Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 89–115

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates☆

Rajesh Ranganath[a], Dan Jurafsky[b,*], Daniel A. McFarland[c]

[a] *Computer Science Department, Stanford University, United States*
[b] *Linguistics Department, Stanford University, United States*
[c] *School of Education, Stanford University, United States*

Received 12 May 2011; received in revised form 6 October 2011; accepted 13 January 2012
Available online 25 January 2012

Abstract

Automatically detecting human social intentions and attitudes from spoken conversation is an important task for speech processing and social computing. We describe a system for detecting interpersonal stance: whether a speaker is *flirtatious*, *friendly*, *awkward*, or *assertive*. We make use of a new spoken corpus of over 1000 4-min speed-dates. Participants rated themselves and their interlocutors for these interpersonal stances, allowing us to build detectors for style both as interpreted by the speaker and as perceived by the hearer. We use lexical, prosodic, and dialog features in an SVM classifier to detect very clear styles (the strongest 10% in each stance) with up to 75% accuracy on previously seen speakers (50% baseline) and up to 59% accuracy on new speakers (48% baseline). A feature analysis suggests that flirtation is marked by joint focus on the woman as a target of the conversation, awkwardness by decreased speaker involvement, and friendliness by a conversational style including other-directed laughter and appreciations. Our work has implications for our understanding of interpersonal stance, their linguistic expression, and their automatic extraction.

Keywords: Paralinguistics; Prosody; Emotion; Dating

1. Introduction

In many of our social encounters we observe the speech and gestures of interactants looking for clues as to their social intention and interpersonal style. Many of these signals are unclear, and ethnomethodologists and conversation analysts have long described how we focus on the most clear and consistent signals as we build our interpersonal perceptions (Garfinkel, 1967; Heritage, 1984). Through a process of demarcated signaling, speech acts build on one another, and persons, even in a brief encounter, signal clear social intentions and interpersonal styles.

In this paper, we focus on distinguishing social interactants who are reported to exhibit (or not exhibit) clear social intentions or styles, detecting what Scherer (2000, 2003) has called *interpersonal stances*, "affective stance[s] taken toward another person in a specific interaction". We focus on the detection of four kinds of interpersonal stances: whether a person is viewed as friendly, flirty, awkward, or assertive.

☆ This paper has been recommended for acceptance by 'Björn Schuller, Ph.D.'.

* Corresponding author. Tel.: +1 650 723 4284.
E-mail addresses: rajeshr@cs.stanford.edu (R. Ranganath), jurafsky@stanford.edu (D. Jurafsky), dmcfarla@stanford.edu (D.A. McFarland).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2012.01.005

Understanding how stances are signaled is central to situational dynamics and language understanding (Goffman, 1967; Jaffe and Anderson, 1979). Successful automatic detection of interpersonal stances in text and audio features of human conversation is also crucial for developing socially aware computing systems and more natural dialog agents (Pentland, 2005; Nass and Brave, 2005; Brave et al., 2005).

We propose to study these interpersonal stances in the context of a particular kind of conversation: romantic dating. Computational studies of conversation have often focused on speech oriented toward problem-solving tasks such as direction-finding or information retrieval. Romantic dating is a quite different domain, offering rich insights into social meaning because of its inextricable social nature, and with the potential of informing our understanding of neighboring areas like human gender or biology. We make use of a new corpus composed of conversations from a particular genre of romantic dating: speed-dates. A speed date is a brief 4-min romantic date. We collected over 1000 such dates, in which participants wore microphones, allowing us to collect approximately 60 h of speech and about 800,000 transcribed words. We also collected pre-test surveys indicating participants general attitudes, preferences, and demographics, and post-test surveys on date perceptions and follow-up interest. In scorecards after each date each person judged his or her own level of flirtatiousness, friendliness, awkwardness, and assertiveness and that of his or her partner.

The task we address is to learn which linguistic cues help detect whether a speaker in a speed-dating conversation is judged by their interlocutor (or by themselves) as particularly *friendly*, *awkward*, *flirtatious* or *assertive*. Our goal is to identify the linguistic signals that distinguish very clear stances; for this reason we focus on features that distinguish the conversation sides with the highest 10% ratings for a stance from the conversation sides with the lowest 10% ratings for a stance.

2. Related literature

Our work draws on previous studies that explored a variety of linguistic cues for detecting emotional and interactional meaning. For example 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 extraversion (Mairesse et al., 2007; Mairesse and Walker, 2008). Lexical cues like the use of positive emotion words signal agreeableness (Mairesse et al., 2007), negative emotion words correspond with deceptive speech (Newman et al., 2003), and speakers that are depressed or under stress use more first person singular pronouns (Rude et al., 2004; Pennebaker and Lay, 2002; Cohn et al., 2004). Dialog features such as the presence of disfluencies (as well as prosodic features) can inform listeners about speakers' confidence (Brennan and Schober, 2001; Pon-Barry and Shieber, 2011). Finally, speakers tend to *accommodate* to their interlocutors, adjusting the parameters of their speech (rate of speech, pronunciation, pitch level, vocabulary) to match their interlocutor in ways that are sensitive to social variables (Natale, 1975; Ireland et al., 2011; Levitan and Hirschberg, 2011).

While linguistic cues have been explored for many kinds of social meaning, there is insufficient work examining the linguistic realization of *interpersonal stances*. The Scherer typology of affective meaning distinguishes interpersonal stances from four other kinds of affective states by functional criteria (their intensity and duration, focus on a particular event, and so on), as described in the following list taken from Scherer (2000, 2003):

Emotion: relatively brief episode of synchronized response of all or most organismic subsystems in response to the evaluation of an external or internal event as being of major significance (*angry, sad, joyful, fearful, ashamed, proud, elated, desperate*)

Mood: diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause (*cheerful, gloomy, irritable, listless, depressed, buoyant*)

Interpersonal stance: affective stance taken toward another person in a specific interaction, coloring the interpersonal exchange in that situation (*distant, cold, warm, supportive, contemptuous*)

Attitude: relatively enduring, affectively colored belief, preference, or predisposition towards objects or persons (*liking, loving, hating, valueing, desiring*)

Personality trait: emotionally laden, stable personality dispositions and behavior tendencies, typical for a person (*nervous, anxious, reckless, morose, hostile, envious, jealous*)

The four affective states that are the object of our study (*friendly*, *awkward*, *flirtatious* and *assertive*) are closest to Scherer's *interpersonal stances*; they are stances taken toward another person in the context of a dating event, and their affective nature colors the date interaction. While these exact interpersonal stances have not been extensively studied, there has been important previous work on other affective classes, including emotions, attitudes, and personality traits that are related to our questions and that have helped shape our own investigation.

First, previous research has shown that speed dating itself can be an important tool for studying interaction (Finkel et al., 2007; Finkel and Eastwick, 2008; Ireland et al., 2011; Madan et al., 2005; Pentland, 2005). Madan et al. (2005) examined 60 5-min speed dates where speakers wore audio recorders. They used signal-processing algorithms to extract four features directly from the audio, without using transcripts: the estimated percentage of speaking time for each speaker, a measure of the turn-taking influence of one speaker on the other, a measure of a speaker's variation in energy and spectrum (the mean-scaled standard deviation of the energy, formant frequency and spectral entropy) and the normalized frequency of short backchannel utterances (very short turns). They then used these features to predict three binary responses from each speaker: whether the speaker was romantically interested in the other, was interested in being friends with the other, or was interested in maintaining a professional relationship with the other. They found that a man's or woman's variation in energy was correlated with their having romantic interest in their partner. They also found that a speaker's having more variation in energy, turn-taking influence, or (for women only) frequency of backchannels was correlated with them wanting to be friends with their partner. Madan et al. (2005) also built SVM classifiers using these features to predict whether one speaker was interested in a future romantic, friendly, or business relationship with the other. They report cross-validation accuracy of between 62% and 82%, although the impact of these rates are difficult to evaluate as no baseline or majority class statistics are reported.

Ireland et al. (2011) showed in a speed-date study that similarity in the dyad's use of function words predicted mutual romantic interest and relationship stability. They selected and transcribed forty 4-min speed dates from a larger speed-date experiment. Speakers were considered a 'match' if they both expressed interest in meeting each other after the date. For each date, a measure of function word similarity between the two speakers was computed. For each of 9 part-of-speech categories (personal pronouns, auxiliary verbs, articles, conjunctions, etc.) they computed the normalized difference between the count of each category used by the two speakers. These differences were then averaged, to give a single measure of how similar the two speakers were in their distributions of parts-of-speech. Thus speakers who used the same number of pronouns, auxiliary verbs, conjunctions and so on as each other would look similar; speakers who use different number of each of these types would look different. Ireland et al. (2011) found that similarity in function-word usage predicted matches; more similar dyads were more likely to match.

The study of cues or signals that correlate with perceived attractiveness has received much attention in human biology. Studies focusing on specifically linguistic cues have tended to examine the role of fundamental frequency (F0). A number of studies with English-speaking participants have suggested that men show a preference for raised pitch in women's voices (Feinberg et al., 2008; Jones et al., 2010) and rate them as more attractive (Collins and Missing, 2003; Puts et al., 2011). Similar studies have shown that women find men with lower fundamental frequency or more closely spaced harmonics (characteristic of longer vocal tracts) more attractive or masculine (Collins, 2000; Feinberg et al., 2005). Despite this work on attractiveness, few studies have specifically examined linguistic cues for the perception of flirtatiousness across sex. Perhaps the closest study is the recent within-sex study of Puts et al. (2011), who found that women perceive (women's) voices with higher pitches or more dispersed formants (characteristic of shorter vocal tracts) as more flirtatious.

Friendliness, by contrast, has received somewhat more attention in the literature, and has been investigated in a wider variety of languages, although our knowledge is still unfortunately quite limited. House (2005) showed in Swedish questions that a raised fundamental frequency (F0), especially if it occurs later in its syllable, is perceived as friendlier than low F0 or a peak early in the syllable. In a study of the perception of friendliness in Chinese statements and questions produced by actors, Chen et al. (2004) and Li and Wang (2004) found that friendly speech had overall higher average F0 mean and that friendly speech was also faster than neutral speech.

In English, Liscombe et al. (2003) investigated the LDC Emotional Prosody Speech and Transcripts corpus of acted speech in 15 different emotional categories. They had new participants relabel the speech into new categories that including *friendliness*. They found friendliness to be positively correlated with higher F0 minimum (correlation

of .32), F0 maximum (.31) and F0 mean (.32), but they also found that all the other positive emotions they considered (happy, encouraging, interested) also correlated with a higher F0. The best feature that distinguished friendliness from other positive emotions was a feature they labeled by hand called *tilt stress*. Tilt stress is the spectral tilt of the vowel with nuclear stress (spectral tilt was computed as the first harmonic subtracted from second harmonic, in dB, over a 30 ms window centered over the middle of the vowel). This and other features used in a classifier achieved an accuracy of 73.9% at friendliness detection in this corpus, compared to a baseline of 59.1%.

Related to friendliness, Gravano et al. (2011) labeled a number of social variables including *likeability* and *trying to be liked* in the Columbia Games Corpus of subjects talking while jointly playing computer games (Gravano and Hirschberg, 2011). Some of the strong features associated with males talking to females who were judged as trying-to-be-liked included more contractions, more words related to activation (particularly *no*), more pleasant words, faster speaking rate, and expanded pitch range. There was a trend for females talking to males to be judged as trying-to-be-liked when they raised their pitch and used fewer contractions. Speakers who were more likeable exhibited higher intensity, lower pitch, and more reduced pitch range, more activation (mainly negation) and imagery, more filled pauses and contractions, and fewer interjections.

