THREE TOBI-BASED MEASURES OF PROSODIC ENTRAINMENT AND THEIR CORRELATIONS WITH SPEAKER ENGAGEMENT

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

[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 & Institute of Informatics, Slovak Academy of Sciences, Slovakia

[4] Department of Computer Science, Columbia University, New York, NY 10027, USA

gravano@dc.uba.ar, rlevitan@cs.columbia.edu, sbenus@ukf.sk, julia@cs.columbia.edu

ABSTRACT

Entrainment is the propensity of conversational partners to align different aspects of their communicative behavior. In this study we present three novel measures of prosodic entrainment based on intonational contours as defined by the ToBI conventions for prosodic description. Each of these measures estimates the similarity of contours used by speakers in different ways: by means of the perplexity of n-gram models, the Levenshtein distance, and the Kullback-Leibler divergence measure. We report significant correlations between each of these measures and manual annotations of 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

Conversational partners tend to coordinate several aspects of their communicative behavior, often adapting their speech to match, or synchronize with, their interlocutors' characteristics. This phenomenon, known as ENTRAINMENT, and sometimes described as ADAPTATION or ALIGNMENT, has been shown to occur in the speakers' choice of referring expressions [1]; linguistic style [2, 3], syntactic structure [4, 5]; speaking rate [6]; acoustic-prosodic features such as fundamental frequency, intensity and voice quality [6]; turn-taking cues [7]; and pronunciation [8].

Motivated by Communication Accommodation Theory [9], which holds that speakers converge their speech behavior to that of their interlocutor in order to minimize social distance, numerous studies have examined links between entrainment and positive conversational attributes, including task success [10, 11], smoothness of interaction [12, 7, 13], speaker attitude [14, 15], cooperation [16], social attractiveness [17], and power relations [18, 19], inter alia.

A number of studies have investigated acoustic-prosodic entrainment and some have examined the relationship between these forms of entrainment and enhanced social values. For example, [13] found that the similarity between partners in pitch, intensity, voice quality and speaking rate was correlated with descriptors of social behavior such as encouragement and trying to be liked, as well as automatically derived measures of dialogue flow. [16] found that partners who converged on speaking rate were more likely to later cooperate in a prisoner's dilemma, although they did not evaluate each other more positively. [14] proposed measures of entrainment on pitch and intensity **slopes** and showed that entrainment on pitch slope was predictive of positive affect in conversations between couples in marital therapy.

In this study we present three novel measures of prosodic entrainment in intonational contours, which we define in the ToBI framework. While other studies have examined low level features such as pitch mean, maximum, minimum or slope, we focus on higher level representations of prosodic variation in this research. To what extent do speakers entrain in terms of their PITCH ACCENTS, PHRASE ACCENTS, and BOUNDARY TONES — in those prosodic elements that, in Standard American English (SAE), contribute meaning variation to utterances? Each of the measures we employ estimates the similarity of pitch contours in different ways: by means of the perplexity of n-gram models, the Levenshtein distance, and the Kullback-Leibler divergence measure. We then examine correlations between each of these measures and manual annotations of a number of social variables related to the level of engagement of speakers, in a corpus of task-oriented dialogues in 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. Players could see only their own screen and were asked to describe card images they saw on it to their partner for various purposes. Each speaker was recorded on a separate channel. Subjects participated in two sessions with two different partners, resulting in 3 female-female, 3 male-male, and 6 female-male sessions. 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. Each session included the same set of 14 placement tasks, with subjects alternating in the Describer and Follower roles.

2.1. Annotation of prosody

Prosodic information was annotated using the ToBI conventions for SAE [20]. These consists of annotations at four time-linked 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 define intonational phrases; and a MISCELLANEOUS TIER, in which phenomena such as disfluencies may be optionally marked. Break indices define two levels of phrasing: level 3 corresponds to an INTERMEDIATE PHRASE in Pierrehumbert's [21] schema for representing SAE and level 4 to her INTONATIONAL PHRASE. This tier is supplemented by a tonal tier in which type of phrase accent and boundary tone is identified. As in [21] 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-**).

