Dialogue scenario classification based on social factors

Yuning Liu[1]*,Di Zhou*[1]*,Masashi Unoki*[1]*,Jianwu Dang*[2]*, Aijun Li*[3]

[1]School of Information Science, Japan Advanced Institute of Science and Technology
[2]College of Intelligence and Computing, Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University
[3]Institute of Linguistics, Chinese Academy of Social Sciences.

liuyuning@jaist.ac.jp

Abstract

The tendency of interlocutors to become more similar to each other in the way they speak, this behavior is known in the literature as entrainment, accommodation, or adaptation. Previous studies indicated that entrainment can be treated as a social factor in human-human conversations. However, previous research suggests that this phenomenon has many subtleties. One of these cues is that entrainment on an acoustic feature might be associated with disentrainment on another in conversation, which means we have to consider these features together. Therefore, we proposed a linear dimensionality-reduction method that combines acoustic features to calculate three entrainment metrics: proximity, convergence, and synchrony. The three entrainment metrics are referred to as social factors hereafter. Our results show these social factors play an important role in a classification task. We also found that these social factors perform a better classification accuracy than combining each individual acoustic feature's entrainment. The proposed social factors can help the human-machine interface to have the ability to adapt to the different scenarios in dialogue.

Index Terms: entrainment, human-computer interaction, social factors

1. Introduction

In everyday interpersonal conversation, people may consciously or unconsciously change their voices to adapt to their partner's voice. Motivated by Communication Accommodation Theory [1], speakers dynamically adjust their communication behaviors, converging to or diverging from their interlocutors to diminish or increase social distance, which is called entrainment. Thus, entrainment can be treated as a social factor and affects many linguistic features such as phonetics [2, 3], linguistic style [4], turn-taking [5], and prosody [6, 7] as well as non-linguistic behavior [8, 9]. For example, people may talk faster and louder when another participant in the argument raises their speaking speed and voice, or may speak more slowly and softly when comforting someone. Previous studies have shown that people who adapt to their partner's speech are considered more socially attractive.

Even when acoustic entrainment has consistently been reported to occur and correlate with so many social outcomes, such as dialogue success [10] or social implications [11], however, previous research suggests that the phenomenon has many subtleties [12]. One piece of evidence indicates that entrainment on an acoustic feature might be associated with disentrainment on another [13]. Therefore, we have to consider these features together in entrainment studies. However, the study of entrainment on acoustic features thus far has been fragmented, with researchers considering numerous individual features and measuring them in various ways. Still, few studies have focused on the relasion among these features.

To make conversation successful or smooth, people usually use different strategies in different types of conversations. Therefore, this research used a small corpus that contains four scenarios from a TV drama. We hypothesize that since entrainment can be treated as a social factor and affect many acoustic features, it may help improve the accuracy of the classification result of dialogue scenarios. We speculate that this social factor in a dialogue affects the changes in all these acoustic features. Thus, it is also necessary to discuss entrainment metrics in combined acoustic features instead of individual ones.

In summary, the present article has a twofold purpose. First, this study proposed a linear dimensionality-reduction method to combine and reduce the dimensionality of acoustic features. With this method, we calculate social factors and investigate whether these social factors can help improve the accuracy for classifying dialogue scenarios. Secondly, to verify whether the social factors perform better than the previous entrainment metrics calculated from individual acoustic features, this study compared the entrainment metrics calculated from individual acoustic features and the proposed social factors in a classification task. The results indicate that the proposed social factors also play an important role in a classification task, and perform better than previous entrainment metrics.

The rest of the paper is structured as follows. Section 2 introduces how we collected and selected the data for our study and presents statistics of the corpus. Section 3 provides details on our proposed method. Section 4 presents results from our empirical evaluation. Sections 5 and 6 provide a discussion and conclusion, respectively.

2. Data collection and selection

2.1. Corpus

To verify our hypothesis, we first extracted four different types of dialogues from the Chinese TV drama "Golden Marriage" and annotated them. The data-selection process is based both on transcription and original audio data. There are three criteria to determine whether a dialogue is eligible: (1) Video quality (The background music in many scenarios is noisy, which leads to low-quality audio. If the noise (background music) is so strong that it covers up the speaker's voice, it is difficult for later acoustic information extraction. This information loss will block the following process); (2) Dialogue length (Six or more turns are required. Note that our corpus is restricted to dialogues between two speakers. If a dialogue contains less than six turns, it will be discarded); and (3) Dialogue content (We were restricted to

Table 1: *Statistics of our corpus*

	Number of dialogue	Duration/per dialogue(second)	Number of turns/per dialogue	Total duration(minute)
Arguing	64	76.7	13	81.8
Comforting	67	82.5	11	91.1
Convincing	41	79.6	11	54.4
Sharing happiness	21	61.0	12	21.3
Total	193	/	/	249.8

four scenarios, i.e., arguing, convincing, comforting, and sharing happiness, we expect to collect conversations that represent the specified scenarios.)

