

Semantic and Acoustic-Prosodic Entrainment of Dialogues in Service Scenarios

Yuning Liu

Japan Advanced Institute of Science and Technology
Japan
liuyuning@jaist.ac.jp

Aijun Li

Institute of Linguistics, Chinese Academy of Social
Sciences
China
liaj@cass.org.cn

Jianwu Dang

Japan Advanced Institute of Science and Technology
Japan
College of Intelligence and Computing, Tianjin University
China
jdang@jaist.ac.jp

Di Zhou

Japan Advanced Institute of Science and Technology
Japan
zhoudi@jaist.ac.jp

ABSTRACT

According to the Communication Accommodation Theory, speakers dynamically adjust their communication behaviors, converging to or diverging from their interlocutors in order to diminish or increase social distance, which is called entrainment. Most of the studies investigated the entrainment of the interlocutors in terms of linguistic and paralinguistic features respectively, but paid less attention to the (dis)entrainment relation between paralinguistic and linguistic ones. In this study, we employed BERT to extract the semantic similarities of turns within dialogues in service scenarios, and found the semantic entrainment. We also found that (dis)entrainments policies were adopted between acoustic-prosodic (paralinguistic) and linguistic (semantic) features. These findings will contribute to fully understanding the mechanism of entrainment in dialogue.

CCS CONCEPTS

• **Computing methodologies** → *Phonology / morphology; Discourse, dialogue and pragmatics.*

KEYWORDS

entrainment, semantic information, prosody

ACM Reference Format:

Yuning Liu, Jianwu Dang, Aijun Li, and Di Zhou. 2021. Semantic and Acoustic-Prosodic Entrainment of Dialogues in Service Scenarios. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion)*, October 18–22, 2021, Montréal, QC, Canada. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3461615.3491105>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMI '21 Companion, October 18–22, 2021, Montréal, QC, Canada

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8471-1/21/10...\$15.00

<https://doi.org/10.1145/3461615.3491105>

1 INTRODUCTION

How speakers adapt to each other has been of great interest to speech researchers for the last two decades. This adaption between interlocutors is referred to as entrainment. In the context of a conversation, it means that two speakers become more similar to each other in speaking styles and representations. Entrainment believed to be present at all levels of communication in various factors, and that this process helps interlocutors to understand each other [11]. As is well known to all, the spoken language conveys linguistic, paralinguistic, and non-linguistic information. In previous studies, the entrainment phenomenon was investigated using linguistic-related parameters [6] or paralinguistic-related parameters [8, 10, 12] independently, but the correlation between the two aspects was less discussed. In the study [13], the authors discussed the entrainment relationship between linguistic information represented by word frequency and acoustic/prosody parameters, but failed to find the relationship between them. In this study, we utilize a more effective way to describe the linguistic information and analyze the relation between the linguistic information and paralinguistic information in the entrainment. To do so, we use BERT [4] to extract the semantic information as the representative of linguistic information, while fundamental frequency (F0), sound intensity and gap between the turns is treated as the representatives of para-linguistic information. Here, we try to find the relation between the linguistic and paralinguistic information. The interaction between them would be useful for address the mechanism of the entrainment rather than only consider linguistic and paralinguistic entrainment individually. The rest of the paper is structured as follows. Section 2 presents the dataset and the method. Section 3 details the relation between the semantic information and prosodic information in entrainment. Section 4 discusses the conclusion and the future work.

2 METHOD

In this section, we introduce the method to extract the semantic and prosodic information in our database, and the calculation of the semantic entrainment metrics.

2.1 Basic layout features

The data we used here were out from the DISCOURSE-CASS [9] – a corpus of Chinese spontaneous telephone conversation recorded in real service scenarios, such as taxi, airport, restaurant, insurance, and so on. 364 dialogues from this corpus were used in this study, which were annotated with multi-leveled information including turns, speaker gender, speech/non-speech labels, and so on.

