Prediction of Social Maladaptation using Emotional Entrainment of Disgust during Comprehensive Psychiatric Interviews

Yokotani Kenji[1], Takagi Gen[2], and Wakashima Kobun[3]
[1]Tokushima University, Graduate School of Science and Technology for Innovation
E-mail: yokotanikenji@tokushima-u.ac.jp
[2] Tohoku Fukushi University, Faculty of Comprehensive Welfare
E-mail: g-takagi@tfu-mail.tfu.ac.jp
[3]Tohoku University, Graduate School of Education
E-mail:k_wakashima@sed.tohoku.ac.jp

Abstract— **Previous speech entrainment studies have shown disagreement in their findings: One group emphasized that acoustic entrainment predicts social adaptation, whereas another group emphasized that it predicts social maladaptation. Our study aims to resolve the disagreement from the perspective of emotional entrainment: the entrainment of positive emotions predicts social adaptation, whereas the entrainment of negative emotions predicts social maladaptation. Using a machine-learned sentiment classifier, we estimated the probability of anger, disgust, fear, happiness, neutrality, and sadness in speech. The corpus consisted of dialogues recorded from 29 comprehensive mental health interviews. The Jensen–Shannon divergence was also calculated to estimate the (dis)entrainment. Results showed that the entrainment of happiness significantly demonstrated the rapport of the participants with their therapist. In contrast, their entrainment of disgust significantly demonstrated their social maladaptation. Our study observed social maladaptation to be contrastingly related to positive and negative emotional entrainment. Classification of speech from an emotional perspective could enrich the study of entrainment and facilitate the analysis of emotional communication.**

Index Terms: emotional entrainment, comprehensive mental health interview, machine-learned sentiment classifier, acoustic synchrony

I. INTRODUCTION

Speech entrainment is the phenomenon wherein each conversational partner adapts his/her speech behavior to that of the other [1]. This phenomenon is considered a key factor in building rapport between conversational partners [2]. Previous studies on acoustic speech entrainment have mainly classified entrainment from the acoustic perspective [2]–[9]. Due to this limited classificatory system, speech entrainment studies have shown disagreement in their findings [4], [10], [11]. Our study aims to classify speech entrainment from the emotional perspective to resolve this disagreement.

Speech entrainment has been observed to encourage conversational partners to be cognizant of their similarities and feel closeness toward each other [6]. Studies showed that improving the acoustic entrainment of a robot with human participants fostered positive emotions during the interaction [7] and motivated them to talk more to the robot [8]. Therapists who displayed a greater degree of pitch entrainment with their clients were more empathic toward them during therapy [2]. Participants who received a greater degree of entrainment with respect to beat and loudness built better collaborative relationships with their therapist [12].

On the other hand, a lack of speech entrainment indicated the social maladaptation of children and adults [3]. A lack of entrainment pertaining to conversational speech rate suggested severe social maladaptation among adult speakers. Furthermore, a lack of pitch entrainment during the interaction of therapists with children indicated the severity of their social maladaptation [13]. These findings indicate that acoustic entrainment is common and relevant to collaborative relationships and social adaptation [1].

However, a comparison of multi-lingual dialogues clarified that speech disentrainment is also common during conversation [11]. Further, a study revealed that speech disentrainment occurred more frequently than entrainment during conversation [10]. Several studies suggested that acoustic disentrainment, rather than entrainment, is predictive of collaborative relationships and social adaptation. For example, acoustic disentrainment during a game is predictive of the collaborative relationship between players [4]. The acoustic disentrainment of a couple during conflict illustrates their high conflict resolution skills [5].

The disagreement in the findings of these previous studies may be due to their classification systems, which only employ acoustic perspectives (e.g., intensity and frequency) and omit emotional perspectives (e.g.,

disgust and happiness). Studies on emotion showed that listeners intuitively perceived negative and positive emotions from speech [14]. Couple communication studies showed that the interaction of negative emotions during communication between couples represents a high risk of divorce and marital distress, whereas the interaction of positive emotions represents a lower risk of divorce and high marital adjustment [15]. These findings indicate that the entrainment of negative emotions during conversations demonstrates destructive relationships and social maladaptation, whereas the entrainment of positive emotions represents collaborative relationships and social adaptation.

