Fakulta informatiky a informačných technológií STU v Bratislave

Meno a priezvisko/*Name and Surname*

Jay Kejriwal
(študenta doktorandského študijného programu/
of the student of the doctorate degree study programme)

Autoreferát dizertačnej práce/*Dissertation Thesis Abstract*

Conversation coordination in human-machine communication.

na získanie akademického titulu/*to obtain the Academic Title of*
„doktor“ („philosophiae doctor“, v skratke/*abbreviated* as „PhD.“)

v doktorandskom študijnom programe/*in the doctorate degree study programme:*

9.2.9. Applied Informatics

v študijnom odbore/*in the field of study*:
18. Computer Science

Forma štúdia/*Form of Study*:
part-time study

Miesto a dátum/*Place and Date*: Bratislava, 15-05-2024

Dizertačná práca bola vypracovaná na/*Dissertation Thesis has been prepared at*
The Institute of Informatics, Slovak Academy of Sciences in Bratislava, Slovakia

(názov vzdelávacej inštitúcie, kde bola vypracovaná dizertačná práca/
Name of External Educational Institution, where the Dissertation Thesis has been prepared)

Predkladateľ/*Submitter*:

Jay Kejriwal
Institute of Informatics, Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava 45
Slovak Republic

(meno a priezvisko doktoranda a adresa jeho pracoviska/
Name and Surname of the Doctoral Candidate and the address of his/her workplace)

Školiteľ/*Supervisor:*

prof. Mgr. Štefan Beňuš, PhD.
Institute of Informatics, Slovak Academy of Sciences

Konzultant/*Consultants:*
NA

Autoreferát bol rozoslaný/*Dissertation Thesis Abstract was sent*: 16-05-2024

Obhajoba dizertačnej práce sa bude konať dňa/*Dissertation Thesis Defence will be held on*
24-06-2024

o/*at*13:00.**h/1:00** *pm* **na/***at* ***Ústave informatiky Slovenskej akadémie vied, Dúbravská cesta 9, 845 07 Bratislava 45***

.....
rektor STU alebo dekan fakulty,
ak sa doktorandský študijný program uskutočňuje na fakulte
(meno a priezvisko s uvedením titulov)/
Rector of STU or Dean of Faculty of STU,
if studies of the doctorate degree study programme were carried out at the Faculty

STU

Table of Contents

1. Introduction:

Entrainment in spoken interaction is the tendency of speakers to adjust some properties of their speech to match the characteristics of their interlocutors. It affects several linguistic dimensions, such as lexical choice [2], syntactic structure [21], acoustic prosodic features [11], or semantic similarity [13]. In addition, it correlates with different social aspects of the conversation, such as task success [22], liking [9], cooperation [15], or naturalness and rapport [14]. Implementing entrainment in Spoken dialogue systems (SDS) has rich potential as it promises to be an important feature for improving the naturalness and effectiveness of Human-Machine interactions (HMI), thus increasing potential application domains.

Existing implementations of entrainment in spoken dialogue systems (SDS) have been relatively simplistic and require further development. For instance, [10] implemented acoustic-prosodic entrainment into an SDS to study its effect in English, Spanish, and Slovak. They examined subjects' willingness to seek advice from an avatar that entrains rather than one that does not. In the English study, SDS entrained on both speech rate and intensity, while studies in Spanish and Slovak SDS entrained only on speech rate. In English, they found entrainment correlated positively with an avatar's perceived reliability and likeability, but in Slovak and Spanish, there was a negative correlation. The design of the experiment can be a contributing factor to the observed variation. For instance, the English SDS entrained on intensity and speech rate, whereas the Slovak and Spanish SDSs entrained on speech rate. Consequently, to develop more advanced models, a considerable amount of theoretically driven, interdisciplinary research with uniform experimental design is required to understand the underlying factors affecting the entrainment behavior of humans in HMI.

The goal of research aimed at improving SDSs is to gain an in-depth understanding of human behavior in Human-Human interactions (HHI) and then apply this understanding to improve HMI. Incorporating findings from HHI to HMI may result in more effective and satisfying communication. This thesis emphasizes this line of research. The thesis is mainly divided into three parts. The first part presents automatic detection models that can be used in existing SDS to detect entrainment. The second part examines the variation of entrainment behavior in different languages at various linguistic dimensions. Lastly, the third part explores the relationship between entrainment and non-verbal social cues, including gaze and emotion, respectively.

