INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Phonetic Entrainment in Cooperative Dialogues: A Case of Russian

Alla Menshikova, Daniil Kocharov, Tatiana Kachkovskaia

Saint Petersburg State University, Saint Petersburg, Russia
menshikova.alla2016@yandex.ru, kocharov@phonetics.pu.ru, kachkovskaia@phonetics.pu.ru

Abstract

It has been shown for a number of languages that speakers accommodate to each other in conversation. Such accommodation, or entrainment, reveals itself in many modalities including speech: interlocutors are found to entrain in intensity, fundamental frequency, tempo and other acoustic features. This paper presents data on speech entrainment in Russian using the standard measures for speech entrainment: proximity, convergence and synchrony. The research uses 49 dialogues from the SibLing speech corpus where speakers played a card-matching game. The list of acoustic features includes various measures of pitch, energy, spectral slope, HNR, jitter, and shimmer. The results for Russian are compared with those published previously for other languages.

Index Terms: phonetics, dialogue, speech entrainment, card-matching game, Russian

1. Introduction

In recent years a lot has been published on speech entrainment—the phenomenon of speakers' adaptation to each other in conversation. The published results encompass a significant number of languages. Among them are Chinese [1], English [2, 3, 4, 5, 6], French [7], German [8, 9], Italian [10], Japanese [11, 12], Polish [13], Portuguese [14], Slovak [15], Spanish [16], Swedish [17] and others. For the scientific community it is crucial to collect data for other languages as well: the more languages are on this list, the more proof we get of the universal nature of speech accommodation, and, on the other hand, the more language-specific traits we describe.

Most of the research on speech entrainment is based on dialogues where interlocutors are solving a common task, as the cooperative nature of such conversations provides ground for hypothesizing that we may expect accommodation. Among the most frequently used tasks are searching for similar objects on the pictures (card-matching game) or arranging pictures in a specific order [18, 15, 1, 16, 19] and describing a route drawn on a map (map task) [3, 20, 5, 21, 14, 22].

So far, Russian speech material has never been used for research in the field of speech entrainment. This is why we recorded a new corpus of Russian dialogue speech [23], where both the card-matching game and the map task were used. However, map tasks with their well-established conversational roles (information giver vs. information receiver) require a more complicated analysis, as interlocutors speak differently depending on their role. This is why for this research we used the recordings of the first type of task, the card-matching game.

Different research groups measure entrainment in different ways and analyse different sets of features. In order to obtain results comparable to at least some of the languages, we used the metrics for entrainment measurement—proximity, convergence and synchrony—described in [24] where data for English, Slovak, Chinese and Spanish are provided. We also suggested some modifications that could help iron out the disputable issues concerning these measurement methods.

In the cross-language analysis presented in [24], the authors came to a number of generalizations. First, synchrony occurs more frequently than proximity. Second, negative synchrony occurs more often than positive synchrony. Third, the features that manifest entrainment can be language-specific (as an example, the authors provide Chinese, where F_0 parameters play a different role, probably because it is a tone language). And last but not least there is much variability across speakers and dialogues. The aim of this paper is to find out whether these observations are also true for Russian or not, and whether Russian speakers show some language-specific behaviour.

2. Method

2.1. Material

The results presented in this paper are based on 49 dialogues from the newly created corpus of Russian dialogue speech SibLing which is now at the final stages of development [23]. The corpus contains recordings of collaborative dialogues where interlocutors play a card-matching game and a map task. The basic set of speakers consists of 10 pairs of same-gender siblings (including 4 pairs of identical twins) aged 23–40. Each of the siblings participated in five dialogue sessions: with the other sibling, with his/her close friend of the same gender and similar age, with a stranger of the same gender and similar age, with a stranger of similar age and different gender, with a stranger of a higher job position, same gender and greater age.

In the card-matching game the interlocutors were searching for similarities in two decks of ten cards. The speakers took turns to describe their picture, but in most cases the role differences (information giver vs. information follower) almost disappeared after the first switch. The map task was a classical task in which the interlocutors explained to each other routes drawn on schematic maps.

Our analysis of this data showed that in map task most speakers' acoustic features differ significantly across their roles in the dialogue (information giver vs. information follower), but card games are more role-neutral. This is why at this stage we limited our analysis by card games only, leaving the map task recordings for our future research.

The total amount of the analyzed material is over 10 hours. The dialogues were recorded in a soundproof studio. The interlocutors were separated by a non-transparent screen to prevent them from seeing each other.