Also related to friendliness is the personality trait *agreeability*. The agreeableness dimension is one of the big five personality dimensions, ranging from agreeable (friendly, cooperative) to disagreeable (antagonistic, faultfinding). Personality traits like *agreeability* are more stable and typical of a person than an interpersonal stance like *friendly*, and also less linked to a particular encounter or event. Nonetheless agreeability and friendliness are clearly related, and so linguistic features associated with agreeability might be expected to also be associated with friendly dates. Pennebaker and King (1999), Mehl et al. (2006), Mairesse et al. (2007), and Mairesse and Walker (2008) studied the linguistic cues correlating with agreeableness (as measured by standard questionnaires) in two corpora, one from student essays (Pennebaker and King, 1999) and one from spoken conversations (Mehl et al., 2006). Linguistic cues were based on the LIWC lexicons of Pennebaker et al. (2007), the standard linguistic tool for the social psychological analysis of lexical features. Mairesse et al. (2007) and Mairesse and Walker (2008) found that the most robust cue to agreeability across these datasets was the lower uses of swearing and anger words, and in some datasets also more frequent use of backchannels, lower uses of negative emotional words, lower maximum voice intensity, less pitch variation, and higher mean pitch.

Assertiveness has not been directly studied, but assertiveness may be related to the personality trait *extraversion*, which had received a significant amount of attention. Extraverts talk more, talk louder, and with a higher rate of speech (Mairesse et al., 2007) and use more positive emotion words and compliments (Pennebaker and King, 1999). Mairesse et al. (2007) also found in conversation that extraverts have greater intensity variation and use more words related to anger, swearing, and positive and negative emotions. If assertiveness in men is related to perceived masculinity, we might expect to see assertiveness associated with features like lower fundamental frequency or more closely spaced harmonics, described above as associated with judgments of masculinity (Feinberg et al., 2005).

By contrast with the other three stances, we know extremely little about the linguistic manifestations of awkwardness. While awkward speech is associated with autism spectrum disorders, and there has been some discussion of the speech of patients with high-functioning autism (Grossman et al., 2010), it's not clear that clinical awkwardness of that sort is related to the awkwardness that participants describe in speed dates.

Finally, a wide variety of studies have looked at accommodation: the adjustment of prosodic, lexical, or grammatical variables by a speaker to be more similar or more distant from the interlocutor's speech, as a means of aligning or disaligning with the interlocutor (Chartrand and Bargh, 1999; Levitan and Hirschberg, 2011; Namdy et al., 2002; Natale, 1975; Nenkova et al., 2008; Niederhoffer and Pennebaker, 2002; Pardo, 2006; Street, 1983). But despite the large number of studies it is still not clear the extent to which the presence or extent of accommodation is conditioned on stances like friendliness, flirtatiousness, awkwardness, or assertiveness. The Ireland et al. (2011) results that found links between romantic matching and accommodation suggests one possible hypothesis: that flirtatious or friendly speakers might accommodate more. We also expect to see less accommodation in awkward speakers.

In summary, previous work on these interpersonal stances seems mainly to have focused on friendliness and agreeableness, and to a lesser extent on flirtation and attractiveness. Attractiveness (and to some extent flirtation) has been associated mainly with higher F0 in women and lower F0 in men. The main linguistic cues that have been suggested for friendliness are higher and later F0 peaks, faster speech, higher spectral tilt, more frequent use of backchannels, and lower use of words related to negative emotions, swearing, or anger.

(4) How often did **you behave** in the following ways on this "date"? (1=never, 10=constantly)

You were *friendly*……..	never	1	2	3	4	5	6	7	8	9	10	constantly
You were *flirtatious*……	never	1	2	3	4	5	6	7	8	9	10	constantly
You were *awkward*……..	never	1	2	3	4	5	6	7	8	9	10	constantly
You were *assertive*……..……..	never	1	2	3	4	5	6	7	8	9	10	constantly

(5) How often did the **other person** behave in the following ways on this "date"? (1=never, 10=constantly)

They were *friendly*……..	never	1	2	3	4	5	6	7	8	9	10	constantly
They were *flirtatious*……	never	1	2	3	4	5	6	7	8	9	10	constantly
They were *awkward*……..	never	1	2	3	4	5	6	7	8	9	10	constantly
They were *assertive*……..……..	never	1	2	3	4	5	6	7	8	9	10	constantly

Fig. 1. The 8 survey questions from the SpeedDate corpus whose answers we attempt to detect.

3. The SpeedDate 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: a large hall, separated by partitions into cubicles with one or two tables each containing one pair of participants. All participants wore audio recorders on a shoulder sash, resulting in two audio recordings of each of the approximately 1100 4-min 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, including the participants self-reported age, height and weight, hobbies and interests, dating background, and self-described attributes. 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, 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'?". In addition they were also asked to rate their own intentions: "How often did you behave in the following ways on this 'date'?" on a scale of 1–10; questions are shown in Fig. 1.

The SpeedDate corpus includes audio data and transcripts. Since both speakers wore microphones, most dates had two recordings, one from each speaker's microphone. The acoustic wave file from each recorder was manually segmented into a sequence of wavefiles, each corresponding to one 4-min date. Each date was then transcribed by a transcription service, producing a diarized transcript that marked words, laughter, filled pauses, speaker overlap, and restarts, and timestamped the beginning and end of each turn at the granularity of a second. Turn boundaries for 10% of the dates were segmented at a finer grain (tenth of a second). Because of the high level of noise, each speaker was much clearer on his/her own recording; transcribers based their transcription on the clearer recording, using the partner's recording when necessary. A sample extract from the transcripts is shown below:

0:01:55.1	0:01:56.8	F:	Well what about you, what are you passionate about?
0:02:05.7	0:02:11.8	M:	Um, I am passionate about probably two things.
0:02:03.2	0:02:03.8	F:	Uh-huh.
0:02:12.4	0:02:15.3	M:	Well, many things, but two that come to mind straightaway. One is travel.
0:02:06.8	0:02:07.3	F:	Okay.
0:02:15.5	0:02:17.2	M:	I like see different parts of the world-
0:02:08.5	0:02:09.2	F:	Uh-huh.
0:02:17.6	0:02:27.9	M:	-experience lots of different things. And I also- recently, I've got into exercise, and, um, just different things, so riding a bike, and swimming, and running.
0:02:18.5	0:02:20.1	F:	Oh, okay. Uh-huh.
0:02:28.3	0:02:30.3	M:	I did my first track run on the weekend.
0:02:21.7	0:02:22.9	F:	Oh, you did? How was it?
0:02:31.9	0:02:33.0	M:	It was hard.
0:02:24.3	0:02:27.0	F:	[laughter] Yeah, I heard it's really hard.
0:02:35.6	0:02:37.1	M:	But I definitely recommend it.

Due to mechanical, operator, and experimenter errors, 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 946 clean dates for which we had complete audio and transcripts. These were on average 812 words long (i.e., on average 406 words from each speaker), and took on average 93 turns. Because some participants did not provide Likert ratings for some traits (38 of the participants, for example, omitted to self-report their own level of flirtation), each of our classifier experiments relies on a subset of the 946 dates for which we had the relevant survey variables as well as the audio and transcripts.

4. Study 1: detection of interpersonal stances and their linguistic features

The goal of our first experiment is to identify friendly, flirtatious, awkward, or assertive speakers, and explore the different linguistic features that are associated with them. Since the stance used by each speaker in each date was labeled twice (self-reported by the speaker and other-reported), we build separate classifiers for detecting self-reported characteristics and other-reported characteristics. One of our goals is to understand the difference in how the linguistic associations of a stance or style may differ depending on who is describing it. To learn about the role of gender, we ran separate experiments for males and females. The combination of 4 stances, two sexes, and self- versus other-reported style results in 16 different experiments. For each experiment we attempt to predict a particular style as perceived by a particular gender (e.g., male self-reported flirtation) from linguistic features both of the speaker and of the interlocutor. In all cases we are interested in understanding the most clear and consistent signals, focusing on distinguishing the conversation sides with the highest 10% ratings for each stance from the conversation sides with the lowest 10% ratings.

This study focuses on speaker-dependent classification: each speaker in the test set is very likely to have been seen in the training set, allowing the system to learn specific patterns of behavior for each speaker. In study 2 we will turn to speaker-independent classification to understand how our system generalizes to unseen speakers. The next section describes the features that we extracted from both wavefiles and transcripts; the following section describes the classifiers and our evaluation metrics.

4.1. Feature extraction

For each *conversation side* (one speaker in one date) we extracted a variety of linguistic features. Prosodic features were extracted from the wavefiles recorded from the speaker's microphone (except for the 130 dates for which we only had one audio file). Lexical and discourse features were extracted from the transcripts.

4.1.1. Prosodic features

Prosodic features characterize F0, energy, and durational properties of the conversation side. We extracted F0 and RMS amplitude features using Praat scripts (Boersma and Weenink, 2005), using the handmarked turn boundary times to extracted each feature over each turn in the conversation side and then took averages and standard deviations over all turns in a side. For example, the feature Fo MIN (minimum F0) for a side was computed by taking the minimum F0 value of each turn in that conversation side (not counting zero values of F0), and then averaging these values over all turns in the side. For F0 we extracted the minimum, maximum and mean values of the conversation side (Fo MIN, Fo MEAN, Fo MAX). No outliers were excluded. An example from the Praat labeling is shown in Fig. 2.

We coded various measures of F0 variation, including the standard deviation of each F0 measure (Fo MIN, MAX, MEAN); Fo MEAN SD is thus the deviation across turns from the global F0 mean for the conversation side. Fo SD is the deviation *within* a turn for the f0 mean, averaged over turns. Fo SD SD measures the variance of Fo SD across turns. The PITCH RANGE is F0 max–f0 min per turn, averaged over turns. For energy, we computed RMS amplitude values for each turn, and then computed RMS min, RMS max, and RMS mean values by averaging over all turns in a conversation side. We then included measure of variation in energy: RMS min SD, RMS mean SD, RMS max SD. Table 1 shows the 18 raw prosodic features.

Using these extracted features, we conducted exploratory factor analysis to see if latent factors organize the pattern with which the features are correlated. Six orthogonal (uncorrelated) factors explain 85% of the variance in our 18 acoustic features. All have eigenvalues over 1 and there is a break between 6 and 7 factors in the scree plot suggesting the use of 6 factors. In addition, nearly identical results hold when oblique (correlated) factors are generated, suggesting

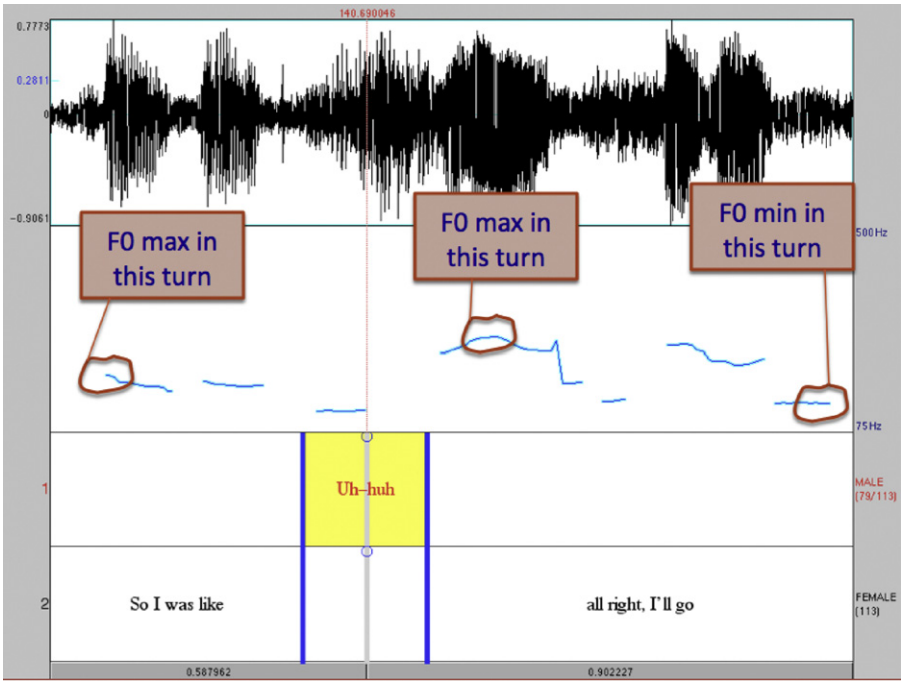

Fig. 2. Example of prosodic features extracted from each turn for each speaker.

Table 1
The 18 raw prosodic features for each conversation side, extracted using Praat from the hand-segmented turns of each side.

