INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Prosodic Facilitation and Interference while Judging on the Veracity of Synthesized Statements

Ramiro H. Gálvez[1], *Štefan Beňuš*[2,3], *Agustín Gravano*[1,4], *Marian Trnka*[3]

[1] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[2] Constantine the Philosopher University in Nitra, Slovakia
[3] Institute of Informatics, Slovak Academy of Sciences, Slovakia
[4] Instituto de Investigación en Ciencias de la Computación, CONICET-UBA, Buenos Aires, Argentina

rgalvez@dc.uba.ar, sbenus@ukf.sk, gravano@dc.uba.ar, Marian.Trnka@savba.sk

Abstract

Two primary sources of information are provided in human speech. On the one hand, the verbal channel encodes linguistic content, while on the other hand, the vocal channel transmits paralinguistic information, mainly through prosody. In line with several studies that induce a conflict between these two channels to better understand the role of prosody, we conducted an experiment in which subjects had to listen to a series of statements synthesized with varying prosody and indicate if they believed them to be true or false. We find evidence suggesting that acoustic/prosodic (a/p) features of the synthesized statements affect response times (a well-known proxy for cognitive load). Our results suggest that prosody in synthesized speech may play a role of either facilitation or interference when subjects judge the truthfulness of a statement. Furthermore, we find that this pattern is amplified when the a/p features of the synthesized statements are analyzed relative to the subjects' own a/p features. This suggests that the entrainment of TTS voices has serious implications in the perceived trustworthiness of the system's skills.

Index Terms: text to speech, trustworthiness, prosodic interference/facilitation, entrainment.

1. Introduction

Human speech provides at least two primary sources of information – a *verbal* channel, encoding linguistic content, and a *vocal* one, which transmits paralinguistic information, mainly through *prosody* [1, 2, 3] (i.e., intonation, tone, stress, rhythm). Acoustic/prosodic (a/p) cues have been found to assist listeners in decoding the vocal channel in the recognition of spoken words, the computation of syntactic structure, the processing of discourse structure, and the detection of emotions [4, 5, 6, 7].

To better understand the effect of prosody, several studies have carried out experiments that induce a conflict between the verbal and vocal channels. For example, [8] study the role of prosody in facilitating and interfering in the resolution of temporary syntactic closure, and find that sentences with cooperating (conflicting) prosody tend to be processed faster (slower) than sentences with baseline prosody. In [9] participants were instructed to identify emotions from either lexico-semantic content or prosody, finding that incongruence between lexico-semantic and prosodic emotion causes cognitive conflict (i.e., higher response times and skin conductance response). Similarly, [10] find that conflict between lexico-semantic and prosodic emotion also activates brain networks previously associated with monitoring cognitive conflict processing.

Text to speech (TTS) systems have possessed the ability of modifying the prosody of synthesized speech for decades [7]. However, the effect of modifying prosody on user interaction with a system has not been extensively investigated (notable exceptions being [11, 12]). As virtual assistants such as Amazon Alexa, Apple Siri, Google Assistant and Microsoft Cortana gain in popularity, understanding which characteristics of their synthesized speech improve the interaction stands as an important research and practical question.

Here we present the results of an experiment in which subjects were asked to judge the veracity of statements synthesized with varying prosody. We find evidence suggesting that a/p features of the synthesized statements affect response times (a well-known proxy for cognitive load). In line with literature exploring conflict between the verbal and the vocal channels as well as with literature on persuasiveness and prosody [13], our results indicate that the same configurations of a/p features associated with faster response times when subjects believe the statements to be true are also associated with slower response times when they do not hold this belief. These results suggest that prosody in synthesized speech may play a role of either facilitation or interference when subjects judge the truthfulness of statements, and that this effect might depend on their prior beliefs. Furthermore, we find that this pattern is amplified when the a/p features of the synthesized statements are analyzed relative to the subjects' a/p features. This suggests that *entrainment* [14, 15, 16, 17, 18] (a tendency of speakers to align features of their speech with those of their interlocutor's) of the TTS voice to the user's voice has serious implications in the perceived trustworthiness of the system.

