INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Prosodic Accommodation in Face-to-face and Telephone Dialogues

Pavel Šturm[1], *Radek Skarnitzl*[1], *Tomáš Nechanský*[1]

[1]Institute of Phonetics, Faculty of Arts, Charles University, Czech Republic
{pavel.sturm}, {radek.skarnitzl}@ff.cuni.cz, tomas.nechansky@seznam.cz

Abstract

The study of phonetic accommodation in various communicative situations is still relatively limited. This paper examines accommodation in spontaneous conversations of eight pairs of Czech young male speakers in two communicative conditions: unconstrained face-to-face conversation and goal-oriented interaction via mobile telephone. Articulation rate and measures of f0 level, range and variability were measured in 40 prosodic phrases per speaker in each condition. Analyses of LME models did not reveal a significant global effect of time throughout the interaction on the distance between speakers (convergence) in any of the examined parameters, or that of preceding phrase value on the subsequent turn-initial value (synchrony). However, more consistent patterns were observed when speaker pairs were examined separately, revealing substantial individual variation on the one hand and non-linear effects on the other. This shows that aggregate analyses can be misleading in the study of phonetic accommodation and that speakers dynamically employ different strategies throughout natural conversations.

Index Terms: accommodation, entrainment, convergence, synchrony, prosody, Czech

1. Introduction

Vocal accommodation (VA), also known as *entrainment*, *alignment* or *convergence*, is a dynamic process in which speakers adapt to their communication partners in some aspects of their speech. This may occur as a result of automatic cerebral processes, or be governed by concerns such as developing social distance throughout an interaction, or by a combination of both [1]. According to Communication Accommodation Theory, interlocutors accommodate to their partner the more they wish to show a positive attitude toward the speaker, while phenomena like under-accommodation and divergence are typically received negatively [2]. VA has been observed especially when cooperation between participants is desirable [3], such as goal-oriented interactions when describing a map or playing a game [4], [5], [6], or in the speech of married couples engaging in marital therapy sessions [7]. Indeed, interactive spontaneous conversation settings, along with social motivation and focus on the task, appear to provide most naturalistic opportunities for VA [1].

It is possible to investigate VA both locally and globally, as speakers may converge at turn exchanges or globally over the course of the interaction. *Convergence* can be conceived as a gradual diminishing of between-talker differences over time, and *synchrony* as a dynamic response pattern with speakers tracking each other closely as they move on.

VA is manifested in various domains. First, speakers may coordinate temporal cues (e.g., speech rate or pause duration). For instance, [8] investigated speech rate adjustments using an ABAB paradigm and found out that participants reduced their speech rate when engaged in a conversation with a "slow" experimenter. In contrast, [9] examined game corpora and did not find consistent convergence in articulation rate (AR), only a weak correlation capturing synchrony at the turn-level. Pitch-related acoustic features seem to yield equally varied findings. In [9], the authors extracted mean and maximum f_0, and although these parameters converged throughout the session, the effect was very small, with synchrony (continuous local adjustments) being more prevalent. In [10], which investigated f_0 mean, median, maximum, and two range parameters in speed-dating settings, the speakers showed significant levels of both global convergence (smaller distances in the end section than the start section) and local convergence (correlation between speaker differences and time). A correlation between the degree of f_0 accommodation and success of the participants in a cooperative card game was demonstrated in [5]. Importantly, it was shown that individual speaker pairs followed different strategies, e.g., synchrony, divergence or a nonlinear pattern over the course of the interaction. Finally, VA is observable in many other linguistic cues, both vocal [11] and non-vocal [12], [13]. With respect to acoustic parameters, speakers have been found to accommodate in intensity [9], [14]; voice quality [9], [15], [16]; VOT [17]; or vowel formants [18], [19]. However, disparate results were found when more measures, languages, types of accommodation, or prosodic features (e.g., phrase and pitch accent characteristics) were compared [6], [16].

Given these findings, our study focuses on interactions in conversational speech and in goal-oriented settings of a picture task. In addition, the picture task was conducted over the phone; given the absence of visual cues, such a setup may prompt a higher need for cooperation than the face-to-face condition.

