Acoustic and Prosodic Correlates of Social Behavior

Agustín Gravano[1], Rivka Levitan[2], Laura Willson[2], Štefan Beňuš[3], Julia Hirschberg[2], Ani Nenkova[4]

[1]Universidad de Buenos Aires, Argentina; [2]Columbia University, USA; [3]Constantine the Philosopher University & Institute of Informatics, Slovak Academy of Sciences, Slovakia; [4]University of Pennsylvania, USA

gravano@dc.uba.ar, rlevitan@cs.columbia.edu, law2142@barnard.edu, sbenus@ukf.sk, julia@cs.columbia.edu, nenkova@seas.upenn.edu

Abstract

We describe acoustic/prosodic and lexical correlates social variables annotated on a large corpus of task-oriented spontaneous speech. We employ Amazon Mechanical Turk to label the corpus with a large number of social behaviors, examining results of three of these here. We find significant differences between male and female speakers for perceptions of attempts to be liked, likeability, speech planning, that also differ depending upon the gender of their conversational partners.

Index Terms: social variables, crowdsourced annotation, dialogue perception.

1. Introduction

There has been much work in the speech community on the acoustic-prosodic and lexical indicators of classic emotions. Similar approaches have also been used to identify other related types of speaker state, including uncertainty, confidence, and deception, as well as less clearly 'emotional' states as charisma, sarcasm, personality, and medical conditions such as depression. More recently researchers have begun to explore the acoustic and prosodic cues that may be correlated with the production and perception of ***social behavior*** in conversation, including flirtation, agreeableness and awkwardness. In this paper we examine the perception of three types of social behavior in conversation: likeability, the atttempt to be liked, and conversational planning. These behaviors represent part of a larger ongoing study of social behavior in task-oriented conversation in the Columbia Games Corpus. In Section 2 we describe previous research in this area. In Section 3 we describe the corpus. Section 4 discusses the annotation of social behavior we elicited using Amazon Mechanical Turk. Our current experiments are described in Section 5 and we discuss our conclusions and future research in Section 6.

2. Previous Research

Previous research has uncovered acoustic/prosodic cues to speaker states such as emotions, charisma, deception, and depression, and personality traits in F0, energy and spectral tilt and in use of pronouns, emotion words, and disfluencies [1,2,3,4,5,6,7]. Additionally, [8] have found acoustic and prosodic correlates of perception of social behavior in a speed-dating corpus in which speakers and listeners rated speakers in terms of flirtatiousness, awkwardness, humor, and assertive-ness. Spontaneous speech features, such as disfluencies, have also been found to influence perception of speakers' problems speech planning [9]. Degree of planning in speech is linked to differences between read and spontaneous speech [10,11,12].

3. The Columbia Games Corpus

The Columbia Games Corpus [13,14] includes 12 spontaneous task-oriented dyadic conversations between native speakers of Standard American English, representing 9h 8m of recorded dialogue. Subjects played a set of computer games using only verbal communication to achieve a common goal – a score which determined overall compensation. Players could see only their own screen and had to describe information on it to their partner. Each speaker was recorded on a separate channel. Subjects participated in 2 sessions with 2 different partners, resulting in 3 female-female, 3 male-male, and 6 female-male sessions. The corpus was transcribed and words were hand-aligned to the speech. Prosodic information was labeled using the ToBI system [15]; other annotations include question types, and *affirmative cue words* (all synonyms of '*yes*'). Acoustic features were extracted from the corpus automatically, using Praat [16]. Speaker turns were labeled for exchange types by trained annotators, including smooth switches, overlaps, interruptions, butting-ins and backchannels (overlapping and non); this annotation scheme is described in [13] and at www.cs.columbia.edu/speech/games-corpus.

A *task* in the Games Corpus corresponds to a simple game played by the subjects, requiring verbal communication to achieve a joint goal of identifying and moving images on the screen. Task boundaries were extracted from the logs collected automatically during the sessions, and subsequently checked by hand. Within these tasks, we define the following segments: we define an *inter-pausal unit* (IPU) as a maximal sequence of words surrounded by silence longer than 50 ms. A *turn* then is defined as a maximal sequence of IPUs from one speaker, such that between any two adjacent IPUs there is no speech from the interlocutor. Boundaries of IPUs and turns were computed automatically from the time-aligned transcriptions.