2. Method

2.1. Participants

A total of 72 adult subjects participated in this study, divided into three rounds of 24 subjects. The first round took place in Buenos Aires (Argentina) in August 2016 (12 F, 12 M; mean age 22.7, stdev 3.25). The second round took place in Nitra (Slovakia) in October 2016 (17 F, 7 M, mean age 21.3, stdev 2.21). The third round took place in Buenos Aires in December 2016 (6 F, 18 M, mean age 22.3, stdev 2.39). All participants were native speakers of the local language (either Argentine Spanish or Slovak). During recruitment, participants were notified that they would be payed for participating (roughly 9 US dollars per hour in local currency) and that they would receive additional money if they achieved a high score in a series of simple tasks.

http://dx.doi.org/10.21437/Interspeech.2017-453

2.2. Stimuli

We selected a series of 20 statements to be presented to participants. These were selected from a larger pool of 29 statements designed for an anchoring bias experiment carried out in a Neuroscience class at Universidad de Buenos Aires [19]. Each statement is a fact that might be true or false, and indicates that something happened before/after some date, weighs more/less than some weight, is formed by more/less than some number of elements, and so on. These values (date, weight, number of elements, etc.) were chosen to make it difficult to judge the correctness of the statements, and this difficulty was validated in a preliminary study. Examples of the statements selected for our experiment are: "*Napoleon Bonaparte was born before the year 1750*", "*The weight of an average iceberg is less than 500 tons*", "*Africa is formed by fewer than 35 countries*".

Statements were synthesized using Slovak and Spanish TTS software as described in [20, 21]. Importantly, each statement was synthesized using all 27 different combinations of three a/p features relative to the TTS systems' default values. The used values were $\{-10\%, +0\%, +10\%\}$ for pitch, $\{-15\%, +0\%, +15\%\}$ for speech rate and $\{-5\%, +0\%, +5\%\}$ for intensity. These values were selected after analyzing the variation observed in the a/p features of subjects who participated in the experiment reported in [20]. Before finally setting these values, we ran preliminary studies checking that they produced speech which was considered natural relative to the baseline a/p features of the synthesizers.

2.3. Procedure

Upon arrival to the lab, subjects were instructed to read and sign an informed consent form, and complete a sociodemographic questionnaire. They sat in front of a desktop computer wearing a headset with microphone, and started playing a series of card games against the computer, as described in [20]. During these games, participants interacted with two avatars by verbally asking for and listening to advice regarding which card to play. These games lasted for approximately 40 minutes; subjects made 45 requests for advice and heard the corresponding 45 responses from the avatars. From the subjects' requests for advice, we were able to detect with high confidence the values of *pitch* (F0 mean), *speech rate* (syllables per second) and *intensity* (ENG mean, mean decibels).

Subjects then proceeded to play one additional shorter game. They were told that they would hear a series of statements, some of which were true facts and others were false; and that, after hearing each statement, they would have to indicate if they considered it to be true or not by pressing designated keys. Instructions mentioned that, even when they might not be completely sure about the veracity of a statement, they should choose the answer for which they leaned the most. Participants then listened to the statements. During the time each statement was being played, they were not able to provide an answer. As soon the statement reproduction ended, the screen displayed a text message indicating them to answer and, once the answer had been provided, the next statement was reproduced in a similar fashion. Each subject heard each of the 20 statements just once. We kept record of each answer given as well as its response time (calculated from the onset of the message asking for the answer). Based on the number of correct answers, a modest extra payment was made.

We randomized the order in which each subject heard the statements, as well as the combination of a/p features used to synthesize each of them. In doing so, we guaranteed that, in each round of 24 subjects: 1) each combination of a/p features would be synthesized at least 15 times, 2) at least 10 different question would be synthesized with each combination of a/p features, and 3) each a/p feature combination would be heard by at least 15 different subjects.