2. Method

2.1. Speakers and recording conditions

We investigated spontaneous conversations between eight pairs of speakers of Common Czech, the nonstandard, supraregional variety of the Czech Republic. They expressed themselves to be friends from school. All the 16 speakers are male (to avoid social dominance phenomena [20]), aged 18 to 20 years, studying at a secondary vocational school in Central Bohemia.

The speakers were recorded in two conditions: face-to-face (F2F) and on-the-phone (MOB) interaction. The experimenter was absent to obtain truly spontaneous speech. In the F2F condition, the subjects were free to talk about whichever topic, the conversations lasted at least 12 minutes and were recorded using the Zoom H1n recorder. In the MOB condition, the speakers were presented with pictures differing in five small details, which they were supposed to locate and characterise (see [21]). The telephone conversations were captured using a call-recording app in the experimenter's telephone and lasted at least 7 minutes. In addition to the task difference, the MOB

http://dx.doi.org/10.21437/Interspeech.2021-130

condition involved the effect of telephone transmission [22] and the absence of the visual component in the interaction [15].

To examine the effect of VA and the effect of the recording condition, the order in which the two conditions were obtained was balanced: four pairs started with the F2F conversation, and four with the MOB condition taking place first.

2.2. Material

Given the spontaneous nature of the conversations, some dialogues were highly asymmetrical, with one speaker uttering long stretches of speech with only minor contributions by the partner (backchannels, short utterances). Inspired by [23], we excluded phrases shorter than four syllables. Forty prosodic phrases (range 4–16 syllables, mean 6.7 syllables) were identified auditorily for each speaker in each condition, yielding 160 phrases per each pair (40 phrases × 2 conditions × 2 speakers). In some telephone dialogues, it was impossible to identify 40 phrases of sufficient length; at the same time, we did not regard lowering the limit to three syllables per phrase as beneficial. The complete dataset is thus based on 1,260 prosodic phrases (160 phrases per each pair minus 20).

A standard procedure is to focus on units (here, prosodic phrases) in the vicinity of turn exchanges [9], [14], [16]. We followed a slightly modified procedure used in [23], with one speaker's prosodic phrase included for analysis only when it follows the other speaker's phrase. This may occur as a "true" turn exchange, or as a "remote" one with a short phrase (< 4 syllables) intervening; we will not differentiate between the two and designate them as "turn exchanges". In effect, turn-medial prosodic phrases and phrases at the beginning or end of the conversation are omitted. Since two speaker pairs yielded a very small number of prosodic phrases around turn exchanges (especially in the F2F condition), they have been excluded, resulting in 12 speakers for analysis. Global analyses are thus based on 948 phrases and local analyses on 316 phrase pairs at turn exchanges.

2.3. Acoustic measures

We examine six prosodic features in this study: *AR* – articulation rate (syll/s); *mean* – mean f_0 (semitones (ST) re 1 Hz); *median* – median f_0 (ST re 1 Hz); *SD* – f_0 standard deviation (ST); *range* – 80-percentile range (ST), which is the difference between the 90th and 10th percentile of f_0 values; *CSI* – cumulative slope index (ST/syll), the sum of absolute frequency differences between subsequent pitch points divided by the number of syllables, thus taking into account possible multiple melodic movements in a phrase [24].

As for the temporal measures, AR was determined by counting the number of syllables in each phrase and dividing that by its duration. Melodic measures of central value and variability were extracted using Praat [25] from the range of 60 to 280 Hz and expressed in ST re 1 Hz. The cumulative slope index was obtained from smoothed and interpolated PitchTier objects in Praat, the remaining measures from "raw" Pitch objects. The most conspicuous extraction errors (most importantly, octave jumps) were manually corrected.

2.4. Analyses

Since accommodative behaviour is highly individual [5], [8], we present several analyses. First, we examine speaker profiles for the six dialogues, separated by condition. Various patterns of VA can be inferred from the visual data (the complete set is provided in the online supplementary materials). In the remaining analyses, the computations are restricted to prosodic phrases at turn exchanges. At each exchange, the turn-initial prosodic phrase is compared to the value of the previous turn-final one (see Section 2.2 for the definition of turns).