FO MIN	minimum (non-zero) F0 per turn, averaged over turns
FO MIN SD	standard deviation from F0 min
FO MAX	maximum F0 per turn, averaged over turns
FO MAX SD	standard deviation from F0 max
FO MEAN	mean F0 per turn, averaged over turns
FO MEAN SD	standard deviation (across turns) from F0 mean
FO SD	standard deviation (within a turn) from F0 mean, averaged over turns
FO 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
TURN DUR SD	standard deviation of turn duration

the results are stable. The same factor pattern arose when run the models separately by gender. Only the order of the factors shifts. Table 2 shows the factor loadings for the six factors.

The resulting 6 factors (max F0, intensity, min F0, variable intensity, long turn, variable F0) were then used instead of the 18 raw prosodic variables in all further analyses. For each speaker we also measured the average duration of the turn in seconds (averaging over all turns in a conversation side), the total time for a speaker in each conversation side (summed over all turns), and the rate of speech (measured in words per second, averaged over turns). Table 3 shows the final set of 7 prosodic features we used in our analyses.

Table 2
Loadings for the 6 prosodic factors.

	Factor1 Max F0	Factor2 Intensity	Factor3 Min F0	Factor4 Var Intensity	Factor5 Long Turn	Factor6 Var. F0
avtndur	20	5	−18	−1	91	−21
sdtndur	6	1	−5	2	95	−7
avpmin	−34	12	84	−3	−23	8
sdpmin	−16	0	91	2	−2	18
avpmax	92	10	1	3	22	−7
sdpmax	−75	−5	−17	6	−2	53
avpmean	59	19	67	−1	−6	−3
sdpmean	12	−4	30	10	−14	73
avpsd	91	−8	−21	1	−3	20
sdpsd	−34	−10	−1	5	−20	76
avimin	−1	51	13	−66	−10	20
sdimin	−5	22	8	66	7	16
avimax	9	91	2	−1	7	−14
sdimax	1	−59	0	70	−6	10
avimean	4	94	12	−23	−1	−1
sdimean	4	−29	−2	89	−9	5
avprange	90	5	−25	4	26	−8
sdprange	−74	−5	15	8	7	51

Table 3
The 7 final prosodic features; 6 factors that collapsed the 18 raw features plus rate of speech.

Max Fo	Higher max F0, mean F0, and pitch range
Intensity	Higher mean, max, and min intensity
Min Fo	Higher min F0 and its variation, and mean F0
Intensity Variation	More variable min, mean, max intensity
Turn Length	Longer turns
Fo Variation	More variable F0 mean, max, and pitch range
rate of speech	Number of words in turn divided by duration of turn in seconds, averaged over turns

Table 4
Lexical features modified from LIWC. Each feature value is a total count of the words in that class for each conversation side; asterisks indicate including suffixed forms (e.g., *eat** = *eat, eats, eating*). Swear, Negemotion, Negate, and Food include more words in addition to those shown.

I	*I'd, I'll, I'm, I've, me, mine, my, myself* (not *I mean*)
you	*you, you'd, you'll, your, you're, yours, you've* (not *you know*)
Sex	*sex, sexy, sexual, stripper, lover, kissed, kissing*
Love	*love, loved, loving, passion, passions, passionate*
Hate	*hate, hates, hated*
Swear	*suck*, hell*, crap*, shit*, screw*, damn*, heck, ass*,...*
Negemo	*bad, weird, crazy, problem*, tough, awkward, worry,...*
Negate	*don't, not, no, didn't, never, haven't, can't, wouldn't, nothing,...*
Food	*food, eat*, cook*, dinner, restaurant, coffee, chocolate, cookies,...*
Drink	*party, bar*, drink*, wine*, beer*, drunk, alcohol*, cocktail,...*

4.1.2. Lexical features

Lexical features have been widely explored in the social-psychological and computational literature. We drew on the insights of the LIWC lexicons of Pennebaker et al. (2007), the standard for social psychological analysis of lexical features, creating the ten modified LIWC features shown in Table 4.[1]

[1] We began by selecting 11 features that the previous work of Mairesse et al. (2007) had found important in detecting personality-related features: Anger, Assent, Ingest, Insight, Negemotion, Sexual, Swear, I, We, Negate, After of preliminary experiments we removed Anger (because

Table 5
New lexical features. Each feature value is a total count of the words in that class for each conversation side; asterisks indicate including suffixed forms (e.g., *work** = *work, works, worked, working*). ACADEMICS include more words in addition to those shown.

HEDGE	*sort of, kind of, I guess, I think, a little, maybe, possibly, probably*
META	*speed date, flirt, event, dating, rating*
ACADEMICS	*work*, program, PhD, research, professor*, advisor, finish*,...*
LIKE	the discourse marker *like* (removing cases of the verb *like*)
IMEAN	the discourse marker *I mean*
YOUKNOW	the discourse marker *you know*
UH	the filled pause *uh*
UM	the filled pause *um*

Table 6
Dialog and discourse features.

TOTALWORDS	total number of words in side
QUESTIONS	number of questions in side
NTRI	clarification question (*Excuse me?*)
INTERRUPT	number of turns in side which one speaker interrupted the other
INITIAL LAUGH	number of instances of turn-initial or whole-turn laughter in side
MEDIAL LAUGH	number of instances of turn-medial or turn-final laughter in side
RESTART	total number of disfluent restarts in conversation side
APPRECIATIONS	number of appreciations in side (*Wow*, *That's true*, *Oh, great*)
SYMPATHY	number of sympathetic negative assessments in side
AGREE	number of agreement turns in side (*That's true*, *For sure...*

We also created 8 new lexical features that captured additional hypotheses, shown in Table 5. Hedges are words or phrases that weaken the force of assertions or indicate uncertainty, marking that some sort of criterion for category membership is weak or lacking (Lakoff, 1973). While hedges can modify adjectives (*a little easier*) or nouns (*a little hiking*), most hedges in our data are verb phrase or sentential modifiers, expressing the speaker's lack of commitment to an entire proposition.

I'm *sort of* just finishing up some work right now...
Yeah, I *kind of* know that area
It was actually *I guess* really nice...
It's going to happen *I think*.

We extracted topical features common in dating conversations, including the topic of dating itself (META), by counting the occurrence of words like *speed date, flirt, event*, and *rating*, and an ACADEMICS topic related to the work and schooling of the graduate student participants. We coded the five common discourse particles *like*, *I mean*, *you know*, *um*, and *uh*. We conjectured based on the work of Schiffrin (1987) that *I mean* and *you know* might be relevant to speaker engagement and other-directed speech. We included *um*, and *uh* because of their link with disfluent speech, which might be associated with awkwardness or flirtation, and *like* because it may play roles as both a disfluency and a hedge. These are extracted with simple regular expressions; we used expressions based on surrounding context to automatically eliminate cases of the verb *like*, leaving only the discourse marker.

4.1.3. Dialog and discourse features

While dialog features are clearly important for extracting social meaning, previous work has focused on prosodic and lexical cues to social meaning, presumably because discourse features are expensive to hand-label and hard to automatically extract. We drew on the conversation analysis and dialog act literature to devise discourse features that could be automatically extracted, shown in Table 6.

---

it was strongly collinear with *Swear*), and three others (*Assent, Ingest, We*) because their use did not correlate with our social variables. We split two of LIWC's categories: INGEST into FOOD and DRINK, and SEXUAL into SEX and LOVE, added a HATE category, and modified others to include words that were common in our sample but not in LIWC's lists, and to remove ambiguous words that were more often used in different senses than suggested by the LIWC categories.

The variable TOTALWORDS includes all words used by the speaker in a side (space-delimited orthographic strings) including filled pauses (uh and um) and fragments (th-, whe-) but not including laughter. About 21% of all turns are questions. The variable QUESTIONS codes the total number of questions used by the speaker in the conversation side, extracted via question marks, and also (for some sentences without questions marks because of interruptions, etc.) by appropriate auxiliary inversion *do you, are you, could you*, etc.). We separately coded *clarification questions* (*repair questions* or NTRIs for 'next turn repair indicators' (Schegloff et al., 1977)), turns in which a speaker signals lack of hearing or understanding:

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

The following regular expression was used to detect NTRIs:

What? | Sorry | Excuse me | Huh? | Who? | Pardon? | Say again?
| Say it again? | What's that | What is that

Interruptions were coded by the transcribers with double dashes at the end of the preceding turn; the female below is coded as having a single interruption. The variable codes the total number of interruptions taken by the speaker in the conversation side:

MALE: Do you really get a lot of information from–
FEMALE: It's really in the development stages so far so I haven't applied it to a field application at all.

Laughter was marked in the transcripts by the transcribers. We extracted two different laughter variables based on the position of laughter in the speaker's turn. Initial laughter occurred at the beginning of the turn (or took the entire turn). We conjectured these would be cases of laughing at the other. Turn-medial/final laughter were laughs that occurred in the middle or the end of a turn. We hypothesized these would be cases of a speaker laughing at themselves.

In addition to filled pauses like *uh* and *um*, we coded disfluent RESTARTS, which were explicitly marked by the transcribers using dashes:

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

Extracting these dialog acts required more complex regular expressions. One such move is an *assessment*, a dialog act that expresses the speaker's sentiment toward the other's recent utterance. Negative assessments, which we call *sympathy*, are phrases like "That must be tough on you":

MALE: . . .consulting, which is the last thing I want to do.
FEMALE: Oh, *that's too bad*.

Positive assessments (or *appreciations*) are phrases like " "Good for you!":

FEMALE: I played in the orchestra.
MALE: *Oh that's cool*.

We designed regular expressions based on the pattern below drawn from the assessment literature (Goodwin, 1996; Goodwin and Goodwin, 1987; Jurafsky et al., 1998), and the hand-labeled Switchboard dialog act corpus (Jurafsky et al., 1997), but augmented with structures to handle expressions like "Awesome", "Good for you!", "Oh no", "I had the same problem", and so on:

```
Pro Term + Copula + (Intensifier) + Assessment Adjective
```

We also coded AGREEMENTS, expressions of strong agreement with the other, e.g., *That's true*, *That's right*, *Definitely*, *For sure*, and so on.

4.1.4. Accommodation features

We discussed in Section 2 the previous literature on speaker's adjusting their linguistic production to be more similar to or more different from that of the other. Drawing on this literature, we investigated four kinds of speaker's accommodation to the other's speech: rate of speech accommodation, function word accommodation, content word accommodation, and laughter accommodation.

We measured rate of speech accommodation by computing whether there were correlations between the rates used by two speakers over time, for example whether a speaker sped up when the other sped up. We followed the turn-correlational methodology of Niederhoffer and Pennebaker (2002), considering the sequences of rates for each speakers

Table 7
The 195 function words used to compute function word mimicry.

Auxiliary and copular verbs
able am are aren't be been being can can't cannot could couldn't did didn't do don't get got gotta had hadn't hasn't have haven't is isn't may should should've shouldn't was were will won't would would've wouldn't
Conjunctions
although and as because 'cause but if or so then unless whereas while
Determiners, Predeterminers, and Quantifiers
a an each every all lot lots the this those
Pronouns and Wh-words
anybody anything anywhere everybody's everyone everything everything's everywhere he he'd he's her him himself herself his I I'd I'll I'm I've it it'd it'll it's its itself me my mine myself nobody nothing nowhere one one's ones our ours she she'll she's she'd somebody someone someplace that that'd that'll that's them themselves these they they'd they'll they're they've us we we'd we'll we're we've what what'd what's whatever when where where'd where's wherever which who who's whom whose why you you'd you'll you're you've your yours yourself
Prepositions
about after against at before by down for from in into near of off on out over than to until up with without
Discourse Particles
ah hi huh like mm-hmm oh okay right uh uh-huh um well yeah yup
Adverbs and Negatives
just no not really too very

Table 8
Features for modeling forms of accommodation.

RateAcc	correlation between turnwise rates of speech across speakers
FuncWordAcc	number of function words also used in other's prior turn.
ContWordAcc	number of content words also used in other's prior turn.
LaughAcc	number of laughs immediately preceded by other laugh.

turn as a vector and computing the Pearson's correlation between the two turn vectors. A high correlation indicates that the changes over time in speakers rates of speech are correlated; as one speaker gets faster or slower, so does the other.

