Backward mimicry and forward influence in prosodic contour choice in Standard American English

Agustín Gravano [1,2], Štefan Beňuš [3], Rivka Levitan [4], Julia Hirschberg [5]

[1] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[2] National Scientific and Technical Research Council (CONICET), Buenos Aires, Argentina
[3] Constantine the Philosopher University in Nitra & II-Slovak Academy of Sciences, Slovakia
[4] Dept. of Computer and Information Science, Brooklyn College CUNY, New York, NY, USA
[5] Dept. of Computer Science, Columbia University, New York, NY, USA

gravano@dc.uba.ar, levitan@sci.brooklyn.cuny.edu, sbenus@ukf.sk, julia@cs.columbia.edu

Abstract

Entrainment is the tendency of speakers engaged in conversation to align different aspects of their communicative behavior. In this study we explore in more detail a measure of prosodic entrainment defined in previous work, which uses a discrete parametrization of intonational contours defined by the ToBI conventions for prosodic description. We divide this measure into two asymmetric variants: *backward mimicry* (in which a speaker uses a contour used previously by the interlocutor) and *forward influence* (in which a speaker's contour appears later in the speech of the interlocutor). This distinction sheds new light on significant correlations with a number of social variables related to the level of engagement of speakers in a corpus of task-oriented dialogues in Standard American English.

Index Terms: Dialogue, entrainment, prosody, ToBI, social variables.

1. Introduction

When engaged in conversation, speakers tend to coordinate different aspects of their communicative behavior, often adapting their speech to match, or synchronize with, their interlocutors' behavior. This phenomenon is known as ENTRAINMENT, ADAPTATION, MIMICRY or ALIGNMENT, and has been shown to occur in pronunciation [1]; choice of referring expressions [2]; linguistic style [3, 4], syntactic structure [5, 6]; speaking rate [7]; acoustic-prosodic features such as fundamental frequency, intensity and voice quality [7]; turn-taking cues [8, 9]; and choice of intonational contour [10].

Entrainment arising through spoken interactions is closely linked to creating, negotiating and maintaining relationships betweeen interlocutors in several social dimensions and reflects speakers' need for social integration or identification with another [11]. For example, entraining interlocutors tend to be more successful in completing tasks [12, 13] and to be perceived as more competent, socially attractive, or likeable [14, 8, 15]. It has been already shown that entrainment improves the perceived naturalness and effectiveness of human-machine interactions [16, 17, 18].

In [10] we presented three measures of prosodic entrainment that take advantage of the descriptions of prosodic contours annotated using the ToBI labeling conventions [19]. We focused on these higher level representations of prosodic variation, such as sequences of PITCH ACCENTS, PHRASE ACCENTS, and BOUNDARY TONES. We found significant correlations between each of these measures of prosodic entrainment and manual annotations of a number of social variables related to the level of engagement of speakers. In this study we further explore one measure of prosodic entrainment described in our previous study, based on the similarity of previous and subsequent contours produced by interlocutors. We divide this metric into two asymmetric variants: backward and forward entrainment – or, *backward mimicry* and *forward influence*, and analyze the utility of each measure at a wider temporal range. This new approach provides novel and finer-grained information on the correlation of speakers' pitch contours with social variables in our corpus of task-oriented dialogues in Standard American English (SAE).

2. Corpus

Our experiments were conducted on a subset of the **Columbia Games Corpus**, a collection of 12 spontaneous task-oriented dyadic conversations between 13 native speakers of SAE, comprising 9h 8m of recorded dialogue. In this corpus, subjects played a set of computer games using only verbal communication to achieve a common goal — a score which determined their overall compensation. Each speaker was recorded on a separate channel. The corpus was transcribed and words were manually aligned to the speech. In this study we examine a portion of the Games Corpus that has complete ToBI annotations, the Objects Games, which comprises just under half of the corpus (4h 18m). In these exercises, one player (the Describer) described the position of an object on his/her screen to the other (the 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; subjects were later paid a bonus based on the number of points they earned. Each session included the same set of 14 placement tasks, with subjects alternating in the Describer and Follower roles.

