Relationship between Speaker/Listener Similarity and Information Transmission Quality in Speech Communication

Bohan Chen*, Norihide Kitaoka† and Kazuya Takeda*
* Nagoya University, Nagoya, Japan
E-mail: {bohan.chen, takeda}@g.sp.m.is.nagoya-u.ac.jp
† Tokushima University, Tokushima, Japan
E-mail: kitaoka@is.tokushima-u.ac.jp

***Abstract*—We investigate the correlation between similarity in speaker characteristics and information transmission quality using a map task dialogue corpus. Similarity between the prosodic features and lexical styles of different speakers are analyzed, and most of these similarity measurements are shown to have significant correlations with information transmission quality as measured by a direction following task. We also combine these similarity measurements using a linear regression prediction model and assess information transmission quality. Prediction scores show a significant 0.37 correlation coefficient between the combined similarity measurement and information transmission quality scores.**

I. Introduction

As a result of communication technology improvement, nowadays, it is more and more convenient for us to talk with each other. It is also not hard to imagine in the future days, that robots will join our everyday conversation and become one of our conversation partners. Therefore, developing an automatic estimation of dialogue success would be a helpful work. On one hand, in human-human dialogues like call center, prediction of dialogue success can help us to find out problems at early stage and do something to avoid guests' disappointment (e.g. change the operator). On the other hand, in human-computer dialogues, dialogue success can be considered as a general goal of our system. Machine learning techniques with the prediction can then be used for searching an optimal dialogue strategy.

In this study, we attempt to solve this problem by using the interlocutors' information presenting style similarity as the measure of their dialogue success. One reason is that there are evidences to show that familiarity (in most of the time can be understood as similarity) has an apparent facilitative effect on information transmission quality. Another reason is that conversational alignment has also been shown to positively correlated with task success. It means that interlocutors who share similarity in information transmission styles would increase their transmission quality. Moreover, the dialogue would be aided if they try to imitate their partner's information transmission styles during the dialogue.

The relationship between dialogue success and interlocutors' presenting style familiarity is suggested by the schema theory. Because when a receiver has relevant background knowledge, he or she can free up more working memory space for analysis and interpretation of the current message [2]. Evidences on several linguistic and conceptual levels have already been found. Use of familiar topics can help foreign language learners improve their performance on reading interpretation tasks, no matter which second language they are learning [3] or what their native language is [4]. Moreover, the facilitative effect of comprehension is revealed by simple nativization processes, such as the changing of character and location names into native ones (e.g., when a Japanese English learner replaces "Barack Obama of Washington D.C." with "Shinzo Abe of Tokyo" [5].

Interactive Alignment Model also suggests that dialogue success link to the similarity between interlocutors[6]. The author claim that, "the linguistic representations employed by the interlocutors become aligned at many levels, as a result of a largely automatic process. This process greatly simplifies production and comprehension in dialogue." Several studies have shown that conversational alignment is an effective predictor of dialogue success in different language and task conditions [1] [7]. As conversational alignment is a general communication skill, the similarity level between interlocutors would additionally represent the skillfulness or even the motivation level of the speakers, which are obviously relate to the dialogue success.

In this study, we define the dialogue success as the information transmission quality. And defined speaker characteristics as features which can be used to identify a particular speaker. The research question which guides this study is as follows: During a dialogue, can the degree of similarity in speech characteristics between the information sender and the information receiver be used to predict the quality of information transmission between them?

II. HCRC map task corpus [8]

The HCRC Map Task Corpus is a set of 128 direction sharing dialogues which have been recorded, transcribed, and annotated, and released for research on a wide range of behaviors [8]. There are 64 speakers featured in the corpus, all

of whom were born in Scotland. Each of the participants took part in four conversations. Sets of two participants take turns playing the roles of a giver and a follower of directions. During the dialogue, the direction giver describes a route that appears on his or her own map to the direction follower, using speech communication only, and the direction follower then tries to reproduce the same route on his or her own map. After the dialogue is finished, both the direction giver's and directions follower's A3 sized maps are covered with a grid of 1 cm squares, and the difference between their routes, measured in squares, is then calculated. This path deviation value is then used to measure the successfulness of the dialogue.

