INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Relationship between auditory and semantic entrainment using Deep Neural Networks (DNN)

Jay Kejriwal[1,2], *Štefan Beňuš*[1,3]

[1]Institute of Informatics, Slovak Academy of Sciences, Slovakia
[2]Faculty of Informatics and Information Technology, Slovak Technical University, Slovakia
[3]Constantine the Philosopher University, Nitra, Slovakia

jay.kejriwal@savba.sk, sbenus@ukf.sk

Abstract

The tendency of people to engage in similar, matching, or synchronized behaviour when interacting is known as entrainment. Many studies examined linguistic (syntactic and lexical structures) and paralinguistic (pitch, intensity) entrainment, but less attention was given to finding the relationship between them. In this study, we utilized state-of-the-art DNN embeddings such as BERT and TRIpLet Loss network (TRILL) vectors to extract features for measuring semantic and auditory similarities of turns within dialogues in two comparable spoken corpora of two different languages. We found people's tendency to entrain on semantic features more when compared to auditory features. Additionally, we found that entrainment in semantic and auditory linguistic features are positively correlated. The findings of this study might assist in implementing the mechanism of entrainment in human-machine interaction (HMI).

Index Terms: entrainment, alignment, semantic information, DNN embeddings, TRILL vectors

1. Introduction

Entrainment is the tendency of a speaker to adjust some properties of a speaker's features to match the interlocutor's characteristics. It has been found to correlate with positive social attributes such as likeability [1], task success [2], and even rapport with a robot [3]. According to the psycholinguistics literature, entrainment affects various linguistic dimensions, such as lexical choice [4], syntactic structure [5], or acoustic-prosodic features [6].

Several studies have investigated the effects of entrainment utilizing different modalities and implemented it in Spoken Dialogue Systems (SDS) [3, 7, 8]. In SDS, speech entrainment functionality would enable machines to dynamically entrain and disentrain on various auditory features, which might result in more efficient, successful, natural, and pleasing interactions. Similarly, implementing semantic entrainment functionality would enable machines to align semantically with humans resulting in more meaningful conversations. An essential first step toward effectively implementing entrainment in SDS is understanding how entrainment works at different linguistic levels and what their relationships are. Understanding these variations will allow us to weigh them meaningfully when they are combined to develop SDS systems equipped with effective entrainment functionalities.

Entrainment has previously been studied independently using linguistic-related parameters [9] or paralinguistic-related parameters [10, 11]. Additionally, researchers have started exploring the correlation between entrainment at different linguistic levels. For instance, [12] explored the relationship between prosodic, lexical, semantic, and syntactic entrainment among individuals with autism spectrum disorder (ASD). The results revealed distinct patterns of prosodic and lexical entrainment. Similarly, [13] explored the correlation between acoustic-prosodic and syntactic entrainment within a dialogue. They reported speakers entrain on some but not all features within a linguistic level. Furthermore, [14] reported correlations between acoustic-prosodic and lexical entrainment in group conversations. On the contrary, [15] found that none of the acoustic-prosodic and lexical entrainment measures were meaningfully correlated, clustered, or exhibited principal components. Hence, the results of studies exploring the relationship of entrainment at different levels are inconclusive.

