Entrainment in Slovak Collaborative Dialogues

Štefan Beňuš
Constantine the Philosopher University in Nitra, Institute of Informatics, Slovak Ac. of Sciences, Bratislava, Slovakia
sbenus@ukf.sk

Rivka Levitan, Julia Hirschberg
Columbia University New York, USA
{rlevitan,julia}@cs.columbia.edu

Agustín Gravano
University of Buenos Aires and CONICET, Buenos Aires, Argentina gravano@dc.uba.ar

Sakhia Darjaa
Institute of Informatics, Slovak Ac. of Sciences, Bratislava, Slovakia
darjaa@savba.sk

***Abstract*—Entrainment is a natural inclination of people who interact with each other to develop similar, matching, or synchronized forms of behavior. In spoken interactions, this was observed at many linguistic (syntactic structures, referring expressions, dialectal features) and para-linguistic (intensity, speech rate) levels. Numerous studies have shown that the degree of spoken entrainment correlates with several social conversational attributes such as task success, likability, attractiveness, power distribution, and others. This feature of human-human interactions promises great potential for improving the naturalness and social acceptability of voice applications in human-machine interactions. However, despite many findings of various forms of entrainment in various languages and cultures, there is very little work on analyzing the effect of language or culture on speech entrainment in comparable spoken corpora. This is a significant gap since many applications utilizing spoken interaction and entrainment will be deployed to be used by native and non-native speakers of many cultures and languages. A welcome exception filling this gap is a recent paper by Xia et al. (2014) comparing entrainment in Standard American English and Mandarin Chinese. In this paper we provide a first analysis of speech entrainment in Slovak based on comparable corpus to the ones analyzed in Xia et al. This work contributes to our understanding and future modeling of cross-cultural aspects of behavioral entrainment.**

Keywords—entrainment; Slovak; prosody; human-human dialogues; human-computer interaction

I. INTRODUCTION

Entrainment is a natural inclination of people who interact with each other visually, through spoken modality, or both, to develop similar, matching, or synchronized forms of behavior. It has been observed for example, in gestures and mannerisms [7] or postural sways [17]. In the spoken modality, which is the focus of this paper, entrainment has been documented at many linguistic levels, such as syntactic structures, referring expressions, or the usage of function words, and in para-linguistic aspects of speech, such as intensity, pitch range, or speech rate. A review of this research can be found in [2].

The importance of entrainment lies in its potential to tap into the development and constant negotiation of social relationships throughout spoken interactions. Multiple studies have shown that the degree of spoken entrainment correlates with positive social attributes such as likability, attractiveness, power distribution, task success, and others; see for example [2] or [11] for a recent reviews.

Entrainment has also proved to be a useful feature for modeling human-machine interaction and thus increasing potential domains of its application. For example, a successful deployment of entrainment in human-machine interaction has been shown for speech rate: human users tend to entrain to the speech rate of the machine, and thus the system might 'elicit' speech that better matches its current capabilities in speech recognition [1]. In this way, speech recognition errors of the system can be reduced and the overall quality of the human-machine interaction improved. Similarly, improved predictability of the speech recognition system stemming, for example from entrainment in the type of syntactic phrases, might improve the performance of the system. Finally, when a spoken tutoring system entrained to its users, thus increasing the cohesion of the interaction, learning gains surpassed those of un-entrained student-tutor interactions [16].

In this respect, entrainment is a relevant concept for Coginfocom research agenda since it is closely linked to its speechability domain [6] that attempts to link cognitive linguistics with verbal and non-verbal social communicative signals and explore the potential usability of these links in speech technologies.

Given that many applications in human-machine interaction will have to be implemented in different languages and for speakers of different cultures, it is vital to understand how social and interactional features such as entrainment are realized in various languages and cultures. However, despite many findings of various forms of entrainment in various languages and cultures (see for example [11] for a review), there is very little work on analyzing the effect of language or culture on speech entrainment in comparable spoken corpora. A welcome exception that starts filling this gap is a recent paper [18] comparing entrainment in Standard American English and Mandarin Chinese using a comparable speech corpora in style and pragmatic intentions if not in size; the Mandarin data were more extensive than the English ones.