2. DNN-based entrainment detection systems:

In the first part of the thesis, a study is presented which utilizes two human-human (HH) corpora (The Fisher Corpus English Part 1 [3], Columbia games corpus [7]) and one human-machine (HM) corpus (Voice Assistant Conversation Corpus (VACC) [23]). First, the study presents a framework that allows extracting information from textual features to compute entrainment distances at lexical, syntactic and semantic linguistic levels. Secondly, utilizing the existing acoustic framework by [18], DNN models are trained using TRIpLet Loss network (TRILL) vectors [8]. The study presents five different analyses, as shown in Table 1.

Table 1. Analysis for comparing the performance of baseline and DNN-based entrainment models.

Sr.No.	Comparison	Similarity measure
1	Performance of baseline models	L1 distance and cosine similarity
2	Performance of DNN-entrainment models	L1 distance and cosine similarity
3	Comparing baseline and DNN-based entrainment models	L1 distance and cosine similarity
4	Performance of DNN-entrainment models for classifying HHI and HMI datasets	Cosine similarity
5	Robustness of DNN-entrainment models	Cosine similarity

Table 2 shows the results of the five analyses presented. The impact of similarity measures on the performance of both baseline and neural-based entrainment models is investigated. Tables 2 (a) and (b) show the results that indicate that the similarity measure, i.e., cosine similarity and absolute distance, affects model performance. Further, the findings suggest that DNN-based entrainment models can be useful in detecting entrainment across four linguistic levels. However, the performance of the models varies, with neural-based acoustic models outperforming text-based models, as shown in Table 2 (b). Furthermore, the acoustic-based DNN models can differentiate between HH and HMI, where a performance drop was noticed in a VAC corpus, as shown in Table 2 (c). This indicates that Alexa does not adjust its features, and the model does not learn entrainment information. Finally, the robustness of the DNN-trained entrainment models has been assessed. Table 2 (d) shows a slight decrease in the performance of acoustic-based models when the entrainment distance was compared to the mean of ten random turns. Conversely, the performance of text-based models increased by 1-2%. The evaluation results will be valuable for integrating a DNN-based entrainment detection system into existing SDS. Employing an entrainment detection system in SDS enables machines to detect whether human interlocutors align with them.

Table 2. Summary of Baseline and DNN trained entrainment models with classification accuracy for different acoustic and textual features on Columbia games corpus, Voice Assistant Conversation Corpus (VACC), and The Fisher Corpus English Part 1 (standard deviation shown in parentheses)

a) Baseline accuracy in three datasets

	Columbia games corpus		VAC corpus		Fisher corpus	
Feature	Baseline 1	Baseline 2	Baseline 1	Baseline 2	Baseline 1	Baseline 2
Acoustic (LLD)	**76.17 (**±6.03)	54.38 (±11.45)	**74.29 (**±4.96)	57.27 (±11.24)	**86.33 (**±5.11)	54.30 (±9.91)
Acoustic (TRILL)	56.33 (±11.13)	56.32 (±11.84)	55.23 (±11.24)	55.93 (±12.09)	56.64 (±17.22)	**85.70 (**±7.98)
Lexical	50 (±21.13)	25 (±2.88)	26.32 (±6.45)	53.46 (±6.74)	54.92 (±17.84)	41.64 (±1.73)
Syntactic	45.70 (±9.57)	46.25 (±8.17)	52.89 (±13.49)	54.84 (±10.52)	45 (±7.50)	45.31 (±6.74)
Semantic	50.62 (±12.22)	**52.34** (±10.64)	62.50 (±8.55)	**65.77 (**±6.93)	56.80 (±17.88)	**57.81** (±14.99)