We measured how often a speaker uses a function word that was also used in the other's previous turn. For each function word w (Table 7) in a given speaker's turn, w is an accommodated function word if it also occurs in the immediately preceding turn of the other. The variable function-accommodation is the total number of accommodated function words over the side.

Content words are defined as any word that is not a function word. For each content word w in a given speaker's turn, if w also occurs in the immediately preceding turn of the other, we count w as an accommodated content word. The raw count of accommodated content words is be the total number of these accommodated content words over every turn in the conversation side. Because content words vary widely in frequency, we normalized our counts by the frequency of each word. We combined our entire speed-date corpus with the Switchboard corpus of conversational American English (Godfrey et al., 1992) and used this larger corpus to compute the frequency of each word in the speed-date corpus. We then computed the variable content-accommodation by summing (over all turns in a side) the inverse frequency of each content word that also occurred in the other's immediately preceding turn. We computed laughter-accommodation by summing over all turns in which a speaker laughed and the other also laughed in the immediately preceding turn (Table 8).

4.1.5. Feature normalization

Before building classifiers, we preprocessed the data. For all studies some features were normalized by the total word count in the conversation side: HATE, ACADEMICS, LOVE, SEX, DRINK, FOOD, NEGEMO, SWEAR, NEGATE, HEDGE, META, NTRI, LIKE, IMEAN, YOUKNOW, YOU, RESTART, LAUGHACCOM, INITLAUGH, MIDLAUGH, I, UH, UM. We follow common procedure in logging rate of speech.

We then standardize all the features to have zero mean and unit variance globally across the training set before training any model. We did this to avoid imposing a prior on any of the features based on their numerical values.[2] We ran our experiments both speaker-normalizing all prosodic feature and without speaker normalization; we saw no average differences in accuracy and so we give results without speaker normalization.

4.2. Classification

Our goal in this work is to learn to detect and characterize clear cases of a conversation style. For this reason we selected our training and test sets from a subset of all the dates, choosing only highly indicative individuals whose stance falls in either the global top ten percent of ratings or the global bottom ten percent of ratings.

For each of the 16 experiments, then, we first combined all three of the SpeedDate events into a single corpus. For the experiment on detecting *self-reported male flirtation*, for example, we sorted all the 946 dates by the self-reported male flirtation score, and then took the top ten percent of date sides to form our positive class examples and the bottom ten percent to form the negative class examples. Due to missing ratings and audio, the corpus for self-report of male flirtation consists of 174 examples. The other 15 experiments had similar sizes.

Because of the very small datasets, rather than break out a fixed training and test set, we use five-fold cross validation to train and evaluate our model. We split the data into five folds of equal size, using four of the folds for training and one for test in round robin, so every example ends up in a test set. This yields a data split with 80% of the data in the training set and 20% in the test set. Each fold was split 50/50 between the positive and negative class. To help ensure that we learned features that were general, particularly given the small size of our datasets, we randomized our data ordering and repeated the test 30 times.

For classification we trained both a linear C-SVM and an L_1 regularized logistic regression. Our goal in the C-SVM is to solve, in primal form:

$$\begin{aligned} &\min_{\gamma,w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ &\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ &\xi_i \geq 0, \quad i = 1, \ldots, m \end{aligned} \tag{1}$$

where m is the number of training examples, $x^{(i)}$ is the ith training examples, and $y^{(i)}$ is the ith class (1 for the positive class, -1 for the negative class). The ξ_i are the slack variables that allow this algorithm to work for non linearly separable datasets.

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]:

$$p(y|x;\theta) = \frac{1}{1 + \exp(-\theta^T x)} \tag{2}$$

In regularized logistic regression we find the weights θ which maximize the following optimization problem:

$$\underset{\theta}{\text{argmax}} \sum_i \log p(y^i|x^i;\theta) - \alpha * R(\theta) \tag{3}$$

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

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

$$\underset{\alpha}{\text{argmax}}\ \text{accuracy}(\theta_\alpha, S_{\text{val}}) \tag{4}$$

[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 the SVM and the L_1 regularized logistic regression both minimize a norm of the weight vector, there is a bias to put weight on feature A (the weight on feature B would need to be 1000 times larger to carry the same effect). The reduction to unit variance was performanced for the same reasons.

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

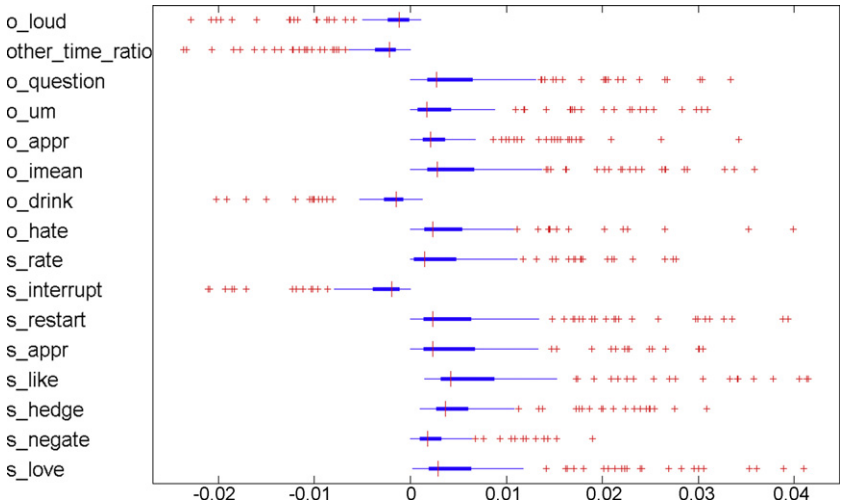

Fig. 3. A boxplot for self-report of awkwardness in women showing the significant features, with median values (central red line), first quartile, third quartile, outliers (red +'s) and interquartile range (filled box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

where for a given sparsity parameter α

$$\theta_\alpha = \underset{\theta}{\operatorname{argmax}} \sum_i \log p(y^i|x^i;\theta) - \alpha * R(\theta) \tag{5}$$

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. The search space over the sparsity parameter α is bounded around an expected sparsity to prevent overfitting.

Both the C-SVM and the L_1-regularized logistic regression's optimization problems have hyperparameters (C for the SVM and α for regression). We learn these hyperparameters by taking one of the folds from the training set to form a validation set and evaluating the model trained on the rest of the folds in the training set for a specific hyperparameter on the validation set. We chose the hyperparameter that is best on the validation set. We searched the hyperparameter space using a log-spaced line search with a base of 2. To produce the final model we train on the entire training set using the learned hyperparameter. Our models were trained with the LIBLINEAR software package (Fan et al., 2008).

We evaluated both of our classifiers on the test set by computing the accuracy, the number of correct examples divided by the total number of test examples. For feature analysis we used weights generated by L_1 regularized logistic regression, whose sparse feature weights are particularly useful in feature analysis (Ng, 2004). In order to select important features for analysis, we first computed the median of the weights for a given feature across all 30 runs of each classifier with randomized ordering. Any feature whose median weight was non-zero was selected. A feature having a non-zero median will have the majority of its weights as either positive or negative, which will be rare when using a sparse algorithm. We then ran a two tailed t-test at the .05 significance level with a null hypothesis of "the feature weight has a mean of 0" and verified that our non-zero median criteria was a stricter criteria than the t-test for all 16 of our models (there was no feature that had a non-zero median that wasn't significant according to the t-test.) The boxplot in Fig. 3 demonstrates the weight distributions for one stance: self-reported female awkwardness.

4.3. Results

We report in Table 9 the accuracies for the 16 classifiers trained for both the SVM and the logistic regression. The first two rows shows the results from SVM on each of the 4 conversation styles, for each gender, and for both self-reported and other-reported styles. The next two rows show the results from the L_1 regularized logistic regression. The SVM outperformed L_1 regularized logistic regression, and was always significantly better than the baseline (paired t-test,

Table 9
Accuracy of binary classification of each conversation side, where chance is 50%. The top two rows describe the SVM classifier, the next two the logistic regression. In each case the first row predicts the stance as labeled by the speaker; the second predicts the stance as labeled by the later. These accuracy results were aggregated from 30 randomized runs of 5-fold cross validation.

		Friendly		Flirt		Awk		Assertive	
		M	F	M	F	M	F	M	F
SVM	Self	.76	.71	.66	.74	.63	.67	.73	.64
SVM	Other	.71	.64	.65	.78	.67	.67	.65	.69
LR	Self	.75	.69	.62	.75	.66	.68	.73	.64
LR	Other	.71	.62	.60	.76	.65	.66	.62	.66

Table 10
Accuracy of binary classification on distinguishing top and bottom quartile of each conversation side, where chance is 50%. The top two rows describe the SVM classifier, the next two the logistic regression. In each case the first row predicts the stance as labeled by the speaker; the second predicts the stance as labeled by the other. These accuracy results were aggregated from 30 randomized runs of 5-fold cross validation. All accuracies are significantly better than the baseline (paired t-test, $p < .0001$).

		Friendly		Flirt		Awk		Assertive	
		M	F	M	F	M	F	M	F
SVM	Self	.62	.62	.59	.66	.58	.63	.64	.59
SVM	Other	.63	.68	.61	.64	.61	.62	.57	.63
LR	Self	.62	.63	.60	.67	.59	.64	.65	.60
LR	Other	.62	.68	.62	.64	.62	.63	.59	.65

$p < .0001$). At least part of the reason for the these high accuracies is the reliance on speaker specific characteristics in the training set; we will see lower scores in our speaker-independent study below.

4.4. Study 1B: quartile experiments

Although our goal to study the clearest examples of each stance led us to choose the top and bottom decile (10%) of each stance as the positive and negative instances, we also ran each experiment using the top and bottom quartiles (25%). We did this for two reasons. First, ten is an arbitrary cutoff; a different cutoff might capture consistent clear behavior. Second, the quartile experiment allowed us to include a much larger training and test set; on average 473 dates instead of 189 dates to crossvalidate through. Finally, the larger data set makes the feature analysis we want to run in the next section more robust and helps avoid overfitting features of individuals.

The quartile experiment does run the risk of including observations in the middle of the stance range which are not clear examples of a stance. The label cutoffs for the top/bottom quartile can also be skewed; the cutoff for self-assessed male friendliness is 7 for negatives and 10 for positives, while for self-assessed male assertiveness it's 4 for negatives and 7 for positives. Nonetheless we felt the extra robustness and significantly more data makes it worth examining this second set of classifiers.

As with our primary experiments to study the clearest examples, we ran 16 experiments stemming from the combinations of 4 stances, two sexes, and self versus other reported style, but where the positive and negative examples were determined by the top and bottom quartiles.

We report in Table 10 the accuracies for the 16 classifiers trained for both the SVM and the logistic regression on top and bottom quartiles (25%). The first two rows shows the SVM results on each of the 4 conversation styles, for each gender, and for both self-reported and other-reported styles. The next two rows show the results from the L_1 regularized logistic regression. For the quartile experiments L_1 regularized logistic regression generally outperforms the SVM, and was always significantly better than the baseline (paired t-test, $p < .0001$).

As expected, the task of distinguishing the top quartile from the bottom quartile is harder than distinguishing the top and bottom decile. The mean accuracy of the SVM classifier on the top decile/bottom decile (averaging all scores in the first two rows of Fig. 9) is .69. By contrast, the mean accuracy of the SVM classifier on the top quartile/bottom quartile (Table 10) is .62.

Table 11
Feature weights for flirtation detection in decile study (median weights of the randomized runs) for the non-zero predictors for each classifier. Boldfaced features hold across self and other ratings, and boldfaced features that were also significant in the quartile study are underlined.