Based on these studies, we formulate two hypotheses: the entrainment of negative emotions would demonstrate low social collaboration and low social adaptation (Hypothesis 1), whereas the entrainment of positive emotions would demonstrate high social collaboration and high social adaptation (Hypothesis 2). Our study utilized comprehensive mental health interview data [12], [16], [17]. As the interviews focused on negative emotions [18], [19], the entrainment of negative emotions could be easily observed through them. Further, we conducted sentiment analysis using a machine-learned sentiment classifier, which labeled the emotions expressed in the speech with high efficiency [20]. The negative emotions included anger, disgust, fear, and sadness, whereas the positive emotion was happiness. Subtle emotions were labeled as neutral to avoid polarized emotional labels.

II. METHOD

A. Machine-learned Sentiment Classifier

Based on a previous study [20], we used convolutional neural networks, bidirectional long short-term memory networks, attention layer, and soft max layer as our classifiers (Fig. 1). The current databases were the Ryerson Audio-Visual Database of Emotional Speech and Song (speech only) [21] and the Surrey Audio-Visual Expressed Emotion Database [22]. The former consists of emotional speech from 12 adult male and 12 female actors (e.g., speech about "kids talking by the door" in happy and disgusted tones). The latter consists of emotional speech from 4 adult male researchers. All the speakers were native English speakers. The length of the speech was adjusted to 3.0 s. The Mel-frequency cepstral coefficients (MFCCs) were calculated for all the frames where the number of Mel-filter banks was set as 40. The sample rate and hop length were set as 44100 and 512, respectively. Hence, each speech has 259 samples consisting of 40 dimensions. When the length of a speech was less than 3 s, the speech was repeated until the length exceeded 3 s.

Similarly, the deltas of the MFCCs and the delta-deltas of the MFCCs were calculated. The MFCCs, the deltas of the MFCCs, and the delta-deltas of the MFCCs were used as the input data of the classifier (Fig. 1). Eighty percent of the data were training data, whereas the remaining were test data. All data were tied with six emotion labels: anger, disgust, fear, happiness, neutral, and sadness. Our classifier continued to learn its own 2,952,960 parameters (Fig. 1) based on the training data, and repeated the learning for 250 epochs (one epoch involves 2,309 speeches) to classify emotion. The accuracy of the classifier was 0.9090 (Fig. 2, left), slightly outperforming that of the previous study (0.8282 + 0.0499).

B. Speech Corpus

The corpus was the same as that of the previous study [12]. The corpus was obtained from the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision Axis I Disorders, Non-Patient Edition [18], using the Japanese version [19]. The data were recorded at 48,000 Hz in a private room (Fig. 4). The interviewer was a Japanese male clinical psychologist, whereas the interviewees were 29 female participants (Fig. 4). We utilized the first 903 s of the recorded data. The first 3 s of the recorded data were used for noise profiles, whereas the remaining 900 s were selected as the spoken dialogue. The dialogues were automatically segmented into the speeches of the therapist and participants (Fig. 3).

Noise was removed from the spoken dialogue according to the noise profiles and the speech was then normalized (Fig. 3). The data were then segmented by silences longer than 100 ms [23], [24] as inter-pausal units (IPUs).

The IPUs involve the voices of both the therapist and the participants; hence, they were separated by speaker through a neural networking model [25] that was the same as in a previous study [26]. The training data were the LibriSpeech dataset [27] with 460 h of clean speech. After the dataset was normalized, our model learned to separate mixed speech data into two separated speech data 644,000 times and showed a speech disturbance ratio of 7.69. Our model then referred to the sound sources for the individual voices of the participants and therapist and separated the IPUs into therapist-voice IPUs and participant-voice IPUs (Fig. 3).