Prosodic information was annotated using the ToBI conventions for SAE [19]. These consist of annotations at four levels of analysis: an ORTHOGRAPHIC TIER of time-aligned words; a BREAK INDEX TIER indicating degrees of juncture between words, from 0 'no word boundary' to 4 'full intonational phrase boundary'; a TONAL TIER, where pitch accents, phrase accents and boundary tones describing targets in the F0 contour are annotated; and a MISCELLANEOUS TIER, in which phe-

nomena such as disfluencies may be optionally marked. Break indices define two levels of phrasing: level 3 corresponds to an INTERMEDIATE PHRASE in Pierrehumbert's [20] schema for representing SAE, and level 4 corresponds to her INTONATIONAL PHRASE. This tier is supplemented by the tonal tier in which type of phrase accent and boundary tone is identified. As in [20], level 4 phrases consist of one or more level 3 phrases, plus a high or low BOUNDARY TONE (**H%** or **L%**) at the right edge of the phrase. Level 3 phrases consist of one or more pitch accents, aligned with the stressed syllable of lexical items, plus a PHRASE ACCENT, which also may be high (**H-**) or low (**L-**).

Several aspects of speakers' **social behavior** in the Objects Games were annotated using Amazon's Mechanical Turk (AMT) crowdsourcing.[1] Annotators listened to an audio clip of an Objects Games task and were asked to answer a series of questions about the dialogue and about each speaker, including *Does Person A 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? Doing a good job contributing to successful completion? Encouraging his/her partner? Making him/herself clear? Planning what s/he is going to say? Polite? Trying to dominate the conversation?*, inter alia. Each task was rated by five unique annotators who answered 'yes' or 'no' to each question, yielding a score ranging from 0 to 5 for each social variable, representing the number of annotators who answered 'yes.' A fuller description of the annotation for social variables and prosodic information may be found in [10].

3. Measure of prosodic entrainment and social variables

In [10] we presented three measures of prosodic entrainment derived from an analysis of sequences of ToBI tone labels in the corpus. Here we focus on one of those metrics, based on the similarity of the neighboring contours produced by the interlocutors. In [10] we termed that measure $\mathcal{E}_2$; here we call it simply $\mathcal{E}$.

We define an INTONATIONAL CONTOUR as a sequence of ToBI tone labels corresponding to an intermediate phrase. For example, given the sequence of ToBI labels "L* L-H% L* L-L% H* !H* H- H* !H- L* H-H%", the corresponding list of contours is ["L* L-H%", "L* L-L%", "H* !H* H-", "H* !H-", "L* H-H%"]. Further, we define a similarity function sim between contours c_1 and c_2 as $sim(c_1, c_2) = (m - l)/m$, where $m = \max(length(c_1), length(c_2))$, and l is the Levenshtein distance [21] between contours c_1 and c_2 (defined such that $l \leq m$ always holds). In these calculations, c_1 and c_2 are considered to be simple strings. Following this definition, $sim(c_1, c_2)$ ranges from 0 when c_1 and c_2 are completely different, to 1 when they are identical.

Next, we extract the list of contours produced by each speaker in an entire Objects Games session, and compute the $\mathcal{E}(A, B)$ measure of prosodic entrainment between speakers A and B using Algorithm 1. Figure 1 illustrates this procedure. For each contour c_1 from speaker B, we look in its near vicinity (a window of radius k sec around c_1) for the most similar contour from speaker A, and record the similarity score between the two contours. We then average all such similarity scores to define our measure of overall prosodic entrainment between A and B. As noted in [10], $\mathcal{E}$ is asymmetric, and a high value of $\mathcal{E}(A, B)$ suggests roughly that $contours(B) \subseteq contours(A)$, since for each contour from B, speaker A also produces a similar contour shortly before or after.

[1] http://www.mturk.com

Algorithm 1 Computation of $\mathcal{E}(A, B)$.

1: $L \leftarrow$ new list
2: **for** each contour c_1 from B **do**
3: $\quad C \leftarrow$ contours from A at most k sec before/after c_1
4: $\quad$ append$\Big(\max_{c_2 \in C} sim(c_1, c_2)\Big)$ to L
5: **end for**
6: return mean(L)

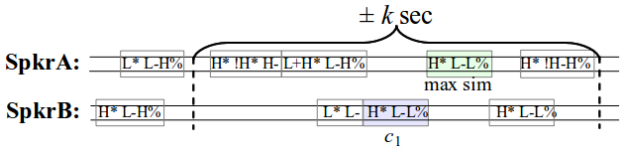

Figure 1: Illustration of the $\mathcal{E}$ measure presented in [10].