MALE				FEMALE			
FLIRT SELF		FLIRT OTHER		FLIRT SELF		FLIRT OTHER	
s_academic	−0.00277	o_midlaugh	−0.00185	o_i	−0.00129	s_question	−0.00210
s_uh	−0.00237	o_question	−0.00165	o_like	−0.00125	**o_loud**	−0.00178
o_academic	−0.00123	o_food	−0.00132	o_academic	−0.00124	s_negemo	−0.00121
s_pitchvar	−0.00121	**s_academic**	−0.00105	**o_loud**	−0.00114	s_maxpitch	0.00130
s_negate	−0.00111	o_imean	0.00109	o_uh	−0.00105	**s_negate**	0.00143
o_uh	0.00117	**o_negemo**	0.00112	s_rateaccom	0.00111	s_ntri	0.00152
s_longturn	0.00121	**s_you**	0.00117	s_love	0.00116	s_um	0.00161
s_youknow	0.00125	o_meta	0.00124	s_midlaugh	0.00157	**s_i**	0.0016
o_sex	0.00141	o_drink	0.00134	o_ntri	0.00159	**o_funcaccom**	0.00180
s_varinten	0.00156	**s_imean**	0.00134	s_hedge	0.00204	**s_longturn**	0.00186
s_swear	0.00191	o_appr	0.00139	**o_appr**	0.00251	s_varinten	0.00188
s_you	0.00195	**s_um**	0.00193	**s_negate**	0.00261	o_varinten	0.00189
s_initlaugh	0.00200	**s_youknow**	0.00273	**s_like**	0.00270	o_midlaugh	0.00194
o_like	0.00206	s_question	0.00341	**o_funcaccom**	0.00298	o_meta	0.00201
o_negemo	0.00213	o_longturn	0.00356	**s_longturn**	0.00327	**o_appr**	0.00201
s_hate	0.00220			**o_interrupt**	0.00330	s_rate	0.00311
o_love	0.00225			**s_i**	0.00370	**s_like**	0.00411
s_imean	0.00231			**o_rate**	0.00547	**o_interrupt**	0.00459
s_sympathy	0.00252					o_question	0.00549
s_um	0.00358					**o_rate**	0.00588

4.5. Feature analysis

Tables 11–14 show feature weights for each stance from the decile results. Although we used the feature selection algorithm described in Section 4.2 that only considers features that were significant across the 30 randomized test runs, we were concerned that the large number of features may still overfit the small test set. This is particularly true with the decile classifier, which uses less than 200 conversations. We therefore focus in this discussion on features that are particularly robust. In the tables below, features that hold across self- and other-ratings are boldfaced. Boldfaced features that were also significant in the quartile study (averaging just under 500 conversations) are further underlined. Our discussion will highlight the boldfaced features that hold for both the self and other labels, and especially those that are also significant in the quartile study.

A number of features were consistently associated with flirtation. Women rated as flirting (whether by self or other, generally whether in the decile or quartile studies) tend to use negation (especially the word *don't* but also *no* and *not*), use the word *like*, and use medial laughter. In the decile but not quartile studies, flirting women also use *I*; in the quartile but not decile studies, flirting women use more appreciations and *you know*. Men rated as flirting (generally either by self or other, in both the decile and quartile studies) tend to use the words *you*, *you know*, and *um*, and are less likely to use words about their *academic* work. In the quartile studies, flirting men also talked about sex. There are also strong characteristics of the speech of men talking to flirting women; in both the decile and quartile studies these men use appreciations and accommodate the woman's function words. Some flirtation features differed between self- and other-rating; in both the decile and quartile studies, men who self-reported as flirting use less negation, but this feature was not associated with other-assessed flirtation.

Turn-medial/final laughter is the most robust cue for friendliness, associated with both self- and other-rating for both men and women in both the decile and quartile studies. Lower use of negative emotion was also robustly associated in the quartile study, and to a lesser extent in the decile study, with friendly men and women, both self- and other-assessed. Across the studies and labelers, friendly men tend to use less hedges, less *uh*, less *you know*, and have more varied intensity, while friendly women tend to use clarification questions. In the quartile but not decile studies, friendly women (both self- and other-assessed) have higher maximum pitch and greater pitch variance, and are more likely to use negation and initial laughter. In both studies, men who use agreement or swearing are perceived as, but do not

Table 12
Feature weights for friendliness detection (median weights of the randomized runs) for the non-zero predictors for each classifier. Boldfaced features hold across self and other ratings, and boldfaced features that were also significant in the quartile study are underlined.

MALE				FEMALE			
FRIENDLY SELF		FRIENDLY OTHER		FRIENDLY SELF		FRIENDLY OTHER	
s_youknow	−0.00299	**s_hedge**	−0.00328	**o_i**	−0.00214	**o_i**	−0.00292
s_hedge	−0.00185	s_negemo	−0.00276	o_hedge	−0.00160	s_loud	−0.00177
s_uh	−0.00173	s_interrupt	−0.00251	o_loud	−0.00135	o_you	0.00101
o_interrupt	0.00112	s_drink	−0.00207	s_like	0.00113	**s_ntri**	0.00105
s_restart	0.00123	o_imean	−0.00172	s_sentence	0.00115	o_sex	0.00113
o_hedge	0.00130	o_like	−0.00144	**s_midlaugh**	0.00118	o_youknow	0.00132
o_rate	0.00131	s_i	0.00118	s_i	0.00121	**o_mimicccont**	0.00147
o_love	0.00172	s_agree	0.00137	**s_ntri**	0.00151	s_you	0.00159
o_midlaugh	0.00176	**o_interrupt**	0.00148	o_agree	0.00186	o_love	0.00200
o_sentence	0.00181	o_you	0.00172	**o_mimicccont**	0.00229	s_youknow	0.00246
o_appr	0.00202	o_appr	0.00174	s_varinten	0.00247	s_funcaccom	0.00262
s_sex	0.00250	s_swear	0.00181	o_rate	0.00277	**o_varinten**	0.00298
s_sympathy	0.00412	**s_restart**	0.00187	**o_varinten**	0.00438	**s_midlaugh**	0.00456
s_midlaugh	0.00570	s_like	0.00190				
s_varinten	0.01070	o_sex	0.00195				
		o_sentence	0.00219				
		s_love	0.00234				
		s_sympathy	0.00243				
		o_longturn	0.00420				
		s_varinten	0.00554				
		s_midlaugh	0.00573				

Table 13
Feature weights for awkwardness detection (median weights of the randomized runs) for the non-zero predictors for each classifier.

MALE				FEMALE			
AWK SELF		AWK OTHER		AWK SELF		AWK OTHER	
o_initlaugh	−0.00307	s_academic	−0.00269	o_time_ratio	−0.00216	o_totalwords	−0.00709
o_food	−0.00301	**s_imean**	−0.00258	s_interrupt	−0.00195	o_laughaccom	−0.00401
o_love	−0.00232	s_like	−0.00228	**o_drink**	−0.00148	s_meta	−0.00275
s_imean	−0.00222	o_ntri	−0.00224	**o_loud**	−0.00115	o_agree	−0.00264
o_negate	−0.00210	o_midlaugh	−0.00209	**s_rate**	0.00150	**o_drink**	−0.00261
s_restart	−0.00174	o_agree	−0.00116	o_um	0.00173	s_initlaugh	−0.00240
s_appr	−0.00162	o_like	0.00106	**s_negate**	0.00181	**o_loud**	−0.00180
s_food	−0.00150	s_love	0.00141	**o_appr**	0.00210	o_sympathy	−0.00172
o_funcaccom	−0.00122	o_hate	0.00156	s_appr	0.00231	s_rateaccom	−0.00142
s_totalwords	−0.00108	o_restart	0.00159	o_hate	0.00233	s_i	−0.00137
s_you	−0.00106	o_pitchvar	0.00217	s_restart	0.00237	s_varinten	0.00117
s_sex	0.00110	s_youknow	0.00252	**o_question**	0.00276	o_hedge	0.00124
s_meta	0.00136	s_swear	0.00256	o_imean	0.00281	o_rate	0.00128
o_meta	0.00142	**o_academic**	0.00257	s_love	0.00290	s_question	0.00194
o_academic	0.00174	s_negemo	0.00286	**s_hedge**	0.00365	**s_negate**	0.00196
s_interrupt	0.00221	o_um	0.00336	s_like	0.00421	o_uh	0.00239
s_midlaugh	0.00250	**s_hedge**	0.00398			**o_appr**	0.00263
s_agree	0.00263	**s_question**	0.00893			**s_rate**	0.00307
s_varinten	0.00285					**o_question**	0.00309
s_um	0.00290					s_academic	0.00320
s_hedge	0.00386					**s_hedge**	0.00358
s_question	0.00892					s_laughaccom	0.00374
						s_longturn	0.00405

Table 14
Feature weights for assertiveness detection (median weights of the randomized runs) for the non-zero predictors for each classifier. Boldfaced features hold across self and other ratings, and boldfaced features that were also significant in the quartile study are underlined.

MALE				FEMALE			
ASSERT SELF		ASSERT OTHER		ASSERT SELF		ASSERT OTHER	
s_negate	−0.00654	o_midlaugh	−0.00495	o_youknow	−0.00176	o_negate	−0.00285
s_uh	−0.00343	**s_imean**	−0.00400	**o_midlaugh**	−0.00168	o_varinten	−0.00232
o_love	−0.00303	**s_uh**	−0.00325	s_youknow	−0.00120	s_pitchvar	−0.00156
s_imean	−0.00236	s_funcaccom	−0.00318	o_agree	−0.00104	s_negemo	−0.00147
s_i	−0.00223	**o_question**	−0.00267	s_love	0.00110	s_sympathy	−0.00129
s_academic	−0.00212	o_negate	−0.00224	s_ntri	0.00117	o_minpitch	−0.00128
o_question	−0.00167	o_imean	−0.00223	**s_academic**	0.00122	s_uh	−0.00120
s_ntri	−0.00156	**s_negate**	−0.00219	s_rate	0.00125	**o_midlaugh**	−0.00117
s_minpitch	−0.00148	**s_minpitch**	−0.00218	s_rateaccom	0.00149	s_you	−0.00110
o_negemo	−0.00137	o_um	−0.00152	o_like	0.00159	**s_academic**	0.00100
o_initlaugh	0.00110	**o_love**	−0.0010	**s_like**	0.00164	o_swear	0.00103
s_rateaccom	0.00200	s_pitchvar	0.00120	s_question	0.00178	**s_i**	0.00120
o_like	0.00201	o_laughaccom	0.00141	**s_i**	0.00202	**s_varinten**	0.00146
s_totalwords	0.00215	o_longturn	0.00160	**s_varinten**	0.00205	o_rate	0.00164
o_laughaccom	0.00227	**s_hedge**	0.00164	o_initlaugh	0.00206	o_appr	0.00172
o_rate	0.00241	s_laughaccom	0.00172	**s_mimicccont**	0.00237	s_imean	0.00179
s_hedge	0.00243	o_you	0.00183	s_appr	0.00340	o_question	0.00183
s_question	0.00316	s_appr	0.00198	**s_restart**	0.00436	o_love	0.00210
s_youknow	0.00330	s_you	0.00201	**s_negate**	0.00465	**s_mimicccont**	0.00211
o_totalwords	0.00332	s_agree	0.00218			**s_like**	0.00237
s_negemo	0.00745	o_swear	0.00229			o_loud	0.0027
		s_totalwords	0.00247			**s_restart**	0.00366
		o_rate	0.00258			**s_negate**	0.00378
		s_youknow	0.00310				
		o_totalwords	0.00318				
		s_um	0.00318				
		s_like	0.00388				
		s_negemo	0.00429				

self-report as, friendlier, and women who use more *I* self-report as, but are not perceived as, friendlier. There are also strong characteristics of the speech of the friendly person's interlocutor; in both the decile and quartile studies men accommodate friendly woman's content words and have more variable intensity, while women use more appreciations with friendly men.

The most robust cue to awkwardness was hedges, which are more frequent in both studies in awkward men and women, whether self- or other-rated. Awkward women in both studies, and awkward men in the quartile study, were also more likely to use negation. Awkward women tended to speak faster, while awkward men ask more questions and are less likely to use *I mean* and more likely to use negative emotion. In the quartile but not decile studies, awkward men are also likely to speak less, speak softer and with less variation in intensity, and use less clarification questions.

Assertive women (self- and other-assessed, decile and quartile) use more negation, more disfluent restarts, more *I*, more function word accommodation, and tended to use more appreciations and clarification questions. Assertive men used more total words, more negative emotion, less negation, and tended to use less *uh*. In the decile study assertive men dropped their minimum pitch; in the quartile study assertive men raised their maximum pitch, and were more likely to swear and use *you*. Self-assessed (but not other-assessed) assertive men asked more questions. We defer deeper discussion of features to Section 8.