Pitch accents make words intonationally prominent and are realized by F0 peaks or valleys, increased loudness, and longer duration of accented syllables. A given word may be accented or DEACCENTED and, if accented, may bear different types of pitch accents. Five types of pitch accent are distinguished in the ToBI system for SAE: two simple accents **H*** and **L***, and three complex ones, **L*+H**, **L+H***, and **H+!H***; the asterisk indicates which tone of the accent is aligned with the stressable syllable of the accented lexical item. Some pitch accents may be DOWNSTEPPED, such that the pitch range of the accent is compressed in comparison to a previous accent. Downsteps are indicated by the '!' diacritic.

ToBI annotations were performed by three experts, who first practiced together on a small portion of corpus until they reached an agreement comparable to that reported in [22]. Subsequently, each annotator worked separately on different files, and had regular meetings to discuss and agree on difficult cases. Each annotator labeled roughly one third of the corpus. Only in 4 out of 12 sessions were both sides of the conversation labeled by the same annotator.

2.2. Annotation of social variables

To annotate the Objects Games with aspects of speakers' social behavior, we used 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: *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? Doing a good job contributing to successful completion? 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?* 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 can be found in [23].

3. MEASURES OF PROSODIC ENTRAINMENT AND SOCIAL VARIABLES

Using the annotations described above, we first model the degree of higher level prosodic entrainment in the Objects Games, using three different metrics. For each measure, we then examine the relationship of entrainment and the social variables annotated in the corpus.

3.1. N-gram perplexity

To measure the occurrence of prosodic entrainment by n-gram perplexity, we first extract the entire sequence of ToBI labels from the tonal tier produced by each speaker in a full Objects Games session. We include pitch accents (e.g., H*, L+H*), single phrase accents associated with level 3 break indices (e.g., L-, !H-), phrase accent and boundary tone combinations for level 4 breaks (e.g., L-H%, !H-L%), and pauses longer than 50ms (denoted by the symbol "#"). For example,

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

such a sequence for one of the speakers in our corpus begins with "L* L-H% L* L-L% # H* !H* H- H* !H- L* H-H% #".

Next we use each of these sequences as training data to estimate a simple trigram model with Good-Turing discounting and Katz backoff for smoothing.[2] We thus end up with 24 trigram models – two for each session in our corpus. We evaluate the trained model by computing its *perplexity* on test data. For each speaker A and his/her interlocutor B, we define the $\mathcal{E}_1(A, B)$ measure as the **negated** perplexity of A's model on B's productions (i.e., on B's sequences of prosodic labels). This measure is asymmetric, and captures how well A's model fits B's productions: a high value of $\mathcal{E}_1(A, B)$ corresponds to a low perplexity, and thus indicates that B's productions are well represented by A's model. In simpler terms, higher values of $\mathcal{E}_1(A, B)$ would roughly correspond to the case in which $contours(B) \subseteq contours(A)$.

To estimate how the $\mathcal{E}_1$ measure of prosodic entrainment correlates with our social variables, we build a vector with the value of $\mathcal{E}_1$ for each member of each speaker pair. Since there are 12 sessions in our corpus, this is a 24-dimensional vector, $\overrightarrow{\mathcal{E}_1} = \langle \mathcal{E}_1(A_1, B_1), \mathcal{E}_1(B_1, A_1), \mathcal{E}_1(A_2, B_2), \mathcal{E}_1(B_2, A_2), \ldots, \mathcal{E}_1(A_{12}, B_{12}), \mathcal{E}_1(B_{12}, A_{12}) \rangle$, 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).

We apply Pearson's correlation tests between $\overrightarrow{\mathcal{E}_1}$ and each of the $\overrightarrow{v}$ vectors. Table 1 summarizes the significant results obtained for five of our social variables; the remaining social variables showed non-significant correlations ($p > 0.05$).

	r	p
contributes-to-successful-completion	0.59	$< .005$
making-self-clear	0.56	$< .005$
bored-with-game	-0.49	$< .05$
planning-what-to-say	0.48	$< .05$
engaged-in-game	-0.41	$< .05$

Table 1. Significant correlations between the $\mathcal{E}_1$ measure of prosodic entrainment and several social variables.

