INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement

Juan M. Pérez[1], *Ramiro H. Gálvez*[1], *Agustín Gravano*[1,2]

[1] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
{jmperez,rgalvez,gravano}@dc.uba.ar

Abstract

Synchrony is a form of entrainment which consists in a relative coordination between two speakers, who throughout conversation simultaneously vary some properties of their speech. We describe two novel measures of acoustic-prosodic synchrony that are derived from a time-series analysis of the speech signal. Both of these measures reward positive synchrony (entrainment) and, while one penalizes negative synchrony (disentrainment), the other one rewards it. We describe significant correlations between the second measure and a number of positive social characteristics of the conversations, such as degree of speaker engagement, in a corpus of task-oriented dialogues in Standard American English. Since these correlations are not found to be significant for the first measure, our results suggest that disentrainment may sometimes have a positive effect on the development of conversation.

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

1. Introduction

ENTRAINMENT in spoken dialogue is commonly defined as a tendency of a speaker to adapt some properties of her speech to match her interlocutor's. This phenomenon has been shown to occur along several dimensions, including pronunciation [1]; choice of referring expressions [2]; syntactic structure [3]; turn-taking cues [4]; acoustic-prosodic (a/p) features [5, 6]; and choice of intonational contour [7]. Entrainment has been associated to multiple social aspects of conversations, such as degree of success in completing tasks [8, 9], perception of competence and social attractiveness [10, 11, 12], or degree of speaker engagement [13, 7]. Recently, effective manipulation of entrainment has been shown to improve the naturalness and competence of spoken dialogue systems [14, 15, 16].

Three forms of entrainment are distinguished in the literature [17, 6, 13]. PROXIMITY is the similarity of particular features over an entire conversation, "the product of a single coordination step at the start of the dialogue" [6, page 3081]; CONVERGENCE is the gradual increase in proximity over time; and SYNCHRONY (the matter of the present study) is a relative coordination between partners as the dialogue advances (e.g., both speakers tend to raise and lower their pitch levels in a coordinated fashion).

How to measure synchrony in spoken dialogue is an open problem that has been addressed through several means, which may be grouped into two types of methods. The first type considers single utterances or conversational turns as the units of analysis, and studies the evolution of a/p features along such units (e.g., [18, 7]). The second type of methods consists in extracting a/p features from fixed-length windows (or frames) that are sled along the speech signal, thus creating a series of values for each speaker, which are subsequently compared (e.g., [19, 20]; see [13] for a detailed literature review).

The literature provides evidence of the existence of a seemingly opposite phenomenon, often referred to as DISENTRAINMENT or ANTIMIMICRY, in which a speaker adjusts *away* from their interlocutor, rather than towards them. For example, Healey et al. describe a systematic divergence found in speakers in their use of syntactic constructions [21]. Recently, Levitan et al. report negative synchrony of a/p features to be more prevalent than positive synchrony in four comparable corpora of task-oriented dialogue in Slovak, Argentine Spanish, American English and Mandarin Chinese [22].

Understanding the social connotation and significance of the disentrainment phenomenon remains an open research question. In an experiment involving an actor who either mimicked or antimimicked postures and gestures of subjects, Dabbs describes a decrease in liking caused by antimimicry when the two people were initially similar to one another [23]. In another experiment, Bourhis and Giles show that both phonetic entrainment and disentrainment can occur as a consequence of complex social factors, such as the sense of belonging to a national group or the need to further career prospects [24]. Further, Giles et al. claim that "convergence is a strategy of identification with the communication patterns of an individual internal to the interaction, whereas divergence is a strategy of identification with linguistic communicative norms of some reference group external to the immediate situation" [25, page 27]. De Looze et al. recently present evidence linking *anti*-synchrony in pitch and intensity to an "unbalance" in dialogue participation, with one speaker being more engaged or dominant than the other [13].

In this work we define two synchrony measures based on sliding windows as defined in [19] and [13], to examine how a/p entrainment correlates with social aspects of conversations as judged by independent raters. We show that one of these measures, which equates positive and negative synchrony (i.e., entrainment and disentrainment), correlates more strongly with speaker engagement, suggesting that negative synchrony may serve a positive function in dialogue.