As discussed in [10], to determine how the $\mathcal{E}$ measure of prosodic entrainment correlates with our social variables, we build a vector with the value of $\mathcal{E}$ for each member of each speaker pair. Since there are 12 sessions in our corpus, this is a 24-dimensional vector, $\overrightarrow{\mathcal{E}} = \langle \mathcal{E}(A_1, B_1), \mathcal{E}(B_1, A_1), \mathcal{E}(A_2, B_2), \mathcal{E}(B_2, A_2), \ldots, \mathcal{E}(A_{12}, B_{12}), \mathcal{E}(B_{12}, A_{12}) \rangle$, where A_i, B_i are the two speakers from session i. Similarly, we build a 24-dimensional vector for each social variable v (such as *bored-with-game* or *making-self-clear*), $\overrightarrow{v} = \langle v(A_1), v(B_1), v(A_2), v(B_2), \ldots, v(A_{12}), v(B_{12}) \rangle$, where again A_i, B_i are the two speakers from session i, and $v(A_i)$ is the mean value of v for speaker A in session i (likewise for speaker B_i).

In [10] we report significant positive correlations between $\overrightarrow{\mathcal{E}}$ (using $k = 30$ seconds – i.e., one minute around each target contour) and the $\overrightarrow{v}$ vectors for six social variables: *contributes-to-successful-completion* ($r = 0.73$), *engaged-in-game* (0.71), *making-self-clear* (0.63), *gives-encouragement* (0.59), *difficult-for-partner-to-speak* (0.48) and *planning-what-to-say* (0.47), as well as negative correlations with two variables: *bored-with-game* ($r = -0.75$) and *dislikes-partner* (-0.54). Generally we found that, when $\mathcal{E}(A, B)$ is high, speaker A is more likely to be perceived as making well-planned, clear contributions to the dialogue, being engaged in the game, and giving encouragement to their partner, and is less likely to be perceived as disliking their partner or being bored with the game [10].

3.1. Backward mimicry and forward influence

In this study, we divide this measure of prosodic entrainment into backward mimicry and forward influence. To accomplish this, in step 3 of Algorithm 1, we now consider speaker A's contours that lie either k seconds *before* c_1 (backward mimicry, or $\mathcal{E}_{back}$) or *after* target contour c_1 (forward influence, or $\mathcal{E}_{fwd}$).

Figure 2 illustrates the computation of $\mathcal{E}_{back}(A, B)$. For each contour c_1 from speaker B, we look in its preceding k seconds for the most similar contour from speaker A, and save the similarity score between the two contours. We then define $\mathcal{E}_{back}(A, B)$ as the mean of all such similarity scores.

Thus, $\mathcal{E}_{back}(A, B)$ attempts to capture the degree to which speaker B mimics (a subset of) the contours produced by speaker A shortly before speaker B's production. Similarly, $\mathcal{E}_{fwd}(A, B)$ attempts to capture the extent to which speaker B

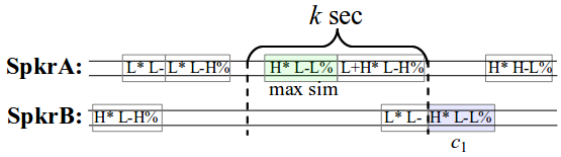

Figure 2: Illustration of the $\mathcal{E}_{back}$ measure

influences the subsequent choice of (a subset of) speaker A's prosodic contours.

As an example, we show how these measures are computed, given the following contours taken from a short conversation, for a large enough value of k:

SpkrA: H* L-L% L+H* !H* L-H%
SpkrB: L* H- H* L-L%

$$\begin{aligned}\mathcal{E}_{back}(A,B) &= (sim(\text{“L* H-”}, \text{“H* L-L%”}) + \\ &\quad sim(\text{“H* L-L%”}, \text{“H* L-L%”}))/2 \\ &= (0.429+1)/2 = 0.714\end{aligned}$$

$$\begin{aligned}\mathcal{E}_{fwd}(A,B) &= (sim(\text{“L* H-”}, \text{“L+H* !H* L-H%”}) + \\ &\quad sim(\text{“H* L-L%”}, \text{“L+H* !H* L-H%”}))/2 \\ &= (0.385+0.462)/2 = 0.423\end{aligned}$$

$$\begin{aligned}\mathcal{E}_{back}(B,A) &= \max(sim(\text{“L+H* !H* L-H%”}, \text{“L* H-”}), \\ &\quad sim(\text{“L+H* !H* L-H%”}, \text{“H* L-L%”})) \\ &= \max(0.385, 0.462) = 0.462\end{aligned}$$

$$\begin{aligned}\mathcal{E}_{fwd}(B,A) &= \max(sim(\text{“H* L-L%”}, \text{“L* H-”}), \\ &\quad sim(\text{“H* L-L%”}, \text{“H* L-L%”})) \\ &= \max(0.429, 1) = 1\end{aligned}$$