Five annotators were in charge of transcription and annotation for each utterance and turn in a dialogue. The four dialogue scenarios were selected manually. The five annotators discussed at length during the selection stage, followed by a majority voting to decide the final dialogues scenario. To obtain the timestamp of each utterance, we extracted their corresponding audiovisual clips from the source episode and extracted audio content from these clips. We formatted the audio files as 16-bit PCM WAV files for further processing. Multi-layer annotations, including segmental, were manually transcribed into orthographic texts (Chinese characters), and scoring of the dialogue and emotion information were done in praat toolkit [14] in accordance with the dialogue scenarios.

After strict selection, we collected 193 dialogues of the four scenarios. The average length of a dialogue was about 75 s, and on average, a dialogue contained 11.75 turns. All the selected transcriptions were further manually corrected and cut into turns by our annotators. Table 1 lists the statistics of this corpus. Table 1 listes the statistics of this corpus.

The annotation was designed with a framework based on previous annotation systems [15]. Each level is time-aligned to the audio data. Specifically, each level is defined as follows. (1) Turn: to take count of the speaker's changes in the conversation. (2) Speaker property: dominant or follower. (3) Orthographic Transcription: manually corrected text with subtitle file. (4) Scenario: arguing, comforting, convincing, and sharing happiness.

2.2. Feature extraction

According to previous research [16], we used the VOICESAUCE [17] toolkit to extract the following acoustic features: fundamental frequency (f0), energy, harmonic-to-noise ratio (HNR), and subharmonic-to-harmonic ratio (SHR). Since the envelope information can show the rhythmic features of the speech [18], we used a window length of 25 ms with a frameshift of 1 ms to extract the envelop information in a dialogue. Each feature is determined at the turn level in a dialogue.

3. Method

3.1. Linear dimensionality-reduction method for social factor's extraction

We present our proposed method for combining and reducing the dimensionality of the acoustic features to further calculate social factors.

Given data matrices $\{X_i\}$ of the ith dialogue which defined as (1), where C and K are the numbers of acoustic features(in our cases, is ten(five from follower speaker, five from dominator speaker)) and sampling points,

$$X_i = [x_{i,1}, x_{i,2}, ..., x_{i,K}] \in \mathbb{R}^{C \times K} \tag{1}$$

$\{\mathbf{P}_n\}_{n=1}^{2}$ (e.g., $n = 1$ means the data from one scenario, $n = 2$ means other scenarios data) are covariance matrices of data matrix $\{X_i\}$. We assume that there is a linear transcription w to make the variance ratio between $P1$ and $P2$ to be maximum, which is expressed as (2).

$$\max_{\mathbf{w} \in \mathbb{R}^C} \frac{\mathbf{w}^T \mathbf{P}_1 \mathbf{w}}{\mathbf{w}^T (\mathbf{P}_1 + \mathbf{P}_2) \mathbf{w}}, \tag{2}$$

where the P_n can be calculated as:

$$\boldsymbol{P}_n = \frac{1}{|M_n|} \sum_{i \in M_n} \mathbf{X}_i \mathbf{X}_i^T. \tag{3}$$

Here, M_n denotes the index set of the data associated with the nth scenario.

The generalized eigen value decomposition (EVD) obtains the solution to Eq.(2). Therefore, to maximize this equation, we should set w equal to the (generalized) eigenvector to the largest (generalized) eigenvalue. The training process for w is shown in Fig. 1.