b) DNN-based entrainment models accuracy using different entrainment distance

	L1 distance	Cosine similarity	L1 distance	Cosine similarity	L1 distance	Cosine similarity
Acoustic (LLD)	**74.98 (**±3.83)	72.34 (±4.27)	**77.26 (**±3.97)	56.48 (±5.92)	**84.14 (**±0.03)	70.16 (±4.76)
Acoustic (TRILL)	54.22 (±3.90)	53.98 (±3.76)	**54.22 (**±3.90)	31.17 (±3.67)	89.14 (±3.26)	**94.14 (**±0.01)
Lexical	49.77 (±4.02)	53.20 (±5.89)	52.50 (±4.10)	55.17 (±4.25)	58.28 (±4.57)	64.29 (±3.90)
Syntactic	46.67 (±6.16)	47.34 (±4.94)	52.58 (±4.50)	55.94 (±6.28)	47.66 (±3.45)	54.06 (±5.60)
Semantic	53.36 (±4.90)	55.05 (±4.67)	64.79 (±6.40)	66.46 (±6.28)	58.36 (±2.39)	60.30(±0.04)

c) Classification accuracy by splitting into groups

	A to B	B to A	Spkr to Alexa	Alexa to Spkr	A to B	B to A
Acoustic (LLD)	72.29 (±3.25)	74.29 (±3.23)	58.25 (±4.98)	**78.87 (**±2.12)	84.13 (±0.12)	80.23 (±0.07)
Acoustic (TRILL)	53.89 (±3.25)	54.19 (±3.13)	15.82 (±4.19)	**30.87** (±3.12)	87.83 (±3.12)	86.12 (±2.07)
Lexical	54.45 (±8.52)	51.72 (±9.92)	59.94 (±5.83)	54.86 (±7.27)	63.98 (±4.08)	61.41 (±3.32)
Syntactic	49.60 (±6.35)	47.34 (±5.15)	50.41 (±8.98)	57.23 (±9.32)	50 (±9.47)	56.81 (±10.21)
Semantic	53.67 (±4.25)	55.19 (±3.89)	66.14 (±5.85)	67.94 (±6.50)	55.27 (±1.91)	57.77(±2.51)

d) Classification accuracy for selecting different random turns (RT)

	One RT	Ten RT	One RT	Ten RT	One RT	Ten RT
Acoustic (LLD)	**74.98 (**±3.83)	71.23 (±3.45)	**77.26 (**±3.97)	76.83 (±2.11)	84.14 (±0.03)	80.16 (±0.08)
Acoustic (TRILL)	53.98 (±3.76)	56.17 (±5.82)	31.17 (±3.67)	29.79 (±3.16)	**94.14** (±0.01)	**93.75 (**±0.02)
Lexical	53.20 (±5.89)	54.41 (±7.64)	55.17 (±4.25)	56.06 (±5.68)	64.29 (±3.90)	**66.48 (**±7.22)
Syntactic	47.34 (±4.94)	47.17 (±4.82)	55.94 (±6.28)	55.98 (±11.63)	54.06 (±5.60)	56.25 (±4.59)
Semantic	55.05 (±4.67)	**58.31 (**±4.67)	66.46 (±6.28)	**69.89 (**±5.21)	60.30 (±0.04)	61.66 (±0.04)

3. Entrainment in different languages

In the second part of the thesis, the evaluation of lexical, syntactic, semantic, and acoustic entrainment is performed in four comparable spoken corpora of four typologically different languages (English [7], Slovak [1], Spanish [4], and Hungarian [17]) using comparable tools and methodologies based on DNN embeddings. A cross-linguistic comparison was made where entrainment was compared in four different languages at each linguistic level. In addition, entrainment was compared across four different linguistic levels in each language and in each session, respectively, and the relationship between them was explored.

The findings suggest that Hungarian speakers entrain the most on all four linguistic levels when compared to English, Slovak, and Spanish speakers, as shown in Figure 1. Further, speakers in all four languages entrain more on the acoustic level, followed by semantic, lexical, and syntax, as shown in Figure 2. The relationship between entrainment at lexical, syntactic, semantic, and acoustic entrainment was positively correlated in all four languages. The results obtained from this analysis will facilitate the identification of the most suitable combination of features at different linguistic levels that must be emphasized to fine-tune language-based features in SDS for particular languages.

4. Entrainment and non-verbal social cues

Understanding the underlying relationship between entrainment and non-verbal social cues is important. Gaze behavior and emotional expression belong among prominent and highly researched cues affecting spoken interactions, but their relationship to speech entrainment is still not fully understood. A better understanding of this relationship will enable us to determine how the gaze behavior of a robot affects the entrainment of human speakers toward the robot and, secondly, which prosodic features should be adjusted based on the emotional state of the interlocutor.

