Comparison of Emotional Understanding in Modality-Controlled Environments using Multimodal Online Emotional Communication Corpus

Yoshiko Arimoto[†§], **Kazuo Okanoya**[‡§]
†College of Engineering, Shibaura Institute of Technology, Saitama, Japan,
‡Graduate School of Arts and Science, the University of Tokyo, Tokyo, Japan,
§JST, ERATO, Okanoya Emotional Information Project, Saitama, Japan
ar@shibaura-it.ac.jp, cokanoya@mail.ecc.u-tokyo.ac.jp

Abstract

In online computer-mediated communication, speakers were considered to have experienced difficulties in catching their partner's emotions and in conveying their own emotions. To explain why online emotional communication is so difficult and to investigate how this problem should be solved, multimodal online emotional communication corpus was constructed by recording approximately 100 speakers' emotional expressions and reactions in a modality-controlled environment. Speakers communicated over the Internet using video chat, voice chat or text chat; their face-to-face conversations were used for comparison purposes. The corpora incorporated emotional labels by evaluating the speaker's dynamic emotional states and the measurements of the speaker's facial expression, vocal expression and autonomic nervous system activity. For the initial study of this project, which used a large-scale emotional communication corpus, the accuracy of online emotional understanding was assessed to demonstrate the emotional labels evaluated by the speakers and to summarize the speaker's answers on the questionnaire regarding the difference between an online chat and face-to-face conversations in which they actually participated. The results revealed that speakers have difficulty communicating their emotions in online communication environments, regardless of the type of communication modality and that inaccurate emotional understanding occurs more frequently in online computer-mediated communication than in face-to-face communication.

Keywords: Emotional Understanding, Empathic Accuracy, Multimodal Dialog Corpus, Computer-mediated Communication

1. Introduction

Communication has shifted from face-to-face communication to online computer-mediated communication. People often send an e-mail to deliver and receive important business information, exchange various messages per day using a messenger service, talk over the (Internet) phone, and conduct conferences with other people using a video chat conferencing system, thus, rarely meeting together. This computer-mediated communication is reasonable, convenient and indispensable for us, and there are no problems in performing our daily tasks rationally. The problem lies in the difficulty of communicating our emotion. Compared to face-to-face communication, speakers were considered to have experienced difficulties in catching their partner's emotions and in conveying their own emotions in an online computer-mediated communication environment because of the limited communication modality. In face-to-face communication, our emotions can be easily shared and communicated via non-verbal cues, i.e., facial expressions, vocal expressions, gestures, or other various attributions of the speaker. On the contrary, online computer-mediated communication restricts the use of our communication modality, including non-verbal cues for emotional communication. Therefore, conveying our emotion in computer-mediated communication is a more effortful task than in face-to-face communication.

Even in a face-to-face communication environment, speakers did not understand their partner's emotion well. Several attempts were made to investigate how accurately the listeners understood their partner's emotion (Ickes et al., 1990; Stinson and Ickes, 1992; Verhofstadt et al., 2008; Zaki et al., 2008; Arimoto and Okanoya, 2015). According to Ickes et al. (1990), listeners understand speakers' complex emotion (written descriptions of how they feel or think) with 21.7% accuracy ($SD = 12.1$) and understand speakers' valence (positive, neutral and negative) with 40.1% accuracy ($SD = 17.1$). The literature on the studies of emotional understanding is briefly reviewed in Decety and Ickes (2011).

There is contradicting evidence on computer-mediated emotional communication. The most famous and intuitive theory is the media richness theory, which argues that the lack of communication channels results in misunderstandings of the partner's emotion, feelings or thoughts (Daft and Lengel, 1986). Harada (1997) supported this theory and reported that the three different communication devices, i.e., video conference, telephone and text message (e-mail and chat), changed the speaker's subjective evaluation on online communication using these devices. However, Dennis and Kinney (1998) concluded that the new media (i.e. computer-mediated media in their paper) is rich enough to enable users to successfully communicate for these tasks. It was found that the speaker's induced emotion (not the acted one) can be sensed by the listener in text-based communication (Hancock et al., 2008). In a meta-analysis on the comparison of face-to-face and computer-mediated communication, Derks et al. (2008) concluded that emotion can be found as frequently in a computer-mediated situation as in a face-to-face setting and that emotion in a computer-mediated situation is more controllable than in face-to-face situations. Moreover, some studies reported that the interpersonal perceptions can be exaggerated in computer-mediated communication. The listener more strongly inferred the partner's feelings because of a limited number of cues in computer-mediated communication (Hancock and Dunham, 2001; Boucher et al., 2008).