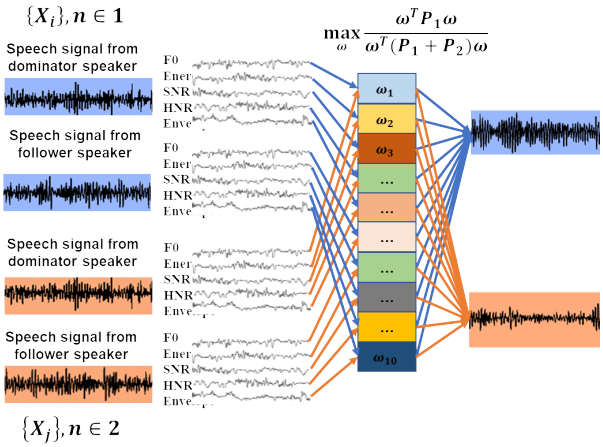

Figure 1: *Training process of proposed method*

As shown in Fig. 1, $\{X_i\}$ denotes one scenario and $\{X_j\}$ denotes the others, and by maximize the variance ratio between $P1$ and $P2$, we can obtain the optimal linear transcription vector w. In our case, since we have four scenarios of dialogue, we need to construct the optimal linear transcription vectors (w_1, w_2, w_3, w_4) for each pair of classes.

Fig. 2 shows how to consider the relation of these features by using the trained w. $\{X_t\}$ denotes the data that were not in

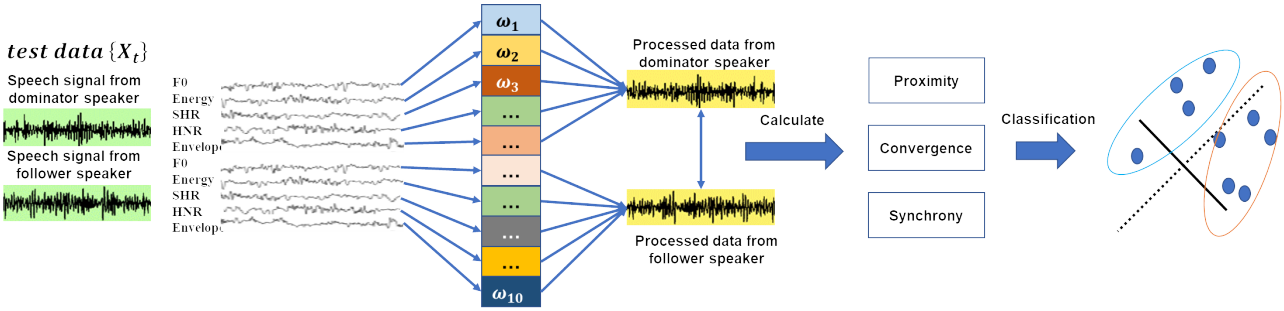

Figure 2: *Calculation procedure of the processed social factors*

the training process. By the linear transformation of w, we can reduce the original dimension of the dominant speaker and follower speakers' data to a low dimension (in our case, 1), and use these processed low dimension data to calculate the social factors. The calculation for the entrainment metrics is introduced in the next subsection.

3.2. Calculation of entrainment metrics

Previous studies [19] found that entrainment can be represented with three metrics: proximity, synchrony, and convergence. Proximity shows the features of different speakers having similar mean values across partners over the entire conversation. Synchrony means that speakers are consistently behaving in a similar way, whereas convergence indicates that speakers progressively become similar over a certain period.

3.2.1. Proximity

We define f^A and f^B as the processed data from dominant speaker and follower speakers, respectively. The proximity between f^A and f^B $\left(prox^{A,B}\right)$ can be measured as the negated absolute difference of the mean values of f^A and f^B [16], that is,

$$-\left|\bar{f^A} - \bar{f^B}\right| \tag{4}$$

where $\bar{f^A}$ and $\bar{f^B}$ stand for the mean value of f^A and f^B, respectively. When $\left(prox^{A,B}\right)$ is close to zero, f^A and f^B are on average close to each other, while when it is far from zero, they are distant.

3.2.2. Convergence

Convergence $\left(conv^{A,B}\right)$ between f^A and f^B can be measured as the Pearson correlation coefficient between $-\left|\bar{f^A} - \bar{f^B}\right|$and time t, which can be respectively calculated as (5)(6) [16]

$$D(t) = -\left|f^A - f^B\right| \tag{5}$$

$$conv^{A,B} = \frac{\int_{t_{\text{st}}}^{t_{\text{end}}} (D(t) - \bar{D}) \cdot (t - \bar{t}) dt}{\sqrt{\int_{t_{\text{st}}}^{t_{\text{end}}} (D(t) - \bar{D})^2 dt \cdot \int_{t_{\text{st}}}^{t_{\text{end}}} (t - \bar{t})^2 dt}} \tag{6}$$

where $\bar{D}$ and $\bar{t}$ denote the mean values of $D(t)$ and t, respectively. Positive/negative values of this metric indicate that f^A and f^B become closer to/further apart from each other as the conversation proceeding.