Fig. 1. Summary of entrainment at (a) lexical, (b) syntactic, (c) semantic, and (d) acoustic levels in English (EN), Spanish (ES), Hungarian (HU), and Slovak (SK).

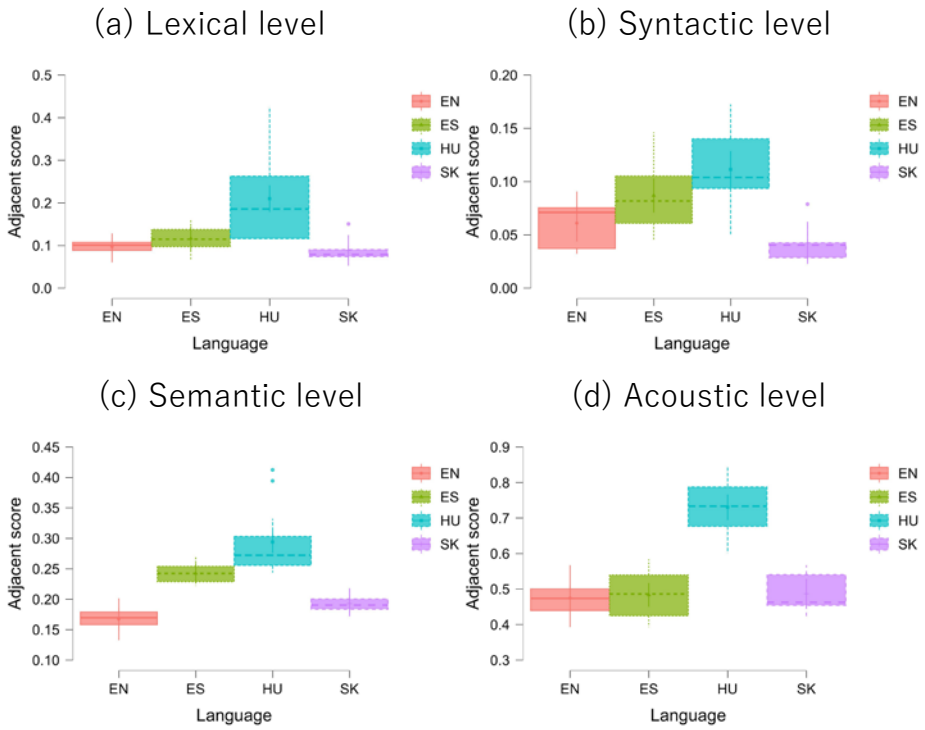
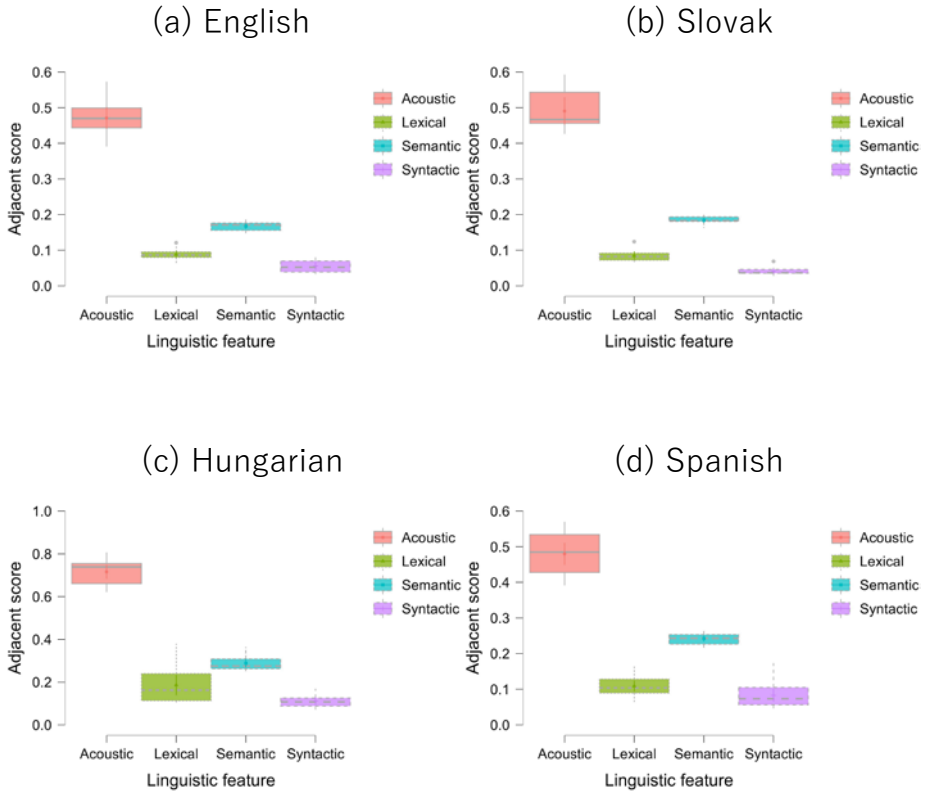