Convergence was examined by means of *speaker distance* measures. At each turn exchange, the distance between the two speakers was computed. Convergence was represented as Pearson correlation between speaker distance and time (i.e., prosodic phrase index/order within the interaction). If speakers converge, the distances should be increasingly smaller with time. This can be captured in an LME model by using PHRASE INDEX (proxy for time) as predictor. Statistical significance of this predictor would indicate convergence (if negative). **Synchrony** between speakers was examined by plotting the values of speaker A with those of speaker B. If a significant correlation is found, then an increase in speaker A's parameters is complemented with an increase in speaker B, and vice versa. This can be evaluated in an LME model by using PRECEDING PHRASE VALUE as a predictor. If it reaches significance, then turn-initial parameter values can be partially predicted from the values in turn-final phrases.

Statistical analyses were performed in *R* [26] using the *lme4* package [27]. The models were fitted with the maximal random effect structure that still allowed convergence, namely including random slopes for CONDITION. The basic model structure is given in (1) for convergence and (2) for synchrony:

$$speaker\ distance \sim phrase\ index + condition + session + (1+condition|speaker\ pair) \quad (1)$$

$$turn\text{-}initial\ value \sim preceding\ phrase\ value + condition + session + (1+condition|speaker\ pair) + (1+condition|speaker) \quad (2)$$

In these, CONDITION (F2F × MOB) and SESSION (session 1 × session 2) were factor variables, whereas PHRASE INDEX and PRECEDING PHRASE VALUE were continuous. As dialogues differed in the number of turn exchanges, PHRASE INDEX was centred on the median to account for different lengths of the vector. Factors included in the model are theoretically motivated and not determined by a model selection process. Statistical evaluation of the predictors was done by comparing the basic model given above to a reduced model lacking the fixed effect in question, using likelihood ratio tests [28]. Interactions between the predictors were evaluated in a similar way. The significance level was set to $\alpha = 0.05$. Graphical outputs were generated using the *ggplot2* package [29].

3. Results

3.1. Speaker profiles

Henceforward, the two speakers in each dialogue and figure will be referred to as *blue* and *red* (the colour was assigned randomly). Figure 1 (and 1_Rate online) shows an illustration of how AR develops in selected speaker pairs in selected conditions (F2F or MOB). The individual data points correspond to the mean AR in each phrase. The data points are also fitted with local polynomial regression [30] to better illustrate overall trends in VA (thick lines). Not surprisingly, AR was quite variable, both between and within speakers, but some trends emerge. The MOB condition of Pair 2 indicates global divergence (with the blue speaker faster by nearly 2 syll/s at conversation end), as well as locally-based synchronous behaviour throughout most of the dialogue. Another example of

synchrony may be observed in the MOB condition of Pair 5, with the nonlinear fits essentially overlapping. In most pairs, there are relatively large oscillations, with the regression curves diverging and again converging, but interesting traces of what may be considered accommodation in AR appear more locally, as seen in the synchronous development at the onset of both conditions of Pair 6.