3.2.3. Synchrony

Synchrony between f^A and f^B $\left(sync^{A,B}\right)$ can be measured as the Pearson correlation coefficient between f^A and f^B. We calculated $\left(sync^{A,B}\right)$ as (7) [16]

$$sync^{A,B} = \frac{\int_{t_{\text{st}}}^{t_{\text{end}}} \left(F^A(t)\right) \cdot \left(F^B(t)\right) dt}{\sqrt{\int_{t_{\text{st}}}^{t_{\text{end}}} \left(F^A(t)\right)^2 dt \cdot \int_{t_{\text{st}}}^{t_{\text{end}}} \left(F^B(t)\right)^2 dt}} \tag{7}$$

where $F^A(t) = \left(f^A(t) - \bar{f^A}\right)$, $F^B(t) = \left(f^B(t) - \bar{f^B}\right)$. Positive values of $\left(sync^{A,B}\right)$ indicate that f^A and f^B behave in synchrony with each other, while negative values indicate the opposite directions.

After calculating these three metrics, as shown in Fig. 2, we use these metrics to conduct a classification task for dialogue scenarios.

4. Results

We conducted a series of experiments to investigate whether the social factors calculated with the proposed method are effective in classifying different dialogue scenarios.

The support vector machine (SVM) was employed as the classification method in our study. The kernel function of the SVM was set to a polynomial kernel. To deal with unbalanced classes, we randomly chose 15 clips from each type of scenario. We used 80% of these data to train the model, and the rest to test it. The training and testing procedure was repeated 100 times.

4.1. Combination of acoustic features and social factor

According to our hypothesis, entrainment has different patterns in different dialogue scenarios, which means the social factors should be effective in the classification task for different dialogue scenarios. Therefore, we used our social factors to conduct a classification test on the corpus we mentioned in 2.1. The performance of multi-classification is shown in the Fig. 3.

We conducted the Permutation Test (Bonferroni) to determine the chance level of the classification task. As shown in Fig. 3, we found that both the acoustic features and our social factors have higher classification accuracy than the chance level. The difference in the accuracy of our social factors and the acoustic features was insignificant, and the combined features have the highest classification accuracy than others, which proves that our social factors have information different from the acoustic features and play an important role in classification tasks.

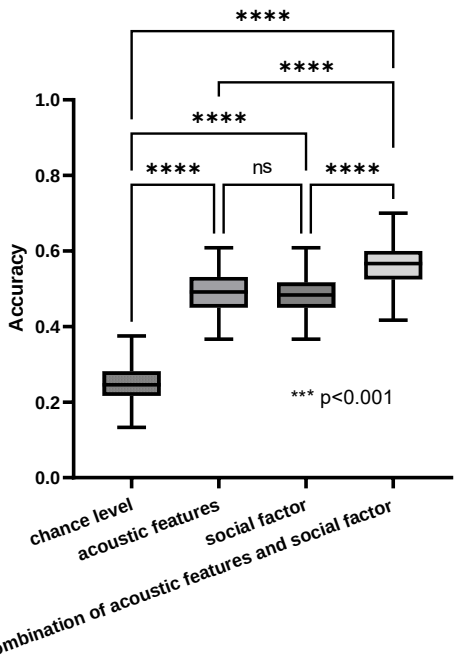

Figure 3: *Classification results with acoustic features and social factors*

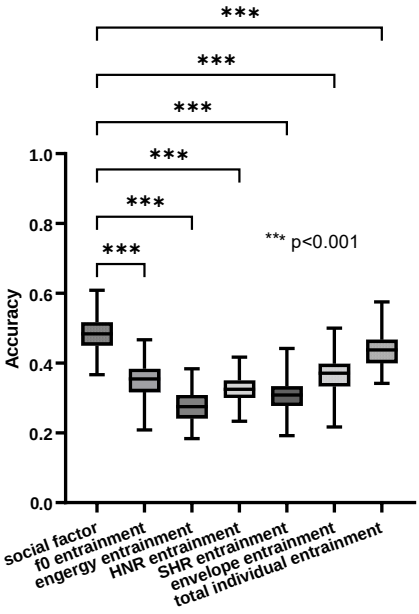

Figure 4: *Comparison of social factors and individual entrainment metrics*

4.2. Comprehensive social factors vs. individual entrainment metrics

In a dialogue, the two interlocutors' social behaviors would comprehensively affect each acoustic feature. Therefore, our social factors should perform better in a classification experiment.