2. Corpus

In the present study we use the 'Objects Games' portion of the COLUMBIA GAMES CORPUS [26], a collection of 12 spontaneous dialogue between pairs of subjects playing a series of computer games. In each game task, the players saw identical collections of objects on their screens. One player (the Describer) had a target object positioned among the other objects, while the other (the Follower) had the same object at the bot-

http://dx.doi.org/10.21437/Interspeech.2016-587

tom of her screen. The Describer was instructed to describe the position of the target object so that the Follower could place it in exactly the same location on her screen. Points (up to 100) were awarded based on how well the Follower's target location matched the Describer's. Each pair of subjects completed 14 such tasks, alternating roles, and were always separated by a curtain to ensure that all communication was oral. The average duration of game tasks was 31 seconds.

The entire corpus has been orthographically transcribed and words aligned with the speech source. We used the Praat toolkit [27] to extract a number of acoustic-prosodic (a/p) features from relevant portions of the speech signal and its orthographic transcriptions. These include F0 maximum and mean values; intensity max and mean; noise-to-harmonics ratio (NHR); jitter and shimmer (computed over voiced frames), and speaking rate (measured in phonemes per second, using dictionary-based phoneme counts).

Several social aspects in the Objects Games were annotated using a crowdsourcing platform. Annotators listened to an audio clip of a game 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? Doing a good job contributing to successful completion? Encouraging his/her partner? Making him/herself clear? Planning what s/he is going to say?*, inter alia. Each task was rated by five unique annotators who answered 'yes' or 'no' to each question, which yielded 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 may be found in [28]. Following [7], in this study we only consider the eight social variables listed above.

3. Method

3.1. Constructing time series of acoustic-prosodic features

As a first step towards constructing a measure of a/p entrainment, we generate time-series data of several a/p features for each game task, by applying the TIME-ALIGNED MOVING AVERAGE (TAMA) method developed by Kousidis et al. [19]. The TAMA method first divides each speaker's speech into overlapping frames of fixed length. We empirically adjust two method parameters, frame length at 16s and time step at 8s, so as to minimize the number of speech-free frames and maximize the frame count for our corpus (Figure 1).

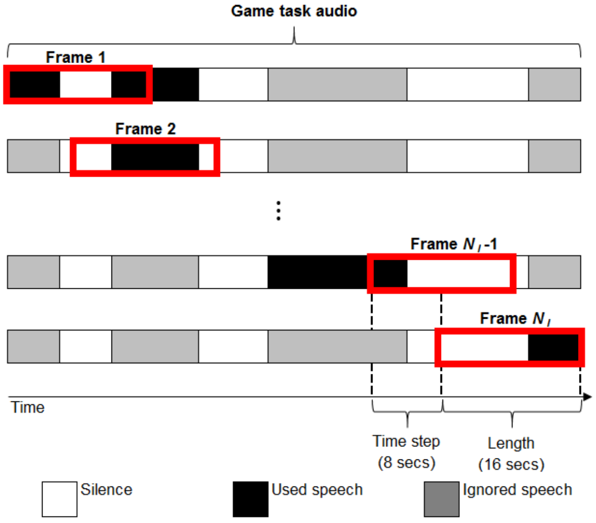

Figure 1: *Schematic diagram of the TAMA method.*

Once the speech has been divided into frames, we compute for each frame the value of the a/p features under consideration. Note that, as Figure 1 shows, it is possible for an utterance to fall either entirely or only partially within a frame. In the latter case, a/p features could be extracted only from the portion of the utterance that falls within the frame, or alternatively from the entire utterance, which De Looze et al. call a HYBRID version of the TAMA method [13]. In this study we use this hybrid method, and refer to utterances which fall at least partially inside a frame as RELEVANT FRAME UTTERANCES (RFU).

All of the a/p features described in Section 2 may be automatically extracted from the RFU speech signal and its corresponding orthographic transcription. For the l-th frame in game task t for speaker s and a/p feature ϕ, this leaves us with a sequence ϕ_1, ϕ_2, ... ϕ_{N_l}, where ϕ_i is the value for the i-th RFU of feature ϕ, and N_l is the number of RFUs that fall at least partially within frame l.

Noting that each RFU has a specific duration (we refer to these durations as d_1, d_2, ... d_{N_l}); the value of feature ϕ for frame l is computed as the duration-weighted mean of the ϕ_i values. Formally,

$$\mu_l = \sum_{i=1}^{N_l} \frac{\phi_i \cdot d_i}{\sum_{h=1}^{N_l} d_h} \tag{1}$$

By repeating this process for all frames in the recording, we end up with a time-series representing the evolution of a/p feature ϕ as dialogue progresses. Let us call this series $U = \{\mu_1, \mu_2, ..., \mu_M\}$, where M is the number of TAMA frames for the current game task and speaker.