5. Study 2: speaker-independent detection of interpersonal stances

Our main study suggests that all four interpersonal stances are detectable. In study 1, speakers could appear in both the training set and the test set; each test set individual appeared about 5 times in the training set (averaged across

Table 15
The number of positive and negative examples in the training and test for the self-report of male flirtation in one round of the event-based cross validation.

Training Set		Test Set	
# Positive	# Negative	# Positive	# Negative
47	62	40	25

all randomizations of our 16 tests). Although the speaker identity was not a feature available to the classifier, having a test speaker in training nonetheless means the algorithm has seen characteristics of that speaker. In order to see to what extent the classification accuracy is due to this kind of speaker-specific knowledge, in this section we describe our speaker-independent detection of stances.

5.1. Methods

The SpeedDate corpus was created by three separate speed dating events a few weeks apart with completely different participants. Study 1 lumped together all three events. In this study we use distinct events for training and testing, cross-validating by training on two of the events and testing on the third event. This ensures that every example will be tested in each of the three sub-experiments and that no speaker will be in both the training and test set.

Recall that for each of our 16 classification tasks from Experiment 1, we took the top ten percent of Likert rating across all the events as positive examples for a style and we took the bottom ten percent as negative examples for a style. Given our small dataset, doing event-based cross validation with this data split can lead to very uneven distributions of positive class and negative class examples, since one event could contain very few examples of either particular class. An example of the imbalance for an iteration of event-based cross validation can been seen in Table 15.

The imbalance in the training sets poses an issue for using an SVM. We can see in the SVM optimization problem in Eq. 1 that each training example contributes equally to the objective, causing a bias towards the majority class in our imbalanced training setting. To remove this bias in the objective function we trained a weighted SVM (Huang and Du, 2005), which are useful if one of the classes appears with low frequency in training. Infrequent classes can be made more important in a weighted SVM by given them larger weights, generally improving the classification accuracy of that class. Since we have one class that appears at a lower rate than the other and we want to accurately classify both of our classes we used weights of (# positive training examples)/(# training examples) on the negative examples and weights of (# negative training examples)/(# training examples) on the positive examples. This choice of weights ensures that both classes contribute to the objective function of the weighted SVM Eq. (6) equally; the weight on training examples of one class is proportional to the number of training examples in the other class.

$$\begin{aligned} &\min_{\gamma,w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} weight_i * \xi_i \\ &\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ &\xi_i \geq 0, \quad i = 1, \ldots, m \end{aligned} \tag{6}$$

We trained a weighted SVM for each of our 16 tasks mentioned in Study 1. We used the same feature transformations as Study 1, but used an alternative validation set (from holding out 20% of the training set) to learn the hyperparameters since we no longer have balanced folds. The top decile and bottom decile define a cutoff at a particular Likert rating; at this rating there are generally many more conversations than needed at that rating to build the top decile or bottom decile. In this experiment, we randomly select data from these 2 pools to build the top and bottom deciles and run across 5 such randomizations.

Accuracies of our SVM model, along with a majority class baseline, are shown below. Note that because we chose the top 10% and bottom 10% as positive and negative examples globally, the labels in the test event will always be

Table 16
Accuracy of binary classification of each conversation side, where chance is 50%. The top two rows describe the stance as labeled by the speaker; the second predicts the stance as labeled by the alter. Each pair of rows show the model against a baseline. Accuracies are aggregated from 5 randomized runs of event-based cross validation and are all significantly better than the baseline (McNemar's test $p < .05$), except self-reported friendliness in females, marked with a dagger.

		Friendly		Flirt		Awk		Assertive	
		M	F	M	F	M	F	M	F
wSVM	Self	.53	.41†	.57	.53	.47	.50	.58	.46
Baseline	Self	.44	.42	.48	.37	.34	.40	.45	.41
wSVM	Other	.58	.54	.59	.52	.50	.53	.57	.53
Baseline	Other	.44	.46	.48	.34	.39	.41	.42	.36

tilted towards the minority class of the training set.[4] This issue is mitigated by the fact that our SVM learning does not take advantage of this class bias. Instead, our training algorithm is affected by the same majority class issue as the baseline because it has no knowledge of the constraint.

5.2. Results

We report the results for the 16 speaker-independent classifiers in Table 16. In every case except for self-report of female friendliness the SVM classifier is significantly better than the majority class baseline (McNemar's test, 2-tailed, $p < .05$).

5.3. Discussion

Our results show that for 15 of our 16 stances, even with an unseen speaker it is possible to extract interpersonal styles with performance better than the baseline majority classifier. The performance improvement over the baseline, however, is not as large as in Study 1. This fact, together with the poor performance on self-report of friendliness in females suggests that allowing the classifier to see the range of traits from test speakers in Study 1 was a useful cue to accurate classification.

6. Study 3: linguistic versus non-linguistic features

We have shown in Studies 1 and 2 that linguistic factors are useful in predicting a conversational style both when we have seen the speaker before and when we haven't. In these studies, however, we did not consider any non-linguistic factors that might play a role in extraction of interpersonal style. It is possible that the predictive power of our linguistic features might already by captured by non-linguistic traits of a person like *homophily* (the similarity between the speakers) or physical characteristics like height or age difference. Our goal in this study is to explore the relevance of non-linguistic features to the interpersonal stances, and see whether non-linguistic and linguistic features can be combined.

6.1. Methods

We chose three stances to explore the role of non-linguistic features: self-report of flirtation for both male and females, and self-report of assertiveness for men. We chose flirtation because some of the non-linguistic features were designed to be related to romantic attraction, pairing it with assertiveness, a stance for which we have no reason to expect the non-linguistic cues to have any predictive power as it is not directly related to romantic attraction.

[4] This is a consequence of doing cross validation. Overall the number of positive class examples is the same as the number of negative class examples, so if we are training on two of the events, and these two events have an excess of one class, then the remaining event which is the test set must have an excess of the other class because the number of examples in each class is equal globally.

Table 17
Non-linguistic features, calculated from a survey that each participant filled out at the beginning of the event.

ORDER	Order of the date (*n*th date the speaker has participated in tonight)
SPEAKER HEIGHT	The height of the speaker
SPEAKER BMI	The body mass index of the speaker
OTHER HEIGHT	The height of the interlocutor
OTHER BMI	The BMI of the interlocutor
AGE DIFFERENCE	The age difference between the speaker and the interlocutor
INTEREST SIMILARITY	The interest similarity between the speaker and the interlocutor.


(17) How interested are you in the following activities? (1=very uninterested, 10=very interested)

Playing sports/ athletics	1	2	3	4	5	6	7	8	9	10
Watching sports	1	2	3	4	5	6	7	8	9	10
Body building/exercising	1	2	3	4	5	6	7	8	9	10
Dining out	1	2	3	4	5	6	7	8	9	10
Museums/galleries	1	2	3	4	5	6	7	8	9	10
Art	1	2	3	4	5	6	7	8	9	10
Hiking/camping	1	2	3	4	5	6	7	8	9	10
Video Gaming	1	2	3	4	5	6	7	8	9	10
Dancing/clubbing	1	2	3	4	5	6	7	8	9	10
Reading	1	2	3	4	5	6	7	8	9	10
Watching TV	1	2	3	4	5	6	7	8	9	10
Theater	1	2	3	4	5	6	7	8	9	10
Movies	1	2	3	4	5	6	7	8	9	10
Going to concerts	1	2	3	4	5	6	7	8	9	10
Music	1	2	3	4	5	6	7	8	9	10
Shopping	1	2	3	4	5	6	7	8	9	10
Yoga/meditation	1	2	3	4	5	6	7	8	9	10

Fig. 4. The registration survey with the list of interests filled out by each speed date participant.

For each of the styles we trained two SVMs: one using just the non-linguistic features and one using both the linguistic and non-linguistic features. We also trained an L_1-regularized logistic regression classifier using just the non-linguistic features to help in feature weight analysis. We use the feature normalization scheme, hyperparameter learning, and the cross-validation scheme outlined in Study 1 to train and evaluate our models.

6.2. Non-linguistic features

Table 17 shows the non-linguistic features, each calculated from a survey that was handed out at the beginning of the event. Participants gave their height and weight (from which we computed their BMI), their age, and filled out a set of hobby interests.

The registration survey with the list of interests are shown in Fig. 4. Interests were labeled from 1 to 10 with 10 indicating very interested in the topic. We calculated interest similarity between two speakers using the L_2 norm, i.e. by taking the square root of the sum of squares of interest difference across several interest categories. The L_2 norm assumes that lacking a little similarity on a couple of categories isn't as different as being very unsimilar in one category.[5]

6.3. Results

Table 18 shows results for the three stances. Non-linguistic features perform better than chance for all 3 tasks (paired t-test, 2-tailed, $p = .0001$) but not nearly as well as the linguistic features by themselves (paired t-test, 2-tailed, $p = .0001$).

For flirtation, models with both the linguistic and non-linguistic features perform better than either feature set by itself.[6] For assert, however, the non-linguistic features hurt performance when added to the linguistic features.

[5] We found that the L_2 norm outperformed the L_1 norm on classification accuracy for calculating interest similarity.

[6] Using a paired 2-tailed t-test, the models for men are significantly better ($p = .0001$) as is the model for women ($p = .05$).

Table 18
Accuracy of binary classification of each conversation side, using linguistic features, non-linguistic features (traits), or both, where chance is 50%. These accuracy results were aggregated from 30 randomized runs of 5-fold cross validation.

	Linguistic features	Traits	Linguistic + Traits
Male Flirt (self)	.66	.64	.72
Female Flirt (self)	.74	.55	.76
Male Assert (self)	.73	.55	.72

Table 19
Feature weights for non-linguistic features (median weights of the randomized runs) for the non-zero predictors for each classifier.

MALE FLIRT SELF		FEMALE FLIRT SELF		MALE ASSERT SELF	
o_bmi	−0.00476	o_bmi	0.00115	s_bmi	0.00200
s_height	0.00102	s_height	−0.00208	s_height	0.00102
s_order	0.00106	s_bmi	0.00210		
s_bmi	0.00217				

6.4. Feature analysis

Table 19 shows the significant non-linguistic features. For both male and female flirtation, physical traits are important. Men or women with higher BMI are more likely to report themselves as flirting, as are both taller men and shorter women. Men tend to flirt with low-BMI women, while women, by contrast, flirt with high-BMI men. As men progress through more dates, they report themselves as getting more flirtatious. Males that report themselves as assertive tend to be taller and have a higher BMI. Interest homophily was not a significant predictor of any of the stances.

6.5. Discussion

Linguistic features clearly offer information beyond what is captured in non-linguistic variables like height, BMI, and homophily. Non-linguistic features are mainly useful in predicting flirtation, confirming our intuition that these particular non-linguistic features, which were designed to predict romantic attachment, are not appropriate for predicting assertiveness. Flirtation, by contrast, is a more likely proxy for romantic interest than is assertiveness.

7. Study 4: systematic differences between self- and other-perceptions

Within the Speed Date corpus, the speaker's interpersonal stance is labeled by both the speaker and the hearer. In the prior section, we focused on features that were robust indicators of a stance when labeled by the speaker and when labeled by the hearer. In many cases, however, speakers and hearers draw on different signals in this labeling task. For example, it is possible that men may think they are indicating friendliness with certain linguistic signals while their female interlocutors may identify male friendliness in alternative signals.

The Brunswikian lens is one model of how the relationship between a surface linguistic feature and a kind of affective meaning like a stance may be different for the producer of that affect than for the perceiver of the affect (Scherer, 1978). There is some evidence for this Brunswikian difference in our data. Some features have stronger associations for the speaker than the perceiver. For example men who self-report as flirting use less negation, but women do not seem to rely on negation as a cue for labeling flirtation. Similarly, women who use more *I* self-report as friendlier, but they are not perceived as friendlier when they use *I*. Other features have stronger associations for the perceiver than the speaker; men who use agreement or swearing are perceived as, but do not self-report as, friendlier.

In addition to differences in the linking of features to stances, we also found systematic differences between the labels that the two participants in a conversation gave to their stances in that conversation. We noticed that speakers in general tend to assign similar values to their own stance as they do to their partner's stance. This effect seemed to be much stronger than the tendency of the two speakers to agree on the stance label for a participant.