In this project, to explain why online emotional communication is so difficult and to investigate how this problem should be solved, multimodal online emotional communication corpus were constructed by recording approximately 100 speakers' emotional expressions and reactions in a modality-controlled environment. In this environment, three types of online chat systems were adopted. Speakers communicated over the Internet using video chat, voice chat or text chat; their face-to-face communication was used for comparison purposes. The corpora incorporated emotional labels by evaluating the speaker's dynamic emotional states, measuring the speaker's facial expression and vocal expression, and measuring autonomic nervous system activity (ANS) (i.e., electromyogram (EMG), electrocardiogram (ECG) and electrodermal activity (EDA)).

For the initial study of this project using a large-scale emotional communication corpus, the accuracy of online emotional understanding was assessed to demonstrate the emotional labels evaluated by the speakers and to summarize the speaker's answers on the questionnaire regarding the difference between the online chat and the face-to-face conversations in which they actually participated. This paper reports the results on 1) whether speakers really experienced difficulty in exchanging their emotions accurately with their partners while they were communicating over the Internet and 2) whether the lack of communication modality actually caused the inaccurate emotional understanding.

2. Multimodal Online Emotional Communication Corpus

2.1. Participants

One hundred speakers participated in 7-minute dyadic dialogs with a friend of the same sex. Although fifty pairs talked with each other for the recording, ten pairs were excluded for subsequent analyses because of recording problems. The remaining eighty speakers had a mean age of 21.2 years ($SD = 2.03$). Forty-six of the speakers were female, and the remaining thirty-four were male.

The closeness between interlocutors varied based on pairs. Many of them were friends who belonged to the same club or were the closest classmate at university. The relationship of a few pairs were senior–junior who belonged to the same club at university or were friends from childhood.

2.2. Task

Each pair performed one task in two different environments. One of the environments was a face-to-face communication environment, and the other was an online communication environment using video chat (VD), voice chat (VC), or text chat (TX) (Fig. 1). The numbers of speakers were 30, 24 and 26 for VD, VC, and TX, respectively. The online chat services adopted for our recording were either Microsoft Skype or Google Hangout. The speakers were allowed to use emoticons (the icons express emotion) provided by the chat services. They had time to practice using those chat services before recording.

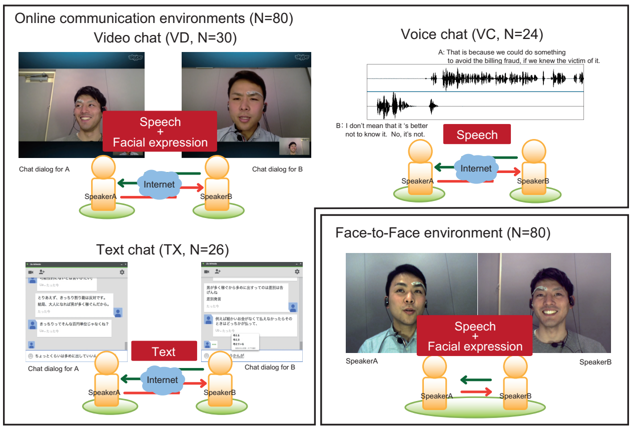
Online communication environments (N=80)
Video chat (VD, N=30)
Speech + Facial expression
Chat dialog for A
Chat dialog for B
Internet
SpeakerA
SpeakerB
Voice chat (VC, N=24)
A: That is because we could do something to avoid the billing fraud, if we knew the victim of it.
B: I don't mean that it 's better not to know it. No, it's not.
Speech
Internet
SpeakerA
SpeakerB
Text chat (TX, N=26)
とりあえず、きっちり割り勘は反対です。
結局、大人になれば男が多く稼ぐんだから。
たった今
きっちりってそんな百円単位じゃなくね？
Ua . たった今
ちょっとくらいは多めに出していいん
男が多く稼ぐから多めに出すってのは差別は告げんね
差別発言
たった今
例えば細かいお金がなくて払えなかったらそのときはどっちかが払って、
Ub . たった今
Text
Chat dialog for A
Chat dialog for B
Internet
SpeakerA
SpeakerB
Face-to-Face environment (N=80)
Speech + Facial expression
SpeakerA
SpeakerB