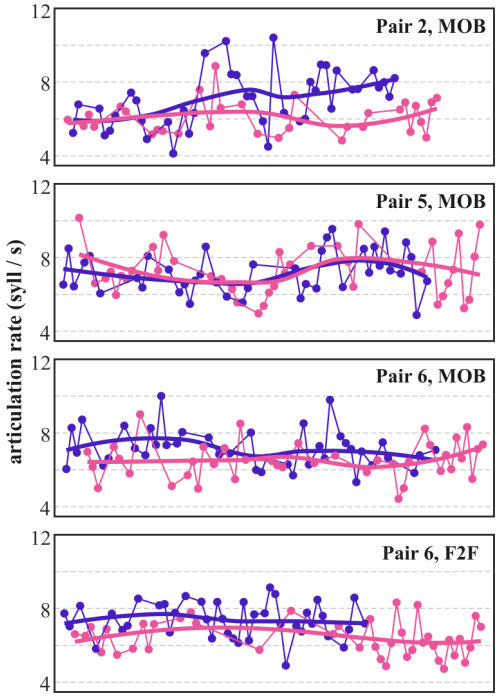

Figure 1: *Articulation rate in individual phrases in selected dialogues.*

The development of the central value of f_0, expressed using the median, is shown in Figure 2. The data point to the importance of a speaker's habitual pitch: in nearly all the speakers analysed here, the median f_0 in the prosodic phrases oscillates around the same value, as illustrated by Pairs 3 and 6 in the MOB condition. The only case which may to some degree be regarded as global convergence in melodic level occurs in Pair 5 (F2F).

There does not seem to be much happening in terms of VA in melodic variability; some synchronous tendencies may be observed in the central portion of the MOB condition of Pair 2 or 7. The supplementary materials feature the complete set of f_0-related figures (2_med, 2_mean, 2_CSI, 2_perc, 2_SD).

3.2. Accommodation at turn exchanges

We will first consider **convergence**. If two speakers converge over the course of the interaction, we should observe a negative correlation between time and speaker distance at a given turn exchange. Our data do not provide a single pattern; speaker pairs differ even within the same condition. Some dialogues did yield a negative slope (e.g., Pair 5 in the F2F condition for both melodic measures), in other dialogues this predicted relationship was reversed, and speakers became more distant as the interaction unfolded (e.g., Pair 2 in the MOB condition, especially for AR). Pearson correlation coefficients (3_tab1 online) showed that convergence reached significance only rarely. Few significant cases of local convergence were also confirmed by LME models which were constructed to verify any generalizable accommodation effects in our data. The crucial predictors of speaker distance were PROSODIC PHRASE INDEX (reflecting time within a condition) and SESSION (reflecting the order of conditions, and thus global time).

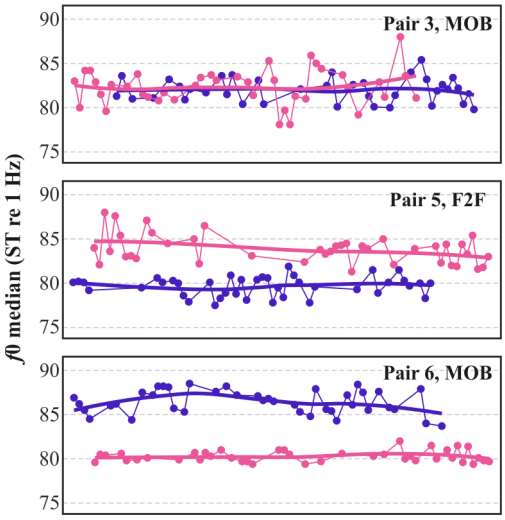

Figure 2: *Median of f0 in individual phrases in selected dialogues.*

The lack of consistent convergence across speakers over time can be attributable to an alternative strategy: speakers might be aligned in a dynamic, **synchronous** manner. It is therefore necessary to examine how measures develop over time even if the distance between speakers remains practically the same. Figure 3 correlates a given measure of one speaker (henceforth Speaker A) with that of his partner (Speaker B). Although the two speakers may have their own distribution of values, if they do entrain into synchrony, a positive slope should be observed. AR is particularly prone to be synchronized between the speakers, as only the MOB condition of Pair 7 yielded a case of negative synchrony (when Speaker A speaks faster, Speaker B slows down in response). The f_0 median and CSI measures (see all correlation plots online, 4_correl) behave differently in the two conditions, and are less consistent across speaker pairs.

LME modelling can test for synchrony by including a predictor specifying the value of the preceding prosodic phrase produced by the dialogue partner; statistical significance of the predictor would suggest synchrony. Parameters of the models are provided in the appendix online (5_append). With regard to AR, CONDITION turned out to be a significant predictor (χ^2 (1) = 6.38, p = 0.012); speakers spoke more slowly in the mobile condition. However, PRECEDING PHRASE VALUE did not reach significance (χ^2 (1) = 1.67, p = 0.196), although there was a trend for the turn-initial tempo to be positively correlated with the turn-final phrase articulated by the partner, and there was no significant interaction with either CONDITION (χ^2 (1) = 0.95, p = 0.330) or SESSION (χ^2 (1) = 0.06, p = 0.811).