In a recent study [16], DNN embeddings were used to explore the relationship between acoustic-prosodic and semantic entrainment. The authors proposed measures of "semantic similarity" of dialogues using BERT embeddings trained on a Chinese spoken corpus. They reported an inverse relationship between them: interlocutors did not adjust prosodic features when their semantics were closer to their partners. However, these results and their wider impact on SDS applications should be interpreted with caution since there were three limitations to the given study. First, the question-response system in Chinese conversations was analyzed. The authors did not provide a cross-linguistic comparison, which would allow observing the trends and underlying patterns by comparing auditory and semantic entrainment in different languages. Second, the authors introduce *convergence* and *synchrony* as entrainment metrics. *Convergence* implies people become more similar over the period of time. *Synchrony* means people are consistently behaving in similar way. The authors did not consider proximity as an entrainment metric which is helpful in understanding if two people are getting semantically closer to each other at a given time. In a session that displays proximity, the speaker turns are more similar to the immediately adjacent turns of the interlocutor than to other random interlocutor's turns [11]. Information about proximity might be valuable for turn-to-turn implementation of entrainment into automatic SDS. Last, the authors did not report if BERT embeddings were normalized or not. Usually, BERT embeddings are not normalized and utilizing Pearson's correlation can provide inconsistent results. There is a high degree of sensitivity in Pearson's *r* to even minor deviations from normality, where an outlier can hide an underlying association [17]. Using a novel approach in this study, we describe linguistic information and analyze the entrainment relationship between two different linguistic levels by utilizing different entrainment metrics (proximity, convergence, and synchrony) on two different spoken corpora using different languages.

Empirical studies exploring the entrainment relationship between different linguistic levels have found variable results so far. There might be three possible reasons associated with

10.21437/Interspeech.2023-1947

it. First, entrainment in linguistic levels has been analyzed using different methods. For example, in [13], the authors measured acoustic-prosodic entrainment using the metrics proposed in [6], which measures correlations among adjacent turns. In contrast, they analyzed syntactic entrainment with generalized logit mixed-effect models (GLMM) [18]. Second, different toolkits are utilized for feature extraction. For example, in [13], researchers used the PRAAT toolkit [19] for extracting 323 temporal and acoustic-prosodic features, whereas [12] derived pitch, intensity, and rhythm-related features using the contour-based, parametric, and super positional intonation stylization (CoPaSul) toolkit that uses some different feature extraction and manipulation approaches [20]. Lastly, researchers measured similarity using different units of analysis. In [15], authors measure acoustic-prosodic entrainment on the inter-pausal unit (IPU)[1], whereas they measured lexical entrainment using *n*-gram sequences. In this study, we will extract features and measure entrainment using the same methodology in an effort to limit the mentioned sources of variability in results.

Finally, empirical findings on entrainment suggest it is a complex phenomenon where people entrain/dis-entrain on different para-linguistic features [6]. Earlier studies on entrainment have utilized paralinguistic features that incorporate spectral, temporal, and acoustic-prosodic features. A DNN embedding can solve the problem of fragmentation in para-linguistic features. DNN embedding is a method used to represent discrete variables as continuous vectors. DNN embedding using textual modality such as transformer [21] is immensely popular and has broader applications in NLP applications. Similarly, DNN embedding using auditory modality has provided promising outcomes in improving the performance of automatic speech recognition and other applications. In [22], the TRILL vector was proposed, which creates embeddings based on a CNN architecture that uses triplet-loss representation. This approach maps audio segments that appear nearer in time to be nearer in the embedding space. A comparison of different auditory features such as Low-level descriptors (LLD) features, spectral features, and DNN audio embeddings (x-vectors, TRILL vectors) was presented in [23]. In comparison to different auditory features, TRILL vectors provided greater classification accuracy in this study. Hence, we employ this method in our work to compare the acoustic and semantic entrainment.

In sum, research into speech entrainment has so far been fragmented, with numerous individual features and measures of similarity being used, but no attempts have been made prior to our knowledge that measures auditory similarity using DNN embeddings. With a long-term goal to develop an effective SDS, we analyze in this study auditory and semantic entrainment in comparable corpora of conversational speech in English and Slovak. Our paper makes three main contributions. First, we measured entrainment in conversational corpora using state-of-the-art DNN embeddings on semantic and auditory levels. Second, we explore the relationship between the two levels using the same methodology. Finally, the experimental result shows that entrainment in both levels is correlated positively in both spoken corpora.

2. Data and features

In this section, we describe two task-oriented spoken language corpora we analysed in the current study, how we extracted semantic and auditory features from them, and how we calculated metrics for measuring auditory and semantic entrainment.