The approach of [18] follows the line of research in [14] that construes entrainment as realized through one of several possible strategies – such as proximity, convergence, or synchrony – and further, as possibly occurring at a global level of conversation or locally in shorter units of analysis, for example at turn-exchanges. In this framework, [18] found

many commonalities in the realization of entrainment in Standard American English and Mandarin Chinese. This suggests that entrainment may be profitably utilized in applications across radically different languages. The main findings regarding global proximity and convergence of [18], slightly refined in [13], can be summarized in the following table.

TABLE I. GLOBAL PROXIMITY AND CONVERGENCE OF [18], "☑" REFERS TO OBSERVED ENTRAINMENT, "SAE" STANDS FOR STANDARD AMERICAN ENGLISH, AND "MC" FOR MANDARIN CHINESE

	Global entrainment			
	Proximity		**Convergence**	
Feature	**SAE**	**MC**	**SAE**	**MC**
F0_mean				
F0_max		☑	☑	
F0_min				
Int_mean	☑	☑		
Int_max	☑	☑		
Int_min				
Speaaking rate	☑	☑	☑	

Entrainment was found consistently for intensity (mean and max) and for speaking rate. Pitch maximum showed proximity in Mandarin only while SAE speakers were found to converge on F0 maximum and speaking rate during their sessions. Moreover, the robustness of entrainment (both the number of entraining features and the degree) were influenced by speakers' gender: mixed gender pairs showed the most robust entrainment while male pairs showed the weakest results.

The primary goal of this paper is to add breadth to this line of research by examining spoken entrainment in yet another language from a different language family: the west-Slavic language of Slovak. In this aspect, this work contributes to our understanding and future modeling of cross-cultural aspects of behavioral entrainment. Another goal of the study is to build a ground work in the description of general entrainment in this corpus so that previous or future studies of entrainment in more specific domains may be properly compared. For example, in a preliminary study, [3] investigated partial aspects of entrainment such as in the prosody of Slovak discourse marker '*no*' finding entrainment, absence of entrainment, as well as dis-entrainment. By providing this groundwork, the relationship between various domains of entrainment might be explored in more meaningful ways.

The paper starts exploring the global measures of proximity and convergence in a Slovak corpus of conversational dialogues, presents initial observations regarding the relationship between entrainment and gender, compares Slovak data to English and Mandarin, but leaves the local and more dynamic measures of entrainment for future work.

II. CORPUS AND METHODOLOGY

A. Sk-Games

Data for this study come from a subset of the Sk-Games corpus described in more detail for example in [4]. It is a corpus of collaborative dyadic conversations following the design of Columbia Games Corpus [10] in which interlocutors do not see each other and perform a series of communicative games requiring mutual spoken interaction. One of the players, referred to as Describer, describes the position of a target object in relation to various other objects on her screen. The task of the other player, Placer, is to position the target object on his screen in exactly the same location as it is on the Describer's screen. This scenario yields semi-spontaneous spoken interactions offering rich and very natural variability in spoken behavior. Speakers were recorded with head-mounted close-talking microphones to separate audio channels.

The audio signal was downsampled to 22050, manually transcribed, and the transcripts were automatically aligned to the signal using the SPHINX toolkit adjusted for Slovak [9], which forces the alignment of both words and individual phonemes. This forced alignment was then manually corrected.

The corpus is still under construction and the analyzed data in this paper are from nine sessions. There were 11 speakers and 7 of them participated in two sessions. Hence, one of the unique features of the corpus is that, for these seven speakers, we can compare their behavior in identical communicative situations when they are paired with a different interlocutor. These nine sessions make roughly 6.3 hours of speech, and consist of 35,729 words.