2.2 Information extraction

2.2.1 Semantic information.

To parameterize the semantics of each turn in the dialogue, we calculated a “semantically similarity” between the same speaker using his/her two adjacent turns [2]. For quantitatively measuring this kind of semantic distance, our approach was based on the BERT model, we represent each turn into a fixed-length vector (in our case 768 dimensions). Each element of the vector encodes the semantics of the original turn. We then calculated the “semantic similarity” of each turn by comparing Pearson’s correlation between the 768-dimensional vector of the current turn with the next turn. The correlation provides a semantic similarity measure for each turn. This similarity is calculated for the same speaker using his/her two adjacent turns, turn by turn.

2.2.2 Gap information.

The gap in dialogue is defined as the silence between two adjacent turns and is attributed to the following speaker. All gap durations were calculated in the log scale. It is necessary to point out that the distribution of gap durations is typically positively skewed [3], which makes arithmetic means over estimators of central tendency. Mean durations in the log domain (or geometric means) may be better suited to describe the gap distribution. According to the study [5], we use the Time-Aligned Moving Average (TAMA)-method to quantitatively analyze the dynamic changes of the gaps in dialogues. This method is based on a moving window average, where the window length is nine points (turns) with a shift of one point.

2.2.3 Other prosodic information.

We used the Praat toolkit [1] to extract the following prosodic features: pitch (fundamental frequency in Hz), intensity (loudness in dB), the arithmetic max and mean for each feature are determined at the turn level in a dialogue. All of these features were also processed using the TAMA-method based on nine points (turns) moving window to describe the dynamic changes.

2.3 Semantic entrainment calculation

The entrainment in communication means the process in which two (or more) people speaking to each other tend to become more similar linguistically and prosodically. Entrainment has been investigated in numerous studies from different aspects, for example, lexical entrainment [6], phonetic entrainment [8], and acoustic-prosodic entrainment [10, 12]. These studies found that entrainment shows two basic tendencies: synchrony and convergence. Synchrony means that speakers are consistently behaving in a similar way, whereas convergence indicates that speakers progressively become more and more similar over a certain period. Both processes can present at the same time: two speakers may exhibit synchronous behavior

that deviates from what we would expect would happen by chance, and at the same time become more similar.

2.3.1 Convergence.

Convergence ($conv^{A,B}$) between f^A (service staff) and f^B (customer speaker) can be measured as the Pearson correlation coefficient between $-|f^A - f^B|$ and time t , which can be calculated as [7]:

$$D(t) = -|f^A - f^B| \quad (1)$$

$$conv^{A,B} = \frac{\int_{t_{st}}^{t_{end}} (D(t) - \bar{D})(t - \bar{t}) dt}{\sqrt{\int_{t_{st}}^{t_{end}} (D(t) - \bar{D})^2 dt \int_{t_{st}}^{t_{end}} (t - \bar{t})^2 dt}} \quad (2)$$

Where f^A or f^B is referred to as the extracted semantic information from the service staff or the customer speaker. t_{st} and t_{end} means the start and the end time of the dialogue. Positive/negative values of this metric indicate that f^A and f^B become closer to/further apart from each other as the conversation is proceeding.

2.3.2 Synchrony.

Synchrony between f^A (service staff) and f^B (customer speaker) ($sync^{A,B}$) can be measured as the Pearson correlation coefficient between f^A and f^B . We calculated $sync^{A,B}$ as [7]:

$$sync^{A,B} = \frac{\int_{t_{st}}^{t_{end}} (f^A(t) - \bar{f}^A)(f^B(t) - \bar{f}^B) dt}{\sqrt{\int_{t_{st}}^{t_{end}} (f^A(t) - \bar{f}^A)^2 dt \int_{t_{st}}^{t_{end}} (f^B(t) - \bar{f}^B)^2 dt}} \quad (3)$$

Positive values of $sync^{A,B}$ indicate that f^A and f^B behave in synchrony with each other, while negative values for the cases in opposite directions.

3 RESULTS AND DISCUSSION

3.1 Semantic entrainment

According to the calculation of entrainment, Figure 1 shows examples of conversations with high (a) and low convergence (b), and with high (c) and low synchrony (d). For high convergence, two semantic curves of the speakers gradually become more and more similar during a dialogue, while the high synchrony shows that speakers are consistently behaving in a similar way.