In particular, note that $\mathcal{E}_{back}(A,B)$ (i.e., how B mimics A's recent contours) is not equal to $\mathcal{E}_{fwd}(B,A)$ (how A influences B's future contour choices). At first sight, these two expressions might seem equivalent, but they are actually different. In the former, for each of B's contours we look for the most similar contour from A's previous speech – i.e., we compare each of B's contours against *a set* of contours from A. In the latter expression, we do the opposite – we compare each of A's contours against *a set* of contours from B's succeeding speech. Thus, while these two expressions are related, due to this asymmetry they in fact compute different things.

Since we have no prior knowledge of how far into the past or future we should look, we analyze the behavior of $\mathcal{E}_{back}$ and $\mathcal{E}_{fwd}$ over increasing window sizes k. Figure 3 shows, for each speaker pair in our corpus, the value of $\mathcal{E}_{back}$ when $k = 5, 10, 15, ..., 180$ seconds. We observe that, for most speakers, this measure remains fairly constant after two minutes. We therefore restrict our subsequent experiments to values of k between 5 and 120 seconds. The results for our forward influence measure, $\mathcal{E}_{fwd}$, are essentially identical, but are omitted due to space constraints.

4. Results

In this section we examine how our $\mathcal{E}_{back}$ and $\mathcal{E}_{fwd}$ measures correlate with the eight social variables listed in Section 3. Additionally, we explore different values for the width of the window in step 3 of Algorithm 1, to study the span of backward mimicry and forward influence in prosodic contour choice. The eight plots shown in Figure 4 summarize our results. In each plot, the horizontal axis corresponds to the window width k, in seconds, and the vertical axis shows Pearson's correlation coefficient r. The statistical significance of each test is indicated with a star when $p < 0.01$, or with a dot when $0.01 \leq p < 0.05$.

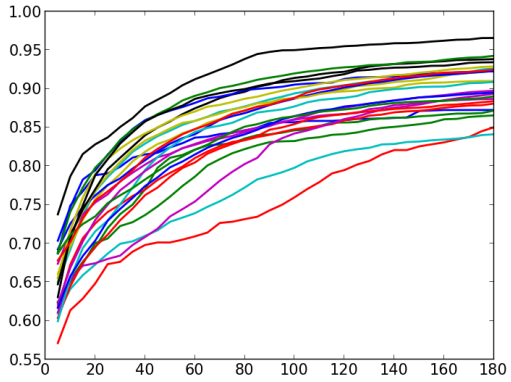

Figure 3: $\mathcal{E}_{back}$ measure for different window widths (in seconds), for the 24 speaker pairs in our corpus.

Plots (a) through (d) in Figure 4 exhibit a similar pattern. These plots correspond to four social variables directly related to the degree of speaker engagement in the task – *contributes-to-successful-completion*, *making-self-clear*, *engaged-in-game* and *planning-what-to-say*. In these plots we observe that these four social variables are positively correlated with $\mathcal{E}_{back}$ at around 15 seconds, and also with $\mathcal{E}_{fwd}$ for 25 seconds or greater values of k. This indicates that, when subjects are perceived to have a positive view of the game and to make clear, well-planned contributions, they exhibit a short-range backward mimicry (they use contours similar to those produced by their interlocutors shortly before) and they also have a longer-range forward influence on their interlocutors' contour choice.

Plot (e) in Figure 4 corresponds to *gives-encouragement*. It shows a strong positive correlation with backward mimicry, for k values of 40 seconds and longer. This seems to suggest that, the more encouragement subjects give to their partners, the more they use prosodic contours similar to the ones used by their interlocutors earlier in the conversation.

Plot (f) corresponds to *difficult-for-partner-to-speak*. This variable correlates positively in our data with *trying-to-dominate-the-conversation* ($r = 0.73$, $p < 0.001$) and with *directs-the-conversation* ($r = 0.51$, $p < 0.05$), and negatively with *is-polite* ($r = -0.46$, $p < 0.05$); correlations with all other social variables are not significant. In other words, this variable describes the behavior of one who is trying to take control of the conversation. It is interesting, then, to find that such a behavior correlates positively with a short-range entrainment (k between 10 and 35 seconds), both in backward and forward directions. These correlations are weaker for wider windows, suggesting that this social dimension is related to shorter- rather than longer-range entrainment of intonational contours.