2.4. Analysis

2.4.1. Measuring absolute behavior of a/p features

As a first step in analyzing existing associations between participants' behaviour and a/p features of the synthesized statements, we needed to extract reliable measures of the variations in the a/p features of the stimuli. We measured a/p feature values in the following way: 1) we chose to measure the a/p features directly from the synthesized speech, rather than trusting the target values specified in the TTS input. We used the Praat toolkit [22] to estimate a/p feature values. For the case of speech rate, given that we knew in advance the number of syllables of each statement, we took as speech rate the ratio between the number of syllables of the synthesized statement and the length of the file containing it; 2) to better compare across languages, we normalized the estimated a/p values by computing their z-scores by language. Concretely, if $\phi_{k,i,j}$ is the value of a/p feature k for subject i (who speaks language l) in trial j, the normalized value $\tilde{\phi}_{k,i,j}$ was calculated as

$$\tilde{\phi}_{k,i,j} = \frac{\phi_{k,i,j} - \mu_{k,\phi,l}}{\sigma_{k,\phi,l}} \quad (1)$$

where $\mu_{k,\phi,l}$ is the mean value of a/p feature k, estimated over all statements synthesized in the experiment for a given language (20×24 cases for Slovak, 20×48 cases for Spanish) and $\sigma_{k,\phi,l}$ its estimated standard deviation. In this way a/p features of a statement are expressed as the signed number of standard deviations above or below their mean values in the experiment.

2.4.2. Measuring behavior of a/p features relative to the listener's voice

Given that the perceived level of an a/p feature may depend on the listener's way of speaking (e.g., a slow talker may perceive the baseline speech rate of a TTS system as fast [23]), we also computed a/p TTS features relative to a/p features of the users. Specifically, we estimated the a/p feature values for every request for advice in the cards game using the same procedure used for the synthesised statements. Then, for every subject we calculated the mean value of their a/p features (let us call these values $\psi_{k,i}$). We then estimated z-scores for $\psi_{k,i}$ as in Equation (1), taking into consideration the speakers' language and, in the case of pitch, the speakers' gender (let us call these values $\tilde{\psi}_{k,i}$). Finally, we estimated the values of the synthesized a/p features relative to the subjects' voices as $\tilde{\phi}_{k,i,j} - \tilde{\psi}_{k,i}$ (let us call these values $\tilde{\phi'}_{k,i,j}$). Note that $\tilde{\phi'}_{k,i,j} > 0$ captures the idea that the value of a/p feature k of the synthesized statement is high relative to the typical realization of the a/p feature by the subject listening to the statement.

Figure 1 plots kernel density estimates of the distribution of $\tilde{\phi'}_{k,i,j}$ separately for the three a/p features and two languages. Note that all distributions are clearly centered around 0 and that distributions for both languages are heavily superimposed, suggesting that our normalizations by language and gender worked well.

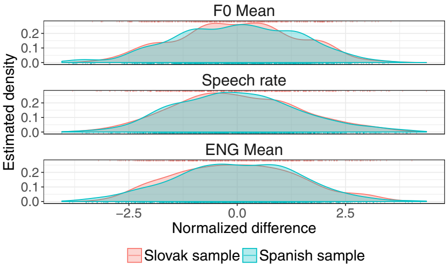

Figure 1: *Distribution of the a/p features measures relative to the listener's voice.*

