | But-and also obviously– |
| FEMALE: | It sounds bigger. |
| MALE: | –people in the CS school are not quite as social in general as other– |

## 5 Classifier Training

Before performing the classification task, we preprocessed the data in two ways. First, we standardized all the variables to have zero mean and unit variance. We did this to avoid imposing a prior on any of the features based on their numerical values.[2] Second, we removed features correlated greater than .7. One goal of removing correlated features was to remove as much colinearity as possible from the regression so that the regression weights could be ranked for their importance in the classification. In addition, we hoped to improve classification because a large number of features require more training examples (Ng, 2004). For example for male flirt we removed *f0 range* (highly correlated with *f0 max*), *f0 min sd* (highly correlated with *f0 min*), and *Swear* (highly correlated with *Anger*).

For each classification task due to the small amounts of data we performed $k$-fold cross validation to learn and evaluate our models. We used a variant of $k$-fold cross validation with five folds where three folds are used for training, one fold is used for validation, and one fold is used as a test set. This test fold is not used in any training step. This yields a datasplit of 60% for training, 20% for validation, and 20% for testing, or 120 training examples, 40 validation examples, and 40 test examples. To ensure that we were not learning something specific to our data split, we randomized our data ordering and repeated the $k$-fold cross validation variant 25 times.

We used regularized logistic regression for classification. Recall that in logistic regression we train a vector of feature weights $\theta \in \mathbb{R}^n$ so as to make the following classification of some output variable $y$ for an input observation $x$:[3]

$$p(y|x;\theta) = \frac{1}{1+\exp(-\theta^T x)} \quad (1)$$

In regularized logistic regression we find the

[2]Consider a feature A with mean 100 and a feature B with mean .1 where A and B are correlated with the output. Since regularization favors small weights there is a bias to put weight on feature A because intuitively the weight on feature B would need to be 1000 times larger to carry the same effect. This argument holds similarly for the reduction to unit variance.

[3]Where $n$ is the number of features plus 1 for the intercept.

weights $\theta$ which maximize the following optimization problem:

$$\underset{\theta}{\operatorname{argmax}} \sum_i \log p(y^i|x^i;\theta) - \alpha * R(\theta) \quad (2)$$

$R(\theta)$ is a regularization term used to penalize large weights. We chose $R(\theta)$, the regularization function, to be the $L_1$ norm of $\theta$. That is, $R(\theta) = ||\theta||_1 = \sum_{i=1}^{n} |\theta_i|$.

In our case, given the training set $S_{\text{train}}$, test set $S_{\text{test}}$, and validation set $S_{\text{val}}$, we trained the weights $\theta$ as follows:

$$\underset{\alpha}{\operatorname{argmax}} \text{accuracy}(\theta_\alpha, S_{\text{val}}) \quad (3)$$

where for a given sparsity parameter $\alpha$

$$\theta_\alpha = \underset{\theta}{\operatorname{argmax}} \sum_i \log p(y^i|x^i;\theta) - \alpha * R(\theta) \quad (4)$$

We chose $L_1$-regularization because the number of training examples to learn well grows logarithmically with the number of input variables (Ng, 2004), and to achieve a sparse activation of our features to find only the most salient explanatory variables. This choice of regularization was made to avoid the problems that often plague supervised learning in situations with large number of features but only a small number of examples. The search space over the sparsity parameter $\alpha$ is bounded around an expected sparsity to prevent overfitting.

Finally, to evaluate our model on the learned $\alpha$ and $\theta_\alpha$ we used the features $X$ of the test set $S_{test}$ to compute the predicted outputs $Y$ using the logistic regression model. Accuracy is simply computed as the percent of correct predictions.

To avoid any data ordering bias, we calculated a $\theta_\alpha$ for each randomized run. The output of the runs was a vector of weights for each feature. We kept any feature if the median of its weight vector was nonzero.[4] A sample boxplot for the highest weighted ego features for predicting male flirt can be found in Figure 1.

[4]We also performed a t-test to find salient feature values significantly different than zero; the non-zero median method turned out to be a more conservative measure in practice (intuitively, because $L_1$ normed regression pushes weights to 0).