Figure 1: *Recording environments.*

As a task for the recording, they were instructed to discuss a topic about which they had opposing opinions. The topics were chosen by the experimenter with respect to the speaker's answers regarding what they thought about thirty-four social issues and topics. The top five topics used for the recording were as follows:

- Children should grow up receiving compliments without scolding. (18)
- Costs of dating should be equally split. (16)
- Educational background is important for success in society. (16)
- It is rude not to care about your grooming and appearance. (12)
- Social Networking Services (SNS) promote youth bullying and crime. (12)

The numbers in parentheses are the times used for the recording.

2.3. Recording and Measurements

The speakers' behaviors (speech and facial expressions) were recorded as a measure of their external reactions during the dialogs. Simultaneously, physiological reactions were recorded using electromyogram (EMG), electrocardiogram (ECG) and electrodermal activity (EDA) as measures of the speakers' internal reactions. Each speaker in the pair sat in a soundproofed room separated from the other person by soundproofed glass. They faced their partners through the glass in the face-to-face session, and they talked with each other using microphones and headphones. In the chat session, they talked over the Internet using a notebook computer. Their speech was recorded as separate channels of an audio stream, with a sample rate of 48 kHz and 16-bit precision. To record each speaker's behavior, a CCD camera and a built-in camera acquired images of the speaker. Both cameras were focused on the speaker's face. In the VD and TX session, the chat window operated by the speaker was captured using QuickTime 7 Pro. Each speaker's behavior was recorded at 30 frames per second in images of 640 × 480 resolution. For synchronous audio and visual recording, the Potato system (Library Co. Ltd., Tokyo, Japan) was used. EMG (the left corrugator supercilii and the zygomaticus major muscle region), ECG and EDA were recorded at a sampling rate of 1 kHz using a Biopac MP150 system (Biopac Systems, CA, USA). To ensure synchronous recording, the Biopac MP150 system began to record signals after it received a sync signal sent by the Potato at the time of the first camera shutter activity.

2.4. Questionnaire on Online Communication

After the recording, each speaker filled out a questionnaire about his or her communication by comparing the online session and the face-to-face session. To assess our belief that emotional communication over the Internet is considered to be more difficult than in a face-to-face environment, the forced-choice questions to choose either one of environments (face-to-face or online) were prepared. They answered the following three questions:

Question1(Q1): Which is easier for delivering your message? **Question2(Q2):** Which is better for understanding your partner's emotions? **Question3(Q3):** Which is easier for conveying your emotions to your partner?

These forced-choice questions compare three key aspects: Q1 compares the ease of delivering the message, Q2 compares emotional understanding, and Q3 compares the ease of conveying emotional information with a partner.

2.5. Emotional Labeling

Speakers also performed subjective evaluations for dynamic emotional state annotation using GTrace (Cowie et al., 2012). Speakers dynamically rated (1) their own internal emotional state and (2) their partner's emotional state during the recorded audio-visual video sequences of the 7-minute dialogs. Speakers were instructed to evaluate (1) how they experienced their own emotions while they were talking and (2) how they experienced their partner's emotion while they were talking. The target emotional states were pleasantness (pleasant–unpleasant), arousal (aroused–sleepy) and dominance (dominant–submissive). Pleasantness, arousal and dominance are the three dimensions of the psychological emotional theory of Mehrabian (1980). The mean evaluated values of each emotional state were calculated in 10-second intervals as a measure of the dynamic emotional states of the speakers.