The Objects games comprise just under half of the corpus (4h 18m). In these, one player (Describer) described the position of an object on his/her screen to the other (Follower), whose task was to position the same object on his/her own screen; neither could see the other's screen; the closer the Follower's object to the Describer's, the higher the score. Each session included the same set of 14 placement tasks, with subjects alternating within the session in the roles of Describer and Follower.

4. Mechanical Turk Annotation

To annotate our corpus with aspects of speakers' social behavior, we used Amazon's Mechanical Turk (AMT). In AMT, a requester is allowed to specify a set of HITs (Human Intelligence Tasks), which are posted to AMT workers who can accept or reject each HIT and are paid a small sum (in our case, US$ 0.30 per hit). Annotators are first required to complete a survey to establish that they are native English

speakers and have no hearing impairment. Only annotators with a 95% success rate on previous AMT HITs and who are located in the United States are accepted for HITs.

Each of our HITs requires the annotator to listen to one task lasting up to 3 minutes from our Objects games, while watching a video simulating the conversation to differentiate speech from Describer and Follower. The screen shows a blue circle when the speaker in one ear is talking and a green circle when the speaker in their other ear is talking. The audio clip is available for replay to annotators throughout the task. After the annotator listens to the complete clip (they must answer a question about the dialogue to ensure that they have completed it), they are asked to answer a series of questions about the dialogue and about the individual speakers: *Is the conversation awkward? Does it flow naturally? Are the participants having trouble understanding each other? Which person do you like more? Who would you rather have as a partner? Does Person A believe s/he is better than his/her partner? Make it difficult for his/her partner to speak? Seem engaged in the game? Seem to dislike his/her partner? Is s/he bored with the game? Directing the conversation? Frustrated with his/her partner? Encouraging his/her partner? Making him/herself clear? Planning what s/he is going to say? Polite? Trying to be liked? Trying to dominate the conversation?* (these questions are repeated for Person B)

A set of check questions to which there is only one correct answer (e.g. "Which speaker is the Describer?") are scattered among the regular questions to ensure that the annotators were attending to the task. Our annotations to date were completed by 94 annotators, each of whom completed between 1 and 62 tasks. Over half of the annotators completed fewer than five hits, and only four completed more than twenty. Each task had between 5 and 7 unique annotators. Of the 168 tasks (4h 19m) in our Objects games, 99 (1h 50m) have been annotated; their durations are shown in Table 1.

Seconds:	0-30	30-60	60-90	90-120	120-150	150-180
#Tasks:	19	29	22	18	8	3

Table 1. Distribution of task durations.

As a sanity check on our annotations, we ran Pearson's correlation tests on each pair of social variables to see whether the associations we were finding seemed intuitively plausible. For example, we hypothesized we would find **positive** correlations between *contributes-to-successful-completion*, *engaged-in-game*, *making-self-clear*, *planning-what-to-say* and *trying-to-be-liked*. Table 2 shows the Pearson coefficient of each variable pair, verifying these expected correlations. Note however that *liked-more-by-rater* is reasonably well correlated with all of the social variables except for *trying-to-be-liked.* We had no initial intuitions about the relationship between

	Eng	*Clear*	*Plan*	*Trying*	*Liked*
Contr	.71	.78	.47	.38	.40
Eng		.65	.35	.37	.30
Clear			.44	.34	.38
Plan				.29	.40
Trying					.15

Table 2. Pearson coefficient of variable pairs.

these two variables. However, post hoc, it seems plausible that speakers perceived as trying to be liked might be considered "annoying" and therefore be less liked by raters. For the current paper we consider only annotations for three of our questions, *trying-to-be-liked*, *planning-what-to-say*, and *liked-more-by-rater*, since we had sufficient data for these questions to require majority agreement of 4 or 5 out of 5 or more judgments for annotated tasks.

5. Experiments

In their study of flirting, friendliness and awkwardness in speed dating between mixed gender partners, [8] found that men labeled as friendly or perceived to be flirting used more second person pronouns, laughed more, and avoided backchannels and signs of appreciation; prosodically, both spoke more quietly than other men. Men perceived as friendly also produced shorter utterances with lower minimum pitch and tended to overlap their partner's speech and to produce more collaborative completions. Men labeled as flirting asked more questions, avoided overlapped speech, and used more sexual and negative emotion words; their speech was higher in pitch and faster. Women labeled as friendly produced more collaborative completions, repair questions, laughter, appreciations, and disfluencies, using more words over all. Their F0 was higher and their intensity varied. Women labeled as flirting also spoke more and were somewhat more disfluent; they spoke faster and louder, with higher and more variable pitch. They asked few questions, used few indications of assent, and used more first person singular pronouns. Men labeled as awkward were more disfluent, with more restarts and filled pauses. They used few appreciations, collaborative completions, second person pronouns, and overlaps, and took fewer turns over all. We compare our results to these, examining similar features.