2.1. Dataset

2.1.1. Columbia Games Corpus

The Columbia Games Corpus [24] consists of 12 spontaneous dyadic conversations between native Standard American English (SAE) speakers. Participants included thirteen individuals (six females and seven males); eleven participated in two sessions on different days and with other partners. Each dyad played four computer games of two kinds: Cards games and Objects games involving communication and teamwork. The subjects did not have visual contact due to a curtain placed between them ensuring verbal communication only. Twelve sessions were recorded, totaling 9 hours and 13 minutes. The subset of the Columbia Games Corpus most closely resembling spontaneous task-related conversations, namely the Objects game, was used for the current study, which roughly comprises 4.3 hours of speech data.

2.1.2. SK-Games Corpus

The SK-games corpus [25] is identical to the Objects games of the Columbia Games Corpus for SAE, except for changes in some screen images and their locations. The corpus contains nine dyadic conversations recorded by native speakers of Slovak. Eleven speakers (five females and six males) participated in the study; seven participated in two sessions, each with a different partner. The corpus involves 6.3 hours of spoken dialogue.

2.2. Feature extraction

The semantic and auditory linguistic levels of entrainment are analysed in each corpus. To extract semantic features, each turn in the dialog is encoded into a fixed-length vector (embeddings). For the Columbia games corpus, we used a neural network-trained model (SBERT) [26], representing 768 one-dimensional semantic features for each turn. Similarly, for the SK-Games corpus, we used the Slovak masked language model called SlovakBERT [27] where each turn is encoded into 768 one-dimensional semantic features. Furthermore, to extract auditory features for each turn, the TRILL vector [28] is used, representing 512 one-dimensional auditory features per turn. Since the TRILL vector model is language-independent, we used the same model on both the spoken corpora.

2.3. Entrainment metrics

In [10], the authors introduced three measures of entrainment: *Proximity* describes the similarity of interlocutor's speech at turn exchanges. *Convergence* quantifies the tendency when two speaker's speech becomes more similar throughout the conversation. *Synchrony* describes the entrainment by direction where speaker's prosodic features become correlated to his/her interlocutor. Based on the definition of the given metrics we used the same metrics for the current study. In earlier studies, absolute values were used to measure entrainment on acoustic-prosodic features. Since we are using DNN embeddings in the current study, the metrics are re-defined.

Proximity is measured using paired t-tests on two sets of differences: a set of adjacent distance (Eq.1) and another corresponding set of non-adjacent distance (Eq.2). Adjacent distance is the cosine distance between speaker's embeddings and

[1] IPU is a pause-free unit in turn separated by at least 50 ms. of silence

his/her conversational partners adjacent embeddings. On the other hand, non-adjacent distance is the cosine distance between the embeddings of a speaker and other random non-adjacent embeddings of his/her conversational partner. For ten random turns of another speaker, we measured the non-adjacent distance and calculated the mean. If the cosine distance of the adjacent distance is greater than the non-adjacent distance, we can infer that speakers are getting closer to each other.

$$adjacent\ distance = \cos(A, B) = \frac{A \cdot B}{|A||B|} \tag{1}$$

$$non - adjacent\ distance = \sum_{i=1}^{10} \cos(A, B_{rand}) \tag{2}$$

Convergence is measured by Pearson's correlation between cosine distance between adjacent turns and turn number (time). *Synchrony* is measured using Pearson's correlation on two sets of self-distance of speaker A and B. Self-distance (Eq.3) of a speaker is measured using cosine similarity between two consecutive turns of the same speaker.