Fig. 2. Summary of entrainment in each individual language in (a) English, (b) Slovak, (c) Hungarian, and (d) Spanish corpus at lexical, syntactic, semantic, and acoustic levels.



4.1. Entrainment and Gaze

Gaze behavior is an important aspect of social spoken interaction that is realized in part through non-verbal signaling. Researchers have explored the relationship between gaze and acoustic-prosodic entrainment. In a recent study conducted by [19], speech entrainment was analyzed by measuring the mean pitch of subjects interacting with a robotic head in two modes of the robot's gaze behavior (fixed vs. variable) described in [16]. The authors reported no significant differences in entrainment between the two conditions. In earlier acoustic-prosodic entrainment studies, researchers analyzed eight different prosodic features [12, 14], whereas [19] focused on only one feature, mean pitch. Employing various acoustic-prosodic features and examining entrainment at different linguistic levels can yield a better understanding of the relationship between entrainment and gaze.

The third part of the thesis presents a study to investigate entrainment in Gaze Aversion Corpus (GAC) [16] on the four linguistic levels, including acoustic-prosodic, lexical, syntactic, and semantic levels under two different gaze conditions (fixed vs. variable).

The observations suggest that speakers tend to entrain more towards a robot in gaze aversion conditions on acoustic-prosodic (mean pitch and Noise to harmonics (NHR)) and lexical linguistic levels. The findings from the study can help us determine which gaze behavior should be adopted by a robot in HMI.

4.2 Entrainment and Emotion

Emotional cues are important in conveying the speaker's intent and can greatly impact the effectiveness of the communication. The relationship between entrainment and emotion was earlier investigated using textual features [5, 6]. However, these studies are limited to textual modality, and it is still unclear which prosodic features are relevant and how they should be adjusted based on the interlocutor's emotional state.

The fourth study analyzed entrainment in the Multimodal EmotionLines Dataset (MELD) [20] and explored eight acoustic-prosodic features, including pitch mean and max, intensity mean and max, jitter, shimmer, noise-to-harmonics ratio (NHR), and speech rate. Entrainment distance was measured between adjacent turns using absolute distance. Later, the distances were compared using an unpaired *t*-test on each prosodic

feature in two different ways. First, to understand how speakers behave differently when their interlocutors are in different emotional states. The two sets of distances were compared, namely, entrainment distance on dyads where a speaker has the same emotional state in both sets, and the interlocutor has a different emotional state in both sets. For instance, positive-negative and positive-neutral. Secondly, to examine which prosodic features are affected by the directionality of the emotional states of the dyads, a set of entrainment distances on the emotional state of dyads and another set of inverse emotional states of dyads were compared. For instance, positive-negative and negative-positive.

Table 3. Summary of prosodic features affected when speaker under one emotional state and interlocutors under different emotional state (bold emotional state refers more dis-entrainment).

Emotional state	Feature
negative-neutral and **negative-positive**	Mean Pitch and NHR
positive-negative and **positive-neutral**	Max Pitch and Max Intensity
neutral-negative and **neutral-positive**	Mean and Max Pitch, Mean Intensity, Jitter and Speech rate

Table 4. Summary of prosodic features affected when emotional state of dyads are compared interchangeably. (bold emotional state refers more dis-entrainment)

Emotional state	Feature
negative-neutral and **neutral-negative**	Shimmer
negative-positive and **positive-negative**	Max Pitch
neutral-positive and **positive-neutral**	Mean Intensity

The study revealed that people tend to dis-entrain with each other on various acoustic and prosodic features when experiencing different emotional states. Further, the directionality of emotions plays an important role, and dyads dis-entrain different features when the directionality of their emotional states is interchanged, i.e., positive-negative to negative-positive. The results obtained from the study can be applied to develop emotional entrainment functionality in existing SDSs, which positively influences entrainment.