By our knowledge, the BERT-based semantic similarity feature has not been used for linguistic entrainment. Therefore, we first investigate the validity of the proposed similarity on convergence and synchrony. In this study, we use the Permutation Test (Bonferroni) to find out the reliable region for convergence and synchrony. It is found that the absolute value of the Pearson coefficient is larger than 0.38, the result would fall in region with the confidence values ($p < 0.05$). The results of synchrony and convergence are presented in Table 1. Over half of the dialogues are found with high synchrony (Person coefficient >0.38) between two interlocutors, while 41% of dialogues show a tendency of divergence (Person coefficient <-0.38) in the entrainment analysis.

3.2 Correlation between semantic information and prosodic information

Now we investigate the relationship between semantic and prosodic factors by computing Pearson correlation coefficient between the

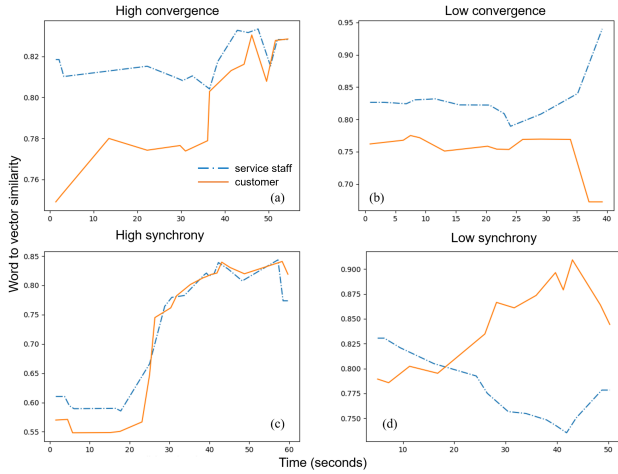


Figure 1: Sample conversations with different values of the semantic entrainment metrics (Notes: Dotted lines are used for service staffs; solid lines are for customer speakers)

Table 1: The number of dialogues that have significant semantic convergence and synchrony

Convergence		Synchrony	
High	Low	High	Low
91 (25%)	148 (41%)	201(55%)	70(19%)

semantic and prosodic entrainments from same interlocutors in a dialogue. By the Permutation Test (Bonferroni), we find that in most dialogues from our database, the semantic entrainment has significant correlation with prosodic (gap, F0 max and mean, intensity max and mean) entrainment produced by the same speaker (service staff or customer). Firstly, we investigate the significant correlations between semantic entrainment and specific prosodic factors that are negative, which can be further divided into two groups. Each group contains several dialogues from our database. The length of dialogues has been normalized. The values of semantic similarity and gap duration of each group have been averaged. As shown in Figure 2, the left panel represents the changes of the mean value of semantic similarity with time, and the right panel shows the changes of the mean gap duration with time. Figure 2(a) and (b) demonstrate correlation of one group, revealing that the similarity increases with time for a given speaker, his/her the gap duration shortened. It indicates that as the semantic similarity becomes closer, the speakers do not need more time to plan their conversation, so that the duration of the gap reduces as the dialogue goes on. As for the other group, conversely, Figure 2(c) and (d) shows a decline tendency of semantic similarity corresponding to an increasing gap duration for service staff, which yields a different type of negative correlation.