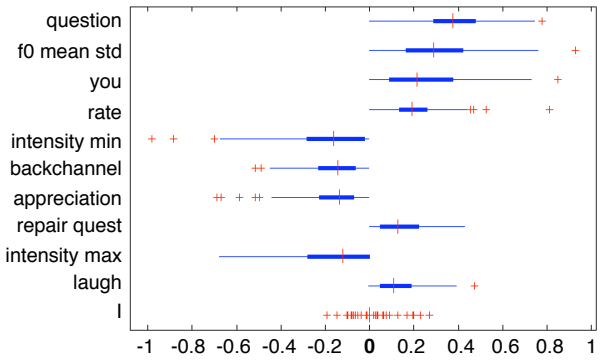


Figure 1: An illustrative boxplot for flirtation in men showing the 10 most significant features and one not significant ('I'). Shown are median values (central red line), first quartile, third quartile, outliers (red X's) and interquartile range (filled box).

## 6 Results

Results for the 6 binary classifiers are presented in Table 5.

| | Awk | | Flirt | | Friendly | |
|---|---|---|---|---|---|---|
| | M | F | M | F | M | F |
| Speaker | 63% | 51% | 67% | 60% | 72% | 68% |
| +other | 64% | 64% | 71% | 60% | 73% | 75% |

Table 5: Accuracy of binary classification of each conversation side, where chance is 50%. The first row uses features only from the single speaker; the second adds all the features from the interlocutor as well. These accuracy results were aggregated from 25 randomized runs of 5-fold cross validation.

The first row shows results using features extracted from the speaker being labeled. Here, all conversational styles are easiest to detect in men.

The second row of table 5 shows the accuracy when using features from both speakers. Not surprisingly, adding information about the interlocutor tends to improve classification, and especially for women, suggesting that male speaking has greater sway over perceptions of conversational style. We discuss below the role of these features.

We first considered the features that helped classification when considering only the ego (i.e., the results in the first row of Table 5). Table 6 shows feature weights for the features (features were normed so weights are comparable), and is summarized in the following paragraphs:

- Men who are labeled as friendly use *you*, col-

| MALE FRIENDLY | | MALE FLIRT | |
|---|---|---|---|
| backchannel | -0.737 | question | 0.376 |
| you | 0.631 | f0 mean sd | 0.288 |
| intensity min sd | 0.552 | you | 0.214 |
| f0 sd sd | -0.446 | rate | 0.190 |
| intensity min | -0.445 | intensity min | -0.163 |
| completion | 0.337 | backchannel | -0.142 |
| time | -0.270 | appreciation | -0.136 |
| Insight | -0.249 | repair question | 0.128 |
| f0 min | -0.226 | intensity max | -0.121 |
| intensity max | -0.221 | laugh | 0.107 |
| overlap | 0.213 | time | -0.092 |
| laugh | 0.192 | overlap | -0.090 |
| turn dur | -0.059 | f0 min | 0.089 |
| Sexual | 0.059 | Sexual | 0.082 |
| appreciation | -0.054 | Negemo | 0.075 |
| Anger | -0.051 | metadate | -0.041 |
| **FEMALE FRIENDLY** | | **FEMALE FLIRT** | |
| intensity min sd | 0.420 | f0 max | 0.475 |
| intensity max sd | -0.367 | rate | 0.346 |
| completion | 0.276 | intensity min sd | 0.269 |
| repair question | 0.255 | f0 mean sd | 0.21 |
| appreciation | 0.253 | Swear | 0.156 |
| f0 max | 0.233 | question | -0.153 |
| Swear | -0.194 | Assent | -0.127 |
| wordcount | 0.165 | f0 min | -0.111 |
| restart | 0.172 | intensity max | 0.092 |
| uh | 0.241 | I | 0.073 |
| I | 0.111 | metadate | -0.071 |
| past | -0.060 | wordcount | 0.065 |
| laugh | 0.048 | laugh | 0.054 |
| Negemotion | -0.021 | restart | 0.046 |
| intensity min | -0.02 | overlap | -0.036 |
| Ingest | -0.017 | f0 sd sd | -0.025 |
| Assent | 0.0087 | Ingest | -0.024 |
| f0 max sd | 0.0089 | | |
| **MALE AWK** | | | |
| restart | 0.502 | completion | -0.141 |
| f0 sd sd | 0.371 | intensity max | -0.135 |
| appreciation | -0.354 | f0 mean sd | -0.091 |
| turns | -0.292 | Ingest | -0.079 |
| uh | 0.270 | Anger | 0.075 |
| you | -0.210 | repair question | -0.067 |
| overlap | -0.190 | Insight | -0.056 |
| past | -0.175 | rate | 0.049 |
| intensity min sd | -0.173 | | |

Table 6: Feature weights (median weights of the randomized runs) for the non-zero predictors for each classifier. Since our accuracy for detecting awkwardness in women based solely on ego features is so close to chance, we didn't analyze the awkwardness features for women here.

laborative completions, laugh, overlap, but don't backchannel or use appreciations. Their utterances are shorter (in seconds and words) and they are quieter and their (minimum) pitch is lower and somewhat less variable.