2.2. Acoustic features

A set of 16 acoustic features was used to explore speech entrainment, including those related to pitch, energy, spectral slope and voice quality. The features are listed in Table 1.

This set of acoustic features was calculated for each IPU (inter-pausal unit). F_0, jitter, and shimmer were calculated over

http://dx.doi.org/10.21437/Interspeech.2020-2696

Table 1: *A list of acoustic features used to explore phonetic entrainment.*

Features	Measures
F_0	mean, max, min, median, range, standard deviation
energy	mean, max, median, standard deviation
harmonics-to-noise ratio (HNR)	mean, max
spectral slope	mean, max
jitter	mean
shimmer	mean

the whole IPU. Energy, HNR and spectral slope were extracted frame-wise using non-overlapping 10 ms frames. The acoustic analysis was performed by means of our own tools implemented in Python 3 using librosa library to extract spectral features [25].

F_0 values were extracted using the REAPER algorithm [26]. At the next step we eliminated microprosody and added new F_0 values for the voiceless parts of the signal using linear interpolation. Then the contour was smoothed using Savitzky-Golay filtering with a third order polynomial in 5-sample windows [27]. The smoothed F_0 values were converted into semitones with the reference frequency of 100 Hz.

Energy was calculated as the arithmetic mean of squared amplitude values and was converted into dB scale.

Harmonics-to-Noise ratio (HNR) was calculated as defined in [28] as the log-ratio of the relative power of the periodic component of the signal and the relative power of the noise component. The periodic component was calculated using the autocorrelation function (AC) at the lag corresponding to the F_0 period, while the noise component was calculated as the difference between the 0^{th} AC coefficient and the periodic component estimate.

Spectral slope was extracted by the procedure described in [29] as the ratio of the strongest spectral energy peak in the 0–2 kHz region to the strongest spectral energy peak in the 2–5 kHz region. The spectrum was calculated by means of FFT using Hann window.

Jitter and shimmer values were calculated using extracted F_0 periods by the procedure described in [30].

Feature values for turns were calculated as arithmetic means across all IPUs within the turn.

2.3. Measuring speech entrainment

In this research we calculated three common metrics of phonetic entrainment: proximity, convergence and synchrony (see, e.g. [24]). T-tests and Pearson's correlation were calculated by means of Python's SciPy library [31].

For the turn-level (local) **proximity** the difference between two speakers' adjacent IPUs was compared with the averaged absolute values of the differences between feature values in the first IPU of the given speaker and the values of the given feature in the last IPUs of 50 % of the other speaker's turns chosen randomly [1]. The random choice of turns to compare with was motivated by [24]. However, our experiments showed that such proximity estimate varied significantly from one round of calculations to the other. Examples of such inconsistency are given in Table 2, where we present proximity values calculated thrice.

[1] In our material the average amount of turn exchanges in a dialogue is 76

Table 2: *Three rounds of proximity calculations with averages based on a random set of turns (for 7 dialogues of 49). Empty cell: no significant proximity in all rounds; '+': significant positive proximity; '-': significant negative proximity; '.': no significant proximity (e.g. '- . .' for max energy in dialogue d11 should read as "negative proximity was found in round 1 and no proximity in rounds 2 and 3". Significance threshold is 0.05.*

Features	Dialogues						
	d11	d12	d13	d14	d16	d17	d18
E max	- . .				. . -	. - .	- . -
E mean	- . .	- - .		. . -	- - -		. . -
F0 max	- - -		- . .			..-	- -.
F0 mean	- . -	.-.	- - -			-..	+++
HNR max	. - -						- - -
HNR mean				- . +		- - -	
Sp.slope max	- . -			- . .		+ . +	- - -
Sp.slope mean		+ . .		. - .	. . +	+ + .	
jitter		. - -			- - -	- - -	- - -
shimmer		. - +			- - .		- - -

As a result, we explored two modified versions of this method. First, we tried to use not 50 % of the interlocutor's turns but *all* of them (except for the current turn). Second, we ran the original method three times, and as significant we considered only those dialogues where all three rounds showed consistent proximity. The difference between all the measures are presented in Table 4.

There are several ways to measure phonetic **convergence**. Probably the easiest way is to divide the recording into two halves and compare the average values for a feature between these parts [5]. Another similar approach is to divide the recording into three parts and compare the 1st and the 3rd ones [32]. Then, data across all the dialogues is summarized, and for each feature, convergence is estimated using a paired t-test. These methods estimate global (session-level) convergence.