A few things should be noted about U. A particular frame could contain no RFUs, in which case its a/p feature values are considered 'missing'. Also, since each task consists of a recording from two speakers, each time-series U from speaker s has a paired time series U' associated to s's partner. U and U' have exactly the same length and, if they have missing values, there is no reason for these to coincide in position in both series. To illustrate, Figure 2 plots two time-series obtained from the two speakers in a particular game task.

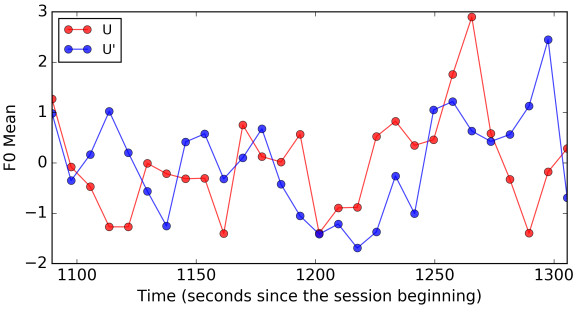

Figure 2: *Time-series from two speakers in a game task.*

3.2. Sample cross-correlation as a proxy for entrainment

The sample cross-correlation is a measure which aims at capturing the correlation between two series as one of them is lagged (i.e., its points are shifted a number of positions). Intuitively, it can be interpreted similarly to Pearson's correlation coefficient between a time-series and a lagged version of another one, which means that its value varies from -1 to 1. Formally, if $h \geq 0$ is the number of lags, then for a/p feature ϕ, for a given speaker and task, the sample cross-correlation r_h is defined as

$$r_h = \frac{\sum_{l=1+h}^{M} (\mu_l - \overline{U}) \cdot (\mu'_{l-h} - \overline{U'})}{\sqrt{\sum_{l=1}^{M} (\mu_l - \overline{U})^2 \cdot \sum_{l=1}^{M} (\mu'_l - \overline{U'})^2}} \qquad (2)$$

where $\overline{U}$ and $\overline{U'}$ are the arithmetic means of non-missing elements of time-series U and U', respectively. In our context, positive (negative) values of r_h can be interpreted as an indication of how much a speaker converged (diverged) in a task in terms of the behavior of a/p feature ϕ to the behavior her partner had h frames before. For illustration purposes, Figure 3 shows the estimated sample cross-correlation for the two time-series shown in Figure 2 as h varies from 0 to 6.

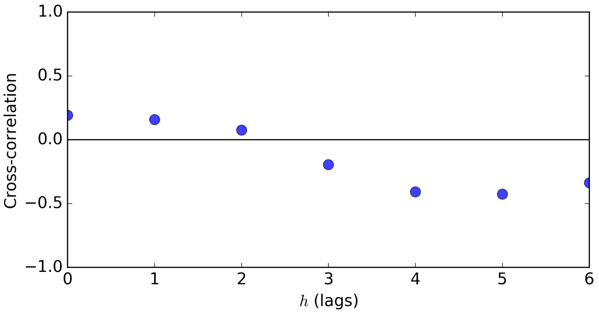

Figure 3: *Example of lagged cross-correlations.*

Two relevant considerations should be done regarding missing values. First, if a term in any summation has a missing value, we simply ignore that term. Second, for a value of r_h to be considered reliable, the numerator in (2) must have at least four non-missing terms; when this does not happen, we simply ignore that particular value of r_h for the rest of the analysis.

We are now ready to define our two measures of acoustic-prosodic synchrony as follows.

Definition 1. The SIGNED SYNCHRONY MEASURE $\mathcal{S}$ is defined as the value of r_h for which $abs(r_h)$ is maximized as h varies from 0 to 6. Formally,

$$\mathcal{S} = \underset{r_h, 0 \leq h \leq 6}{\text{argmax}}\ abs(r_h).$$

Note that positive values of $\mathcal{S}$ represent positive synchrony (or entrainment) in a straightforward way, and negative values represent negative synchrony (disentrainment). Hereafter, we refer as $\mathcal{S}_{s,t,\phi}$ to the signed synchrony measure for speaker s in task t regarding a/p feature ϕ.