2.4.3. Identifying associations between a/p features and subjects' behavior

We will focus on associations between the a/p features of the synthesized statements and two aspects of participants' behaviour (which we treat as dependent variables): 1) whether subjects believed the statement to be true (*Answered true*), and 2) response time. To quantify these associations we estimated regression models with the following specification

$$y_{i,j} = \beta_0 + \sum_{k \in K} \beta_k \cdot I(\tilde{\phi}_{k,i,j}) + \epsilon_{i,j} \qquad (2)$$

where $y_{i,j}$ is the observed value of the relevant dependent variable for subject i in trial j, $K = \{F0\,mean,\ Speech\,rate,\ ENG\,mean\}$, $I(x)$ is a function that takes the value 1 if $x \geq 0$ and 0 in any other case, $\tilde{\phi}_{k,i,j}$ are the a/p features values of the synthesized statements, and $\epsilon_{i,j}$ is an error term. Our parameters of interest are $\beta_{F0\,mean}$, $\beta_{Speech\,rate}$ and $\beta_{ENG\,mean}$. It should be mentioned that Equation 2 uses $I(x)$ for ease of interpretation of the results; nevertheless, all conclusions hold when regressing $y_{i,j}$ against the raw values of $\tilde{\phi}_{k,i,j}$ and $\tilde{\phi}'_{k,i,j}$ instead of passing them to this function.

3. Results

3.1. Estimated associations between a/p features behavior and subjects' veracity judgements

Table 1 presents the estimated coefficients obtained when regressing the answers subjects gave regarding the truthfulness of statements against their a/p features. Given that this is a binary outcome (1 if the subject answered true, 0 if not), we estimated Equation 2 through a logistic regression model. Each column in each panel presents the relevant estimated coefficients for individual regressions. Regression results are presented for the pooled Slovak and Argentinian data (*SK* + *ES*), as well as for regressions which just consider Slovak data (*SK*) and Argentinian data (*ES*). The left panel presents estimated coefficients for regressions were a/p features behavior is introduced in an absolute way as detailed in Section 2.4.1, the right one presents estimations when it is introduced relative to the subjects voices as detailed in Section 2.4.2.

We do not find significant relations between subjects response and neither pitch nor speech rate. For intensity, we observe a negative relationship between high intensity and the probability of answering true, which is captured only when intensity is measured relative to subjects' voices. This last result is mainly driven by the Slovak sample (although Argentinian results point toward the same direction).

Table 1: *Estimated associations between believing the statement and the synthesized a/p features.*

	Absolute			Relative		
	SK + ES	*SK*	*ES*	*SK + ES*	*SK*	*ES*
F0 mean	.10	.17	.04	−.03	−.09	−.01
Speech rate	.11	.20	.06	−.06	−.15	−.03
ENG mean	−.02	−.09	.00	−.19*	−.35*	−.12
Obs.	1,440	480	960	1,440	480	960

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

3.2. Estimated associations between a/p feature behavior and response time

Table 2 presents estimated coefficients when regressing response times (in log-transformed milliseconds) against the a/p features behavior. As suggested in [24], response times are expressed in logarithms to correct high positive asymmetries in their distribution. Additionally, to exclude possible outliers, we dropped trials with response times lower than 250 ms (18 trials) and higher than 8.000 ms (182 trials). We set these thresholds trough exploratory analysis of the response time distribution. Our results are robust to the inclusion of these trials.[1] As the observed effect an a/p feature may depend on the prior belief a subject had regarding the veracity of a statements (vocal cues that sounds natural or facilitating when being told a truth may interfere when being told a lie), in Table 2 we divided the analysis into samples considering only trials in which subjects believed the statement to be true (left panels) and false (right panels). Additionally, in this table the top panels present the coefficients estimated for regressions using absolute a/p features (as detailed in Section 2.4.1), and the bottom panels for regressions using a/p features relative to the subjects voices (as detailed in Section 2.4.2). All regressions were estimated through ordinary least squares.