B. Measures and analysis

In this paper we are following the descriptions of measures for global entrainment proposed in [13] and employed also in [18]. Using Praat [5], we extracted standard prosodic features for pitch and intensity (mean, median, maximum, minimum, and standard deviation) from the entire session (or its half) by concatenating inter-pausal-units (IPUs) separately for each speaker.

The number of syllables was algorithmically estimated based on grapheme-to-phoneme conversion for Slovak [7] and utilizing the transparent syllable definition that the number of syllables equals the number of syllabic phonemes, which in Slovak correspond to all vowels, including diphthongs, and syllabic consonants /r/ and /l/. Using the alignment, grapheme to phoneme information, and syllable definitions, mean syllable and phoneme rates were calculated for the session and its two halves.

Entrainment was assessed as global similarity in these features (see Section I for discussion). First, following [13] and [18], for each speaker in a session we calculated the difference between the speaker and her partner, referred to as *partner difference*, and the mean difference between the speaker and all non-partners, *non-partner difference*. Speaker's non-partners are defined as the speakers with whom the current speaker did not interact in any session, excluding also the interlocutor of the current partner. All differences were converted to absolute values. Hence, entrainment is operationalized as smaller

partner difference than non-partner difference. We used paired-samples t-test to assess the significant difference between partner and non-partner samples.

We experimented with several ways of gender normalization for the raw extracted features. Especially the pitch features must be normalized due to physiological differences between the genders, First, all features were normalized by gender using z-scores calculated over the means of the session features: $z = (x - mean_{GENDER})/stdev_{GENDER}$ in which x is the feature value for the session (or its half), and $mean_{GENDER}$ and $stdev_{GENDER}$ corresponds to means and standard deviations over all session features for the speakers of the same gender. Second, we employed the strategy used in [13] and adjusted extracted raw average pitch values from female speakers to match the range of the male speakers. Since pitch range is approximately [75,500] for female speakers and [50,300] for male speakers, female values were transformed as follows: new_pitch_value = K * original_pitch_value + D with K, D such that 500K + D = 300, 75K + D = 50. Finally, the strategy used for the Mandarin Chinese corpus of [18] was used so that the set of non-partners, defined in the previous paragraph, was further restricted to non-partners of the same gender as the interlocutor and raw features were used for statistical comparisons.

Finally, another way of assessing how speaker's behavior changes as a factor of the identity of the interlocutor is to compare the degree of entrainment between the partners in a session and between the features extracted from the speech of the same speaker if she participated in two sessions. Hence, it is natural to expect that the speech of one speaker will be extremely similar and thus less different than when the speech of two interlocutors is compared. However, intensity features might show extreme entrainment in that speakers show more similarity to their partners than to themselves [13]. Seven of our eleven speakers in the corpus participated in two sessions, each time with a different partner. A paired t-test compared the differences between each speaker and her partner with the differences between each speaker and herself in another session. The adjustment of pitch features only (Option #2) was used for gender normalization in these tests.

III. RESULTS

Our initial results give a complex picture for entrainment in this corpus. The findings for similarity and convergence are summarized in Table I and are discussed below. We report the values from the most conservative normalization (option #1 in Section II B above: gender-based z-scores) and discuss results from other normalizations in the text.

For global proximity, i.e. testing whether a speaker is more similar to her partner than her non-partners over the entire session, the only feature showing significant entrainment is minimum intensity ($t[17] = -2.44$, $p = 0.019$); a tendency was observed for maximum intensity ($t[17] = -1.98$, $p = 0.066$); negative t-values show entrainment: smaller differences in the partner sample than in the non-partner sample. Similar results were obtained from other normalizations. In addition to Int_max and Int_min, Int_mean also showed significant entrainment in #2 normalization (pitch adjustment) and only Int_mean and Int-max were significant in #3 normalization (non-partners of the same gender). In summary, speakers show rather strong entrainment on intensity, irrespective of the normalization approach. On the contrary, neither the pitch features nor the speech rate features produced speaker entrainment as operationalized here.