Definition 2. The UNSIGNED SYNCHRONY MEASURE $|\mathcal{S}_{s,t,\phi}|$ for speaker s in task t regarding a/p feature ϕ is defined as the absolute value of $\mathcal{S}_{s,t,\phi}$. Formally,

$$|\mathcal{S}_{s,t,\phi}| = abs(\mathcal{S}_{s,t,\phi}).$$

By taking the absolute value, this second measure gives equal treatment to positive and negative synchrony values. In other words, high values of $|\mathcal{S}|$ are indicative of high levels of either entrainment or disentrainment; and low values correspond to a total lack of coordination in either direction.

3.3. Identifying associations between synchrony measures and social variables

Next, we try to identify associations between our two synchrony measures and the eight social variables listed in Section 2. For this purpose, we conduct a linear regression analysis.

Since the way people perceive social variables of particular speakers may be correlated across tasks (e.g., a person may sound consistently friendlier than another), and since our data contain multiple observations from each speaker, we exploit the PANEL STRUCTURE of the data [29], by proposing the following regression model that controls for speaker identity.

$$\gamma_{s,t} = \alpha_{\gamma,s,\phi} + \beta_{\gamma,\phi} \cdot \mathcal{S}_{s,t,\phi} + \varepsilon_{\gamma,s,t,\phi} \qquad (3)$$

Here, $\gamma_{s,t}$ represents the value assigned to social variable γ for speaker s in task t; $\alpha_{\gamma,s,\phi}$ is a fixed effect for each speaker s in the regression studying social variable γ which focuses on a/p feature ϕ;[1] $\mathcal{S}_{s,t,\phi}$ is the value of our entrainment measure for speaker s in task t regarding a/p feature ϕ (subsequently, we also estimate this model for $|\mathcal{S}_{s,t,\phi}|$); $\varepsilon_{\gamma,s,t,\phi}$ is the error term.

Our main interest relies in the estimated values of $\beta_{\gamma,\phi}$ and their statistical significance.[2] A positive (negative) and statistically significant estimate of $\beta_{\gamma,\phi}$ would indicate a positive (negative) correlation between our synchrony measure for a/p feature ϕ and social variable γ.

4. Results

Table 1 presents the estimated values of the $\beta_{\gamma,\phi}$ coefficients for all combinations of social variables and a/p features, considering both signed (Panel A) and unsigned (Panel B) versions of our synchrony measures. The level of statistical significance is signalled for each coefficient with one star (*) for 10% level, two stars (**) for 5% level, and three stars (***) for 1% level.

Inspection of Panel A of Table 1 reveals that the signed synchrony measure generally fails to capture meaningful associations with the social variables. Only five coefficients approach statistical significance; for the rest, the estimated values of $\beta_{\gamma,\phi}$ are remarkably low and non-significant. Overall, no clear pattern arises from these results. In fact, given the number of tests considered and the low confidence for these five coefficients, it is very likely that they are simply statistical artifacts.

Panel B reproduces Panel A's results by using the unsigned version of the synchrony measure, $|\mathcal{S}_{s,t,\phi}|$. In this case, we can observe a clear pattern in these results. The three leftmost columns show social variables presumed to be associated with a positive perception of the speakers' interaction: *contributes-to-successful-completion*, *making-self-clear* and *engaged-in-game*. For these variables, the estimated values of $\beta_{\gamma,\phi}$ tend to be positive and, many times, statistically significant. This tendency seems to hold especially for features related to F0 and intensity.

[1] Coefficient $\alpha_{\gamma,s,\phi}$ captures any constant-through-tasks difference across speakers in any observable or unobservable predictors.

[2] Given that $\varepsilon_{\gamma,s,t,\phi}$ could be correlated within tasks of the same speaker, we calculated standard errors clustered at the speaker level in order to avoid overestimating the precision of our results [30].

Table 1: *Estimated associations between social variables and entrainment of acoustic-prosodic features.*