These results indicate a strong positive correlation between how well speaker A's model fits B's productions on the one hand, and A's propensity to be perceived by annotators as contributing to the success of the conversation, making themselves clear and planning their contributions on the other. Roughly speaking, when $contours(B) \subseteq contours(A)$, speaker A is perceived to make clearer, better-planned contributions to the task at hand.

The $\mathcal{E}_1$ measure also correlates **negatively** with A's tendency to be perceived as bored or as engaged in the game. These two results together are quite surprising and the second (engaged) appears to run counter to the other significant correlations as well as findings for measures $\mathcal{E}_2$ and $\mathcal{E}_3$ described below. We have found no reasonable explanation for this finding, and we attribute it either to a statistical artifact (especially given its high p-value at 0.04496) or to a weakness of the $\mathcal{E}_1$ measure.

A crucial weakness of the n-gram models on which the $\mathcal{E}_1$ measure relies is that they only capture local features of the prosodic contours produced by the speakers. In the next sections we consider two additional measures of prosodic entrainment that take full intonational contours into account.

3.2. Levenshtein distance

The second measure of prosodic entrainment presented here is based on an analysis of the sequence of prosodic contours produced by speakers, rather than the sequence of tone labels. We define a CONTOUR as a sequence of 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

$$sim(c_1, c_2) = \frac{m - l}{m}$$

where $m = \max(length(c_1), length(c_2))$, and l is the Levenshtein distance [25] between contours c_1 and c_2. In these calculations, c_1 and c_2 are considered as simple strings. Following this definition, $sim(c_1, c_2)$ ranges from 0 when c_1 and c_2 are completely diferent, to 1 when they are identical.

Next, we extract the list of contours produced by each speaker in an entire Objects Games session, and define the $\mathcal{E}_2(A, B)$ measure of prosodic entrainment between speakers A and B using the following algorithm.

$L \leftarrow$ new list
for each contour c_1 from B:
  $C \leftarrow$ contours from A at most k seconds before/after c_1
  append$\Big(\max_{c_2 \in C} sim(c_1, c_2) \Big)$ to L
return mean(L)

In other words, for each contour from speaker B, we look in its near vicinity ($\pm$ k seconds) for the most similar contour from speaker A, and use the mean of such similarity scores as a measure of overall prosodic entrainment between A and B. Note that $\mathcal{E}_2$ is asymmetric, and a high value of $\mathcal{E}_2(A, B)$ suggests roughly that $contours(B) \subseteq contours(A)$, since for each contour from B, speaker A also produces a similar one shortly before or after.

To study how the $\mathcal{E}_2$ measure of prosodic entrainment correlates with our social variables, we build another 24-dimensional vector, similar to the $\overrightarrow{\mathcal{E}_1}$ vector described above:

[2] We used the SRILM toolkit to perform this task [24].

$\overrightarrow{\mathcal{E}_2} = \langle \mathcal{E}_2(A_1, B_1),\ \mathcal{E}_2(B_1, A_1),\ \mathcal{E}_2(A_2, B_2),\ \mathcal{E}_2(B_2, A_2),\ \ldots,\ \mathcal{E}_2(A_{12}, B_{12}),\ \mathcal{E}_2(B_{12}, A_{12}) \rangle$, where A_i, B_i are the two speakers from session i.

Again, we run Pearson's correlation tests between $\overrightarrow{\mathcal{E}_2}$ and each of the $\overrightarrow{v}$ vectors for our social variables. Table 2 and Figure 1 summarize the significant results obtained for eight of our social variables, using $k = 30$ seconds in the algorithm shown above, or approximately a minute-wide window around each target contour c_1 ($k = 15$ and 60 seconds lead to almost identical results). For the remaining variables the correlations were non-significant ($p > 0.05$).

	r	p
bored-with-game	-0.75	$< .0001$
contributes-to-successful-completion	0.73	$< .0001$
engaged-in-game	0.71	$< .0001$
making-self-clear	0.63	$< .001$
gives-encouragement	0.59	$< .005$
dislikes-partner	-0.54	$< .01$
difficult-for-partner-to-speak	0.48	$< .05$
planning-what-to-say	0.47	$< .05$

Table 2. Significant correlations between the $\mathcal{E}_2$ measure of prosodic entrainment and several social variables.