The last two plots in Figure 4 show the correlations for *bored-with-game* and *dislikes-partner* – variables that reflect negative aspects of the speakers' social behavior. Note that the correlation coefficients on the vertical axes are negative. Plot (g) suggests that, the greater the tendency of a speaker to be bored with the game, the less likely they are to mimic their interlocutor's recent contours or to have a long-range influence on their interlocutor's prosodic contour choice. Plot (h) indicates that, when a speaker dislikes their partner, they are also unlikely to influence their partner's prosodic contour choice.

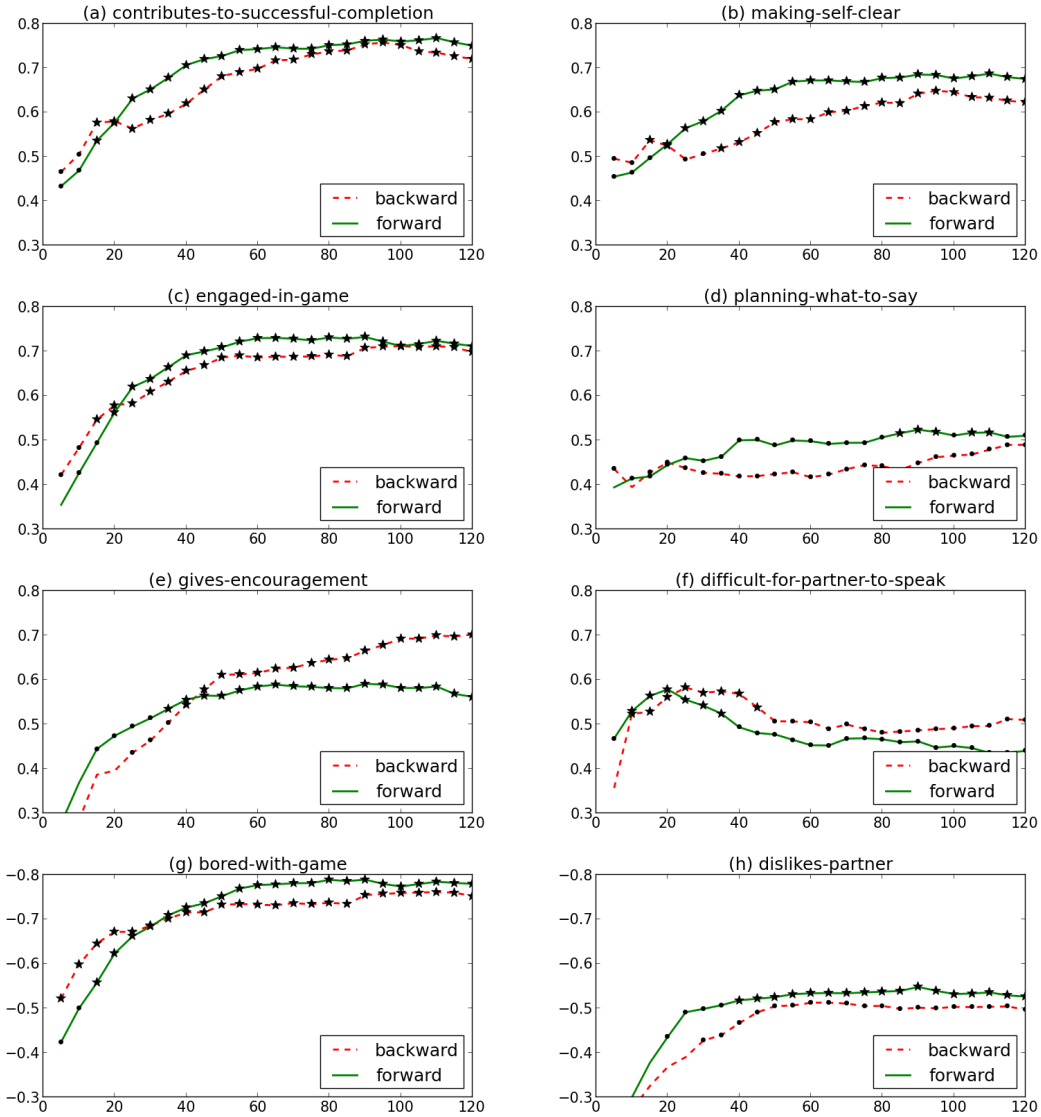

Figure 4: Correlations of the $\mathcal{E}_{back}$ and $\mathcal{E}_{fwd}$ measures of prosodic entrainment with eight social variables, for varying window widths. Horizontal axes: window width k (in seconds); vertical axes: Pearson's correlation coefficient r. Statistical significance is signalled with a star ($\star$) when $p < 0.01$, or with a dot ($\cdot$) when $0.01 \leq p < 0.05$.