With regard to the f_0 median measure, none of the predictors turned out to be significant on their own, and there were no significant two-way or three-way interactions. As for the f_0 mean measure, CONDITION was a significant predictor (χ^2 (1) = 4.32, p = 0.038); the speakers used higher pitch when talking over the phone, whereas PRECEDING VALUE and SESSION did not reach significance (χ^2 (1) = 2.34, p = 0.125; χ^2 (1) = 0.24, p = 0.627, respectively). Again, none of the interaction terms were significant ($p > 0.05$). The CSI measure yielded a significant

predictor of CONDITION (χ^2 (1) = 5.07, p = 0.024) but not of PRECEDING VALUE (χ^2 (1) = 0.26, p = 0.608) or SESSION (χ^2 (1) = 0.34, p = 0.561). Interactions between the factors were not significant (p > 0.05). We can thus only conclude that variability in f_0 was higher in the MOB condition.

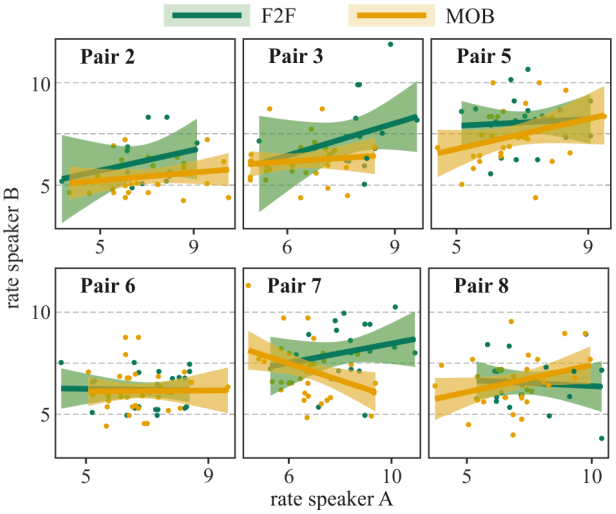

Figure 3: *Correlation of articulation rate between the two speakers in a pair throughout the dialogue. Positive slope = synchrony. Shaded areas = 95% CI.*

4. Discussion

This study explored vocal accommodation in six dyadic interactions. We focused on two types of VA: convergence and synchrony. Although there were some interesting tendencies, LME models did not reveal a significant effect of time on speaker distance (convergence), or that of preceding phrase value on the turn-initial value (synchrony). The only exception was the CSI measure, which showed significant divergence. Several explanations may be proposed for the absence of significant results in the other measures.

First, it should be noted that VA effects reported in literature vary widely in size. For instance, interactive game corpora were examined by [9], who found smaller distances between speakers in the second half of their conversations, but most measures (including those used here) did not reach significance. Similarly, their turn-level analysis yielded significant but very weak correlations in convergence (r < 0.01) and somewhat stronger in synchrony (r = 0.28 for f_0 mean, r = 0.15 for rate). Small effects can be found in many other studies [5], [8], [15], and the examined measures are affected by other factors apart from VA. It should thus not be surprising to find only weak effects of accommodation in our data, especially given the fact that the Czech language is relatively inconspicuous in terms of prosody variation, being characterized by very small habitual f_0 range [31] and lexical stress with non-salient acoustic cues [32]. It is instrumental to compare the illustrations in the supplementary materials: for instance, the 80-percentile range remains below 4 ST for more than 70% of the prosodic phrases examined.

Secondly, it is possible that VA occurs very early in the conversation. In contrast to other studies, our speakers were friends, so it is plausible that their conversational behaviour might be considerably synchronized from the start, or that a few phrases would suffice for this to occur [23].