The $\mathcal{E}_2$ measure presents strong correlations with social variables linked to speaker A's level of engagement. Roughly speaking, when $contours(B) \subseteq contours(A)$, speaker A tends to make better-planned, clearer contributions to the conversation; be more engaged in the game; give more encouragement to their partner (even making it difficult for their partner to speak); like their partner more; and not be bored with the game.

3.3. Kullback-Leibler divergence

Our third measure of prosodic entrainment is based on the *Kullback-Leibler divergence*, an asymmetric measure of the difference of two probability distributions P and Q. It was first defined in [26] as

$$D_{KL}(P \,||\, Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

Note that D_{KL} is asymmetric and $D_{KL}(X, Y) \geq 0$, with $D_{KL}(X, Y) = 0$ iff $X = Y$. Also, low values of $D_{KL}(X, Y)$ correspond roughly to cases in which $X \subseteq Y$. D_{KL} was subsequently adapted to deal with word probabilities in text documents, which allows us to compute the KL divergence for two text documents with respect to their word usage [27].

We define our third measure of prosodic entrainment $\mathcal{E}_3(A, B) = -D_{KL}\big(contours(B),\ contours(A)\big)$ (i.e, the **negated** D_{KL} measure computed over the contours produced by B and A, in that order). In this case we consider each contour (e.g., "H* L-L%" or "H* H* H-") to be a separate 'word' in these computations. Following this definition, high values of $\mathcal{E}_3(A, B)$ correspond roughly to cases in which $contours(B) \subseteq contours(A)$.

We repeat the procedure described above to study how $\mathcal{E}_3$ correlates with our social variables. We build $\overrightarrow{\mathcal{E}_3} = \langle \mathcal{E}_3(A_1, B_1),\ \mathcal{E}_3(B_1, A_1),\ \mathcal{E}_3(A_2, B_2),\ \mathcal{E}_3(B_2, A_2),\ \ldots,\ \mathcal{E}_3(A_{12}, B_{12}),\ \mathcal{E}_3(B_{12}, A_{12}) \rangle$, where A_i, B_i are the two speakers from session i. We run Pearson's correlation tests between $\overrightarrow{\mathcal{E}_3}$ and each of the $\overrightarrow{v}$ vectors for our social variables. Table 3 summarizes the significant results obtained for six of our social variables; for the remaining ones the correlations were non-significant ($p > 0.05$). These results show again strong links between prosodic entrainment and different measures of speaker engagement. Additionally, we find in this case that the $\mathcal{E}_3$ measure correlates positively with speaker A's desire to be liked by their interlocutor.

	r	p
trying-to-be-liked	0.55	$< .01$
bored-with-game	-0.50	$< .05$
difficult-for-partner-to-speak	0.49	$< .05$
contributes-to-successful-completion	0.45	$< .05$
engaged-in-game	0.43	$< .05$
gives-encouragement	0.41	$< .05$

Table 3. Significant correlations between the $\mathcal{E}_3$ measure of prosodic entrainment and several social variables.

4. EFFECT OF PROSODIC CONTOUR USAGE ON PERCEPTION OF ENGAGEMENT

So far, our results indicate that when speaker A uses a superset of B's prosodic contours ($contours(B) \subseteq contours(A)$), A tends to be perceived by annotators as more engaged in the conversation. We hypothesize that this occurs when speaker A **entrains** to B's usage of prosodic contours – i.e., B's contours are added to A's inventory during the conversation.

An alternative, simpler explanation for these findings could be that speakers who use a richer inventory of prosodic contours (i.e., more expressive speech) are more likely to be perceived as more engaged, **independently of their conversational partner's behavior**. In this section we investigate this alternate possibility in our corpus. We measure the prosodic complexity of speaker A using the following indicators of expressiveness:

- **Entropy:** Entropy of A's sequence of tone labels (the entropy of a string S is defined as $-\sum_{i=1..n} P(s_i) \log P(s_i)$, and we consider each tone label as a different character).
- **#Tones:** Number of distinct tone labels used by A.
- **#Contours:** Number of distinct contours used by A.
- **%Rising:** Proportion of rising final intonations used by A (i.e., percentage of -H% over all boundary tones).