Table 2: *Estimated associations between response time and the synthesized a/p features.*

	Answered true			Answered false		
	SK + ES	*SK*	*ES*	*SK + ES*	*SK*	*ES*
Absolute						
F0 mean	.09	.19	.04	−.13**	−.22**	−.05
Speech rate	−.07	−.10	−.06	−.08	−.12	−.05
ENG mean	−.06	−.01	−.09	.02	−.03	.04
Relative						
F0 mean	.01	.02	.01	−.16**	−.23**	−.10
Speech rate	.14**	.27**	.08	.01	−.05	.05
ENG mean	−.04	.05	−.07	−.11*	−.26**	−.05
Obs.	679	215	464	561	188	373

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

In the case of a/p features expressed in absolute terms, response time only shows a strong association with pitch. Results suggest that a high pitch leads to somewhat longer response times when subjects believed the statement to be true and shorter ones when they did not. In the case of a/p features relative to the subjects' voices, results are more evident. The previously mentioned relation of response times with pitch still holds. Additionally, longer response times are associated with high speech rates when subjects believed the statement to be true and almost no effect is found when they did not. For in-

[1] An additional reason for this exclusion is that, if one assumes the effects of prosody to decrease as response time passes, high response time answers might be driven solely by strong conscious reasoning and would hardly be affected by prosodic facilitation/interference.

tensity longer response times are related to lower intensity in statements believed to be false.

Estimating opposite signs when subjects answered true relative to when they did not is reassuring, as it goes in line with the idea of prosodic facilitation and interference. In Table 2 this pattern only seems to hold for pitch. As this facilitation or conflict may depend to a great extent on the strength of subjects' prior beliefs regarding the statements veracity, in Table 3 we further explore how the estimated coefficients presented in the bottom panels of Table 2 vary when the sample is divided into trials correctly answered (*Ans. right*, top panels) and wrongly answered (*Ans. wrong*, bottom panels). Note that even when some of observations might end in the top panels by pure luck (as subjects might have guessed some answers), if a subject had a strong and correct prior belief regarding a statement, its response should fall in the top panels; and, if there is any facilitation/interference effect between beliefs on veracity and prosody, it should be manifested primarily in these panels.

Table 3: *Estimated associations between response time and the synthesized a/p features, for trials correctly and wrongly answered.*

	Answered true			Answered false		
	SK + ES	*SK*	*ES*	*SK + ES*	*SK*	*ES*
Ans. right						
F0 mean	.03	.03	.07	$-.19^{**}$	$-.35^{***}$	$-.08$
Speech rate	$.20^{*}$	$.34^{*}$	.20	$-.03$	$-.07$	.00
ENG mean	.09	$.64^{***}$	$-.13$	$-.10$	$-.26^{**}$	$-.02$
Obs.	254	74	180	382	139	243
Ans. wrong						
F0 mean	$-.02$	$-.03$	$-.05$	$-.12$	.07	$-.14$
Speech rate	$.19^{**}$	$.28^{*}$	.13	.12	$-.03$	.21
ENG mean	$-.11$	$-.26^{*}$	$-.04$	$-.11$	$-.31$	$-.10$
Obs.	425	141	284	179	49	130

* significant at 10%, ** significant at 5%, *** significant at 1%. This table only presents estimated coefficients for a/p features relative to the listeners voice.

In line with the idea of facilitation/interference depending on prior beliefs, estimated coefficients presented in the top panels of Table 3 point in the opposite direction when subjects answered true relative to when they answered false. Estimates are positive for trials which were believed to be true and negative for the ones which were not. Note that this opposing pattern also points toward the idea that our results are not driven by perceived unnaturalness of the synthesized a/p patterns, as one would expect unnaturalness to affect response times in the same direction whether subjects answer true or false. Interestingly, this opposing pattern is not seen in trials answered wrongly (bottom panels), for which signs tend to coincide between the left and right panels. If one assumes a weaker belief for wrongly answered trials, this last result is consistent with a lack of conflict/agreement between beliefs on the veracity of the statement and its a/p features, and therefore with an absence of facilitating or interfering effects.