We performed a series of statistical analyses correlating acoustic/prosodic and lexical features with AMT annotations of our three social variables to identify features which were positively or negatively associated with majority judgments. Our unit of analysis for the experiments is the *intonational phrase* (IP), according to the (manual) ToBI labels in the corpus. An IP is a segment of speech that contains a single prosodic contour and normally ends in a phrase accent and boundary tone (e.g., L-H%, !H-H%). IPs and speaker turns inherit their AMT judgments from the parent Games task.

We extracted the following acoustic/prosodic features from each target IP: mean pitch, intensity, voiced-frames ratio, jitter, shimmer and noise-to-harmonics ratio (NHR); pitch range (estimated with the pitch value at ToBI HiF0 labels); pitch and intensity slopes computed over the final 100, 200 and 300 ms; and speaking rate (syllables and phonemes per second). All features were speaker normalized using *z*-scores: $z = (x - \mu)/\sigma$, where x is a raw measurement, and μ and σ are the mean and standard deviation for a speaker.

We also computed lexical features for each IP. *Fragments*, *repetitions* and a more general category called *self-repairs* were marked in the course of the annotation of the Games Corpus, as were *laughs*, *affirmative cue words* (e.g. "okay", "mm-hm"), and *filled pauses*. *Contractions* and *interjections* were identified by Ratnaparkhi's [17] maxent POS tagger. For each of these, the value for a target IP was the feature's count in the IP. Additionally, each IP received a score for mean *pleasantness, activation* and *imagery* from Whissell's Dictionary of Affect (DOA) [18]. Activation, a measure of the "strength" of the target speech, is primarily associated with the word *no* in our data. Imagery is defined as "how easy it is to form a mental picture of the word." In our corpus, it appears to reflect the prevalence of content words relating to the game.

For each social variable we first define two groups of IPs. The 'yes' group contains all IPs belonging to a task rated 'yes'

by at least k raters; the 'no' group is defined analogously. Likewise, for the *liked-more-by-rater* social variable, the 'yes' group contains all IPs belonging to a task in which at least k raters chose the IP utterer as the person they liked more. Whenever possible, we choose k=5, although in some cases that leaves us with too few data points to analyze; in such cases we use k=4. We then compare the mean value of each numerical feature for the 'yes' and 'no' groups, using ANOVA and Kruskal-Wallis tests. We consider a result to be statistically significant when its p-value is lower than 0.05, and to approach significance when p<0.1. For some of our experiments, we further divide our IP groups with respect to the gender of the speakers: male talking to male (m-m), male talking to female (m-f), and so on. For our turn-taking experiments, we compute the distribution of turn exchange types for each group and run Fisher's Exact tests to assess significant deviations from random distributions.

First we explore the *trying-to-be-liked* variable, searching for differences in our numerical features between the 'yes' and 'no' groups. Since we find somewhat contradictory results in the general case, we further divide our IP groups with respect to the gender of the speakers, following [8]'s approach – male talking to male (m-m), male talking to female (m-f), and so on. Table 3 summarizes the differences for each of the 4 gender pair combinations (for m-f we use k=5; for the rest, k=4).

	...talking to Male	...talking to Female
Male	Lower intensity Slower speaking rate Expanded pitch range Higher final intonation†	Lower intensity† Faster speaking rate Expanded pitch range Higher NHR, jitter, shimmer
Female	Higher pitch† Lower NHR†	Higher intensity Lower pitch† Lower NHR, jitter, shimmer Faster speaking rate

Table 3. *Trying-to-be-liked*. Significant results (or approaching significance, marked †) for each gender pair.

Males tend to lower their intensity and expand their pitch range when perceived as trying to be liked, regardless of their interlocutor's gender; interestingly, they slow down their speaking rate when talking to males but speed it up when talking to females. Females talking to males raise their pitch level when perceived as trying to be liked, but when speaking with females they lower it and increase their intensity and speaking rate. We also observe in most cases a significant effect for three acoustic features that are typically correlated with voice quality: NHR, jitter and shimmer.