A fundamentally different approach which does not rely on averages is to estimate local (turn-level) convergence as in [24]: at each turn exchange we measure the between-speaker difference in feature values in the adjacent IPUs (i.e. the last IPU of speaker A's turn and the first IPU of speaker B's turn), and then calculate Pearson's correlation between these values and time. If the correlation coefficient is negative, the between-speaker difference for the feature diminishes with time, which means convergence; positive correlation means divergence; no correlation means either maintenance or synchrony.

We also tried an alternative way of calculating convergence, comparing adjacent *turns* instead of IPUs. The comparison is presented in Table 4. Overly short turns were filtered out in order to get rid of backchannels, which are known to have significantly different prosodic features, mostly—duration. Duration threshold was taken as the mean turn duration minus standard deviation calculated across the speaker's data in the dialogue.

Synchrony was calculated as suggested in [24]: as Pearson's correlation between adjacent IPUs in the course of the dialogue. If the correlation coefficient is positive, the speakers "mimic" each other's speech features, which means positive synchrony; negative correlation means negative synchrony (asynchrony).

3. Results and discussion

The detailed information about turn-level proximity, convergence and synchrony is presented in Table 3. The detailed in-

Table 3: *Positive and negative entrainment—proximity, convergence and synchrony—of the analyzed features in 49 dialogues. '+': significant positive entrainment; '-': significant negative entrainment; '.': no significant entrainment; empty cell: no significant entrainment by any measure. The markers are given in the following order: proximity, convergence, synchrony (e.g. '.-+' for mean energy in dialogue d03 should read as "no statistically significant proximity, significant negative convergence, and significant positive synchrony"). Significance threshold is 0.05.*

Features	Dialogues																								
	d01	d02	d03	d04	d06	d07	d08	d11	d12	d13	d14	d16	d17	d18	d21	d22	d24	d26	d27	d28	d31	d32	d36	d41	d42
E max		. + .	. - .			. . +	. . +		. . +				. . +	. + .	. - .		- - .								. - .
E mean		. + .	. - +	. . +		. + +		- - .	- . .			- . .		. + .	. - +	. . +	- - +	. + .		. - .	- . .				. - .
E median		. + .		. . +		. + .		- - .	- . .		- . .	- . .		. + .	. - .		- - +	+ + .	- . .	. - .					. - .
E std		. + .	. + .		+ . +	. + +	+ . +	. + .	+ . +			+ . .		+ . -		. . +								- . .		
F0 max			. . +				. . +	- + .				- . .					. - .		- . .	. . +			- . .		
F0 min				- . .		- . .	+ . .		- . .	- . .	- . .			- . .	- . .	- . .		- . .						- . .		- . .
F0 mean						- . .		- . .	- . .	- . +				+ . .	. . +	- . .		- . .		- . .					- . .	
F0 std					- - .	- . .		. . +	- . .	- . .						- . .	- . .		. . +						- - .
F0 median		. - .	. - .			- . .		- . .	- . .	. . +			- . .		. . +	- . +	- . .	- . .	- . .	- . .				- . .	
F0 range	- . .	. + .		- . .	- - .	- . .	. . +		- . .	- . .			- . .	- . .	- . .	- . .	- . .	- . .		. + .	- . .	- . .			- . .
HNR max						. . +		- . .	. . +	. - .		. . +		- . .		- . .				- . .				- . .	- . .
HNR mean	- . .	. - .	. . +			- . .	. . +			. . +	. . -		- . .		. . +					- . .	- . .		- . .		. - .
sp. slope max			- . .	- . .	- . .								. - .	- . .		- . .	- . .	- . .			- . .	- . .			- . .
sp. slope mean			. . +												+ . .	- . .						- - .		- . .	. . +
jitter		- . .			. - +				- . .	. . +		- . .	- . .		. . +	- - .	- . .	. . +				- + .			
shimmer		- . .		. + .	. . +	- . .				. . +			- . .	- . .	- . .	. - .			- . .	- . .					. - .
	d43	d44	d46	d47	d48	d51	d52	d53	d54	d56	d57	d58	d61	d62	d63	d66	d67	d71	d72	d74	d76	d91	d92	d96	
E max		- . .					. . +			- . -	. . +	- . .	. . +	. . +				- . .	- . -					. + .	
E mean	. . +	- . .	. . +				. - +	. . +					. . +	. . +		. - .							- . .	- . .	
E median		- . .					. . +	. . +					. . +			- . .			- . .					- . .	
E std	- . .	- . .					- . .		- . .	- . .				- . .		. - .		. - .	- . .			- + .			
F0 max					- . .			. . +	. . -	- . .			- . .	- . .				- . .				- . .		- . .	
F0 min	- . .			- . .	- . .		- . .	. - .	+ . .				- . .		- . .	- . .		- . .	- . .	- . .		- . .	- . .		
F0 mean					- . .	- . .							- . -				- . .	- + .					- . .		
F0 std	- . .			- . .	- . .	- . .							- . .					- . .	- . .			- . .	- . .		
F0 median								- . .					- . -			. - .		- + .				- . .			
F0 range	- . .			- . .	- . .								- . .		- . .	- . .		- . .	- . .	- . .		- . .	- . .		
HNR max			- . .	. . +				- . .		. . -								- . .		- . .	. + .		. - .		
HNR mean	- . .			- . .			- . .			- . .		. . +						- . .							
sp. slope max		- . .			- . .	- . .	- . .					- . .	. . +	- . .		- . .	. - .					- . .			
sp. slope mean		- . .				- . .	- . .	- . .	- . .			- + .	+ . .	- . .	. - .	- - .	- . .	- . .		- . .					
jitter				- . .	. + +					- . .			- . .				- - .	- . .						. . +		
shimmer		- . +	- . .	- . .		- . .	- - .	+ + +				- . .	- . .		- + .	. . +	. . +								