TABLE II. T-VALUES OF T-TESTS FOR GLOBAL SIMILARITY AND CONVERGENCE IN SLOVAK, DF FOR ALL TESTS = 17, '*' CORRESPONDS TO $P < 0.05$, '**' $P < 0.01$; '?' $P < 0.1$

	Partner vs. non-partner				
Feature	**Session**	**1st half**	**2nd half**	**Session-males**	**Session-females**
F0_mean	NS	NS	NS	-2.82*	NS
F0_med	NS	NS	NS	-2.48*	2.4*
F0_max	NS	NS	NS	NS	NS
F0_min	NS	NS	NS	NS	NS
Int_mean	NS	-1.9?	NS	NS	NS
Int_med	NS	NS	-2.33*	NS	NS
Int_max	-1.98?	NS	NS	NS	-5.26**
Int_min	-2.44*	-2.7*	-2.2*	-2.1?	-2.95*
Syll-rate	NS	NS	-3.46**	NS	NS

The analysis of global convergence offers further insights. Recall, we compare entrainment in the two halves of each session. Table II shows the results for the partner vs. non-partner proximity in the 3rd and 4th column. We see that minimum intensity shows a steady entrainment for the entire session and thus presents a case of rather fast adjustment of speaker behavior. On the other hand, Intensity median and especially speech rate show significant global convergence since no proximity in the first half of the conversation changes to significant proximity in the second half. These two features thus seem to require some time before inter-speaker entrainment develops.

Finally, the last two columns offer a first look at the effect of speaker gender on the tendency for similarity using gender normalization for pitch only (Option #2). The most striking result is the evidence for entrainment in males on pitch mean and median while *dis*-entrainment on F0 median is observed for females given the positive sign of the t-value. On the other hand, females entrain to their partners in intensity much more robustly than the males.

The results from analyzing partner vs. self entrainment are summarized in Table III below. We see that most features show significantly more similarity with self than the partner, as expected. This is evidenced in the positive signs for the t-values. However, intensity maximum shows the opposite: speakers' max intensity values are more similar to their interlocutors than to their own speech with different interlocutors. In this, the Slovak corpus is similar to data in [13] in that intensity (max/mean) showed this pattern for the similar corpus of Standard American English. Hence, intensity is the feature most prone to entrainment since it shows partner

entrainment when compared to both no-partners and another speech of the identical speaker.

TABLE III. T-VALUES OF T-TESTS FOR GLOBAL SIMILARITY AND CONVERGENCE BASED ON PARTNER VS. SELF COMPARISON, '*' CORRESPONDS TO $P < 0.05$, '**' $P < 0.01$; '?' $P < 0.1$

	Partner vs. self				
Feature	**Session**	**1st half**	**2nd half**	**Session-males**	**Session-females**
F0_mean	3.37**	3.84**	2.89*	2.09?	2.65*
F0_med	4.02**	4.78**	3.08**	3.14*	2.65*
F0_max	2.41*	3.73**	3.57**	NS	NS
F0_min	2.46*	2.96*	NS	2.3*	NS
Int_mean	-1.9?	-2.75*	2.95*	-2.06?	NS
Int_med	NS	-2.56*	-1.92?	-1.91?	NS
Int_max	-2.16*	NS	NS	NS	-2.7*
Int_min	NS	NS	NS	NS	NS
Syll-rate	2.64**	2.4*	NS	2.51*	NS

The third and fourth columns of Table III show the results for convergence when analyzing the proximity between the partner and self for the two session halves. Here, we see some indication for weak entrainment in pitch indirectly by comparing the significance of self proximity for the entire session and the two halves. For F0 maximum, median, and minimum, the positive t-values for the first half are greater than for the second half, which suggests a slight change in pitch toward that of the partner rather than the self. Second, we again see partner entrainment on intensity mostly due to this behavior in the first half and attenuation of self-similarity in speech rate over the two halves. These results corroborate the findings of the partner vs. non-partner analysis in Table II; especially for speech rate.