3. Analysis

3.1. Questionnaire

To compare the message and emotional exchange between face-to-face and online communication based on the speakers' opinions, a binomial test was conducted with the speakers' answers to the questionnaire (null hypothesis: $p = 0.5$) regarding the communication environment (Chat or FaceToFace). To compare the answers between each face-to-face communication and online communication (VD, VC, or TX), binomial tests were also conducted with the answers grouped into each chat type (VD, VC, or TX).

3.2. Emotional Labeling

To measure emotional understanding using the evaluation of the emotional state, each pair's correlation coefficients between the listener's evaluation of the speaker's emotion and the speaker's self-evaluation of his or her emotion were calculated. The first and last 30-second emotional labeling periods in each dialog were removed from the data for this correlation coefficient calculation. The correlation coefficients indicate the accuracy of the listener's emotional understanding of the speaker. A positive correlation implies that the listener accurately understood the speaker's emotions throughout the dialog. In contrast, a negative correlation implies that the listener understood the speaker's emotions to be the opposite of what they actually were. No correlation implies that the listener did not understand the speaker's emotion at all. Then, a 2×3 analysis of variance (ANOVA) was performed on the factors of communication environment (Chat or FaceToFace) and chat type (VD, VD or TX) for each emotional dimension. Specifically, multiple comparison tests were performed with the Tukey-Kramer method to investigate interactions across the factors.

4. Results

4.1. Questionnaire

The binomial test revealed no significant difference between the communication environment (Chat vs FaceToFace) in the answers for Q1 (the left panel in Fig. 2). The tests also revealed no significant difference between FaceToFace and either VD or VC for the answers to Q1 (the two middle panels in Fig. 2). However, the binomial test revealed a marginally significant difference between FaceToFace and TX for the answers to Q1 ($p < 0.08$, the right panel in Fig. 2). The binomial test revealed a significant difference between Chat and FaceToFace in the answers for Q2 and Q3 ($p < 0.001$). Moreover, the test revealed a significant difference between FaceToFace and each chat type (VD, VC or TX, $p < 0.05$) for Q2 and Q3.

4.2. Emotional Labeling

Figure 4, 5, and 6 shows the results of the ANOVA and multiple comparison tests for each emotional dimension. The ANOVA for the arousal evaluation (Fig. 5) revealed a significant interaction between the communication environment and chat type ($F(2, 77) = 3.817, p < 0.05$). Multiple comparison tests revealed a significant difference between VC and TX ($p < 0.05$, including chat environment evaluation only) and a marginally significant difference between TX and FaceToFace ($p < 0.07$).

The ANOVA test for the dominance evaluation (Fig. 6) revealed a significant main effect on the communication environment ($F(1, 77) = 5.550, p < 0.05$) and a marginally significant main effect on chat type ($F(2, 77) = 2.526, p < 0.09$). Multiple comparison test revealed a significant difference between VC and TX ($p < 0.05$, including both the chat and face-to-face environment evaluation) among the three levels of chat type. There was no significant interaction across the factors.

The ANOVA test for pleasantness evaluation (Fig. 4) did not reveal any significant differences between the factors.

5. Discussion

According to the result of the binomial test on the answers for Q1 (the left panel in Fig. 2), there was no significant dif-

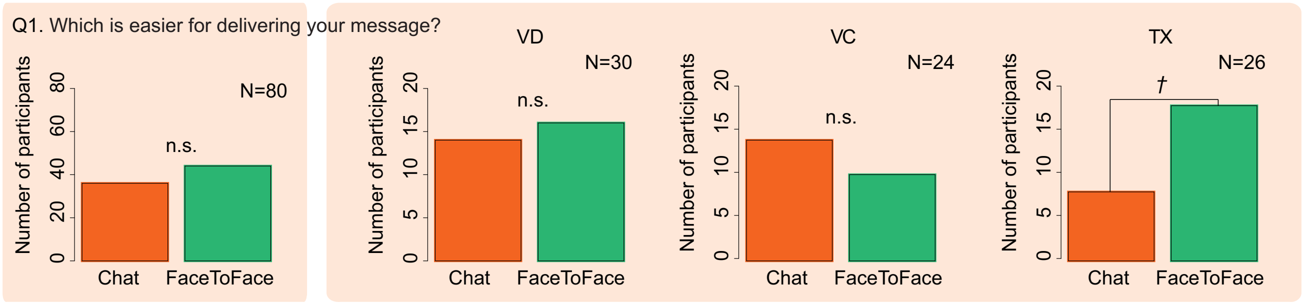

Figure 2: *The results of binomial tests (Chat vs FaceToFace) on the answer for Question 1.* ($^{\dagger}p < 0.08$, $^{***}p < 0.001$)