$$self\ distance = \cos(A_i, A_{i+1}) \tag{3}$$

3. Results

3.1. Auditory and semantic entrainment using DNN

	Proximity				**Convergece**				**Synchrony**			
	Auditory		Semantic		Auditory		Semantic		Auditory		Semantic	
Session	*t*	*Sig.*	*t*	*Sig.*	*r*	*Sig.*	*r*	*Sig.*	*r*	*Sig.*	*r*	*Sig.*
1	4.04	*	3.28	*	0.02		0.03		0.03		0.03	
2	-0.05		4.44	*	0.11	+	-0.02		-0.02		0.08	
3	-0.28		1.05		-0.09		-0.13	+	-0.13	+	-0.07	
4	3.05	*	5.03	*	-0.08		-0.08		-0.08		0.01	
5	3.01	*	3.01	*	-0.13	+	-0.01		-0.01		-0.06	
6	2.36	+	4.28	*	-0.02		-0.04		-0.04		0.16	+
7	1.32		2.87	+	0.03		0.09	+	0.09	+	0.04	
8	-0.02		3.13	*	-0.06		0.05		0.05		0.01	
9	-1.92		0.15		0.10		-0.01		-0.01		0.06	
10	-0.11		2.27	+	0.06		0.10	*	0.10	*	-0.06	
11	-2.19	+	1.84		-0.15	*	-0.11	+	-0.11	+	0.04	
12	2.06	+	3.12	*	-0.03		0.03		0.03		0.00	

(a) Columbia games corpus (CGC)

1	-0.21		0.65		-0.09	+	0.06		-0.01		0.07	
2	0.03		4.39	*	0.15	*	0.09		-0.18	+	0.05	
3	1.33		3.32	*	0.03		0.02		-0.05		-0.04	
4	-0.38		-0.04		0.09	+	0.01		-0.01		0.08	
5	-0.02		0.94		-0.11	+	-0.06		-0.09		0.00	
6	-3.88	*	-3.86	*	0.05		0.00		0.06		0.11	
7	3.14	*	4.62	*	-0.09		-0.18	*	-0.07		0.11	
8	1.86		3.86	*	0.10	*	-0.02		0.07		0.12	*
9	0.67		2.08	+	0.04		0.03		0.01		0.02	

(b) SK Games corpus

Table 1: *Summary of entrainment results on auditory and semantic entrainment in (a) Columbia games corpus and (b) Sk-games corpus with significant results after Bonferroni correction (*) with* $\alpha = 0.004$ *and* 0.005 *for the English and Slovak and without Bonferroni correction (+) with* $\alpha = 0.05$ *entrainment type (proximity, convergence, and synchrony)*

3.1.1. Columbia-Games corpus

Table 1 (a) shows the auditory and semantic entrainment results in the Columbia games corpus.

Proximity: The English dataset shows little evidence of local proximity on auditory features. Only three sessions shows evidence of positive proximity. On the semantic level, in contrast, we found seven sessions that showed positive proximity. In addition, we observed that the distribution of the sessions with positive proximity in two levels is not random and that in all but one case, if people entrain on the auditory level they also entrain on the semantic level.

Convergence: We found little evidence of convergence on both levels in the Columbia games corpus. In auditory features, only one session shows significant evidence of divergence, i.e., differences between partners increase over time. On the contrary, one session shows significant evidence of positive convergence in semantic features.

Synchrony: The auditory features showed little evidence of synchrony. Only one session shows evidence of positive synchrony. Positive synchrony implies both the speakers are moving in the same direction, i.e., if speaker A raises his/her voice, then speaker B also raises his/her voice. On the contrary, we did not find evidence of synchrony on semantic features in the English corpus. Furthermore, before the Bonferroni correction, we found that two sessions showed negative synchrony; one session exhibited positive synchrony in semantic features, and one session exhibited positive synchrony in auditory features.

3.1.2. SK-Games Corpus

Table 1 (b) shows the results of auditory and semantic entrainment with proximity, convergence, and synchrony as entrainment metrics based on the Slovak games corpus.