Table 20
Correlations between speaker intentions and perception for all four styles.

Variable	Self-perceive-Other &Self-perceive-Self	Self-perceive-Other &Other-perceive-Other
Flirting	.73	.15
Friendly	.77	.05
Awkward	.58	.07
Assertive	.58	.09

We tested this observation by computing two sets of Pearson correlations. The first tests how similarly the respondent labels their own stance to their labeling of their partner. The second tests how similarly the respondent labels their partner with how similarly the partner labels their self. As Table 20 shows, for all four styles, speakers' perception of others is strongly correlated with the speakers' perception of themselves, far more so than with what the others actually think they are doing.[7]

Note that although perception of the other does not correlate highly with the other's intent for any of the styles, the correlations are somewhat better (.15) for flirting, perhaps because in the speed-date setting speakers are focusing more on detecting this behavior (Higgins and Bargh, 1987). It is also possible that for styles with positive valence (friendliness and flirting) speakers see more similarity between the self and the other than for negative styles (awkward and assertive) (Krahé, 1983).

Why should this strong bias exist to link their own stance with that of the other of the other? One possibility is that speakers are just not very good at capturing the intentions of others in 4 min. Speakers would instead base their judgments on their own behavior or intentions, perhaps because of a bias to maintain consistency in attitudes and relations (Festinger, 1957; Taylor, 1970) or to assume there is reciprocation in interpersonal perceptions (Kenny, 1998).

We propose an alternative hypothesis, however, which is that our participants describe these interpersonal stances as being fundamentally about the relation between the two actors and the conversation as a whole, and not merely facts about an individual speaker. That is, speakers are to some extent labeling "conversations that felt friendly" or "conversations that felt awkward" instead of just "friendly interlocutors". Evidence for this hypothesis is the large number of cases where the linguistic features of the "other" for a stance bears a strong relation to the stance labeled for the speaker.

If true, this hypothesis suggests that research on interpersonal stance and affective meaning like personality in the dyad need to be careful in teasing apart aspects of the individual from aspects of the conversation. One solution that we are exploring is the use of statistical methods like the Actor-Partner Interdependence Model (Kenny et al., 2006), which distinguish intra-personal actor effects from inter-personal partner effects.

8. General discussion

Our results show that it is possible to use automatically extracted prosodic, lexical, and dialog features to extract interpersonal stances (flirtation, friendliness, awkwardness, assertiveness), with performance of up to 78% compared to a 50% baseline for contrasting prototypes when test speakers occur in training, and up to 53% compared to a 37% baseline with unseen test speakers. The linguistic features we used offer information for classification even beyond non-linguistic features like shared interests (homophily) and physical attributes like weight and height and the basic prosodic features that have traditionally been studied in the bonding literature in sociology and social psychology.

These results also have a number of implications for modeling stances and their associated linguistic features. One implication is the presence of a sort of *collaborative conversational style* (probably related to the *collaborative floor* of Edelsky (1981) and Coates (1996)), cued by the use of turn-medial/final laughter, sympathy (in men), clarification questions (in women), an avoidance of negative emotional language and hedges, and to some extent accommodation, appreciations, and the use of *you*. These collaborative techniques were used by both women and men who were labeled

[7] This was true no matter how the correlations were run, whether with raw Likert values, with ego-centered (transformed) values and with self ego-centered but other raw.

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. Although we found no prosodic correlations of friendliness in our decile study, in our larger (quartile) study, we found that friendly women had higher maximum pitch and more pitch variance, consistent with the literature (Chen et al., 2004; House, 2005; Li and Wang, 2004; Liscombe et al., 2003; Mairesse et al., 2007; Gravano1 et al., 2011).

Our results on flirtation suggested different patterns for men and women. Flirting in women is associated with negation, the word *like*, and collaborative style (appreciations, medial laughter) and in the decile study with the word *I*. Flirting in men is associated with greater use of *you*, *you know*, *um*, and words about sex, as well as less likelihood of talking about work. We did not in general see the increase in F0 maximum suggested by the results of Puts et al. (2011), although it was at least associated with other-reported flirtation in the decile study.

The fact that men tend to use *you* and *you know* (soliciting hearers orientation toward the speaker's talk (Schiffrin, 1987)) while women tend to use *I* suggests a conversational norm in which women are the target or topic of the conversation. Previous research on politeness (Brown and Levinson, 1978), strategic local action (Leifer and Rajah, 2000) and contextualizational cues (Gumperz, 1982) regard this kind of targeting as a key feature of interaction. In the case of politeness, the empowered party is usually the target of communication and their face a matter of mutual concern and protection (Goffman, 1967). Since women are much more selective in these dates and hence generally make the decision on whether to continue the date, her face may be a natural target for both parties. We are currently analyzing other cues and implications of the topicality of women in the conversation.

As for the negation used by flirting women, this may be related to teasing and/or self-deprecation; we return to this issue below in our further discussion of negation.

The linguistic manifestations of an assertive stance do show some commonalities with extraversion, discussed in Section 2. Features of both extraversion and assertive speech include talking more (for men only), variable intensity (for women in the decile study only), faster speech (for self-reported women in both the decile and quartile study), and increased use of negative emotional words (for men only) and swearing (for men only). The decrease in men's pitch floor that we found with assertive men in the decile study also seems to be consistent with the findings of Feinberg et al. (2005) that women find men with lower fundamental frequency more masculine, but we also saw an increased pitch maximum for assertive men in the quartile study, which makes us loath to draw conclusions here.

On the other hand, some of our assertive features were very different from previous results on extraversion. For example we found the use of negatives to be the strongest assertive feature for women but Mairesse et al. (2007) found extraverts to use significantly less negatives, at least in written essays. We saw increased use of hedges by assertive men, but previous research suggest that extraverts use less hedges. Our finding of increased restarts (disfluencies) by assertive women is also not what would be expected for extraversion. In sum, our results suggest that ratings of assertiveness did capture some aspects of the extraverted personality, but that assertiveness and extraversion were not completely aligned and further study is warranted.

For awkwardness, many of the linguistic cues seem to be associated with psychological distancing, the mechanism of psychologically removing oneself from a difficult or traumatic event. Previous research has shown that of first-person singular, longer words, articles, and avoidance of the present tense are signs of distancing (Cohn et al., 2004). The decreased use of *I* in some awkward women, and the strongly decreased use of *I mean* in awkward men is thus consistent with these distancing cues (Schiffrin, 1987). But by far the strongest linguistic association with awkward men and women is the use of hedges. Hedges like *sort of*, *kind of*, *maybe*, and *a little* are used to express the speaker's lack of commitment to a proposition, but our results suggest that hedges are also used metalinguistically to indicate the speaker's psychological distancing from or discomfort with the situation; words which are distancing at the semantic or pragmatic level acquire the metapragmatic connotations of distancing. This use of hedges to indicate discomfort is consistent with research showing that hedges can extend their semantic or pragmatic meaning to metalinguistic meaning (Jurafsky, 1996), and research showing that other kinds of meaning like negation have this metalinguistic extension (Potts, 2011). Our results suggests that hedges may be a useful cue to augment current models for detecting psychological distancing. The fact that awkward men in the quartile study are also likely to speak less, speak more softly, and use less variation in intensity may also be related to a lack of engagement. This association between distancing and awkward speakers may be due to personality factors like low self-esteem, or it may simply be that speakers who are distancing themselves from an uncomfortable date are likely to describe themselves (or be described) as awkward.

In addition to the results on the individual stances, our work also has a number of implications for individual features themselves. We show that the two filled pauses *um* and *uh* function very differently. *Um* marks flirting for both men

and women. *Uh*, on the other hand, has a negative correlation with almost all of the stances. *Uh* is used less by assertive men or women, and less by men who describe themselves as friendly or flirtatious. This suggests that *uh* is somehow dispreferred or marked, an idea that is consistent with recent research on *uh* by Acton (2010), who showed that speakers associate *uh* with connotations of stupidity.

Another implication of our study is the separation of laughter into different classes. Distinguishing turn-initial and turn-medial laughter is one of the contributions of our study. Turn-medial or turn-final laughter is part of friendly collaborative conversational style. In a follow-up investigation, we analyzed a number of example of laughter in the corpus, and found that this kinds of laughter is often used by a speaker to poke fun at themselves, as in the following two examples:

FEMALE: Why are you single?
MALE: Why I'm single? That's a good question. [laughter]

MALE: And I stopped by the–the beer party the other day.
FEMALE: Oh goodness. And you saw me in- [laughter]

Turn-initial laughter, by contrast, seems to play other functions. One common use of initial laughter is laughing at your other's jokes, as the woman does in the following:

MALE: . . ."speed filling out forms" is what this should be called.
FEMALE: [laughter] Yeah.

Initial laughter in the SpeedDate corpus is also used for teasing the other, as in the following example, where the man accuses the woman of being defensive:

MALE: You're on the rebound.
FEMALE: huh–uh.
MALE: [laughter] Defensive.

These corpus investigations suggest that turn-final laughter is self-directed, a way for a speaker to poke fun at himself or herself, while turn-initial laughter is other-directed, a way of laughing at the other's jokes or teasing them. The difference between these two kinds of laughter, cued as they are just by turn position, makes them easy to differentially extract. The roles of these two kinds of laughter should be examined in other domains.

Our results also suggest a number of directions for the study of negation. We found that women's use of negation (especially *don't*, *not*, and *no*) is particularly associated with flirtation, assertiveness, and awkwardness. In a brief corpus investigation we explored the link with flirtation. A number of flirty cases of negation seem to be women teasing men, challenging them by ordering them not to do something, or refusing to do something for them:

MALE: I have to say that that's a great question.
FEMALE: No, don't say that. That's a stalling technique in a business interview. It's not allowed here. [Laughter].

Below are some other negative utterances with these properties:

FEMALE: What three things? Really three things? You're not going to say, like, three best albums or-?
FEMALE: Are we in the same- I don't want to give you my email address.
FEMALE: What if I don't like your plan?

Negation seems also to be used by women for self-deprecation; in particular, note the co-occurrence in both of these examples with turn-medial laughter, another cue for self-deprecation:

FEMALE: Do you know how to do math? [Laughter] I don't. So I'm an English major.
FEMALE: But actually I haven't studied in a while, so I uh, I've forgotten a lot. [Laughter]

Deeper analysis of the relation of negation to flirtation, and understanding the links to awkwardness and assertiveness in women and possibly in men remains an open question for future research, as does further exploration of the linguistic cues to both self-deprecation and teasing.

For accommodation, we found effects, but not exactly where we predicted them. We found no effects of accommodation in friendly or flirtatious speakers. On the other hand, we found strong effects of increased accommodation in assertive women. For flirting and friendly women we find strong effects of accommodations not in their own speech but in that of their partner. That is, men talking with friendly or flirtatious women were more likely to accommodate to their words. Why accommodation might be a response to another's flirtation remains a subject for further research.

Our work also has methodological implications, both for lexical and prosodic cues. Most previous work on the extraction of social meaning has looked at classes of single words or simple disfluencies. Our work suggests that detection of entire dialog acts (sympathy, appreciations, clarification questions) as well as lexical families such as hedges, can help in meaning extraction. For prosody, much prior research has either focused on only one or two prosodic features such as mean F0, or used a very large variety of prosodic features in a regression. The use of F0 alone obviously underdescribes the space, but the use of a large collection of features also has problems, especially since prosodic features tend to be collinear. While the accuracy of regression and SVM classifiers are relatively robust to collinear features, analysis of their feature weights are not. Here, we have adopted an inductive approach based on factor analysis that reveals interpretable acoustic manipulations being done in the course of conversations. Our factors suggest that minimum and maximum pitch are distinctly relevant for social meaning, while mean pitch is not, and min, mean, and max intensity pattern together. These hypotheses can be tested in our further work.

The fact that the two interlocutors are often assigned the same stance suggests that stance is to a great extent a property of the dyadic interaction rather than the individual, a result with important implications for the study of interpersonal stance.