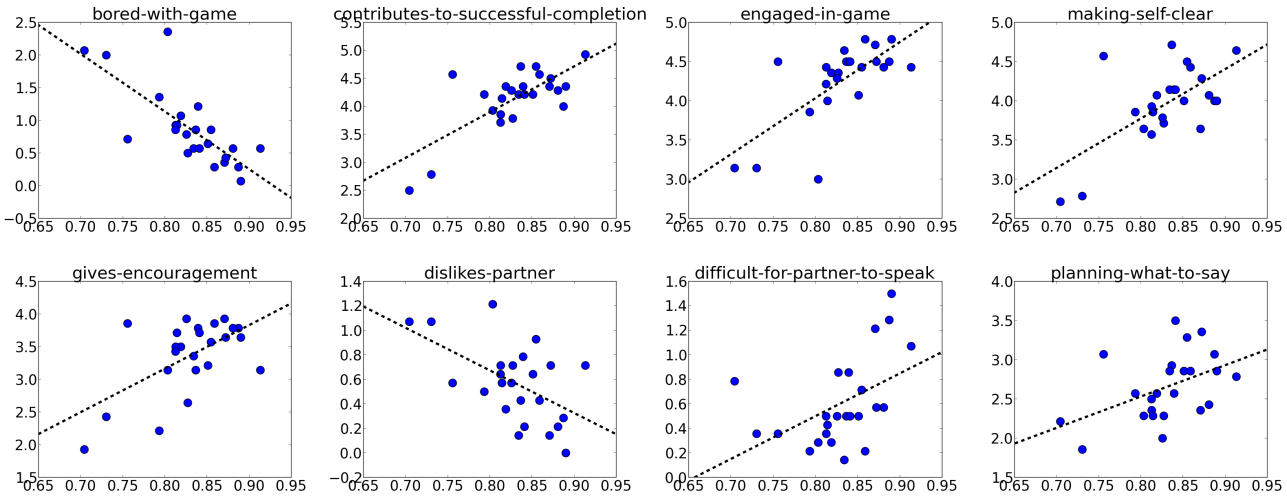

Fig. 1. Scatter plots illustrating the correlations between the $\mathcal{E}_2$ measure (x axes) and the social variables of Table 2 (y axes).

- **%Downstep:** Proportion of downstepped contours used by A (i.e., percentage of contours containing at least one !H* pitch accent).
- **%Complex:** Proportion of complex pitch accents used by A, such as L*+H or H+!H*.

Next we compute Pearson's correlation between each of these measures of expressiveness and each of our social variables, and summarize the results in Table 4. First, we observe a number of significant correlations with social variables related to the speaker's degree of frustration and desire to be liked by the interlocutor, which are mostly unrelated to the variables showing significant correlations with our measures of contour entrainment. At the same time, we find only mild evidence that speaker engagement correlates positively with usage of a richer prosodic variation (or, more expressive speech): the proportion of complex pitch accents shows a correlation coefficient of 0.41 ($p < 0.05$) with *engaged-in-game*. This correlation is weaker and less significant than the ones presented in the previous sections. It seems plausible, then, that it is simply a consequence of a stronger entrainment effect. In other words, speakers who entrain to their interlocutor's usage of prosodic contours will also (necessarily) show a richer prosodic inventory. Therefore, this finding strengthens the hypothesis that the results presented in Sections 3.1, 3.2 and 3.3 are likely to truly represent a strong link between prosodic entrainment and the degree of engagement.

		r	p
Entropy	*frustrated-with-game*	-0.52	$< .01$
Entropy	*more-frustrated-than-B*	-0.41	$< .05$
%Downstep	*frustrated-with-game*	-0.46	$< .05$
%Complex	*engaged-in-game*	0.41	$< .05$
%Complex	*trying-to-be-liked*	0.41	$< .05$
%Rising	*believes-is-better-than-B*	-0.46	$< .05$
%Rising	*frustrated-with-game*	0.60	$< .005$
%Rising	*more-frustrated-than-B*	0.48	$< .05$

Table 4. Significant correlations between our measures of expressiveness and several social variables.