5. Contributions

In sum, this thesis provides four main contributions. First, an unsupervised deep learning framework is presented that can detect entrainment at different linguistic levels using speech and textual features using state-of-the-art DNN embeddings. More generally, the immense potential of neural entrainment measures is highlighted, which can be employed in existing SDS for automatically detecting entrainment in different linguistic dimensions. Secondly, this thesis contributes knowledge towards entrainment at multiple linguistic dimensions in different languages and how linguistic dimensions are related. To the best of current knowledge, this is the first cross-comparative study that concentrates on entrainment in four linguistic dimensions utilizing speech and textual features using comparable tools and methodologies. Thirdly, the relationship between entrainment and emotion was explored using speech features. This dissertation provides a better understanding of how prosodic features participate in entrainment based on the emotional state of speakers, which in turn provides suggestions for how to implement emotional entrainment functionality in SDS meaningfully. Finally, as a part of exploring the relationship between gaze and entrainment, this dissertation provides an understanding of how the gaze behavior of a robot affects the entrainment at different linguistic dimensions. This understanding is an essential step in planning the robot's gaze behavior.

References:

1. Beňuš, Š.: Prosodic forms and pragmatic meanings: The case of the discourse marker 'no' in slovak. In: 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom). pp. 77–81 (2012). https://doi.org/10.1109/CogInfoCom.2012.6421961
2. Brennan, S.E., Clark, H.H.: Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition **22**(6), 1482–1493 (1996)
3. Cieri, C., Miller, D., Walker, K.: The fisher corpus: a resource for the next generations of speech-to-text. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal. European Language Resources Association (ELRA (2004)
4. Gravano, A., Kamienkowski, J.E., Brusco, P.: Uba games corpus. consejo nacional de investigaciones científicas y técnicas (2023), http://hdl.handle.net/11336/191235, dataset).
5. He, S., Zheng, X., Bao, X., Ma, H., Zeng, D., Xu, B., Li, C., Hao, H.: Characterizing emotion entrainment in social media. In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). pp. 642–648 (2014). https://doi.org/10.1109/ASONAM.2014.6921653
6. He, S., Zheng, X., Zeng, D., Luo, C., Zhang, Z.: Exploring Entrainment Patterns of Human Emotion in Social Media. PLOS ONE **11**(3), e0150630 (Mar 2016). https://doi.org/10.1371/journal.pone.0150630,https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150630, publisher: Public Library of Science
7. Hirschberg, J., Gravano, A., Beňuš, Š., Ward, G., German, E.S.: Columbia Games Corpus LDC2021S02. Web Download. Linguistic Data Consortium, Philadelphia (2021)
8. Hoffer, E., Ailon, N.: Deep metric learning using triplet network (2018), arXiv:1412.6622 [cs, stat]. arXiv: 1412.6622.