Finally, the references to self-deprecating and teasing actions above demonstrate how many speech acts are polysemous, deriving their meaning from their situation in sequence. Much prior work that relies solely on turn-internal features like single words or prosody gives little recognition to the dynamic contextual nature of dialog meaning. To advance our understanding of the linguistic realizations of social meaning we need to model more complex acts to better represent the way resources like laughter, for example, can be used in both self-deprecating and other-deferring ways. Our attempts to define acts such as sympathy, appreciations, or initial and medial laughter begin this process, but future work must look at pair-wise ordering and other structures above the individual speech act to capture the nuances of actions, demeaner displays, teasing, and the full richness of dialog maneuvers (Goffman, 1981). Another important goal for future work is to take into count the joint influence of stances; if flirtation and assertiveness are present in the same conversation, for example, they will affect the distribution of features together. Some sort of joint estimation is likely called for to disentangle the influences.

All of the results presented here should be regarded with some caution. The sample is small and not a random sample of English speakers or American adults, and speed dating is not a natural context for expressing every conversational style. 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 s) 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 several types of interpersonal stance have linguistic signals, mainly within gender but also to some extent across genders. Further investigation of these and other new variables at the spectral, prosodic, lexical, grammatical, and discourse level can only help enrich our understanding of social meaning and its linguistic realization, and move the field toward automatic extraction of social meaning for social computing applications.

Acknowledgments

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

References

Acton, E., 2010. Date-uh analysis: On gender differences in the distribution of um and uh, presentation at NWAV-2010.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human–computer dialog. In: INTERSPEECH-02.

Boersma, P., Weenink, D., 2005. Praat: Doing Phonetics by Computer (version 4.3.14), [Computer program]. http://www.praat.org/ (retrieved 26.05.05).

Brave, S., Nass, C., Hutchinson, K., 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.

Brennan, S.E., Schober, M.F., 2001. How listeners compensate for disfluencies in spontaneous speech. Journal of Memory and Language 44, 274–296.
Brown, P., Levinson, S.C., 1978. Politeness: Some Universals in Language Usage. Cambridge University Press.
Chartrand, T.L., Bargh, J.A., 1999. The chameleon effect: the perception-behavior link and social interaction. Journal of Personality and Social Psychology 76 (6), 893–910.
Chen, F., Li, A., Wang, H., Wang, T., Fang, Q., 2004. Acoustic analysis of friendly speech. In: ICASSP-2004, pp. 569–572.
Coates, J., 1996. Women Talk. Blackwell.
Cohn, M.A., Mehl, M.R., Pennebaker, J.W., 2004. Linguistic markers of psychological change surrounding September 11, 2001. Psychological Science 15, 687–693.
Collins, S.A., 2000. Men's voices and women's choices. Animal Behaviour 60 (6), 773–780.
Collins, S.A., Missing, C., 2003. Vocal and visual attractiveness are related in women. Animal Behaviour 65 (5), 997–1004.
Edelsky, C., 1981. Who's got the floor? Language in Society 10, 383–421.
Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008. LIBLINEAR: a library for large linear classification. Journal of Machine Learning Research 9, 1871–1874.
Feinberg, D.R., DeBruine, L.M., Jones, B.C., Perrett, D.I., 2008. The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. Perception 37 (4), 615.
Feinberg, D.R., Jones, B.C., Little, A.C., Burt, D.M., Perrett, D.I., 2005. Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. Animal Behaviour 69 (3), 561–568.
Festinger, L., 1957. A Theory of Cognitive Dissonance. Row, Peterson, Evanston, IL.
Finkel, E.J., Eastwick, P.W., 2008. Speed-dating. Current Directions in Psychological Science 17 (3), 193.
Finkel, E.J., Eastwick, P.W., Matthews, J., 2007. Speed-dating as an invaluable tool for studying romantic attraction: a methodological primer. Personal Relationships 14 (1), 149–166.
Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.
Godfrey, J., Holliman, E., McDaniel, J., 1992. SWITCHBOARD: telephone speech corpus for research and development. In: ICASSP-92, San Francisco, pp. 517–520.
Goffman, E., 1967. Interaction Ritual: Essays on Face-to-Face Behavior. Pantheon Books.
Goffman, E., 1981. Forms of Talk. University of Pennsylvania Press.
Goodwin, C., 1996. Transparent vision. In: Ochs, E., Schegloff, E.A., Thompson, S.A. (Eds.), Interaction and Grammar. Cambridge University Press, pp. 370–404.
Goodwin, C., Goodwin, M., 1987. Concurrent operations on talk: notes on the interactive organization of assessments. IPRA Papers in Pragmatics 1, 1–52.
Gravano, A., Hirschberg, J., 2011. Turn-taking cues in task-oriented dialogue. Computer Speech and Language 25 (3), 601–634.
Gravano1, A., Levitan, R., Willson, L., Beňuš, Štefan., Hirschberg, J., Nenkova, A., 2011. Acoustic and prosodic correlates of social behavior. In: INTERSPEECH 2011.
Grossman, R., Bemis, R., Plesa Skwerer, D., Tager-Flusberg, H., 2010. Lexical and affective prosody in children with high-functioning autism. Journal of Speech, Language, and Hearing Research 53 (3), 778.
Gumperz, J., 1982. Discourse Strategies. Cambridge University Press.
Heritage, J., 1984. Garfinkel and Ethnomethodology. Polity Press, Cambridge.
Higgins, E.T., Bargh, J.A., 1987. Social cognition and social perception. Annual Review of Psychology 38, 369–425.
House, D., 2005. Phrase-final rises as a prosodic feature in wh-questions in Swedish human–machine dialogue. Speech Communication 46 (3–4), 268–283.
Huang, Y.-M., Du, S.-X., 2005. Weighted support vector machine for classification with uneven training class sizes. In: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol. 7, pp. 4365–4369.
Ireland, M.E., Slatcher, R.B., Eastwick, P.W., Scissors, L.E., Finkel, E.J., Pennebaker, J.W., 2011. Language style matching predicts relationship initiation and stability. Psychological Science 22 (1), 39.
Jaffe, J., Anderson, S.W., 1979. Communication rhythms and the evolution of language. In: Siegman, A., Feldstein, S. (Eds.), Of Speech and Temporal Speech Patterns in Interpersonal Contexts. Lawrence Erlbaum Associates, Hillsdale NJ.
Jones, B.C., Feinberg, D.R., DeBruine, L.M., Little, A.C., Vukovic, J., 2010. A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. Animal Behaviour 79 (1), 57–62.
Jurafsky, D., 1996. Universal tendencies in the semantics of the diminutive. Language 72, 533–578.
Jurafsky, D., Shriberg, E., Biasca, D., 1997. Switchboard SWBD-DAMSL Labeling Project Coder's Manual, Draft 13. Tech. Rep. 97-02, University of Colorado Institute of Cognitive Science.
Jurafsky, D., Shriberg, E., Fox, B., Curl, T., 1998. Lexical, prosodic, and syntactic cues for dialog acts. In: Proceedings, COLING-ACL Workshop on Discourse Relations and Discourse Markers, pp. 114–120.
Kenny, D., 1998. Interpersonal Perception: A Social Relations Analysis. Guilford Press, New York, NY.
Kenny, D.A., Kashy, D., Cook, W.L., 2006. Dyadic Data Analysis. Guilford Press, New York.
Krahé, B., 1983. Self-serving biases in perceived similarity and causal attributions of other people's performance. Social Psychology Quarterly 46, 318–329.
Lakoff, G., 1973. Hedges: a study in meaning criteria and the logic of fuzzy concepts. Journal of Philosophical Logic 2 (4), 458–508.
Lee, C.M., Narayanan, S.S., 2002. Combining acoustic and language information for emotion recognition. In: ICSLP-02, Denver, CO, pp. 873–876.
Leifer, E., Rajah, V., 2000. Getting observations: strategic ambiguities in social interaction. Soziale Systeme 6, 251–267.
Levitan, R., Hirschberg, J., 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: ACL 2011.

Li, A., Wang, H., 2004. Friendly speech analysis and perception in standard Chinese. In: INTERSPEECH-2004, pp. 897–900.
Liscombe, J., Venditti, J., Hirschberg, J., 2003. Classifying subject ratings of emotional speech using acoustic features. In: INTERSPEECH-03, pp. 725–728.
Madan, A., Caneel, R., Pentland, A., 2005. Voices of attraction. In: Proceedings of 1st International Conference on Augmented Cognition, HCI International 2005, Las Vegas.
Mairesse, F., Walker, M., 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. In: ACL-08.
Mairesse, F., Walker, M., Mehl, M., Moore, R., 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research (JAIR) 30, 457–500.
Mehl, M.R., Gosling, S.D., Pennebaker, J.W., 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. Journal of Personality and Social Psychology 90 (5), 862–877.
Namdy, L.L., Nygaard, L.C., Sauerteig, D., 2002. Gender differences in vocal accommodation: the role of perception. Journal of Language and Social Psychology 21 (4), 422–432.
Nass, C., Brave, S., 2005. Wired for Speech: How Voice Activates and Advances the Human–Computer Relationship. MIT Press, Cambridge, MA.
Natale, M., 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. Journal of Personality and Social Psychology 32 (5), 790–804.
Nenkova, A., Gravano, A., Hirschberg, J., 2008. High frequency word entrainment in spoken dialogue. In: Proceedings of ACL, 2008.
Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M., 2003. Lying words: predicting deception from linguistic style. Personality and Social Psychology Bulletin 29, 665–675.
Ng, A.Y., 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: ICML 2004.
Niederhoffer, K.G., Pennebaker, J.W., 2002. Linguistic style matching in social interaction. Journal of Language and Social Psychology 21 (4), 337–360.
Pardo, J.S., 2006. On phonetic convergence during conversational interaction. JASA 119 (4), 2382–2393.
Pennebaker, J. W., Booth, R., Francis, M., 2007. Linguistic inquiry and word count: LIWC2007 – operator's manual. Tech. rep., University of Texas.
Pennebaker, J.W., King, L.A., 1999. Linguistic styles: language use as an individual difference. Journal of Personality and Social Psychology 77 (6), 1296.
Pennebaker, J.W., Lay, T.C., 2002. Language use and personality during crises: analyses of Mayor Rudolph Giuliani's press conferences. Journal of Research in Personality 36, 271–282.
Pentland, A., 2005. Socially aware computation and communication. Computer, 63–70.
Pon-Barry, H., Shieber, S.M., 2011. Recognizing uncertainty in speech. EURASIP Journal on Advances in Signal Processing: Special Issue on Emotion and Mental State Recognition from Speech, 2011.
Potts, C., 2011. On the negativity of negation. In: Li, N., Lutz, D. (Eds.), Proceedings of Semantics and Linguistic Theory 20. CLC Publications, Ithaca, NY, pp. 636–659.
Puts, D.A., Barndt, J.L., Welling, L.L.M., Dawood, K., Burriss, R.P., 2011. Intrasexual competition among women: vocal femininity affects perceptions of attractiveness and flirtatiousness. Personality and Individual Differences 50 (1), 111–115.
Rosenberg, A., Hirschberg, J., 2005. Acoustic/prosodic and lexical correlates of charismatic speech. In: EUROSPEECH-05, Lisbon, Portugal, pp. 513–516.
Rude, S.S., Gortner, E.M., Pennebaker, J.W., 2004. Language use of depressed and depression-vulnerable college students. Cognition and Emotion 18, 1121–1133.
Schegloff, E.A., Jefferson, G., Sacks, H., 1977. The preference for self-correction in the organization of repair in conversation. Language 53, 361–382.
Scherer, K.R., 1978. Personality inference from voice quality: the loud voice of extroversion. European Journal of Social Psychology 8, 467–487.
Scherer, K.R., 2000. Psychological models of emotion. In: Borod, J. (Ed.), The Neuropsychology of Emotion. Oxford University Press, pp. 137–162.
Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Communication 40 (1–2), 227–256.
Schiffrin, D., 1987. Information and participation: y'know and i mean. In: Discourse Markers. Cambridge University Press.
Street Jr., R.L., 1983. Noncontent speech convergence in adult-child interactions. In: Bostrom, R. (Ed.), Communication Yearbook 7. Sage, Beverly Hills, CA, pp. 369–395.
Taylor, H., 1970. Balance in Small Groups. Von Nostrand Reinhold Company, New York, NY (Chapter 2).