	Contributes to Successful Completion	Making Self Clear	Engaged in Game	Planning what to Say	Gives Encour-agement	Difficult for Partner to Speak	Bored with Game	Dislikes Partner
Panel A. Entrainment calculated using our *signed* synchrony measure ($\mathcal{S}_{s,t,\phi}$)								
Intensity Max	-0.08	0.23 *	0.03	0.09	0.07	-0.11	0.05	-0.07
Intensity Mean	-0.27	-0.15	0.04	-0.17	-0.04	0.01	0.02	-0.03
F0 Max	-0.10	-0.17	0.02	-0.17	0.02	0.07	0.14	-0.06
F0 Mean	0.11	-0.08	0.09	0.09	0.12	-0.01	0.03	-0.21 *
NHR	-0.06	0.03	-0.08	0.07	-0.07	-0.05	-0.03	-0.28 **
Shimmer	-0.12	0.03	-0.15	-0.02	-0.10	-0.01	0.13	-0.02
Jitter	-0.22 *	-0.01	0.02	-0.27 *	-0.10	-0.10	0.15	-0.02
Phonemes/sec	-0.06	0.09	0.04	0.03	-0.01	-0.08	-0.03	-0.12
Panel B. Entrainment calculated using our *unsigned* synchrony measure ($\lvert\mathcal{S}_{s,t,\phi}\rvert$)								
Intensity Max	0.08	1.40 ***	0.17	0.21	0.52	-0.56 **	0.35	-0.53
Intensity Mean	0.72 **	0.82	0.65 *	0.61	0.32	0.36	0.00	0.31
F0 Max	1.02 **	0.69 *	0.54	0.27	0.41	-0.09	-0.31	0.31
F0 Mean	1.00 ***	0.63 **	0.67 **	0.74	0.07	-0.56	-0.40	-0.03
NHR	0.53	0.83 **	0.34	-0.12	0.38	0.07	0.31	0.19
Shimmer	0.30	0.11	0.11	-0.71	0.23	0.14	-0.19	-0.11
Jitter	0.60 *	0.59	0.54 **	-0.06	-0.16	-0.29	-0.02	0.06
Phonemes/sec	0.36	0.78 *	0.23	0.74	-0.01	-0.75 **	0.02	0.05

* significant at 10%, ** significant at 5%, *** significant at 1%.

On the other hand, for social variables presumed to be associated with a negative perception of speakers' interaction (*difficult-for-partner-to-speak*, *bored-with-game*, *dislikes-partner*), we observe negative values of $\beta_{\gamma,\phi}$. However, in this case only two coefficients for *difficult-for-partner-to-speak* achieve statistically significance.

It could be argued that our unsigned synchrony measure and the social variables may be correlated with the role speakers had in the game tasks (Describer or Follower). Thus, we need to discard the possibility that our results are artifacts coming from a tentative confounding variable [31]. For this purpose, we check the robustness of all regressions presented in Table 1 by controlling for speaker role. We achieve this by adding to Equation (3) an independent dummy variable that takes value 1 if the target speaker was the Describer, or 0 otherwise. All results (which we omit here due to space limitations) remain qualitatively identical; only negligible changes are observed in the estimated coefficients and in the levels of statistical significance.

5. Discussion and conclusions

The two measures of acoustic-prosodic synchrony presented in this work appear to effectively capture two seemingly opposed phenomena of dialogue, entrainment and disentrainment, in different ways. The *signed* synchrony measure distinguishes, as its name suggests, between positive and negative synchrony. That is, it distinguishes between entrainment and disentrainment, rewarding the former with positive scores and penalizing the latter with negative scores. The *unsigned* synchrony measure removes such a distinction by using absolute values, and gives equal importance to entrainment and disentrainment.

Finding positive associations between social variables with positive connotation (such as the degree of speaker engagement) and the *signed* synchrony measure would signal entrainment as a desirable speaker characteristic, as has been repeatedly reported in the literature. Additionally, in this scenario disentrainment would be attributed an effect ranging from neutral to negative on the advancement of dialogue; in other words, it would not necessarily be a desirable trait. This scenario was discarded by the first set of regression tests we conducted, which found practically no relation between social variables and the signed synchrony measure.

On the other hand, our regression analysis *did* find significant positive associations between positive social variables and the *unsigned* synchrony measure of several acoustic-prosodic features. Note that the sole difference between our two measures is the equal treatment of entrainment and disentrainment given by the unsigned measure. Therefore, given the lack of significant associations for the signed measure described in the previous paragraph, we conclude that disentrainment (or at least, some form of it) must have a positive effect on conversation. This hypothesis gains further support from the negative associations found between a negative social variable (*difficult-for-partner-to-speak*) and two acoustic-prosodic features, thus indicating that when neither entrainment nor disentrainment were present, the dialogues had negative characteristics.

In future research, we plan to check the robustness and generalizability of the results presented in this paper, by reproducing the analysis on other spoken dialogue corpora. Additionally, since time-series built with the TAMA method are autocorrelated of order 1 by construction, we plan to further check the robustness of our conclusions by running pre-whitening filters over the pair of speakers in each conversation. Preliminary results points toward no relevant differences in including or not such a filter in the analysis. Furthermore, it might be interesting to analyze some other acoustic-prosodic features, such as the pause duration between speakers.