To estimate the speaker's turn, we estimated the numbers of moras of the IPUs of the therapist and participants as follows [28]. First, we generated a smoothed MFCC line to represent the flow of human voice. According to the MFCCs, we calculated the power of the MFCCs and the delta of the power (Fig. 2 left upper, middle). We then multiplied the power of the MFCCs and the delta of the power to create a single line. To smoothen the line, we filtered it through a Gaussian filter with the window size set to 4 (Fig. 2 left lower). The smoothed line is defined as

$$F = G(p \circ \Delta p, 4)$$

where p is the power of the MFCCs, Δp is the delta of the power, and $^{\circ}$ is the element-wise product. $G(\ ,4)$ indicates the Gaussian filter setting with a window size of 4. To find local maxima, we compared the smoothed MFCC line score among the previous, current, and next times. When the current score is higher than the previous and next scores, the current time has a mora (Fig. 2 left lower):

$$[\{F(n-1) < F(n)\} \wedge \{F(n+1) < F(n)\}] \rightarrow F(n) \in Mora,$$

where $F(n)$ is the smoothed MFCC line score at time n. Through this calculation, all the IPUs of the participants and the therapist were assigned a mora number, including zero.

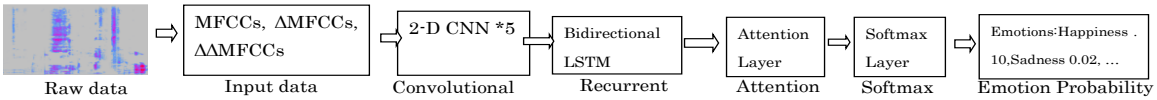

Fig. 1. Attention-based convolutional recurrent network architecture of the sentiment classifier and learning mechanism

Note. MFCCs: Mel-frequency cepstral coefficients, ,2-D CNN: 2-Dimensional Convolutional Neural Networks, LSTM: Long short-term memory

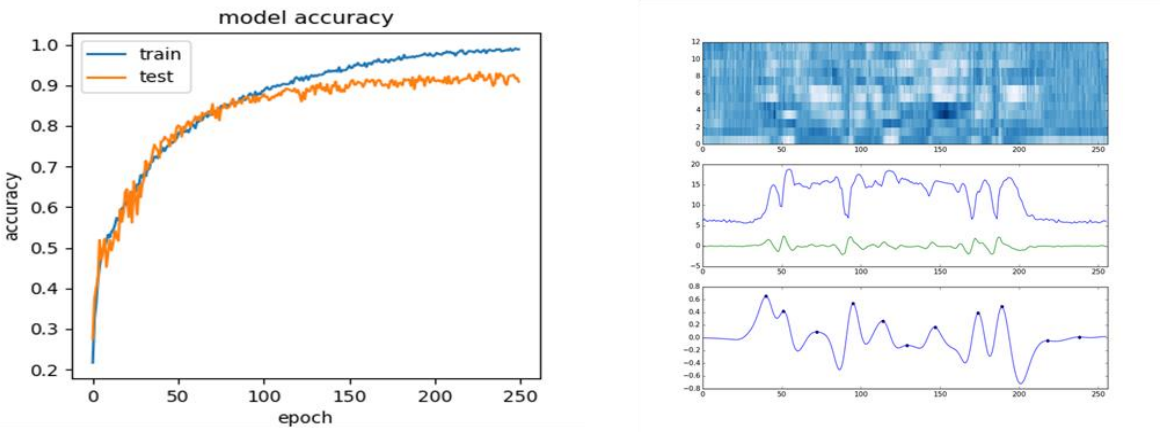

Fig. 2. Results of the model accuracy test and estimation of mora

Note. The right upper figure shows the heat map of 12 MFCCs. The blue line in the right middle figure shows the power of the MFCCs, whereas the red line shows the delta of the power. The right lower figure shows the element-wise product of the power and the delta after the Gaussian filter. The dots indicate the moras.