The measurement of information transmission quality in our study, represented by E in the following equation, can therefore be expressed as: $E = \frac{D}{C}$, where D is path deviation and normalized by the "length" of the correct route C.

III. Measurement of similarity in speaker characteristics

In this paper, we defined speaker characteristics as features which can be used to identify a particular speaker. We investigate the correlation between speaker characteristics at the prosodic and idiolect level and their impact on quality when transmitting information.

A. Prosodic similarity measurement

1) Prosodic distribution similarity: In [9], researchers introduced a method to approximate pitch and intensity contours using a linear combination of Legendre polynomials, which can be expressed as:

$$f(t) = \sum_{i=0}^{M} a_i P_i(t), \tag{1}$$

where $f(t)$ is the pitch/intensity value at time t, a_i is the linear coefficient, and $P_i(t)$ is the *i*-th order of Legendre polynomials. In order for these coefficients in Eq. (1) to be comparable across speech segments, we first scale and map the duration of all of the segments to the same interval $[-1, +1]$. For each segment, we used six coefficients to represent the segment's pitch contour and another six coefficients to represent its intensity contour. These pitch and intensity coefficients, in addition to the segment's duration, produce a 13 dimensional feature vector which we can then use for GMM modeling. Finally, a log-likelihood ratio (LLR) is used to measure the similarity between two GMMs.

2) Similarity in prosodic dynamics: In [10], researchers used the slope of both pitch and intensity contours, in conjunction with segment duration to encode the prosodic dynamic characteristics (or "pitch accent" as the authors claimed) of speech segment. After segmenting the wave file into syllable-like units, we first calculate the slope of both pitch and intensity, and code positive slopes as "+" and negative slope as "-". In addition, segment duration is coded as either **S**, **L**, or **M**, with **S** representing the shortest 33% of segment durations, **L** representing the longest 33% of segment durations, and **M** representing all segment durations in between. This means that for each speech segment, three symbols will be used to encode the slope of the pitch contour, the slope of the intensity contour and the duration of the speech segment (e.g., ++S or +-M). A bigram model is then used to model all of the coded words spoken by a particular speaker. To measure the similarity of two bigram models constructed from the speech of the direction giver and direction follower in a current dialogue, respectively, we use KL2 divergence, which can be expressed as:

$$D_{KL2}(P, Q) = -[\sum_{i=1}^{n} P_i \log \frac{P_i}{Q_i} + \sum_{i=1}^{n} Q_i \log \frac{Q_i}{P_i}], \tag{2}$$

where P_i is the probability of the *i*-th prosodic dynamic bigram (e.g. $Pr(+ + L| + -L)$) of the direction giver, Q_i is the probability of the same *i*-th prosodic dynamic bigram of the miliar , and n is the number of possible bigrams (in this case, $n = (2 \times 2 \times 3)^2 = 144$).

B. Idiolect similarity measurement

1) Relative frequency of keywords: One simple method of capturing a speaker's idiolectal speech is to count certain keywords [11]. Backchannel responses such as "right", "yeah", "really", etc. are often used as keywords. In this study, because we use map task dialogue as our analysis data, we used a set of keywords related to different styles of direction and distance representation. The keywords we counted were categorized into five style groups as follows:

- Absolute direction I: south, north, west, east;
- Absolute direction II: up, down;
- Relative direction: turn right, turn left;
- Imperial length unit: inches;
- Metric length unit: centimeters.

Note that keywords are counted by style group. This means that we used the total relative frequency of keywords to create a 5 dimensional vector as a representation of each speaker's idiolectic direction and distance representation style. For similarity measurement, we used the Euclidean distance between two vectors calculated from the speech of two speakers (using the direction giving parts of the dialogue only). We do not use cosine distance but Euclidean distance here because we consider the origin point would represent some special presenting styles (e.g. like the "u" shape curve we mentioned above).