formation includes results obtained for all the features in all the processed dialogues. One can see that the entrainment strategies differ much in the analyzed material. E.g. maximum energy in dialogue d02 shows positive convergence and no proximity or synchrony ('. + .'). There are cases when different measures show opposite tendencies, e.g. jitter in dialogue d06 shows negative convergence and positive synchrony ('. - +'). The summary over all dialogues is given in Table 4.

3.1. Proximity

As mentioned above, the original method for estimating local (turn-level) proximity showed inconsistent results due to random choice of the set of interlocutor's turns to compare with. Table 2 presents proximity data for 10 features calculated three times on a set of 7 dialogues. One can see that the values are equal in all rounds only in 44 cells (of 70): 33 cases of 'no significant proximity' (empty cells), 10 cases of consistent 'negative proximity' ('- - -') and 1 case of 'positive proximity' ('+ + +'). The other cases are examples of changes in proximity decisions. There are cases when one round showed positive proximity and another one showed negative proximity (see, e.g., results for HNR mean in dialogue d14). Because of this, we only counted the cells with either '+ + +' or '- - -'. These results are presented in Table 4, column "random 50 % of IPUs". However, one may still argue about how to analyse cells with two similar significant values and one non-significant, such as '- - .', as well as how many rounds should be run in order to get consistent results.

The results for the alternative measurement method, when the set of interlocutor's turns is not random, are presented in Table 4, column "all IPUs". We chose this approach as the main method to calculate proximity, and its results for all the dialogues and all the features are presented in Table 3.

Looking at the data presented in the summary (see Table 4) one may see that the two measurement methods for proximity are not much contradictory, as significance tends to appear for roughly the same features. In both cases we conclude that negative proximity prevails and can manifest itself in all acoustic features.

In terms of proximity, more often speakers entrain in F_0 features. A similar result was obtained for Mandarin Chinese in [24], while Slovak and English do not show such tendency.

3.2. Convergence

Convergence enables us to understand whether two speakers become closer to each other in the course of conversation—in terms of the given acoustic feature. The experiments on global (session-level) convergence showed different results when dialogues were split into two parts and when dialogues were split into three parts. We found no convergence for any of the features in the former case. In the latter case convergence was found for spectral slope maximum only ($p < 0.05$, $t = 2.242$).

As mentioned above, local (turn-level) convergence was measured in two ways: with feature values calculated across the IPU (see Table 3 and Table 4, column "IPU") and across the whole turn (see Table 4, column "turn"). For both measurement methods, we observe prevalence of negative convergence, but cases of positive convergence are also found. Among the languages discussed in [24], only Slovak shows much negative convergence.

In terms of the acoustic features, there is more frequent negative convergence in features related to energy. This means that in terms of loudness speakers tend to choose the strategy of divergence, while standard deviation of energy shows more cases of positive convergence.