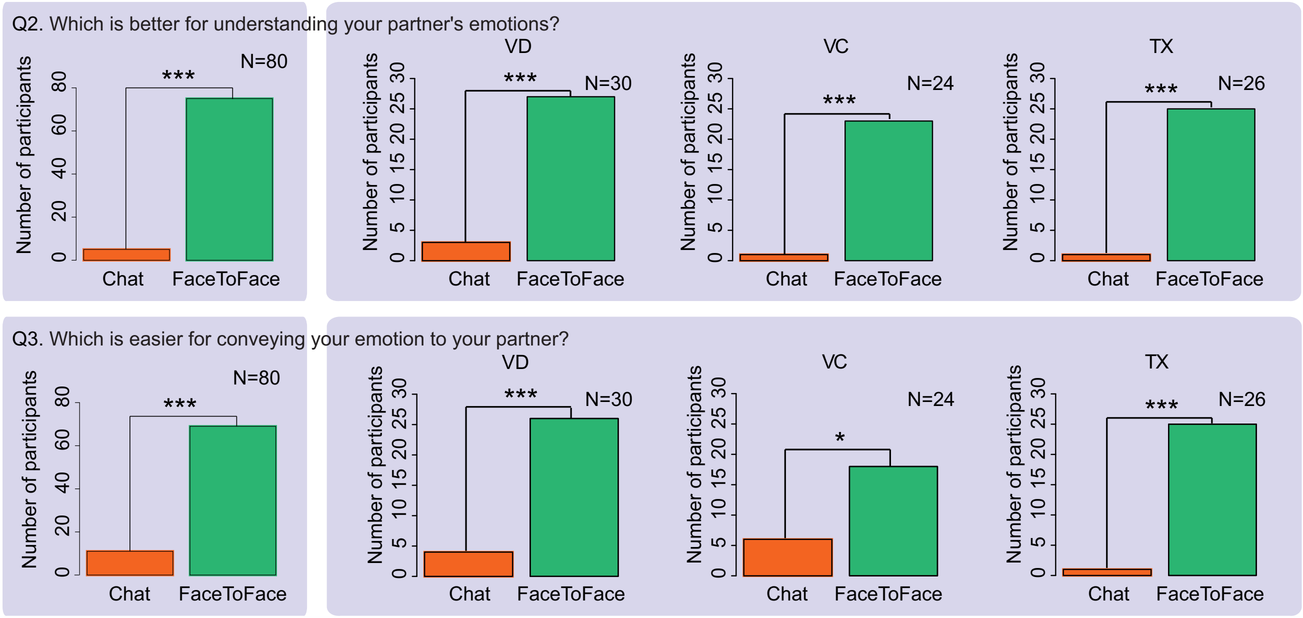

Figure 3: *The results of binomial tests (Chat vs FaceToFace) on the answer for Questions 2 and 3.* ($^{*}p < 0.05$, $^{***}p < 0.001$)

ference between the number of speakers who selected face-to-face communication and the number of speakers who selected online communication. This implies that they did not experience difficulty in delivering their message in either method of communication. However, when the binomial test was performed using the data grouped by chat type, the test revealed a marginally significant difference between TX and FaceToFace ($p < 0.08$, the right panel in Fig. 2). In contrast, there was no difference between FaceToFace and either VD or VC (the two middle panels in Fig. 2). This suggests that the speakers experienced difficulty in delivering their messages in a text chat situation, where the communication modality was restricted only to text information; the speakers did not, however, experience any difficulty in the video chat or voice chat situations, where the communication modalities were less restricted than in the text chat situation. According to the results of the binomial tests on the answers for Q2 and Q3, the test revealed a significant difference between Chat and FaceToFace for both questions ($p < 0.001$, the left two panels in Fig. 3). The test between FaceToFace and each chat type (VD, VC or TX) also revealed a significant difference between these factors ($p < 0.05$, the right six panels in Fig. 3). These results suggest that the speakers experienced difficulty in conveying their own emotions to their partners and in understanding their partners' emotions in online communication situations, regardless of chat type.