As shown in Fig. 4, according to the analysis of variance test (ANOVA-test) we conducted, the proposed social factors shows the best classification result (0.486) even combine each individual acoustic features' entrainment together (0.435). These results are consistent with our hypothesis.

5. Discussion

Although previous studies predicted the naturalness and satisfaction of a dialogue using individual entrainment calculated from acoustic features. The results do not obviously higher than the chance level. The reason may be due to considering these acoustic features individually. On the basis of our hypothesis, we believe that the entrainment in a dialogue affects the changes in all these acoustic features. Therefore, the relation between these features is also important for measuring the entrainment. For this reason, we proposed a linear dimensionality-reduction method for combining and reducing the dimension of acoustic features to calculate social factors. According to our results, the social factors based on our hypothesis performed better classification results than any individual entrainment metrics, even better than combining each individual entrainment metrics. Our results indicate that the proposed method can describe this social factor more accurately.

In addition, this research compared the classification accuracy between acoustic features and social factors calculated from our method. We found that the accuracy for classifying the dialogue scenarios is highest when combining the entrainment metrics and the acoustic features. These results indicate the importance of these social factors in the classification task. It implies these social factors should be considered in future human-machine interface tasks.

6. Conclusions

We proposed a linear dimensionality-reduction method for combining and reducing the dimension of acoustic features to calculate three entrainment metrics: proximity, convergence, and synchrony, which are referred to as social factors in our study. We conducted an experiment involving a corpus containing four different dialogue scenarios to determine whether the social factors calculated with our proposed method provide valuable information for classifying different dialogue scenarios. Our results indicate the proposed social factors have some additional information different from the acoustic features. And by combining them with acoustic features, we can get a better classification result than using only the acoustic feature, which may, to some extent, address questions raised in previous studies that entrainment on an acoustic feature might be associated with disentrainment on another and indicate the potential of the proposed method in future human-machine interface tasks.

7. Acknowledgements

This research was supported by the National Natural Science Foundation of China (No.62276185), a Grant-in-Aid for Scientific Research (B) (No. 21H03463), a Grant-in-Aid for Scientific Fund for the Promotion of Joint International Research [Fostering Joint International Research(B); 20KK0233], Project of Cultural Experts and "Four Batches" of Talents awarded to Aijun LI, and in part by JSPS KAKENHI Grant (20K11883).

8. References

[1] H. Giles, N. Coupland, and I. Coupland, "1. accommodation theory: Communication, context, and," *Contexts of accommodation: Developments in applied sociolinguistics*, vol. 1, 1991.

[2] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[3] 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*, 2012, pp. 11–19.

[4] K. G. Niederhoffer and J. W. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 337–360, 2002.

[5] R. Levitan, A. Gravano, and J. B. Hirschberg, "Entrainment in speech preceding backchannels," 2011.

[6] 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*, 2015, pp. 325–334.

[7] J. M. Pérez, R. H. Gálvez, and A. Gravano, "Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement." in *INTERSPEECH*, 2016, pp. 1270–1274.

[8] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction." *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999.

[9] 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*, 2012, pp. 11–19.

[10] N. Campbell and S. Scherer, "Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[11] Š. Beňuš, "Social aspects of entrainment in spoken interaction," *Cognitive Computation*, vol. 6, no. 4, pp. 802–813, 2014.

[12] R. H. Gálvez, A. Gravano, Š. Beňuš, R. Levitan, M. Trnka, and J. Hirschberg, "An empirical study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars," *Speech Communication*, vol. 124, pp. 46–67, 2020.

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

[14] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat. org/*, 2006.

[15] K. Zhou, A. Li, Z. Yin, and C. Zong, "Casia-cassil: A chinese telephone conversation corpus in real scenarios with multi-leveled annotation," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[16] R. H. Gálvez, L. Gauder, J. Luque, and A. Gravano, "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*, 2020, pp. 215–224.

[17] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," *Energy*, vol. 1, no. H2, pp. H1–A1, 2010.

[18] T. Wu, L. Zhao, and Q. Zhang, "Research on speech synthesis technology based on rhythm embedding," in *Journal of Physics: Conference Series*, vol. 1693, no. 1. IOP Publishing, 2020, p. 012127.

[19] R. Levitan and J. B. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," 2011.

[This paper was published at ISCSLP 2022]