6. Acknowledgements

This material is based upon work supported by CONICET, ANPCYT PICT 2014-1561, UBACYT 20020120200025BA, and the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055. The authors thank Štefan Beňus, Pablo Brusco, Andrés Farall, Luciana Ferrer, Julia Hirschberg, Juan Kamienkowski and Rivka Levitan for valuable suggestions and comments.

7. References

[1] J. S. Pardo, "On phonetic convergence during conversational interaction," *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 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, p. 1482, 1996.

[3] 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.

[4] R. Levitan, Š. Benuš, 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*, 2015.

[5] A. Ward and D. Litman, "Measuring convergence and priming in tutorial dialog," University of Pittsburgh, Tech. Rep., 2007.

[6] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions." in *Proceedings of Interspeech*. ISCA, 2011, pp. 3081–3084.

[7] A. Gravano, Š. Benuš, R. Levitan, and J. Hirschberg, "Backward mimicry and forward influence in prosodic contour choice in Standard American English," in *Proceedings of Interspeech*, 2015.

[8] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of ACL*, 2008, pp. 169–172.

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

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

[11] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels." in *Proceedings of ACL*, 2011, pp. 113–117.

[12] Š. Beňuš, A. Gravano, R. Levitan, S. I. Levitan, L. Willson, and J. Hirschberg, "Entrainment, dominance and alliance in supreme court hearings," *Knowledge-Based Systems*, vol. 71, pp. 3–14, 2014.

[13] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, pp. 11–34, 2014.

[14] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction," in *Proceedings of ICPhS*, 2003, pp. 2453–2456.

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

[16] J. Thomason, H. V. Nguyen, and D. Litman, "Prosodic entrainment and tutoring dialogue success," in *Artificial Intelligence in Education*. Springer, 2013, pp. 750–753.

[17] J. Edlund, J. B. Hirschberg, and M. Heldner, "Pause and gap length in face-to-face interaction," in *Proceedings of Interspeech*, 2009.

[18] M. Heldner, J. Edlund, and J. Hirschberg, "Pitch similarity in the vicinity of backchannels," in *Proceedings of Interspeech*, 2010, pp. 3054–3057.

[19] S. Kousidis, D. Dorran, C. Mcdonnell, and E. Coyle, "Time series analysis of acoustic feature convergence in human dialogues," in *Proceedings of SPECOM*, St. Petersburg, Russian Federation, 2009.

[20] A. V. Barbosa, R.-M. Déchaine, E. Vatikiotis-Bateson, and H. C. Yehia, "Quantifying time-varying coordination of multimodal speech signals using correlation map analysis," *Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2162–2172, 2012.

[21] P. G. Healey, M. Purver, and C. Howes, "Divergence in dialogue," *PloS one*, vol. 9, no. 6, p. e98598, 2014.

[22] R. Levitan, Š. Benuš, A. Gravano, and J. Hirschberg, "Acoustic-prosodic entrainment in slovak, spanish, english and chinese: A cross-linguistic comparison," in *Proceedings of SIGdial*, 2015, pp. 325–334.

[23] J. M. Dabbs Jr., "Similarity of gestures and interpersonal influence," in *Proceedings of the Annual Convention of the American Psychological Association*, vol. 4, no. 1, 1969, pp. 337–338.

[24] R. Y. Bourhis and H. Giles, "The language of intergroup distinctiveness," *Language, Ethnicity, and Intergroup Relations*, pp. 119–36, 1977.

[25] H. Giles, N. Coupland, and J. Coupland, "Accommodation theory: Communication, context and consequence," *Contexts of Accommodation: Developments in Applied Sociolinguistics*, pp. 1–68, 1991.

[26] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[27] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2016. [Online]. Available: http://www.praat.org

[28] A. Gravano, R. Levitan, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic and prosodic correlates of social behavior," in *Proceedings of Interspeech*, 2011.

[29] J. M. Wooldridge, *Econometric analysis of cross section and panel data*. Cambridge and London: MIT Press, 2002.

[30] A. C. Cameron and D. L. Miller, "A practitioners guide to cluster-robust inference," *Journal of Human Resources*, vol. 50, no. 2, pp. 317–372, 2015.

[31] K. A. Frank, "Impact of a confounding variable on a regression coefficient," *Sociological Methods & Research*, vol. 29, no. 2, pp. 147–194, 2000.