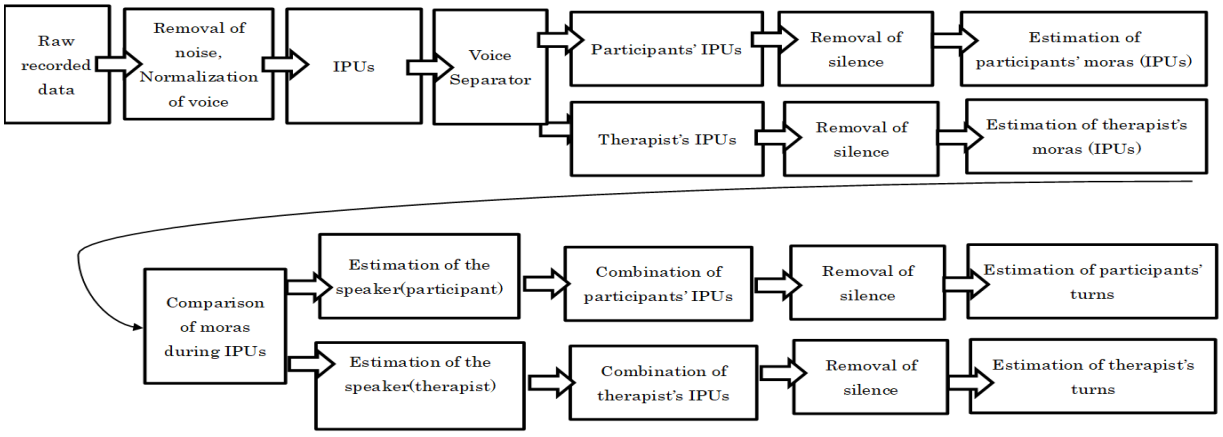

Fig. 3. Segmentation schema of spoken dialogue

Note. IPUs: Inter-pausal units

Fig. 4. Interview setting

To determine the speaker in the IPUs, we compared the numbers of moras between the IPUs of the therapist and the participants. If one speaker's (e.g., the therapist's) mora number is larger than that of the other at the current time (t), the speaker of current IPU was regarded as the therapist. If the number of moras was the same between two IPUs at the current time (t), the numbers of moras at the previous time (t−1) and next time (t+1) were added and compared. When the number of moras was still the same, we also added the numbers of moras from the two previous times (t−2) to the two next times (t+2). If the number of moras was the same even after summing the moras from the five previous times (t−5) to the five next times (t+5), we assigned no speaker at the current IPU(t). Through this calculation, all the speakers of the IPUs were classified as either the therapist, the participants, or none.

When speakers were the same during consecutive times (e.g., t+1, t+2), the IPUs were combined. When speakers were different during consecutive times (e.g., t+2, t+3), we set the boundary at these times. Based on the boundary, we combined the IPUs, and series of the combined IPUs were regarded as turns (Fig. 1). The average number of turns per interview was 138.10. The average length of the turns was 3.01 s. The moras and other acoustic features of the turns were estimated in the same manner as the IPUs.

The first 2.5 s of the speeches were extracted and the first 0.5 s were repeated, because these initial parts of the speeches included emotionally relevant features [29]. Several speeches from the therapist and participants were shorter than 2.5 s. These speeches were repeated until their length exceeded 3 s.

C. Measurement

C-1. Emotions

Each speech was assigned probabilities of six emotions such as anger, disgust, fear, happiness, neutral, and sadness, based on our classifier. The probabilities (e.g., 10.2% of happiness, 20.5% of anger) rather than classes of emotions (happiness or anger) were used to elaborate the interactions between the therapist and the participants.

C-2. (Dis)entrainment

To evaluate the entrainment of the emotions we calculated the Jensen–Shannon divergence (JSD) to measure the disentrainment between two speeches.

JSD is defined as follows:

$$JSD(p||q) = \frac{1}{2}KLD(p||m) + \frac{1}{2}KLD(q||m)$$

$$\text{Where } m = \frac{1}{2}(p+q), \quad (1)$$

where KLD is the Kullback–Leibler divergence [30]. p and q indicate the arrays of the scores divided by the total scores, which are regarded as probabilities. Furthermore, the KLD and JSD cannot be calculated adequately when either p or q is zero; hence, we added a small number (10^{-5}) to all the scores to normalize them. Due to the spike-shaped distribution (Fig. 5A), the overall shape of the distribution did not change after the normalization (Fig. 5B). The higher the JSD score, the lower was the entrainment.