• Women labeled as friendly have more collaborative completions, repair questions, laughter, and appreciations. They use more words overall, and use *I* more often. They are more disfluent (both restarts and *uh*) but less likely to swear. Prosodically their f0 is higher, and there seems to be some pattern involving quiet speech; more variation in intensity minimum than intensity max.

• Men who are labeled as flirting ask more questions, including repair questions, and use *you*. They don't use backchannels or appreciations, or overlap as much. They laugh more, and use more sexual and negative emotional words. Prosodically they speak faster, with higher and more variable pitch, but quieter (lower intensity max).

• The strongest features for women who are labeled as flirting are prosodic; they speak faster and louder with higher and more variable pitch. They also use more words in general, swear more, don't ask questions or use Assent, use more I, laugh more, and are somewhat more disfluent (restarts).

• Men who are labeled as awkward are more disfluent, with increased restarts and filled pauses (*uh* and *um*). They are also not 'collaborative' conversationalists; they don't use appreciations, repair questions, collaborative completions, past-tense, or *you*, take fewer turns overall, and don't overlap. Prosodically the awkward labels are hard to characterize; there is both an increase in pitch variation (f0 sd sd) and a decrease (f0 mean sd). They don't seem to get quite as loud (intensity max).

The previous analysis showed what features of the ego help in classification. We next asked about features of the alter, based on the results using both ego and alter features in the second row of Table 5. Here we are asking about the linguistic behaviors of a speaker who describes the interlocutor as flirting, friendly, or awkward.

While we don't show these values in a table, we offer here an overview of their tendencies. For example for women who labeled their male interlocutors as friendly, the women got much quieter, used 'well' much more, laughed, asked more

repair questions, used collaborative completions, and backchanneled more. When a man labeled a woman as friendly, he used an expanded intensity range (quieter intensity min, louder intensity max). laughed more, used more sexual terms, used less negative emotional terms, and overlapped more.

When women labeled their male interlocutor as flirting, the women used many more repair questions, laughed more, and got quieter (lower intensity min). By contrast, when a man said his female interlocutor was flirting, he used more Insight and Anger words, and raised his pitch.

When women labeled their male interlocutor as awkward, the women asked a lot of questions, used *well*, were disfluent (restarts), had a diminished pitch range, and didn't use *I*. In listening to some of these conversations, it was clear that the conversation lagged repeatedly, and the women used questions at these points to restart the conversations.

## 7 Discussion

The results presented here should be regarded with some caution. The sample is not a random sample of English speakers or American adults, and speed dating is not a natural context for expressing every conversational style. Therefore, a wider array of studies across populations and genres would be required before a more general theory of conversational styles is established.

On the other hand, the presented results may under-reflect the relations being captured. The quality of recordings and coarse granularity (1 second) of the time-stamps likely cloud the relations, and as the data is cleaned and improved, we expect the associations to only grow stronger.

Caveats aside, we believe the evidence indicates that the perception of several types of conversational style have relatively clear signals across genders, but with some additional gender contextualization.

Both genders convey flirtation by laughing more, speaking faster, and using higher and more variable pitch. Both genders convey friendliness by laughing more, and using collaborative completions.

However, we do find gender differences; men asl more questions when (labeled as) flirting, women ask fewer. Men labeled as flirting are softer, but women labeled as flirting are louder. Women flirting swear more, while men are more likely to use sexual vocabulary. Gender differences exist as well for the other variables. Men labeled as friendly use *you* while women labeled as friendly use *I*. Friendly women are very disfluent; friendly men are not.

While the features for friendly and flirtatious speech overlap, there are clear differences. Men speaker faster and with higher f0 (min) in flirtatious speech, but not faster and with lower f0 (min) in friendly speech. For men, flirtatious speech involves more questions and repair questions, while friendly speech does not. For women, friendly speech is more disfluent than flirtatious speech, and has more collaborative style (completions, repair questions, appreciations).

We also seem to see a model of *collaborative conversational style* (probably related to the *collaborative floor* of Edelsky (1981) and Coates (1996)), cued by the use of more collaborative completions, repair questions and other questions, *you*, and laughter. These collaborative techniques were used by both women and men who were labeled as friendly, and occurred less with men labeled as awkward. Women themselves displayed more of this collaborative conversational style when they labeled the men as friendly. For women only, collaborative style included appreciations; while for men only, collaborative style included overlaps.