2) Bigram model of POS tags: Textual bigram models are a common method of modeling text data [12]. As we only have an average of about 900 words of each direction giver's speech, in this study we used part-of-speech (POS) tagging instead of real words to build bigram models for each speaker. The POS tags we use in this study are: verbs, nouns, adjectives, adverbs, auxiliary verbs, determiners, pronouns, prepositions, conjunctions, interjections and punctuation (which means a silence separating POS units). The similarity measurement used in Section III-A2 is also used here.

IV. Experiment

A. Experimental setup

Firstly, silence part of the speech was removed manually. Pitch and intensity contours were extracted using Praat [14] every 10 ms using a 25 ms analysis window. After extraction, we used linear interpolation to fill the zero values, and then took log values of every data point. We used the algorithm proposed by [13] to detect syllable-like unit for prosodic similarity analysis. For our GMMs, we used a five mixture GMM to model the prosodic distribution features (Section III-A1). In order to avoid the sparseness problem, when training the bigram model we used data from all of a particular speaker's speech. When using the GMM to model the distribution of prosodic features, we used the direction giver's part of the current dialogue and the direction follower's part of another dialogue, in which the direction follower in the current dialogue played the direction giver with a previously unknown direction follower in another dialogue, to train our direction giver and direction follower speaker models, respectively. For convenience, we changed all of our measures into similarity measures, which means we multiplied their value by -1 when there was a negative similarity correlation.

B. Correlation between prosodic similarity of speakers and information transmission quality

Correlation between prosodic distribution similarity and information transmission quality (Section III-A1), is shown in Fig. 1, and the correlation between prosodic dynamics similarity and information transmission quality (Section III-A2) is shown in and Fig. 2. The correlation coefficients was 0.13 ($p = 0.14$)for prosodic distribution similarity, and for KL2 similarity of the prosodic dynamics bigram model was 0.31 ($p < 0.01$). These results suggest that prosodic presentation style influences our ability to understand speech, especially when we consider its dynamic transition.

C. Correlation between speakers' idiolectal similarity and information transmission quality

The correlation between similarity in the relative frequency of keywords and information transmission quality (Section III-B1) is shown in Fig. 3, and the correlation between similarity in POS tagger bigram models and information transmission quality (Section III-B2) is shown in Fig. 4. Both of the correlation coefficients are significant ($p < 0.01$). The correlation coefficient for keyword relative frequency similarity was 0.35, and for KL2 similarity of POS tagger bigram models was 0.27.

D. Multiple linear prediction model

Finally, we tried to combine all of the similarity measures mentioned in this paper to construct a prediction model. We used multiple regression as our prediction model. The linear coefficients were calculated using the least square method. We used leave-one-out cross-validation to test the accuracy of our prediction. As in the HCRC map task corpus, the participants in our study took part in eight dialogues (called a "quad" in the corpus); we actually left all eight of these dialogues out when training the prediction model, however. Correlations between prediction scores and information transmission quality under different conditions are shown in Table 1. The highest correlation appeared when we combined the prosodic dynamic similarity measure and the keywords relative frequency similarity measure, which together achieved a correlation coefficient of 0.35. We consider 0.35 is still an acceptable high value (similar to [1]). Because our prediction model uses relatively simple features, some of the factors strongly associated with task successfulness like instruction understanding are not included in our analysis.