The two rightmost columns of Table III list the results for males and females separately when we compare their speech with the speech of their partners on the one hand and their own speech in another session with a different partner on the other hand. First, we see positive signs for pitch features and negative for intensity features in line with the general trend already discussed above: pitch is consistent for speakers in the two sessions while intensity adjusts to that of the interlocutor. Second, males show entrainment on mean and median intensity while females on maximum intensity. Finally, males seem to be more likely to maintain their speech rate (be similar to self) while the data do not show this for females suggesting that the indication of convergence in speech rate in Table II might be attributed more to females than to males.

IV. DISCUSSION AND CONCLUSIONS

The initial results examining two forms of global entrainment (similarity and convergence) in Slovak conversational data presented in this paper suggest a rather complex picture but the one that is comparable to the observations in Standard American English and Mandarin Chinese reported in [18]. First, almost no evidence of entrainment was observed on pitch features; however, with a notable exception of male speakers mean and median pitch. Second, speech intensity seems most prone to entrainment showing not only significant differences between speaking with partners as compared to non-partners, but also significant differences between sessions of the same speaker with a different interlocutor. Third, a similarly robust entrainment in terms of speaking rate was not observed, and only convergence, i.e. greater similarity in speech rate in the second half of the conversation than in the first half, was reported.

In this sense, our results from Slovak are comparable to entrainment observed for English and Mandarin. Results in [18] refined in [13] also show intensity as the most robust and consistent entraining features while pitch did not show global entrainment in English but it did for Chinese. It is important to keep in mind, however, that the current Slovak findings in intensity are less robust than those reported for English and Chinese. Regarding speaking rate, Slovak data show even weaker entrainment than entrainment observed for intensity since only convergence for data pooled from both genders was observed. This contrasts with the other two languages since English showed both similarity and convergence while Mandarin showed similarity.

Regarding the relationship between entrainment and gender, several studies suggest that females should entrain more than males; either due to their increased perceptual sensitivity or less powerful social status. References [18] and [15], however, both reported greatest entrainment in mixed pairs when females talked to males than both same-sex pairs. Nevertheless, females showed more entrainment than males. Our data give mixed initial results regarding the effect of gender on global entrainment. Males were the only ones to show entrainment on pitch while females showed more robust entrainment on intensity than males. Hence, our preliminary observations do not lead to clear differences between the genders but suggest differential utilization of prosodic features of entrainment by the genders.

Little evidence of entrainment in pitch, we might speculate, might have different sources for the three languages. In Chinese, pitch is an important phonemic component participating in cuing word meanings. Informal observations on Slovak intonation suggest a lower general pitch range and pitch variability than English, especially for marking prominent words since Slovak might use flexible word order for cuing focus and other aspects commonly linked to pitch marking in English. Hence, greater functional load of pitch in Chinese and English – in lexical differences for the former and information structure signaling for the latter – than in Slovak might render a somewhat decreased pitch range in Slovak, which in turn might limit the affordance for pitch entrainment in Slovak.

On the other hand, data from the three languages regarding entrainment in intensity and speech rate suggest that intensity is cross-linguistically most prone to entrainment while speech rate presents a true continuum in the propensity for entrainment. This is useful information for application design utilizing human-machine spoken interactions since the language or culture of the potential users should be taken into

consideration for speech rate generation but entrainment in speech intensity might be applied irrespective of the native language of the user.

With these observations and speculations one has to keep in mind, however, that the current Slovak corpus is smaller than the other two in terms of the number of sessions as well as speakers suitable for partner vs. self analysis. It is possible that more comparable data in terms of size would provide a slightly different picture.