4. Discussion and conclusion

We deployed an experiment in which subjects listened to a series of statements synthesized with varying prosody and had to indicate if they believed them to be true or not. Although we did not find strong associations between prosody and the responses given, we did find evidence suggesting that a/p features of the statements affected response times. Our results contribute to at least three open research questions.

First, as far as we know, facilitation and interference effects between lexicosemantics and prosody have not been studied with the focus placed on listeners' judgements on the veracity of statements. Our results provide evidence suggesting that these effects do arise, as certain a/p features configurations facilitate response when subjects believe a statement to be true and interfere when they do not.

Second, the fact that associations are better captured when a/p features are measured relative to the listeners' voices suggests that facilitation in the interaction between virtual assistants and users may be induced by adapting the former a/p features to the latter ones. This closely relates to the idea of entrainment by *proximity* (similarity of an a/p feature over the entire conversation, see [16]) and contributes to a growing literature on the effects of entrainment in dialogue (see, for example, [25, 26, 14]).

Third, our results contribute to a vast literature on the associations that prosody has with credibility, persuasion, deception and trustworthiness. For pitch, our findings indicate that high pitch interferes with judgement when subjects regard the statements to be true and facilitates it when they do not. This goes in hand with the association of deception with higher pitch [27] and of dominance with low pitch. The fact that even for women low pitch is associated to dominance [28, 29], goes in hand with our finding that pitch has an already significant effect even when it is not measured relative to the listeners' voices. On speech rate, our results go in hand with Smith and Shaffer's [13] results, which state that persuasion depends on the message being delivered, finding that for pro-attitudinal/counter-attitudinal messages, slowing down speech seems to improve/lessen persuasion. Finally, regarding intensity, a less studied feature, for which there is not a clear consensus on its effect on trustworthiness and persuasion (see [30]); our results suggest again that the effect depends on subjects' prior beliefs, where high intensity interferes with judgement when subjects' regard the statements to be true and facilitates it when they do not.

Finally, to further place our results in context, the fact that pitch and intensity features tend to pattern in the same way regarding the facilitation/inhibition, might be viewed through the universal biological codes associated with prosody [31, 32, 33]. In the *Effort code*, speakers articulate more precisely and expand more energy (note that physiologically greater subglottal pressure increases both energy and pitch). We speculate that greater prosodic effort is associated with the need to *'boost'* statements' truth value, and subsequent compensation for these effects in subjects' responses leads to the interference/facilitation effects observed in our data. (For example, if a statement is believed to be true/false and the prosody signals great effort in the speaker producing the statement, maybe the statement is not really true/false and the prosody just conveys this boosting effect; this dissonance then increases cognitive load and leads to longer response times.) It is unclear if our speech rate results fall in the same pattern since articulatory effort is increased in both fast and slow rates compared to normal rates. Further research is needed to better understand these effects.

5. Acknowledgements

This material is based upon work supported by CONICET, ANPCYT PICT 2014-1561, and the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055. The authors thank Juan Kamienkowski, Brian Litwak, and our anonymous reviewers for valuable suggestions and comments.

6. References

[1] W. Apple, L. A. Streeter, and R. M. Krauss, "Effects of pitch and speech rate on personal attributions." *Journal of Personality and Social Psychology*, vol. 37, no. 5, p. 715, 1979.

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language – State-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[3] X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[4] K. L. Burns and E. G. Beier, "Significance of vocal and visual channels in the decoding of emotional meaning," *Journal of Communication*, vol. 23, no. 1, pp. 118–130, 1973. [Online]. Available: http://dx.doi.org/10.1111/j.1460-2466.1973.tb00936.x

[5] A. Cutler, D. Dahan, and W. Van Donselaar, "Prosody in the comprehension of spoken language: A literature review," *Language and speech*, vol. 40, no. 2, pp. 141–201, 1997.

[6] G. Fairbanks and W. Pronovost, "An experimental study of the pitch characteristics of the voice during the expression of emotion," *Communications Monographs*, vol. 6, no. 1, pp. 87–104, 1939.