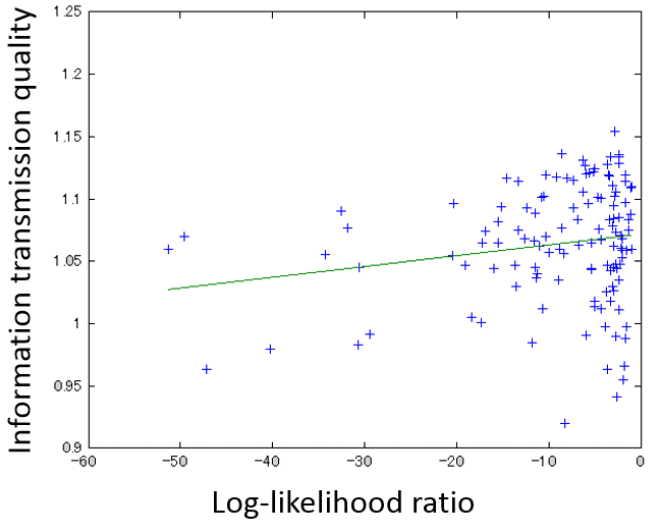

Fig. 1. Correlation between prosodic distribution GMM-LLR similarity and information transmission quality (correlation coefficient = 0.13)

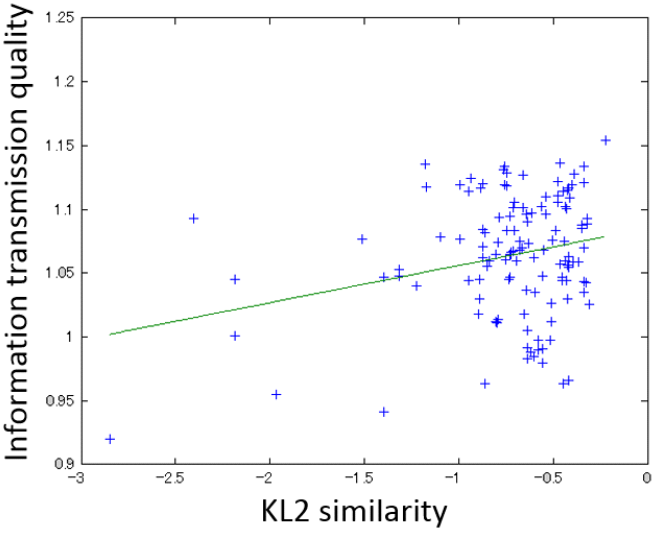

Fig. 2. Correlation between prosodic dynamic bigram KL2 similarity and information transmission quality (correlation coefficient = 0.31)

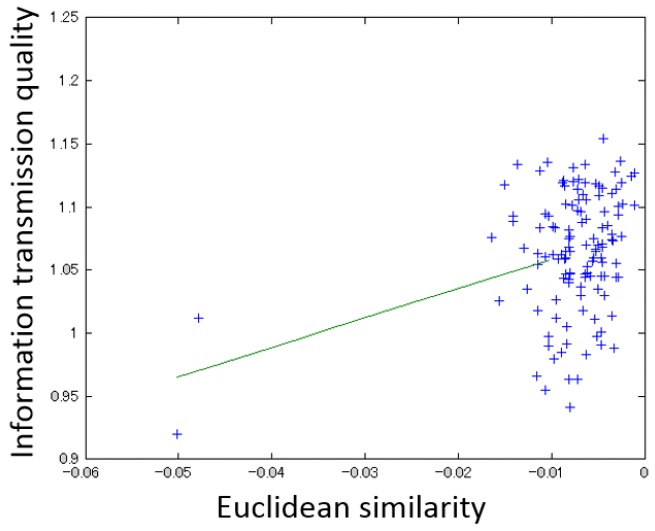

Fig. 3. Correlation between Euclidean keyword similarity and information transmission quality (correlation coefficient = 0.35)