With respect to lexical cues, speakers perceived as trying to be liked used fewer contractions and fewer self-repairs (p<0.1) than other speakers. The exception is the case of males speaking to females, who used more contractions when perceived as trying to be liked; their speech under the same condition was more pleasant and more activated under a Dictionary of Affect analysis. Females perceived as trying to be liked when speaking to females used more affirmative cue words, and their speech was less activated. No significant differences were found in turn-taking behavior.

We note that our findings overlap only slightly with [8] reported findings for friendly and flirting behavior, although we examined many of the same acoustic/prosodic and lexical features. We did find that males labeled as trying to be liked exhibited lower intensity and higher pitch when talking to females, similar to males perceived to be flirting by [8]. Females talking to males also exhibited higher pitch, as did women perceived to be friendly and women perceived to be flirting in [8]. However, we found no significant effects for other acoustic/prosodic and lexical features, although these may emerge with more data.

Speakers who were more likely to be liked by raters exhibited similar acoustic/prosodic characteristics in both genders: higher intensity, lower pitch, lower shimmer, and more reduced pitch range. In terms of lexical cues, their speech included more (DOA) activation and imagery; when addressed to females it was more pleasant, independent of the speaker's gender. These speakers used more filled pauses and contractions, with fewer interjections, affirmative cue words, and fragments (p<0.01). Female speakers who were liked laughed less when speaking to females.

We also examined whether speaker role and turn-taking behavior had an effect on annotator likeability ratings. Table 4 shows the number of turn-taking categories from the target speaker, grouped by speaker role (Describer or Follower) and by the speaker selected by the raters as the one they liked more. We examined backchannels, disruptive turn-taking categories (e.g. interruptions, butting-ins), and non-disruptive categories (e.g. smooth switches, non-disruptive overlaps).

Target speaker's role:	**Describer**		**Follower**	
Speaker liked more:	**Target**	**Other**	**Target**	**Other**
Backchannels	1	0	28	110
Disruptive switches	24	9	5	37
Non-disruptive switches	137	50	65	168
Fisher's Exact Tests:	p = 1 N.S.		**p = 0.039**	

Table 4. Distribution of turn-taking categories from target speakers, tabulated by speakers liked more by raters (target or interlocutor) and by speaker role (describer or follower).

Table 4 shows that Followers who interrupt more often and produce more backchannels tend to be liked less by raters. While it comes as no surprise that people who interrupt tend to be disliked, it is remarkable that the production of backchannels is disliked by third parties, although this is consistent with [8]'s finding that people rated as friendly tend to avoid backchannels. A possible explanation is that backchannels are not really needed by third parties, and thus somehow disrupt the discourse of the speaker holding the floor. Alternatively, annotators may tend to prefer the Describer whenever the Follower limits his/her contributions to just backchanneling. We also see that Followers who overlap more often in smooth turn exchanges are liked better (Fisher, p=0.002), again consistent with [8]'s findings. In other words, third parties seem to prefer slight overlaps over silent gaps between speaker turns, since the former normally lead to swifter conversations.

From our analyses of annotations of the social variable *planning-what-to-say*, we find that speakers who are perceived as planning have significantly longer IPs, slower speech rate (both in terms of syllables and phones per second), and a tendency toward lower maximum intensity (p<0.1). Features related to pitch and voice quality did not affect the perception of planning in speech significantly. When we examine gender influences on these ratings, we find a number of differences between speaker pairs (Table 5). Interestingly, in the three features correlated with perception of planning in the pooled data, the gender of the interlocutor reverses the direction of the effect in female speech: females who plan what to say when they talk to other females have shorter IPs, slower rate, and lower max intensity than when they are not so rated. But when females speak to males, those rated as planning what to say have longer IPs, faster rate, and higher max intensity than when they are not. It may be that when females talk to males,

their planned speech has features used for preventing interruptions from males (longer IP, faster rate, higher maximum intensity), while when they talk to other females, the effort to avoid interruptions is less clear and more canonical planned speech occurs (slower rate, lower intensity). Additionally, elements of flirtatious speech observed in [1] include faster speech rate and higher intensity, which are two features of planned females speech present when interacting with males but not in female-female planned speech (slower rate, lower intensity).