5. Conclusions

In this study we refine the definition of a measure of prosodic entrainment presented in previous work, derived from an analysis of ToBI intonational contours. We divide this measure into two variants: backward mimicry ($\mathcal{E}_{back}$) and forward influence ($\mathcal{E}_{fwd}$). This distinction sheds new light on significant correlations of different forms of prosodic entrainment with a number of social variables in a corpus of task-oriented dialogues in SAE: 1) speakers perceived as engaged in the game exhibit a short-range backward mimicry as well as a longer-range forward influence on prosodic contour usage; 2) when speakers are perceived as encouraging their partners, they tend to mimic their partners' previous prosodic contours; 3) the behavior of speakers thought to be trying to control the conversation correlates positively with a short-range entrainment, both backward and forward; 4) when speakers are believed to be bored with the game and/or to dislike their partners, they are unlikely to mimic or influence their partners' choice of prosodic contours.

In future, we will analyze the feasibility of replacing the manual annotations of prosody with automatic estimates such as AuToBI [22]. We will also investigate whether $\mathcal{E}_{back}$ and $\mathcal{E}_{fwd}$ may be employed with other discrete labels besides ToBI contours, such as part-of-speech labels, to measure other types of entrainment.

6. Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055, and by UBACYT 20020120200025BA.

7. References

[1] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, p. 2382, 2006.

[2] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation," *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 22, no. 6, pp. 1482–1493, 1996.

[3] K. Niederhoffer and J. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language & Social Psychology*, vol. 21, no. 4, pp. 337–360, 2002.

[4] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais, "Mark my words! Linguistic style accommodation in social media," in *Proc. of WWW 2011*, 2011.

[5] D. Reitter, F. Keller, and J. Moore, "Computational modelling of structural priming in dialogue," in *Proceedings of HLT/NAACL*, 2006.

[6] ——, "A computational cognitive model of syntactic priming," *Cognitive Science*, vol. 35, no. 4, pp. 587–637, 2011.

[7] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. of Interspeech*, 2011.

[8] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[9] R. Levitan, S. Benus, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*, Stanford, CA, 2015.

[10] A. Gravano, S. Benus, R. Levitan, and J. Hirschberg, "Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement," in *IEEE Spoken Language Technology Workshop (SLT)*, South Lake Tahoe, NV, 2014.

[11] H. Giles, N. Coupland, and J. Coupland, "Accommodation theory: Communication, context, and consequence," *Contexts of accommodation: Developments in applied sociolinguistics*, vol. 1, 1991.

[12] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of ACL/HLT*, 2008.

[13] D. Reitter and J. Moore, "Alignment and task success in spoken dialogue," *Journal of Memory and Language*, vol. 76, pp. 29–46, October 2014.

[14] R. L. Street, "Speech convergence and speech evaluation in fact-finding interviews," *Human Communication Research*, vol. 11, no. 2, pp. 139–169, 1984.

[15] S. Benus, A. Gravano, R. Levitan, S. Levitan, L. Willson, and J. Hirschberg, "Entrainment, dominance and alliance in Supreme Court hearings," *Knowledge-Based Systems*, vol. in press, 2014.

[16] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction," in *Proc. of International Congress of Phonetic Sciences*, 2003, pp. 2463–2466.

[17] S. Oviatt, C. Darves, and R. Coulston, "Toward adaptive conversational interfaces: Modeling speech convergence with animated personas," *ACM Transactions on Computer-Human Interaction*, vol. 11, no. 3, pp. 300–328, 2004.

[18] J. Thomason, H. Nguyen, and D. Litman, "Prosodic entrainment and tutoring dialogue success," *Artificial Intelligence in Education (LNCS 7926)*, pp. 750–753, 2013.

[19] M. E. Beckman and J. Hirschberg, "The ToBI annotation conventions," *Ohio State University*, 1994.

[20] J. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, MIT, 1980.

[21] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet physics doklady*, vol. 10, p. 707, 1966.

[22] A. Rosenberg, "AutoBI – A tool for automatic ToBI annotation," in *Proc. of Interspeech*, 2010.