9. Ireland, M.E., Slatcher, R.B., Eastwick, P.W., Scissors, L.E., Finkel, E.J., Pennebaker, J.W.: Language style matching predicts relationship initiation and stability. Psychological Science **22**(1), 39–44 (2011)
10. Levitan, R., Beňuš, Š., Gálvez, R.H., Gravano, A., Savoretti, F., Trnka, M., Weise, A., Hirschberg, J.: Implementing Acoustic-Prosodic Entrainment in a Conversational Avatar. In: Proc. Interspeech 2016. pp. 1166–1170 (2016). https://doi.org/10.21437/Interspeech.2016985
11. Levitan, R., Gravano, A., Hirschberg, J.: Entrainment in speech preceding backchannels. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 113–117. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), https://aclanthology.org/P11-2020
12. Levitan, R., Hirschberg, J.: Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: Proc. Interspeech 2011. pp. 3081–3084 (2011). https://doi.org/10.21437/Interspeech.2011-771
13. Liu, Y., Li, A., Dang, J., Zhou: Semantic and acoustic-prosodic entrainment of dialogues in service scenarios. In: Companion Publication of the 2021 International Conference on Multimodal Interaction. pp. 71–74,. Association for Computing Machinery. event-place, New York, NY, USA (2021)
14. Lubold, N., Pon-Barry, H., Walker, E.: Naturalness and rapport in a pitch adaptive learning companion. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU p. 103–110 (2015)
15. Manson, J.H., Bryant, G.A., Gervais, M.M., Kline, M.A.: Convergence of speech rate in conversation predicts cooperation. Evolution and Human Behavior **34**, 419–426 (2013)
16. Mishra, C., Offrede, T., Fuchs, S., Mooshammer, C., Skantze, G.: Does a robot's gaze aversion affect human gaze aversion? Frontiers in Robotics and AI **10** (2023). https://doi.org/10.3389/frobt.2023.1127626
17. Mády, K., Kohári, A., Reichel, U.D., Szalontai, Á., Mihajlik, P.: The budapest games corpus. In: Speech Research conference. p. 75–77 (2023)
18. Nasir, M., Baucom, B., Narayanan, S., Georgiou, P.: Towards an Unsupervised Entrainment Distance in Conversational Speech Using Deep Neural Networks. In: Proc. Interspeech 2018. pp. 3423–3427 (2018). https://doi.org/10.21437/Interspeech.2018-1395
19. Offrede, T., Mishra, C., Skantze, G., Fuchs, S., Mooshammer, C.: Do humans converge phonetically when talking to a robot? In: Skarnitzl, R..J.V. (ed.) Proceedings of the 20th International Congress of Phonetic Sciences, Prague. pp. 3507–3511 (05 2023)
20. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 527–536. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1050, https://aclanthology.org/P19-1050
21. Reitter, D., Moore, J., Keller, F.: Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In: Sun, R. (ed.) Proceedings of the 28th Annual Conference of the Cognitive Science Society. p. 685–690. Vancouver (2006)
22. Reitter, D., Moore, J.D.: Predicting success in dialogue. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 808–815. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https:// aclanthology.org/P07-1102
23. Siegert, I., Krüger, J., Egorow, O., Nietzold, J., Heinemann, R., Requardt, A.: Voice assistant conversation corpus (vacc): A multi-scenario dataset for addressee detection in humancomputer-interaction using amazon's alexa. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (may 2018)

STU

Publications:

[1] J. Kejriwal, Š. Beňuš, M. Trnka, Stress detection using non-semantic speech representation, in: 2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA),2022,pp.1-5. doi:10.1109/RADIOELEKTRONIKA54537.2022.9764916.

[2] Kejriwal, J. Relationship between speech entrainment and emotion. 2022 10th Inter- national Conference On Affective Computing And Intelligent Interaction Workshops And Demos (ACIIW). pp. 1-4 (2022).

[3] Kejriwal, J. & Beňuš, Š. Speech Entrainment and Emotion. 2023 14th IEEE International Conference On Cognitive Infocommunications (CogInfoCom). pp. 000099-000104 (2023).

[4] Kejriwal, J., Beňuš, Š & Rojas-Barahona, L. Unsupervised Auditory and Semantic Entrainment Models with Deep Neural Networks. Proc. INTERSPEECH 2023. pp.2628-2632 (2023).

[5] J. Kejriwal, Š. Beňuš, Relationship between auditory and semantic entrainment using Deep Neural Networks (DNN), in: Proc. INTERSPEECH 2023, 2023,pp. 2623–2627. https://doi.org/10.21437/Interspeech.2023-1947 doi:10.21437/Interspeech.2023-1947.

[6] Kejriwal, J., Mishra, C., Offrede, T., Skantze, G. & Beňuš, Š. Does a Robot's Gaze Behavior Affect Entrainment in HRI? (2024). Submitted to Cognitive Computation Journal (under review). https://www.researchsquare.com/article/rs-3961654/v1

[7] Kejriwal, J. & Beňuš, S. Lexical, syntactic, semantic and acoustic entrainment in Slovak, Spanish, English, and Hungarian: A cross-linguistic comparison (2024). Submitted to Speech Communication Journal (under review). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4729223

[8] Kejriwal, J., Benˇusˇ, S. & M. Rojas-Barahona, L. Entrainment Detection Using DNN (2024). Submitted to Computer Speech & Language Journal (under review).
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4769763159