The ANOVA results regarding the correlation coefficients between the listener's evaluation and the speaker's self-evaluation upon arousal (Fig. 5) revealed a significant interaction between chat type and the communication environment ($F(2, 77) = 3.817, p < 0.05$). The Tukey-Kramer test revealed a significant difference between VC and TX and a marginally significant difference between TX and FaceToFace. These results suggest that it was difficult to understand the partner's arousal using text chat, where the communication modality was restricted to only the text information. The ANOVA results regarding dominance labeling (Fig. 6) revealed a significant main effect of the communication environment ($F(1, 77) = 5.550, p < 0.05$) and a marginally significant main effect of chat type ($F(2, 77) = 2.526, p < 0.07$). These results suggest that dominance understanding is more difficult in an online situation than in face-to-face situations, regardless of the type of chat. The ANOVA results on pleasantness labeling (Fig. 4) did not reveal any differences. Therefore, pleasantness under-

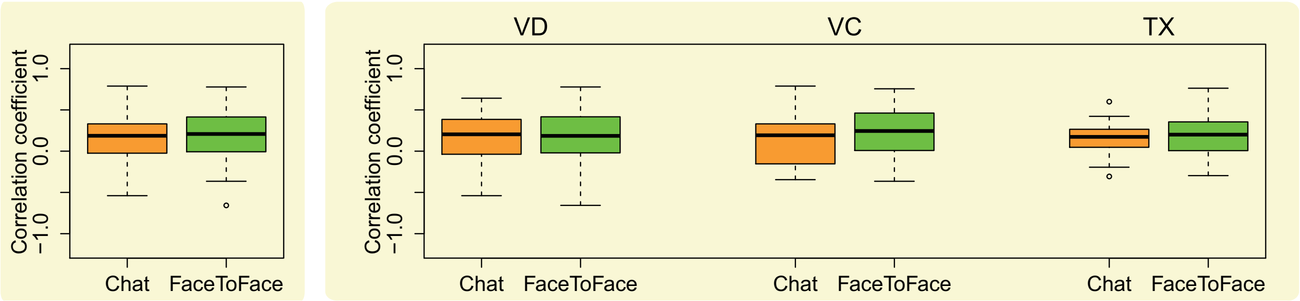

Figure 4: *The result of the ANOVA and multiple comparison tests for pleasantness evaluation.*

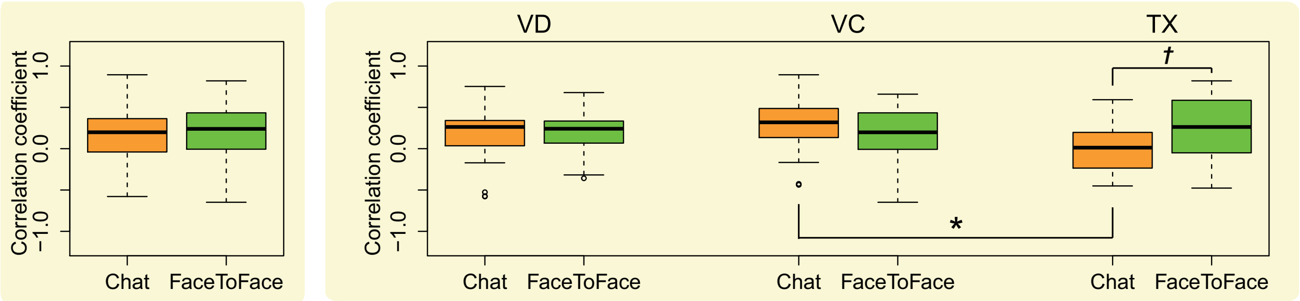

Figure 5: *The result of the ANOVA and multiple comparison tests for arousal evaluation* (${}^{\dagger}p < 0.1, {}^{*}p < 0.05$).