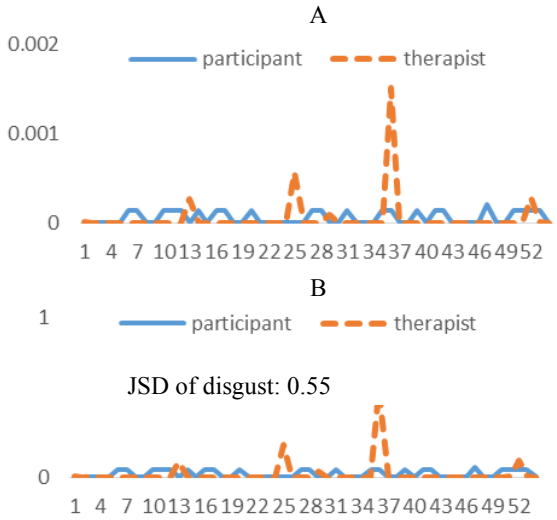

Fig. 5. Probabilities of disgust during a comprehensive mental health interview

Note: Fig. 5A shows the estimated probabilities, whereas Fig. 5B shows the normalized probabilities.

C-3. Rapport during Mental Health Interview

The corpus included data concerning rapport, which were gathered through a questionnaire that used a five-point scale with the following six items [31]: "Did you feel that your point was understood?", "Did you feel that it was easy to talk to the counselor?", "Did you feel like you could speak your mind?", "Did you feel that the counselor was receptive to your feelings?", "Did you feel that the counselor's warmth was conveyed to you?", and "Did you feel like you were familiar with the counselor?". A high score for the participants indicates a high level of rapport with the therapist. The average of this scale was used as the rapport score, similar to the previous study [16].

C-4. Social Adaptation

The corpus also included the Global Assessment of Functioning (GAF) scale [18]. The GAF scale measures how much a participant's symptoms affect his/her daily life on a scale of 0 to 100. For example, a GAF score from 91 to 100 indicates that participants had no symptoms and enjoyed

excellent functioning in a wide range of social activities because of their resources. A GAF score from 81 to 90 indicates that they had minimal symptoms (e.g., mild anxiety before an examination), but they were interested and involved in a wide range of social activities. A GAF score from 71 to 80 indicates that they have transient symptoms, but these are expected reactions to psychosocial events (e.g., difficulty concentrating after a family argument). A GAF score from 61 to 70 indicates that they had mild symptoms (e.g., depressed mood and mild insomnia) and some difficulty in their social functioning (e.g., theft within the household). A GAF score from 51 to 60 indicates that they had moderate symptoms (e.g., flat affect and circumstantial speech, occasional panic attacks) and moderate difficulty in social functioning (e.g., few friends, conflicts with co-workers). A GAF score from 41 to 50 indicates they had serious symptoms (e.g., suicidal ideation) and experienced serious impairment in social functioning (e.g., no friends, unable to keep a job) [18]. The GAF score was measured by the clinical psychologist. A high score indicates high social adaptation [19]. The GAF score was used as the social adaptation score. Further, participants whose GAF scores were 70 or less were regarded as clinical population [18]. Accordingly, we categorized the participants into clinical (n = 14) and non-clinical groups (n = 15).

III. Results

A. Comparison of Emotional Expressions between Participants and Therapist

Before we test our hypotheses, we compared the emotional expressions between the participants and the therapist. The participants' speeches showed significantly more expressions of emotions such as angry, fear, and sadness during the interview than the therapist's speeches (Table I). In contrast, the therapist's speeches showed significantly more expressions of happiness than the participants' speeches (Table I). The emotional expressions in acoustic speech were slightly different from the expressions conveyed through the face [17].