In addition to these implications for social science, our work has implications for the extraction of meaning in general. A key focus of our work was on ways to extract useful dialog act and disfluency features (repair questions, backchannels, appreciations, restarts, dispreferreds) with very shallow methods. These features were indeed extractable and proved to be useful features in classification.

We are currently extending these results to predict date outcomes including 'liking', extending work such as Madan and Pentland (2006).

## Acknowledgments

Thanks to three anonymous reviewers, Sonal Nalkur and Tanzeem Choudhury for assistance and advice on data collection, Sandy Pentland for a helpful discussion about feature extraction, and to Google for gift funding.

## References


J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. 2002. Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog. In *INTERSPEECH-02*.

P. Boersma and D. Weenink. 2005. Praat: doing phonetics by computer (version 4.3.14). [Computer program]. Retrieved May 26, 2005, from `http://www.praat.org/`.

S. Brave, C. Nass, and K. Hutchinson. 2005. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied conversational agent. *International Journal of Human-Computer Studies*, 62(2):161–178.

S. E. Brennan and M. F. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44:274–296.

S. E. Brennan and M. Williams. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34:383–398.

J. Coates. 1996. *Women Talk*. Blackwell.

M. A. Cohn, M. R. Mehl, and J. W. Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, 15:687–693.

C. Edelsky. 1981. Who's got the floor? *Language in Society*, 10:383–421.

F. Enos, E. Shriberg, M. Graciarena, J. Hirschberg, and A. Stolcke. 2007. Detecting Deception Using Critical Segments. In *INTERSPEECH-07*.

D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL Labeling Project Coder's Manual, Draft 13. Technical Report 97-02, University of Colorado Institute of Cognitive Science.

D. Jurafsky, E. Shriberg, B. Fox, and T. Curl. 1998. Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings, COLING-ACL Workshop on Discourse Relations and Discourse Markers*, pages 114–120.

D. Jurafsky. 2001. Pragmatics and computational linguistics. In L. R. Horn and G. Ward, editors, *Handbook of Pragmatics*. Blackwell.

C. M. Lee and S. S. Narayanan. 2002. Combining acoustic and language information for emotion recognition. In *ICSLP-02*, pages 873–876, Denver, CO.

G. H. Lerner. 1991. On the syntax of sentences-in-progress. *Language in Society*, 20(3):441–458.

G. H. Lerner. 1996. On the "semi-permeable" character of grammatical units in conversation: Conditional entry into the turn space of another speaker. In E. Ochs, E. A. Schegloff, and S. A. Thompson, editors, *Interaction and Grammar*, pages 238–276. Cambridge University Press.

J. Liscombe, J. Venditti, and J. Hirschberg. 2003. Classifying Subject Ratings of Emotional Speech Using Acoustic Features. In *INTERSPEECH-03*.

A. Madan and A. Pentland. 2006. Vibefones: Socially aware mobile phones. In *Tenth IEEE International Symposium on Wearable Computers*.

A. Madan, R. Caneel, and A. Pentland. 2005. Voices of attraction. Presented at Augmented Cognition, HCI 2005, Las Vegas.

F. Mairesse and M. Walker. 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. In *ACL-08*, Columbus.

F. Mairesse, M. Walker, M. Mehl, and R. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500.

C. Nass and S. Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT Press, Cambridge, MA.

M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. 2003. Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin*, 29:665–675.

A. Y. Ng. 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *ICML 2004*.

J. W. Pennebaker and T. C. Lay. 2002. Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality*, 36:271–282.

J. W. Pennebaker, R. Booth, and M. Francis. 2007. Linguistic inquiry and word count: LIWC2007 operator's manual. Technical report, University of Texas.

A. Pentland. 2005. Socially aware computation and communication. *Computer*, pages 63–70.

A. M. Pomerantz. 1984. Agreeing and disagreeing with assessment: Some features of preferred/dispreferred turn shapes. In J. M. Atkinson and J. Heritage, editors, *Structure of Social Action: Studies in Conversation Analysis*. Cambridge University Press.

A. Rosenberg and J. Hirschberg. 2005. Acoustic/prosodic and lexical correlates of charismatic speech. In *EUROSPEECH-05*, pages 513–516, Lisbon, Portugal.

S. S. Rude, E. M. Gortner, and J. W. Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18:1121–1133.

H. Sacks, E. A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

E. A. Schegloff, G. Jefferson, and H. Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53:361–382.