Another reason for the lack of straightforward VA effects in our statistical models is that speakers employ different strategies. Entrainment has been shown to be particular to speaker pairs, most likely reflecting factors such as gender, power, liking or personality [14]. Many other studies found considerable variation in the degree and types of accommodation across pairs. For instance, [5] present three types of speaker profiles, namely interactions that suggest divergence, synchrony, or a non-linear relationship. Speakers may alter their speech in various ways, often accommodating in only one of several examined measures, not necessarily the same [16]. The correlation plots and tables for our data (see the supplementary materials) make it clear that although some pairs indeed show no evidence of VA in one or both conditions, most speaker pairs seem to follow a strategy (convergence vs. divergence, synchrony vs. asynchrony); this strategy may further differ between the two conditions (see below). In short, when dialogues are examined separately, one can see more consistent patterns of accommodation over the course of interactions than in the aggregate analysis.

Moreover, since conversation is in principle a dynamic and non-linear process, we may assume that VA strength and strategies will not be constant during a conversation. This is evident from the speaker profiles presented in Figures 1 and 2, where the speakers often converge or diverge briefly in one part of the session (*cf.* also the non-linear patterns in [5]). Such fluctuations in the degree of accommodation may also be attributed to the fact that the speakers' degree of involvement does not remain constant over the course of a conversation [33].

What remains to be discussed is the effect of condition (F2F vs. MOB) on accommodative behaviour. Generally, speakers talked more slowly, in shorter phrases, at higher pitch and with higher variability in the MOB condition, which is in line with [22]. The correlations in Figure 3 between the members of a speaker pair do suggest a difference between the conditions. While the general trend for most pairs and most measures was that of synchrony, this is true especially for the F2F condition; in fact, significant positive correlations emerged only in the F2F condition, while all of the significantly negative correlations (indicating reversed synchrony) were associated with the MOB condition. As [34] investigated convergence at different levels of task difficulty and found that convergence may occur more in contexts of low cognitive load, the tendency toward higher synchrony in F2F conversations might be the product of lower cognitive load compared to the cooperative, problem-oriented diapix task in the MOB condition.

To conclude, our study corresponds to the findings reported by many studies, namely that synchrony and convergence should be treated separately. While our study suggests no evidence for global convergence and synchrony in the data, this can be attributed to the speakers adopting different local strategies. Moreover, profiles capturing the raw data revealed significant non-linearities in speaker distance over time, suggesting that conversations in natural conditions do not follow a linear path.

5. Acknowledgements

This study was supported from the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734) and by CUNI project Progres 4, "Language in the shiftings of time, space, and culture".

6. References

[1] E. H. Coles-Harris, "Perspectives on the motivations for phonetic convergence," *Lang. and Linguistics Compass*, vol. 11, no. 12, 2017.

[2] C. Gallois and H. Giles, "Communication Accommodation Theory," in *The International Encyclopedia of Language and Social Interaction*, K. Tracy, T. Sandel, and C. Ilie, Eds., Hoboken, NJ: John Wiley & Sons, 2015, pp. 1–18.

[3] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evol. Hum. Behav.*, vol. 34, no. 6, pp. 419–426, 2013.

[4] J. S. Pardo, "On phonetic convergence during conversational interaction," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2382–2393, 2006.

[5] O. Ibrahim, G. Skantze, S. Stoll, and V. Dellwo, "Fundamental frequency accommodation in multi-party human-robot game interactions: The effect of winning or losing," in *Proc. Interspeech 2019*, pp. 3980–3984.

[6] U. D. Reichel, Š. Beňuš, and K. Mády, "Entrainment profiles: Comparison by gender, role, and feature set," *Speech Commun.*, vol. 100, pp. 46–57, 2018.

[7] C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proc. Interspeech 2010*, pp. 793–796.

[8] D. Freud, R. Ezrati-Vinacour, and O. Amir, "Speech rate adjustment of adults during conversation," *J. Fluency Disord.*, vol. 57, pp. 1–10, 2018.

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

[10] J. Michalsky and H. Schoormann, "Pitch Convergence as an Effect of Perceived Attractiveness and Likability," in *Proc. Interspeech 2017*, pp. 2253–2256.

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

[12] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 22, no. 6, pp. 1482–1493, 1996.