TABLE I
COMPARISON OF EMOTIONAL EXPRESSIONS BETWEEN PARTICIPANTS AND THERAPIST

	participant		therapist		JSDs			
	M	*SD*	*M*	*SD*	*M*	*SD*	*paired-t*	*p*
Anger	0.77	0.11	0.60	0.08	0.18	0.05	10.13	***
Disgust	0.00	0.00	0.00	0.00	1.08	0.90	-1.23	*n.s.*
Fear	0.09	0.05	0.06	0.02	0.67	0.24	3.09	**
Happiness	0.13	0.09	0.33	0.08	0.64	0.33	-12.00	***
Neutral	0.00	0.00	0.00	0.00	0.76	0.64	1.36	*n.s.*
Sadness	0.01	0.01	0.01	0.00	0.50	0.36	3.05	**
Rapport	4.48	0.41	-	-	-	-		
GAF	70.24	8.35	-	-	-	-		

Note: N = 29, ***: $p < .001$, **: $p < .01$, *n.s.*: not significant; GAF: Global Assessment of Functioning; JSD: Jensen–Shannon divergence. JSD may exceed 1. The higher the JSD score, the lower is the entrainment.

B. Links between Positive Emotional Entrainment and Rapport

Table II lists the correlations of emotional entrainment with rapport. The JSD of happiness was negatively and significantly correlated with rapport. In other words, the participant–therapist pairs with high entrainment of happiness demonstrated high mutual rapport (Table II).

TABLE II
CORRELATIONS OF SPEECH EMOTIONAL ENTRAINMENT WITH RAPPORT AND SOCIAL ADAPTATION

	Rapport	Social adaptation
JSDs of angry	-.226	-.309
JSDs of disgust	.272	**.433***
JSDs of fear	-.227	-.075
JSDs of happiness	**-.565****	-.183
JSDs of neutral	-.197	.060
JSDs of sadness	-.333	-.204

Note: **: $p < .01$, *: $p < .05$

C. Links between Negative Emotional Entrainment and Social Maladaptation

We also compared emotional entrainment between the clinical and non-clinical groups. As expected, the JSDs of disgust were significantly and positively correlated with social adaptation in Table II. Fig. 6 shows the comparison of the JSDs of disgust between the clinical and non-clinical groups. The JSDs of disgust were significantly lower for the clinical group than for the non-clinical group (t = 2.589, df = 16.113, p = .020, N = 29). In other words, the clinical group showed greater entrainment of disgust than the non-clinical group.

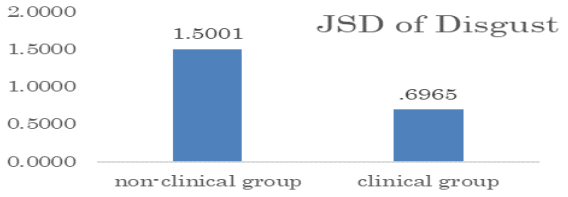

Fig. 6. Comparison of the JSD of disgust between non-clinical and clinical groups

Note: JSD: Jensen–Shannon divergence. JSD may exceed 1. The higher the JSD score, the lower is the entrainment.

IV. Discussion

A. Negative Emotional Entrainment Demonstrating Social Maladaptation

As hypothesized, we observed that the entrainment of disgust during the mental health interview predicted the social

maladaptation of the participants. Consistent with couple communication studies [15], the interaction of disgust predicted social maladaptation. The participants who show the entrainment of disgust with their therapist can show the same among strangers, which would worsen their social adaptation. In the context of disentrainment studies [4], [5], we observed that negative emotional disentrainment predicted social adaptation.

On the other hand, the entrainment of fear and sadness was not significantly related to social maladaptation. In the mental health interview setting, fear and sadness were the main topics in the diagnosis of anxiety and depression [18], [19]. These topics might contaminate the evaluation of fear and sadness, indicating that their entrainment might not have been adequately evaluated. The entrainment of anger was not significantly related to social maladaptation. A previous study indicated that anger was related to seriousness during interviews [17]. Hence, the seriousness of the participants and therapist might have contaminated the evaluation of anger. Another dialogue corpus is required to clarify the association of these negative emotions with social maladaptation.