In general, the two measurement methods show similar ten-

Table 4: *Summary of results on local acoustic-prosodic entrainment as percentages of sessions with significant positive (+) and negative (–) entrainment type (proximity, synchrony, convergence). Convergence is presented calculated on both turns and IPUs. Proximity is presented calculated on both a set of random 50% of IPUs and all IPUs but the current. Grey colour shows entrainment of acoustic features within more than 10 % of sessions (5 sessions of 49).*

Features	Convergence (% of sessions)				Synchrony (% of sessions)		Proximity (% of sessions)			
	turn		IPU		IPU		all IPUs		random 50% of IPUs	
	positive	negative	positive	negative	positive	negative	positive	negative	positive	negative
E max	8	14	6	8	16	4		12	4	6
E mean	4	12	8	16	24			16	2	6
E median	4	12	8	10	10		2	20		16
E std	14	6	10	4	10	2	10	18	2	10
F0 max	2	12	2	2	8	2		22		10
F0 mean	2	6	2		4	2	2	29	4	12
F0 median	2	8	2	6	6	2		31		12
F0 min		8		2			4	47	4	37
F0 range		4	4	2	2			53		41
F0 std		8		4	4			33		24
HNR max	2	8	2	4	8	2		22		6
HNR mean	4	2		4	10	2		22		4
Sp.slope max	2	2		4	2			37		16
Sp.slope mean	4	4	2	6	4		4	29	2	12
jitter	4	6	4	6	12			24		16
shimmer	6	6	6	6	12		2	33	2	8

dencies. However, convergence calculated over the whole turn, as opposed to the IPU-based method, detects entrainment in maximum F_0 in more dialogues. A possible reason for this lies in the fact that some IPUs are too short to comprise the beginning of the declination trend where maximum F_0 can be detected.

3.3. Synchrony

Synchrony enables us to see whether there are local synchronous changes in acoustic features in the interlocutors' speech. In our material positive synchrony prevails and manifests itself mostly in features related to energy and voice quality (see Table 4). Among the languages discussed in [24], positive synchrony is observed for English and Slovak, but in those cases mostly for energy only.

4. Conclusions

In many dialogues we found phonetic manifestations of speech entrainment. Statistical analysis has proved that the way a speaker begins his/her turn indeed depends on the way the interlocutor ended his/her turn; this can be seen in various speech features—related to loudness, melody, voice quality. Entrainment within the dialogue (globally) does not occur in most cases. However, in certain dialogues we found evidence for synchrony (local "copying" of the acoustic features of the interlocutor; e.g., speaking louder after the interlocutor, then speaking quieter as he/she starts to speak quieter); this is mostly manifested in features related to loudness, more rarely—to voice quality.

Still, our data confirms the thesis of high variability across speakers and dialogues. Indeed, speakers show different entrainment strategies, including no entrainment—at least in terms of these features and these metrics. It is known that entrainment can manifest itself many other features—such as lexical and syntactic features, gestures and facial movements etc. The latter, however, we tried to eliminate here by putting a non-transparent screen between the speakers.

There is still much to be done to solve the issues concerning the measurement method. First, averaging within an IPU may add noise to the analysis for short IPUs. But even switching to turns instead of IPUs will not solve the problem, as turns may be short as well. Given that most of the short turns are backchannels, which differ prosodically form other turns, short turns should be completely excluded from this analysis; at the same time, backchannels themselves may manifest speech entrainment—which should be analyzed separately. Second, proximity measurement method can be improved in two ways, as we discussed in this paper: by either running the random-IPU-based analysis several times, or analyzing all of the interlocutor's turns.

In general, our data confirms the universality of speech entrainment. But when we compare our results to those for English, Slovak, Spanish and Chinese, we find that Russian is not similar to any of these languages. In terms of proximity, it resembles Chinese, in terms of synchrony—Standard American English, and in terms of convergence—more ore less, Slovak. But the results are still hugely variable between speakers. The acoustic features that manifest accommodation vary significantly as well, with the only cross-language generalization that the most prominent are energy-related features.

In the recent decades, a lot of research has been done on the social and individual factors influencing speech entrainment (see, e.g. [33]). This is why our next step is to include some of these factors in our analysis—the SibLing corpus that we used here enables to analyse the data with respect to gender, conversational roles, degree of speakers' familiarity, and social hierarchy.

5. Acknowledgements

This research is supported by the Russian Science Foundation (grant 19-78-10046).

6. References

[1] Z. Xia, R. Levitan, and J. Hirschberg, "Prosodic entrainment in Mandarin Chinese and English: A cross-linguistic comparison," in *Proceedings of the 7th Speech Prosody Conference*, 2014, pp. 65–69.

[2] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability," *Journal of Personality and Social Psychology*, vol. 32, no. 5, pp. 790–804, 1975.