Proximity: The Slovak data shows little evidence of proximity on the auditory level. For auditory features, only one session shows evidence of positive proximity, and only one shows negative proximity. On the contrary, four sessions show significant positive proximity for semantic features, while one shows significant negative proximity. In addition, we observed a similar pattern that we observed earlier in the English corpus, i.e., people entrain on both features when they entrain on auditory features.

Convergence: We found little evidence of convergence on both levels in the Slovak data. Two sessions display evidence of positive convergence for auditory features. On the contrary, only one session showed evidence of divergence on the semantic level. Additionally, before applying the Bonferroni correction, we found that people converge on auditory features more when compared to semantic features in the SK-games corpus.

Synchrony: The Slovak data shows little evidence of synchrony on semantic features: One session shows positive synchrony for semantic features. On the contrary, no session shows evidence of synchrony in auditory features. In [11], the authors reported negative synchrony is evident on almost every paralinguistic (auditory) feature of the SK-games corpus. We found similar evidence to be true where 6 out of 9 sessions show negative synchrony; however, they are not statistically significant.

3.2. Relationship between auditory and semantic entrainment

We measured two sets of adjacent distances using (Eq. 1): a set of adjacent distances on auditory features and another set of adjacent distances on semantic features. We measured Pearson's correlation between adjacent distance on auditory and semantic embeddings to investigate the relationship between semantic and auditory features.

Columbia Games Corpus: Table 2 (left panel 1a) shows results for the entrainment relationship between auditory and semantic features using the SBERT model in English Data. We found six sessions out of 12 exhibits a slightly significant positive correlation (mean *r*=0.21). To explore the potential effect

Session	r	p-value	Sig.
1	0.02	0.738	
2	0.13	0.004	+
3	0.12	0.008	+
4	0.17	0.000	*
5	0.23	0.001	*
6	0.17	0.004	*
7	0.19	0.000	*
8	0.12	0.005	+
9	0.11	0.025	+
10	0.30	0.000	*
11	0.20	0.000	*
12	0.11	0.035	+

1) Using SBERT model

r	p-value	Sig.
0.14	0.038	+
0.16	0.000	*
0.17	0.000	*
0.19	0.000	*
0.17	0.012	+
0.22	0.000	*
0.22	0.000	*
0.15	0.000	*
0.15	0.002	*
0.25	0.000	*
0.28	0.000	*
0.16	0.002	*

2) Using USE model

a) Columbia games corpus

r	p-value	Sig.
0.45	0.000	*
0.44	0.000	*
0.45	0.000	*
0.36	0.000	*
0.42	0.000	*
0.50	0.000	*
0.32	0.000	*
0.30	0.000	*
0.23	0.000	*

b) SK-games corpus

Table 2: *Entrainment relationship between semantic and auditory features in Columbia Games Corpus and SK-games corpus after Bonferroni correction (*) with* $\alpha = 0.004$ *and* 0.005 *for the English and Slovak and without Bonferroni correction (+) with* $\alpha = 0.05$

of the selection of language models (semantic model), we also utilized Google's Universal sentence encoder (USE) model [29] for extracting semantic features for each turn. Using the USE model, we measured adjacent distance on semantic features and measured Pearson's correlation between semantic and auditory features. Table 2 (middle panel 2a) shows the entrainment relationship between auditory and semantic features for the USE model. We found ten sessions out of 12 exhibits a slightly positive correlation (mean *r*=0.20).

SK-Games Corpus: Table 2 (rightmost panel b) shows that Slovak data has a stronger positive correlation between entrainment in both linguistic levels than the English data where all the sessions are positively correlated with mean *r* = 0.40.

4. Discussion and conclusion

We analyzed semantic and auditory entrainment using three different entrainment metrics over a total of 21 sessions of collaborative dyadic interactions in two languages. We observed the following patterns that emerged from the analysis.

Firstly, proximity is more prevalent than synchrony and convergence in both semantic and auditory entrainment. In both languages, positive proximity is evident in a greater number of dialogues compared to convergence and synchrony, indicating the tendency of people to get closer to each other in both semantic and auditory space at a given point in time.