B. Positive Emotional Entrainment Demonstrating Collaborative Relationships

We also confirmed that the entrainment of happiness during the interview predicted the rapport between the participants and the therapist. These findings are consistent with several previous findings [3], [6]–[8], [13]. The dialogue data in these studies were emotionally positive because play [3], [13], date [6], game [7], and chat [8] settings were positive. Hence, most of the acoustic entrainment in these studies could be relevant to the entrainment of happiness. The dialogue data from negative settings, such as drug counseling [2] and suicide assessment [9], also illustrated that acoustic entrainment was related to collaborative relationships. Our findings indicate that the entrainment of happiness could be observed in these dialogues and was related to collaborative relationships.

On the other hand, the entrainment of happiness did not predict social adaptation. Our study measured social maladaptation [18], [19], which is more likely to be related to negative emotions. Measurements focused on social adaptation, such as social support and self-esteem, need to be performed in future studies. Further, couple communication studies were helpful to evaluate the daily social adaptation among couples [15].

C. Limitations

Our study has four limitations. First, our study involved only male–female dialogue; hence, our findings cannot be generalized to same-sex dialogue. A previous study suggested that female participants show a higher collaborative relationship with their therapist than their male counterparts do [16]. Future studies need to include same-sex dialogue.

Second, our experimental design involved facial interaction between the therapist and the participants, but the interaction was not controlled. As the interaction affected the collaborative relationship [17], future studies need to regulate the effects of such interactions on the relationship between the therapist and the participants.

Third, our sentiment classifier was trained using a European speech dataset [21], [22], but it was utilized for Japanese speech. Considering the linguistic and cultural differences between the two datasets [11], the sentiment classifier should be trained using an Asian speech dataset.

Fourth, the current speech corpus expressed few feelings of disgust. The results of our study would not be applicable in speech corpus with high rate of disgust expression. The corpus containing more disgust expressions are needed in the future.

V. CONCLUSIONS

We clarified the contrasting links between positive and negative emotional entrainment and social maladaptation. Entrainment studies, especially for close relationships [2], [9], [15], need to classify speech from an emotional perspective, because the dialogue in such communication involves not only information communication but also emotional communication. The classification of speech from an emotional perspective could enrich entrainment studies [2]–[9] and facilitate the analysis of emotional communication.

ACKNOWLEDGMENTS

This study was funded by KAKENHI (19K11206). The authors would like to thank Professor Sato Yutaka of Tokushima University for his insightful comments on the early drafts of this manuscript.

REFERENCES

[1] H. Giles, “Communication Accommodation Theory,” in *The International Encyclopedia of Communication Theory and Philosophy*, American Cancer Society, 2016, pp. 1–7.

[2] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. Narayanan, “Modeling therapist empathy and vocal entrainment in drug addiction counseling.,” in *INTERSPEECH*, 2013, pp. 2861–2865.

[3] C. J. Wynn, S. A. Borrie, and T. P. Sellers, “Speech rate entrainment in children and adults with and without autism spectrum disorder,” *Am. J. Speech Lang. Pathol.*, vol. 27, no. 3, pp. 965–974, 2018.

[4] 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.

[5] S. Weidman, M. Breen, and K. Haydon, “Prosodic speech entrainment in romantic relationships,” May 2016, pp. 508–512, doi: 10.21437/SpeechProsody.2016-104.

[6] J. Michalsky and H. Schoormann, “Pitch Convergence as an Effect of Perceived Attractiveness and Likability.,” in *INTERSPEECH*, 2017, pp. 2253–2256.

[7] J. M. Kory-Westlund and C. Breazeal, "Exploring the Effects of a Social Robot's Speech Entrainment and Backstory on Young Children's Emotion, Rapport, Relationship, and Learning," *Front. Robot. AI*, vol. 6, 2019, doi: 10.3389/frobt.2019.00054.

[8] T. Kawahara, T. Yamaguchi, M. Uesato, K. Yoshino, and K. Takanashi, "Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015, pp. 392–395.