Secondly, we examine the positive correlations between semantic and prosodic factors that are of significance. Similar to the analysis

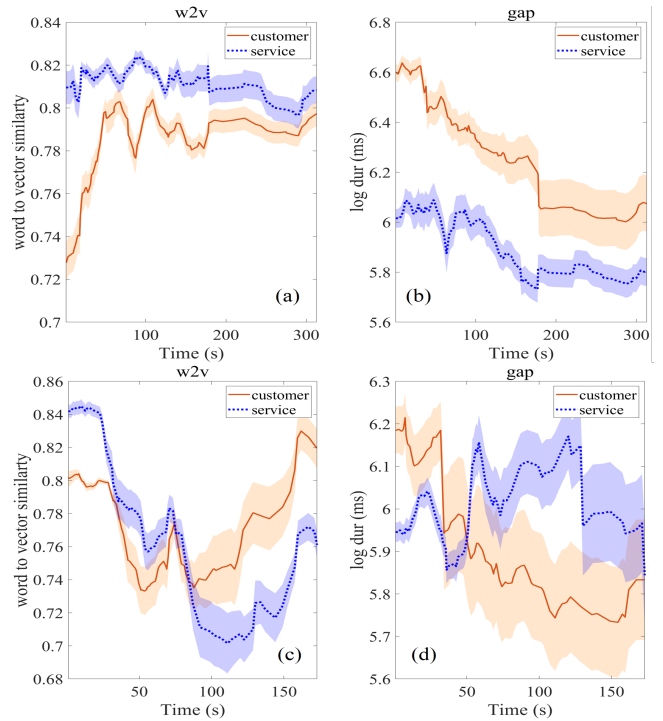


Figure 2: The results with negative correlation between semantic and gap entrainments

of negative correlations, Figure 3(a) and (b) show that the more similar the semantics for a given speaker, the longer the gap duration the speaker took. In (c) and (d), service staff shows another situation of positive correlation in which the semantic become dissimilar, the gap duration becomes shorter.

We also analyzed the correlation between semantic and other prosodic factors (F0 max and mean, intensity max and mean) and present the results in Table 2. The table shows the number of dialogues in which service staff or customer have strong correlation (absolute Pearson coefficient >0.38) between the semantic and prosodic information.

The negative correlation implies that the more similar the semantics, the smaller the F0 values and/or intensity values and vice versa. In general, when two interlocutors' semantics become closer, the interlocutor will be able to catch the content of the conversation more easily. Thus, it is not necessary for speakers to express their intention by emphasizing the prosodic parameters. For the cases with positive correlation, it seems to conflict with the previous explanation. One possible explanation is that one speaker may dominate the conversation while the other acts as a follower. Therefore, the conversation processes depend upon the dominating speaker, causing the positive correlation. In other words, the follower matches his/her semantic/prosodic styles to the dominating speaker. In service scenarios, the service staffs may have to follow the customers' behavior in general. We also noted that, there are some other cases in which an interlocutors show an impatient attitude to another interlocutor, which also causes positive correlation.

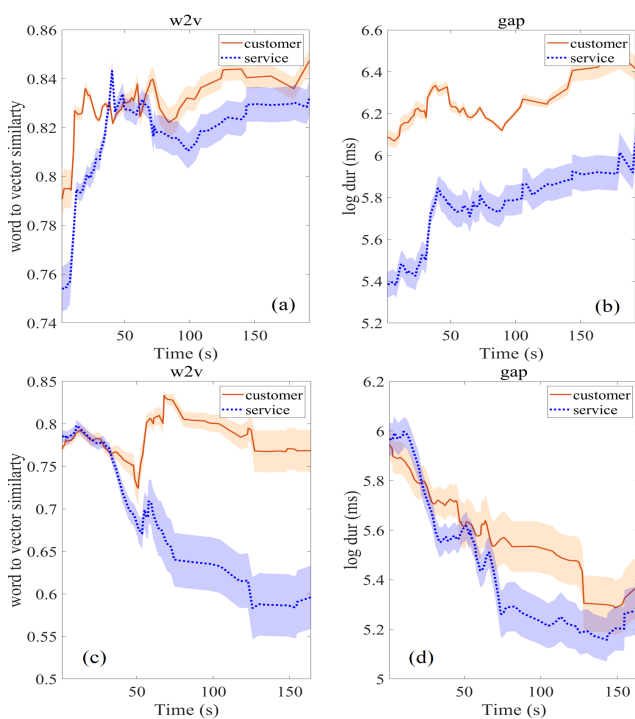


Figure 3: The data with a positive correlation between semantic and gap entrainments