[13] H. P. Branigan, M. J. Pickering, and A. A. Cleland, "Syntactic co-ordination in dialogue," *Cognition*, vol. 75, no. 2, pp. B13-25, 2000.

[14] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison," in *Proc. 16th Annu. Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 325–334, 2015.

[15] A. Schweitzer, W. Wokurek, and M. Pützer, "Convergence of harmonic voice quality parameters in spontaneous dialogues," in *Proc. ICPhS 2019*, pp. 363–367.

[16] A. Weise, S. I. Levitan, J. Hirschberg, and R. Levitan, "Individual differences in acoustic-prosodic entrainment in spoken dialogue," *Speech Commun.*, vol. 115, pp. 78–87, 2019.

[17] M. Bane, P. Graff, and M. Sonderegger, "Longitudinal phonetic variation in a closed system," in *Proc. Annu. Meeting of the Chicago Linguistic Soc.*, vol. 46, no. 1, pp. 43–58, 2009.

[18] J. S. Pardo, "Expressing Oneself in Conversational Interaction," in *Expressing Oneself/Expressing One's Self: Communication, Cognition, and Identity*, E. Morsella, Ed., Psychology Press/Taylor and Francis, 2010, pp. 183–196.

[19] J. S. Pardo, I. C. Jay, and R. M. Krauss, "Conversational role influences speech imitation," *Atten. Percept. Psychophys.*, vol. 72, no. 8, pp. 2254–2264, 2010.

[20] L. L. Namy, L. C. Nygaard, and D. Sauerteig, "Gender differences in vocal accommodation: The role of perception," *J. Lang. Soc. Psychol.*, vol. 21, no. 4, pp. 422–432, 2002.

[21] K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow, "The Wildcat Corpus of native- and foreign-accented English: communicative efficiency across conversational dyads with varying language alignment profiles," *Lang. and Speech*, vol. 53, no. 4, pp. 510–540, 2010.

[22] M. Jessen, "Forensic phonetics and the influence of speaking style on global measures of fundamental frequency," in *Formal Linguistics and Law*, G. Grewendorf and M. Rathert, Eds., Berlin: Mouton de Gruyter, 2009, pp. 115–139.

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

[24] R. Hruška and T. Bořil, "Temporal variability of fundamental frequency contours," *AUC Philologica*, vol. 3, pp. 35–44, 2017.

[25] *Praat: doing phonetics by computer.* (6.1.11), P. Boersma and D. Weenink. Accessed: 6th April 2020. [Computer program]. Available: http://www.praat.org/

[26] *R: A Language and Environment for Statistical Computing*. (3.6.3). R Core Team: R Foundation for Statistical Computing. Accessed: 1st September 2020. [Computer program]. Available: https://www.r-project.org/

[27] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–46, 2015.

[28] R. H. Baayen, *Analyzing linguistic data*. Cambridge: Cambridge University Press, 2008.

[29] H. Wickham, *Ggplot2: Elegant graphics for data analysis*. New York, NY: Springer, 2009.

[30] W. S. Cleveland, E. Grosse, and W. M. Shyu, "Local regression models," in *Statistical models in S*, J. M. Chambers and T. J. Hastie, Eds., Wadsworth and Brooks/Cole, 1992, pp. 309–396.

[31] J. Volín, K. Poesová, and L. Weingartová, "Speech melody properties in English, Czech and Czech English: Reference and interference," *Res. Lang.*, vol. 13, no. 1, pp. 107–123, 2015.

[32] R. Skarnitzl, "Fonetická realizace slovního přízvuku u delších slov v češtině," *Slovo a slovesnost*, vol. 79, pp. 199–216, 2018.

[33] C. De Looze and S. Rauzy, "Measuring Speakers' Similarity in Speech by Means of Prosodic Cues: Methods and Potential," in *Proc. Interspeech 2011*, pp. 1393–1396.

[34] J. Abel and M. Babel, "Cognitive load reduces perceived linguistic convergence between dyads," *Lang. Speech*, vol. 60, no. 3, pp. 479–502, 2017.