Secondly, we found that semantic proximity is more prevalent than auditory proximity. In both datasets, we observed that people entrain on semantic features more when compared to auditory features. In general, when the semantics of two interlocutors become more similar, the interlocutor can understand the content of the conversation more easily. One possible reason for such a result can be traced to the type of corpora utilized for entrainment analysis. We used task-oriented corpora where the objective was to communicate about specific items in order to reach a joint goal. Semantic entrainment is crucial in task-oriented conversations like this since the task cannot be completed successfully without it. In contrast, auditory entrainment is optional and may be used to support semantic entrainment or indicate various aspects of the negotiation in terms of social relationship between the interlocutors. The findings of our study might vary from analyzing entrainment in real-life conversational corpora where semantic and auditory entrainment might weigh differently.

Thirdly, we noticed that semantic and auditory entrainment are positively correlated. A positive relationship between different linguistic levels can be conceptualized as people who entrain on one level are more likely to entrain on other levels. This finding is consistent with the Interactive Alignment Model proposed by [30]. This cognitive theory suggests that alignment at one level leads to alignment at other levels. Our findings suggest entrainment can be considered a single latent behavior or a collection of linked behaviors where people aligning on auditory features are more likely to align on semantic features. It is interesting to note the directionality in our findings: semantic entrainment implies auditory one whereas the reverse is not the case. The results of our study may also inform models dealing with the percolation of entrainment across linguistic levels.

Lastly, we noted that selecting a language model is crucial in identifying the relationship between different linguistic levels. We measured the relationship between auditory and semantic linguistic levels using two different language models for extracting semantic features in the English dataset. We found variance in results where utilizing the SBERT model reported six sessions are significantly positively correlated with mean *r* of 0.20. In contrast, the USE model reported that ten sessions are significantly positively correlated with mean *r* = 0.21. The average results of correlations are almost identical (*r* = 0.2 and 0.21); however, the number of sessions that are significantly positively correlated is different. A language model might account for such variability in results and when considering the entire corpus, differences are smoothed out.

In the Slovak dataset, we found a relatively stronger correlation between auditory and semantic entrainment with mean *r* = 0.40 on all sessions. It remains to be explored if this difference stems from the difference among the patterns of entrainment in Slovak and English or if, in part, it might stem from the selection of the language model as both datasets in the current study are similar. We did not find any other language models trained in Slovak due lower NLP resources compared to English. Extracting semantic features from different language models could allow us to have a more meaningful comparison and understand if such a stronger correlation is due to the language model.

To conclude, in earlier studies researchers used fragmented features and different methods to measure entrainment, which might have contributed to the variation in results. We measured entrainment using the comparable methodology on different levels and in different languages, and our measures captured entrainment patterns that differ from previous studies, e.g. [16]. This further implies that methodology and features utilized for measuring entrainment play an important role in finding the relationship between different levels. In our future work, we plan to investigate entrainment relationships also on other linguistic levels, such as lexical and syntactic, and analyze the entrainment relationships among them. This will allow us to pursue developing SDS whose entrainment functionalities are informed by the relationship among entrainment on different linguistic levels, which could provide a more naturalistic conversational experience in future human-machine spoken interactions.

5. Acknowledgements

This project has received funding from the European Union's Horizon 2020 *research and innovation programme under the Marie Skłodowska-Curie grant agreement No* 859588 and in part from the Slovak Granting Agency grant VEGA2/0165/21 and Slovak Research and Development Agency grant APVV-21-0373.

6. References

[1] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker, "Language style matching predicts relationship initiation and stability," *Psychological Science*, vol. 22, no. 1, p. 39–44, 2011.

[2] D. Reitter and J. D. Moore, "Predicting success in dialogue," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 808–815. [Online]. Available: https://aclanthology.org/P07-1102

[3] N. Lubold, H. Pon-Barry, and E. Walker, "Naturalness and rapport in a pitch adaptive learning companion," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU*, p. 103–110, 2015.