[3] J. S. Pardo, "On phonetic convergence during conversational interaction," *Journal of Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[4] 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 *Proceedings of Interspeech*, 2010, pp. 793–796.

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

[6] M. Babel and D. Bulatov, "The role of fundamental frequency in phonetic accommodation," *Language and Speech*, vol. 55, pp. 231–248, 2012.

[7] G. Bailly and A. Lelong, "Speech dominoes and phonetic convergence," in *Proceedings of Interspeech*, 2010, pp. 1153–1156.

[8] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech," in *Proceedings of Interspeech*, 2013, pp. 525–529.

[9] M. Zellers, "Prosodic convergence with spoken stimuli in laboratory data," in *Proceedings of Interspeech*, 2016, pp. 1021–1025.

[10] M. Savino, L. Lapertosa, A. O. Caffò, and M. Refice, "Exploring prosodic convergence in Italian game dialogues," in *Proceedings of the 7th Tutorial and Research Workshop on Experimental Linguistics (Exling)*, 2016, pp. 151–154.

[11] N. Suzuki and Y. Katagiri, "Prosodic alignment in human-computer interaction," *Connection Science*, vol. 19, no. 4, pp. 131–141, 2007.

[12] T. Kawahara, T. Yamaguchi, M. Uesato, K. Yoshino, and K. Takanashi, "Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening," in *Proceedings of APSIPA Annual Summit and Conference*, 2015, pp. 392–395.

[13] M. Karpiński, K. Klessa, and A. Czoska, "Local and global convergence in the temporal domain in Polish task-oriented dialogue," in *Proc. of 7th International Conference on Speech Prosody*, 2014, pp. 743–747.

[14] V. Cabarrão, I. Trancoso, A. I. Mata, H. Moniz, and F. Batista, "Global analysis of entrainment in dialogues," in *IberSPEECH*, 2016, pp. 215–223.

[15] Š. Beňuš, R. Levitan, J. Hirschberg, A. Gravano, and S. Darjaa, "Entrainment in Slovak collaborative dialogues," in *5th IEEE International Conference on Cognitive Infocommunications*, 2014, pp. 1270–1274.

[16] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *Proceedings of AAAI 2015 Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, 2015, pp. 44–51.

[17] J. Edlund, M. Heldner, and J. Hirschberg, "Pause and gap length in face-to-face interaction," in *Proceedings of Interspeech*, 2009, pp. 2779–2782.

[18] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008, pp. 169–172.

[19] 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," in *Interspeech*, 2016, pp. 1270–1274.

[20] D. Reitter, J. D. Moore, and F. Keller, "Priming of syntactic rules in task-oriented dialogue and spontaneous conversation," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 2006, pp. 685–690.

[21] J. Pardo, I. C. Jay, R. Hoshino, S. M. Hasbun, C. Sowemimo-Coker, and R. M. Krauss, "Influence of role-switching on phonetic convergence in conversation," *Discourse Processes*, vol. 50, no. 4, pp. 276–300, 2013.

[22] V. Cabarrão, F. Batista, H. Moniz, I. Trancoso, and A. I. Mata, "Acoustic-prosodic entrainment in structural metadata events," in *Proceedings of Interspeech*, 2018, pp. 2176–2180.

[23] T. Kachkovskaia, T. Chukaeva, V. Evdokimova, P. Kholiavin, D. Kocharov, N. Kriakina, A. Mamushina, A. Menshikova, and S. Zimina, "SibLing corpus of Russian dialogue speech designed for research on speech entrainment," in *Proceeding of LREC (in press)*, 2020.

[24] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison," in *Proceedings of the SIGDIAL*, 2015, pp. 325–334.

[25] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proceedings of the 14th Python in Science Conference*, 2015.

[26] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR," https://github.com/google/REAPER, 2015.

[27] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[28] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proceedings of Institute of Phonetic Sciences*, vol. 17, pp. 97–110, 1993.

[29] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Oto-Laryngologica*, vol. 90, no. 5–6, pp. 441–451, 1980.

[30] F. Eyben, K. R. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[31] P. Virtanen, R. Gommers, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[32] J. Michalsky and H. Schoormann, "Pitch convergence as an effect of perceived attractiveness and likability," in *Proceedings of Interspeech*, 2017, pp. 2253–2256.

[33] N. Lewandowski and M. Jilka, "Phonetic convergence, language talent, personality and attention," *Frontiers in Communication*, no. 4, 2019.