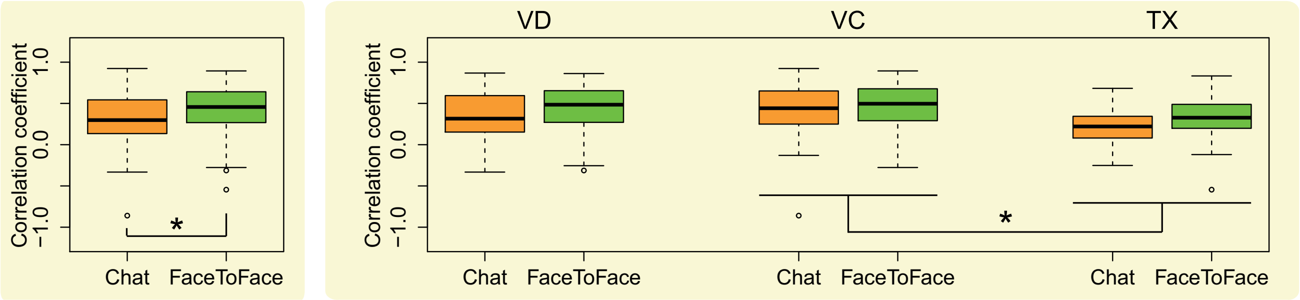

Figure 6: *The result of the ANOVA and multiple comparison tests for dominance evaluation* (${}^{*}p < 0.05$).

standing in the online situation is equivalent to face-to-face communication.

The media richness theory is partially supported by our results. According to the results of the analysis of the questionnaire (Q2 and Q3), the speakers felt difficulties in communicating their emotions in the modality-regulated chat environments (text chat and voice chat). Moreover, the analysis of arousal and dominant labeling revealed the worst emotional understanding is in the most modality-restricted environment (text chat).

However, two contradicting results were also obtained. First, the analysis of Q2 and Q3 also revealed that the speakers felt difficulty in communicating their emotions in the video chat environment, where the cues used for the communication were almost equivalent to the face-to-face environment. This result is inconsistent with the media richness theory. The only difference between the video chat environment and the face-to-face environment is the physical presence of the partner. According to Derks et al. (2008), the physical co-presence enables us to make bodily contact (touching or hitting), and those were important for emotional expression. The lack of physical co-presence leads to less emotional expression of bodily contact in the video chat environment. Therefore, the video chat environment obtained less emotional understanding than the face-to-face environment. Second, the analysis of pleasantness labeling revealed that the emotional understanding in any online situation is equivalent to that in face-to-face communication. This result is also inconsistent with the media richness theory. One explanation for this discrepancy is that interpersonal perceptions can be exaggerated in computer-mediated communication because of the limited number of cues in computer-mediated communication (Hancock and Dunham, 2001; Boucher et al., 2008). Only pleasantness of the partner may be more strongly inferred by the speaker because of the limited number of cues in computer-mediated communication. Another explanation for this result is the use of the alternative cues, i.e., clearly writing emotional words or phrases or the use of emoticons, to express pleasantness in computer-mediated communication. The effect of those words and emoticons were reported by some researchers (Walther and D'Addario, 2001; Ganster et al., 2012). This result indicates that, pleasantness is easier to express with those alternative cues than other emotions in computer-mediated communication. Therefore, it was easy for the speaker to understand the partner's emotion. For example, in the text chat environment the speaker can more clearly express his or her pleasantness than in the face-to-face environment by writing feelings with emotional words (e.g., "happy", "hate", or "like") and with the emoticons provided by the chat service. Those emotional words and emoticons can be helpful for the speaker's understanding of the partner's pleasantness. On the other hand, arousal and dominance were considered to be difficult to express with words and emoticons. Therefore, it was not understandable for the speakers.