[4] 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. 1482–1493, 1996.

[5] D. Reitter, J. 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*, R. Sun, Ed., Vancouver, 2006, p. 685–690.

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

[7] R. Levitan, Štefan Beňuš, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing Acoustic-Prosodic Entrainment in a Conversational Avatar," in *Proc. Interspeech 2016*, 2016, pp. 1166–1170.

[8] R. Levitan, "Developing an integrated model of speech entrainment," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Yokohama, Japan, 2021.

[9] A. Ward and D. J. Litman, "Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora," in *SLaTE Workshop on Speech and Language Technology in Education*, 2007. [Online]. Available: http://d-scholarship.pitt.edu/23210/

[10] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels." in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 113–117. [Online]. Available: https://aclanthology.org/P11-2020

[11] 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 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, Czech Republic: Association for Computational Linguistics, Sep. 2015, pp. 325–334. [Online]. Available: https://aclanthology.org/W15-4644

[12] S. Patel, J. Cole, J. Lau, G. Fragnito, and M. Losh, "Verbal entrainment in autism spectrum disorder and first-degree relatives," *Scientific Reports*, vol. 12, p. 11496, 07 2022.

[13] R. Ostrand and E. Chodroff, "It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue," *Journal of Phonetics*, vol. 88, no. 101074, 2021.

[14] Z. Rahimi, A. Kumar, D. Litman, S. Paletz, and M. Yu, "Entrainment in Multi-Party Spoken Dialogues at Multiple Linguistic Levels," in *Proc. Interspeech 2017*, 2017, pp. 1696–1700.

[15] A. Weise and R. Levitan, "Looking for structure in lexical and acoustic-prosodic entrainment behaviors," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ser. Short Papers, vol. 2. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 297–302,.

[16] Y. Liu, A. Li, J. Dang, and Zhou, "Semantic and acoustic-prosodic entrainment of dialogues in service scenarios," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*. New York, NY, USA: Association for Computing Machinery. event-place, 2021, pp. 71–74,.

[17] V. Zhelezniak, A. Savkov, A. Shen, and N. Hammerla, "Correlation coefficients and semantic textual similarity," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ser. Long and Short Papers, vol. 1. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 951–962,.

[18] N. Breslow and D. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 9–25,, 1993. [Online]. Available: http://www.jstor.org/stable/2290687.

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[20] U. D. Reichel, "Copasul manual - contour-based parametric and superpositional intonation stylization," *ArXiv*, vol. abs/1612.04765, 2016.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[22] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. Chaumont Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," *Interspeech*, vol. 2020, p. 140–144, 2020, arXiv: 2002.12764.

[23] J. Kejriwal, Beňuš, and M. Trnka, "Stress detection using non-semantic speech representation," in *2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA)*, 2022, pp. 1–5.

[24] J. Hirschberg, S. B. Gravano, G. Ward, and E. S. German, *Columbia Games Corpus LDC2021S02. Web Download.* Philadelphia: Linguistic Data Consortium, 2021.

[25] Beňuš, "Prosodic forms and pragmatic meanings: The case of the discourse marker 'no' in slovak," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 2012, pp. 77–81.

[26] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019, arXiv:1908.10084 [cs]. arXiv:.

[27] M. Pikuliak, Grivalský, M. Konôpka, M. Blšták, M. Tamajka, V. Bachratý, M. Šimko, P. Balážik, M. Trnka, and F. Uhlárik, 2021, slovakBERT: Slovak Masked Language Model. arXiv:2109.15254 [cs]. arXiv: 2109.15254.

[28] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," 2018, arXiv:1412.6622 [cs, stat]. arXiv: 1412.6622.

[29] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal sentence encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. [Online]. Available: https://aclanthology.org/D18-2029

[30] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, no. 2, p. 169–190, 2004.