5. CONCLUSIONS AND FUTURE WORK

Using the perplexity of n-gram models, Levenshtein distance, and Kullback-Leibler divergence as metrics, we introduced three novel approaches to examining prosodic entrainment on intonational contours annotated within the ToBI framework in a subset of the Columbia Games Corpus. In these task-oriented dialogues, we find correlations of entrainment on intonational contours with perceived levels of engagement of speakers and positive partner-oriented features of social behavior such as giving encouragement, making self clear, and contributing to successful completion of a task. Importantly, we find that these correlations cannot be attributed to the level of expressiveness of the speakers in terms of the variability of the contours they used. Our results extend previous findings by showing that prosodic entrainment has a robust correlation with positive social traits and that discrete descriptions of prosody provide a useful tool for modeling the relationship of speech prosody to their social functions.

We plan to extend this research in several directions. We will conduct a fine-grained analysis of the specific contours most frequently entrained on by speakers. We will consider using tools such as AuToBI [28] to automate the computation of our measures of prosodic entrainment. We will also enrich our analysis with part-of-speech tags, to study how

different contours are used. Finally, we plan to experiment with dynamic measures of prosodic entrainment, to analyze its progress throughout a conversation.

6. ACKNOWLEDGEMENTS

This work was supported in part by UBACYT 200201202000-25BA, NSF IIS-0803148, the MVTS GAMMA project of the Slovak Academy of Sciences and the VEGA 1/0547/14 grant.

7. REFERENCES

[1] Susan E. Brennan and Herbert 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.

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

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

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

[5] D. Reitter, F. Keller, and J.D. Moore, "A computational cognitive model of syntactic priming," *Cognitive Science*, vol. 35, no. 4, pp. 587–637, 2011.

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

[7] Rivka Levitan, Agustín Gravano, and Julia Hirschberg, "Entrainment in speech preceding backchannels," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

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

[9] Howard Giles, Nikolas Coupland, and Justine Coupland, "Accommodation theory: Communication, context, and consequence," *Contexts of accommodation: Developments in applied sociolinguistics*, vol. 1, 1991.

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

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

[12] TL Chartrand and JA Bargh, "The chameleon effect: the perception-behavior link and social interaction," *J Pers Soc Psychol*, vol. 76, no. 6, pp. 893–910, 1999.

[13] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of NAACL/HLT*, 2012.

[14] C.C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P.G. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proc. of Interspeech*, 2010.

[15] C.C. Lee, A. Katsamanis, M.P. Black, B. Baucom, P.G. Georgiou, and S.S. Narayanan, "An analysis of PCA-based vocal entrainment measures in married couples' affective spoken interactions," in *Interspeech*, 2011.

[16] J.H. Manson, G.A. Bryant, M.M. Gervais, and M.A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evolution and Human Behavior*, vol. 34, pp. 419426, 2013.

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

[18] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg, "Echoes of power: language effects and power differences in social interaction," in *Proc. of WWW*, 2012.

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

[20] Mary E. Beckman and Julia Hirschberg, "The ToBI annotation conventions," *Ohio State University*, 1994.

[21] J.B. Pierrehumbert, *The phonology and phonetics of English intonation*, Ph.D. thesis, MIT, 1980.

[22] John F. Pitrelli, Mary E. Beckman, and Julia Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," in *Proc. of ICSLP*, 1994.

[23] A. Gravano, R. Levitan, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic and prosodic correlates of social behavior," in *Interspeech*, 2011.

[24] Andreas Stolcke, "Srilm – an extensible language modeling toolkit," in *Proc. of Interspeech*, 2002.

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

[26] S. Kullback and R.A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 7986, 1951.

[27] Brigitte Bigi, "Using Kullback-Leibler distance for text categorization," in *European Conference on Information Retrieval (ECIR)*, F. Sebastiani, Ed., vol. 2633 of *LNCS*, p. 305319. Springer, Heidelberg, 2003.

[28] Andrew Rosenberg, "Autobi-a tool for automatic tobi annotation.," in *Proc. of Interspeech*, 2010.