Table 2: The number of the dialogues that have strong correlation between the semantic and different prosodic information for each interlocutor

	Service stuff		Customer	
	Pos	Neg	Pos	Neg
Semantic vs. Gap	119	129	97	114
Semantic vs. F0 max	113	109	112	126
Semantic vs. F0 mean	107	114	103	131
Semantic vs. Int max	112	114	94	139
Semantic vs. Int mean	91	126	86	143
Semantic vs. speech rate	133	100	138	89

4 CONCLUSION

In this study, we proposed “semantic similarity” to quantify the semantic information of dialogues, which was the Pearson’s correlation between the BERT-based word embedding of two adjacent turns. The semantic similarity gave a good measure for the entrainment of the conversations. It is found that in most cases, the entrainment relation between linguistic information and paralinguistic information is almost consistent. Therefore, in a smooth conversation, the prosodic features (gaps duration, F0 max and mean, intensity max and mean) of the interlocutors will probably show a tendency to decline as they can easily catch the content of the conversation. In the future, we will explore which factors influence the positive and negative correlations between linguistic

and paralinguistic, then investigate the mechanism of the entrainment based on the interaction between linguistic and paralinguistic information.

ACKNOWLEDGMENTS

This study is supported in part by JSPS KAKENHI Grant (20K11883), and in part by National Natural Science Foundation of China (No.61876126) and by the National Key R&D Program of China (2017YFE0111900) and “Four Batches” Talent Project “Chinese Intonational Typology”.

REFERENCES

- [1] P Boersma and D Weenink. 2019. Praat: doing phonetics by computer (6.1.08)[Computer program].
- [2] Michael P. Broderick, Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, and Edmund C. Lalor. 2018. Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology* 28, 5 (2018), 803–809.e3. <https://doi.org/10.1016/j.cub.2018.01.080>
- [3] Estelle Campione and Jean Véronis. 2002. A large-scale multilingual study of pause duration. *Speech Prosody 2002. Proceedings of the 1st International Conference on Speech Prosody* (2002), 199–202. http://www.isca-speech.org/archive/sp2002/sp02_199.html
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [5] Jens Edlund, Julia Bell Hirschberg, and Mattias Heldner. 2009. Pause and gap length in face-to-face interaction. (2009).
- [6] Heather Friedberg, Diane Litman, and Susannah B F Paletz. 2012. LEXICAL ENTRAINMENT AND SUCCESS IN STUDENT ENGINEERING GROUPS Department of Computer Science and 2 Learning Research and Development Center University of Pittsburgh, Pittsburgh, PA 15260. (2012), 404–409.
- [7] Ramiro H Gálvez, Lara Gauder, Jordi Luque, and Agustín Gravano. 2020. A unifying framework for modeling acoustic/prosodic entrainment: definition and evaluation on two large corpora. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 215–224.
- [8] Rivka Levitan, Štefan Beňuš, Agustín Gravano, and Julia Hirschberg. 2015. Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison. *SIGDIAL 2015 - 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference* September (2015), 325–334. <https://doi.org/10.18653/v1/w15-4644>
- [9] Aijun Li. 2018. Response Acts in Chinese Conversation: the Coding Scheme and Analysis. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 478–482.
- [10] Juan Manuel Pérez, Ramiro H Gálvez, and Agustín Gravano. 2016. Disentrainment may be a Positive Thing: A Novel Measure of Unsigned Acoustic-Prosodic Synchrony, and its Relation to Speaker Engagement. In *INTERSPEECH*. 1270–1274.
- [11] Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27, 2 (2004), 169–190. <https://doi.org/10.1017/s0140525x04000056>
- [12] Uwe D. Reichel, Katalin Mády, and Jennifer Cole. 2018. Prosodic entrainment in dialog acts. *arXiv* (2018), 1–19. [arXiv:1810.12646](https://arxiv.org/abs/1810.12646)
- [13] Andreas Weise and Rivka Levitan. 2018. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 2* (2018), 297–302. <https://doi.org/10.18653/v1/n18-2048>