	...talking to Male	...talking to Female
Male	Longer IPs in ms.	Lower pitch Lower jitter Expanded pitch range Lower final pitch slope †
Female	Longer IPs Faster speech rate Higher intensity Lower jitter, shimmer	Shorter IPs Slower speech rate Lower intensity

Table 5. *Planning-what-to-say.* Significant results (or approaching significance, marked †) for each gender pair, k=4.

We also examined lexical cues to ratings of *planning-what-to-say.* Here we find an overall negative correlation of planning ratings with speakers' use of affirmative cue words, fragments (except when males talk to females), (DOA) activation, and (DOA) interjections; there is a positive correlation only with DOA-scored imagery. Finally, we investigated the effect of filled pauses on the perception of planning. To test this effect, we divide the speech in each rated dialogue as containing a) no filled pause vs. b) at least one filled pause, and use a Pearson chi-square test to assess if the presence of filled pauses affects the rating of social variables. We find that speakers who produce at least one filled pause are rated significantly more highly with respect to planning. This contrasts with [8]'s finding that men labeled as awkward tend to be more disfluent, although our definition of the term is different here. Interestingly, ***partners*** of speakers who are perceived as planning what to say use fewer filled pauses than partners of those who do not plan what to say. Hence, our data support an intuitive prediction that on-line planning, signalled by longer IPs and slower speech rate, provides more time for the interlocutor to plan and leads to fewer disfluencies in her/his speech.

6. Conclusions and Future Work

Our study of annotated social variables in a corpus of spontaneous speech provides new information on characteristics of speech rated as indicating that a speaker is trying to be liked, is actually likeable, and is planning what to say. We expand upon related studies of likeability and awkwardness, examining additional acoustic/prosodic features such as pitch slope and voice quality, and additional lexical features such as turn-taking behaviors and DOA emotion words. We also find differences in behavior between single and cross-gender pairings in realization of social variables, examining the gender or both speaker and hearer. In terms of our task-oriented corpus, we also find differences between speakers who play the more active and more passive roles, in terms of likeability. In future work we will expand our analysis of our current set of social variables and examine others from our annotations as well, to further our study of how social information is realized and perceived in conversation.

7. Acknowledgements

This material is based upon work supported in part by the National Science Foundation under Grants Nos. IIS-0803148 and 0803159; UBACYT No. 2002009030008701; VEGA No. 2/0202/11; and the EU project CRISIS ITMS 26240220060.

8. References

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

[2] C. M. Lee and S. S. Narayanan. 2002. Combining acoustic and language information for emotion recognition. ICSLP-02.

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

[4] A. Rosenberg and J. Hirschberg. 2005. Acoustic/prosodic and lexical correlates of charismatic speech. EUROSPEECH-05.

[5] F. Mairesse and M. Walker. 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. ACL-08.

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

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

[8] D. Jurafsky, R. Ranganath, and D. McFarland, 2009. Extracting social meaning: Identifying interactional style in spoken conversation. NAACL HLT-09.

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

[10] E. Blaauw, 1994, The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech, Speech Communication 14(4): 359-375.

[11] R. Martin, J. Crowther, M. Knight, F. Tamborello, and C. L. Yang, Planning in sentence production: Evidence for the phrase as a default planning scope, Cognition 116(2): 177-192

[12] J. Krivokapic, 2010. Speech Planning and Prosodic Phrase Length. Speech Prosody 2010.

[13] A. Gravano and J. Hirschberg, 2011. Turn-taking cues in task-oriented dialogue, *Computer Speech and Language*, 25(3): 601-634.

[14] A. Gravano, J. Hirschberg, and Š. Beňuš. Affirmative cue-words in task-oriented dialogue, *Computational Linguistics*. In press.

[15] M. E. Beckman and J. Hirschberg, 1994. The ToBI Annotation Conventions. Ohio State University.

[16] P. Boersma and D. Weenink, 2001. Praat: Doing phonetics by computer, *http://www.praat.org*.

[17] A. Ratnaparkhi, E. Brill, and K. Church, 1996. A maximum entropy model for part-of-speech tagging, in Proc. of EMNLP, pp. 133-142.

[18] C. Whissel, 1989. The Dictionary of affect in language. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, Research and Experience* (pp. 113-131). Academic Press.