This first analysis of general and static measures of entrainment in Slovak only scratches the surface of possible ways of how entrainment might be realized. We plan further analysis of local entrainment at turn exchanges as well as more dynamic measures gauging the temporal development of entrainment. Moreover, the completion of the Slovak games corpus development in near future will allow for the analysis of the relationship between gender and entrainment by dividing interlocutor pairs to male-male, female-female, and mixed ones, which follows previous research.

Keeping these limitations in mind, our results nevertheless suggest that entrainment might not be applicable to all domains of spoken behavior identically for different languages and cultures. This further supports argument that entrainment is not a mechanistic feature of human-human interactions but that it might interact with social and cultural aspects. Therefore, entrainment can be characterized as a feature of social cognition [12] and exploring its relation to other aspects of spoken behavior improves our understanding of human cognitive abilities and improves models of these abilities needed for application utilizing human-machine spoken interactions.

ACKNOWLEDGMENT

This work was supported by the MVTS GAMMA project of the Slovak Academy of Sciences.

REFERENCES

[1] L. Bell, J. Gustafson and M. Heldner, "Prosodic adaptation in human-computer interaction," In: Proceedings of International Congress of Phonetic Sciences, 2003. pp. 2463-2466.

[2] Š. Beňuš, "Social aspects of entrainment in spoken interaction," Cognitive Computation, in press.

[3] Š. Beňuš, "Conversational Entrainment in the Use of Discourse Markers," in: Recent Advances of Neural Networks Models and Applications, Smart inovations, systems, and technologies 26, S. Basis et al., Eds. Berlin: Springer, 2014, pp. 345-352.

[4] Š. Beňuš, "Prosodic forms and pragmatic meanings: the case of the discourse marker 'no' in Slovak," Proceedings of 3rd CogInfoCom conference, 2012.

[5] P. Boersma and D. Weenink, "Praat: doing phonetics by computers", [http://www.fon.hum.uva.nl/praat/]

[6] N. Campbell, "Social Aspects & Speechability *in* CogInfoCom System," talk presented at CogInfoCom 2012, Kosice, Slovakia.

[7] M. Cernak, M.Rusko,M. Trnka, and S. Darjaa, "Data-Driven Versus Knowledge-Based Approaches to Orthoepic Transcription in Slovak," in Proceedings of ICETA 2003, pp 95-97.

[8] T. Chartrand and J. Bargh, "The chameleon effect: The perception-behavior link and social interaction," Journal of Personality and Social Psychology, 1999; 76: 893-910.

[9] S. Darjaa, M. Cerňak, M. Trnka, M. Rusko, R. Sabo, "Effective triphone mapping for acoustic modeling in speech recognition," in Proceedings of Interspeech, 2011.

[10] A. Gravano, "Turn-taking and affirmative cue words in task-oriented dialogue," Ph.D. dissertation, Columbia University, 2009.

[11] J. Hirschberg, "Speaking more like you: Entrainment in conversational speech. Proceedings of Interspeech, 2011, 27-31.

[12] H. De Jaegher, E. Di Paolo and S. Gallagher, "Can social interaction constitute social cognition?"Trends in Cognitive Sciences, 2010; 14: 441-447.

[13] R. Levitan, "Acoustic-Prosodic Entrainment in Human-Human and Human-Computer Dialogue," PhD dissertation, Columbia University, 2014.

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

[15] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montreal, Canada: Association for Computational Linguistics, June 2012, pp. 11–19.

[16] D. Litman D, H. Friedberg and K. Forbes-Riley, "Prosodic Cues to Disengagement and Uncertainty in Physics Tutorial Dialogues," In: Proceedings of Interspeech, 2012.

[17] K. Shockley, M. V. Santana, and C. A. Fowler, "Mutual interpersonal postural constraints are involved in cooperative conversation," Journal of Experimental Psychology: Human Perception & Performance, 2003; 29: 326-332.

[18] Z. Xia, R. Levitan and J. Hirschberg, " Prosodic Entrainment in Mandarin and English: A Cross-Linguistic Comparison," in Proceedings of 7th Speech Prosody, 2014.