[7] J. Crumpton and C. L. Bethel, "A survey of using vocal prosody to convey emotion in robot speech," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 271–285, 2016. [Online]. Available: http://dx.doi.org/10.1007/s12369-015-0329-4

[8] M. M. Kjelgaard and S. R. Speer, "Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity," *Journal of Memory and Language*, vol. 40, no. 2, pp. 153–194, 1999.

[9] R. L. Mitchell, "Does incongruence of lexicosemantic and prosodic information cause discernible cognitive conflict?" *Cognitive, Affective, & Behavioral Neuroscience*, vol. 6, no. 4, pp. 298–305, 2006.

[10] M. Wittfoth, C. Schrder, D. M. Schardt, R. Dengler, H.-J. Heinze, and S. A. Kotz, "On emotional conflict: Interference resolution of happy and angry prosody reveals valence-specific effects," *Cerebral Cortex*, vol. 20, no. 2, p. 383, 2009. [Online]. Available: + http://dx.doi.org/10.1093/cercor/bhp106

[11] C. Nass, I.-M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama, "Improving automotive safety by pairing driver emotion and car voice emotion," in *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '05. New York, NY, USA: ACM, 2005, pp. 1973–1976.

[12] S. D'mello and A. Graesser, "Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 4, pp. 23:1–23:39, Jan. 2013.

[13] S. M. Smith and D. R. Shaffer, "Celerity and cajolery: Rapid speech may promote or inhibit persuasion through its impact on message elaboration," *Personality and Social Psychology Bulletin*, vol. 17, no. 6, pp. 663–669, 1991. [Online]. Available: http://dx.doi.org/10.1177/0146167291176009

[14] 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.

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

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

[17] 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.

[18] 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.

[19] B. Litwak, personal communication, 2016-08-05.

[20] R. Levitan, Š. Benuš, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar," *Interspeech 2016*, pp. 1166–1170, 2016.

[21] L. Violante, P. R. Zivic, and A. Gravano, "Improving speech synthesis quality by reducing pitch peaks in the source recordings." in *HLT-NAACL*, 2013, pp. 502–506.

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

[23] C. Jones, L. Berry, and C. Stevens, "Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners," *Computer Speech & Language*, vol. 21, no. 4, pp. 641–651, 2007.

[24] A. Heathcote, S. J. Popiel, and D. Mewhort, "Analysis of response time distributions: An example using the stroop task." *Psychological Bulletin*, vol. 109, no. 2, p. 340, 1991.

[25] J. M. Pérez, R. H. Gálvez, and A. Gravano, "Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement," *Interspeech 2016*, pp. 1270–1274, 2016.

[26] 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.

[27] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception." *Psychological bulletin*, vol. 129, no. 1, p. 74, 2003.

[28] J. T. Cheng, J. L. Tracy, S. Ho, and J. Henrich, "Listen, follow me: Dynamic vocal signals of dominance predict emergent social rank in humans." *Journal of Experimental Psychology: General*, vol. 145, no. 5, p. 536, 2016.

[29] C. A. Klofstad, R. C. Anderson, and S. Peters, "Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 279, no. 1738, pp. 2698–2704, 2012.

[30] P. Rockwell, D. B. Buller, and J. K. Burgoon, "The voice of deceit: Refining and expanding vocal cues to deception," *Communication Research Reports*, vol. 14, no. 4, pp. 451–459, 1997.

[31] J. J. Ohala, "Cross-language use of pitch: an ethological view," *Phonetica*, vol. 40, no. 1, pp. 1–18, 1983.

[32] J. Hirschberg, "The pragmatics of intonational meaning," in *Proceedings of Speech Prosody*, 2002.

[33] C. Gussenhoven, "Intonation and interpretation: Phonetics and phonology," in *Proceedings of Speech Prosody*, 2002.