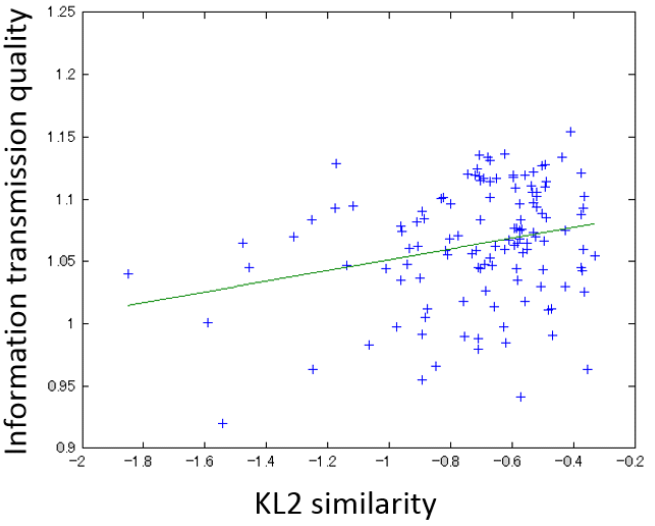

Fig. 4. Correlation between POS tag bigram KL2 similarity and information transmission quality (correlation coefficient = 0.27)

V. Conclusion

In this study we investigated relationships between the voice characteristics of speakers, such as prosodic and idiolectal similarity, and information transmission quality when speakers communicated using spoken language. Our results showed that speakers who used similar words, while exhibiting similar prosodic behaviors, tended to achieve higher levels of information transmission quality. Prosodic dynamic similarity, which is considered to represent intonation information, and keywords relative frequency similarity achieved two of the highest scores in our prediction model.

Since the participants in the HCRC map task were all born in Scotland, this implies that they shared similar prosodic dynamic features and vocabularies. The correlation coefficient between these similarity measures and information transmission quality seem to support that our prosodic processing of intonation is more sensitive than we generally believe.

In this paper, the keywords set we used is simple and only about 3% of the total words spoken are counted as keywords. We plan to increase our keyword list and improve our similarity measurement method in the future for catching more lexical information. Furthermore, our multiple regression model, combining different similarity measurement methods, did not achieve good performance. As a result, our combination method is another aspect of our approach which needs to be improved.

TABLE I
Correlation between prediction score and information transmission quality

Similarity measurement	correlation coefficient
prosodic_GMM	0.02
prosodic_bigram	0.25
keywords relative frequency	0.27
POS_bigram	0.14
keywords+POS_bigram	0.35

References

[1] D. Reitter, and J. D. Moore, "Predicting success in dialogue." Annual Meeting - Association for Computational Linguistics, 45(1), 808-815, 2007
[2] H. Nassaji, "Schema theory and knowledge based processes in second language reading comprehension: A need for alternative perspectives," Language Learning, 52(2), 439-481, 2002.
[3] M. J. Leeser, "Learner based factors in l2 reading comprehension and processing grammatical form: Topic familiarity and working memory," Language Learning, 57(2), 229-270., 2007.
[4] S. K. Lee, "Effects of textual enhancement and topic familiarity on korean efl students' reading comprehension and learning of passive form," Language learning, 57(1), 87-118, 2007
[5] I. H. Erten, and S. Razi, "The effects of cultural of arity on reading comprehension." Reading in a Foreign Language, 21(1), 60-77., 2009.
[6] M. J. Pickering, and S. Garrod, "Toward a mechanistic psychology of dialogue," Behavioral and brain sciences, 27(02), 169-190, 2004
[7] R. Nishimura, N. kitaoka, and S. Nakagawa, "Analysis of factors to make prosodic change in spoken dialog," Journal of the Phonetic Society of Japan,13(3), 66-84, 2009
[8] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The hcrc map task corpus," Language and speech, 34(4), 351-366., 1991.
[9] E. Grabe, G. Kochanski, and J. Coleman, "Quantitative modelling of intonational variation," Proc. of SASRTLM, 45-57, 2003.
[10] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. God- frey, "Modeling prosodic dynamics for speaker recognition," In Acoustics, Speech, and Signal Processing, 2003. (Vol. 4, pp. IV-788). IEEE., 2003.
[11] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for au- thorship identification of online messages: Writing style features and classification techniques," Journal of the American Society for Information Science and Technology, 57(3), 378-393., 2006.
[12] G. R. Doddington, "Speaker recognition based on idiolectal dif- ferences between speakers," INTERSPEECH (pp. 2521-2524)., 2001.
[13] A. Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," ICASSP 2003
[14] http://www.fon.hum.uva.nl/praat/.