6. Conclusion

The present study, using large-scale emotional communication corpus, aimed to investigate 1) whether speakers really experienced difficulty in exchanging their emotions accurately with their partners while they were communicating over the Internet and 2) whether the lack of communication modality actually caused the inaccuracy of emotional understanding. The results revealed that speakers have difficulty communicating their emotions in online communication environments, regardless of the type of communication modality, and that inaccurate emotional understanding more frequently occurs in online computer-mediated communication than in face-to-face communication.

For future research with this corpus, it will be demonstrated how the speaker's behavior and physiological reactions contribute to emotional understanding.

7. Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 25871139. We would like to thank Mr. Kotaro Eda from Waseda University, for his help recording our large-scale and complex multimodal data.

8. Bibliographical References

Arimoto, Y. and Okanoya, K. (2015). Mutual emotional understanding under face-to-face communication environments: How speakers understand and react to listeners' emotion in a game task dialog. *Acoustical Science and Technology*, 36(4):370–373.

Boucher, E. M., Hancock, J. T., and Dunham, P. J. (2008). Interpersonal Sensitivity in Computer-Mediated and Face-to-Face Conversations. *Media Psychology*, 11(2):235–258.

Cowie, R., McKeown, G., and Douglas-Cowie, E. (2012). Tracing Emotion. *International Journal of Synthetic Emotions*, 3(1):1–17, January.

Daft, R. L. and Lengel, R. H. (1986). Organizational Information Requirements, Media Richness and Structural Design. *Management Science*, 32(5):554–571, may.

Decety, J. and Ickes, W. (2011). *The Social Neuroscience of Empathy*. The MIT Press, Cambridge, Massachusetts.

Dennis, A. R. and Kinney, S. T. (1998). Testing media richness theory in the new media: The effects of cues, deedback, and task equivocality. *Information Systems Research*, 9:256–274.

Derks, D., Fischer, A. H., and Bos, A. E. R. (2008). The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, 24(3):766–785.

Ganster, T., Eimler, S. C., and Krämer, N. C. (2012). Same Same But Different!? The Differential Influence of Smilies and Emoticons on Person Perception. *Cyberpsychology, Behavior, and Social Networking*, 15(4):226–230.

Hancock, J. T. and Dunham, P. J. (2001). Impression Formation in Computer-Mediated Communication Revisited: An Analysis of the Breadth and Intensity of Impressions. *Communication Research*, 28(3):325–347, jun.

Hancock, J. T., Gee, K., Ciaccio, K., and Lin, J. M.-h. (2008). I'm sad you're sad: emotional contagion in CMC. *Proceegs of CSCW '08: ACM Conference on Computer Supported Cooperative Work*, pages 295–298.

Harada, E. (1997). *Hito no shiten kara mita jinkoubutu kenkyu*. Kyoritsu Shuppan Co., Ltd. (in Japanese).

Ickes, W., Stinson, L., Bissonnette, V., and Garcia, S. (1990). Naturalistic Social Cognition : Empathic Accuracy in Mixed-Sex Dyads. *Journal of Personality and Social Psychology*, 59(4):730–742.

Mehrabian, A. (1980). Basic dimensions for a general psychological theory: implications for personality, social, environmental, and developmental studies. In *Basic dimensions for a general psychological theory: implications for personality, social, environmental, and developmental studies*, pages 38–53. Oelgeschlager,Gunn & Hain Inc.,U.S. (August 1980).

Stinson, L. and Ickes, W. (1992). Empathic accuracy in the interactions of male friends versus male strangers. *Journal of personality and social psychology*, 62(5):787–797.

Verhofstadt, L. L., Buysse, A., Ickes, W., Davis, M., and Devoldre, I. (2008). Support provision in marriage: the role of emotional similarity and empathic accuracy. *Emotion*, 8(6):792–802.

Walther, J. B. and D'Addario, K. P. (2001). The Impacts of Emoticons on Message Interpretation in Computer-Mediated Communication. *Social Science Computer Review*, 19(3):324–347, aug.

Zaki, J., Bolger, N., and Ochsner, K. (2008). It takes two: the interpersonal nature of empathic accuracy. *Psychological science*, 19(4):399–404.