[9] M. Nasir, B. R. Baucom, C. J. Bryan, S. S. Narayanan, and P. G. Georgiou, "Complexity in Speech and its Relation to Emotional Bond in Therapist-Patient Interactions During Suicide Risk Assessment Interviews.," in *INTERSPEECH*, 2017, pp. 3296–3300.

[10] P. G. T. Healey, M. Purver, and C. Howes, "Divergence in Dialogue," *PLoS ONE*, vol. 9, no. 6, Jun. 2014, doi: 10.1371/journal.pone.0098598.

[11] 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.

[12] K. Yokotani, G. Takagi, and K. Wakashima, "Acoustic entrainment at utterance turn level predicts confidence during psychiatric interviews," presented at the The 82nd National Convention of IPSJ, Kanazawa, Japan, Mar. 2020.

[13] D. Bone *et al.*, "Acoustic-prosodic, turn-taking, and language cues in child-psychologist interactions for varying social demand.," in *INTERSPEECH*, 2013, pp. 2400–2404.

[14] C. F. Lima, S. L. Castro, and S. K. Scott, "When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1234–1245, Dec. 2013, doi: 10.3758/s13428-013-0324-3.

[15] H. J. Markman, G. K. Rhoades, S. M. Stanley, E. P. Ragan, and S. W. Whitton, "The premarital communication roots of marital distress and divorce: The first five years of marriage," *J. Fam. Psychol.*, vol. 24, no. 3, pp. 289–298, 2010, doi: 10.1037/a0019481.

[16] K. Yokotani, G. Takagi, and K. Wakashima, "Advantages of virtual agents over clinical psychologists during comprehensive mental health interviews using a mixed methods design," *Comput. Hum. Behav.*, vol. 85, pp. 135–145, Aug. 2018, doi: 10.1016/j.chb.2018.03.045.

[17] K. Yokotani, G. Takagi, and K. Wakashima, "Nonverbal Synchrony of Facial Movements and Expressions Predict Therapeutic Alliance During a Structured Psychotherapeutic Interview," *J. Nonverbal Behav.*, vol. 44, no. 1, pp. 85–116, Mar. 2020, doi: 10.1007/s10919-019-00319-w.

[18] M. B. First, R. L. Spitzer, M. Gibbon, and J. B. W. Williams, *Structured Clinical Interview for DSM-IV Axis I Disorders*. Washington, DC: American Psychiatric Publishing, Inc., 1997.

[19] M. B. First *et al.*, *SeishinkashindanmensetsumanualSCID: shiyonotebiki/tesutoyoshi*, 2nd edition. Nihonhyoronsha, 2010.

[20] M. Chen, X. He, J. Yang, and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.

[21] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)." Zenodo, Apr. 05, 2018, doi: 10.5281/zenodo.1188976.

[22] "Surrey Audio-Visual Expressed Emotion (SAVEE) Database." http://kahlan.eps.surrey.ac.uk/savee/Download.html (accessed Feb. 20, 2020).

[23] Y. Den *et al.*, "Two-level Annotation of Utterance-units in Japanese Dialogs: An Empirically Emerged Scheme.," 2010.

[24] H. Koiso and Y. Den, "How is the Smooth Transition between Speakers Realized?," *Cogn. Stud. Bull. Jpn Cogn. Sci. Soc.*, vol. 7, no. 1, pp. 93–106, 2000.

[25] Q. Wang *et al.*, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," *ArXiv E-Prints*, vol. 1810, p. arXiv:1810.04826, Oct. 2018, Accessed: Dec. 19, 2019. [Online]. Available: http://adsabs.harvard.edu/abs/2018arXiv181004826W.

[26] P. Seungwon, *mindslab-ai/voicefilter*. MINDs Lab, 2019.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[28] ikuo0, "Detection of voice valid intervals and mora," *Qiita*, Mar. 19, 2019. https://qiita.com/ikuo0/items/0d5798db824f3df074af (accessed Jul. 31, 2020).

[29] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003, doi: 10.1016/S0167-6393(03)00099-2.

[30] R. Levitan, S. Benus, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," 2015.

[31] T. Kakii, "Characteristics of multimedia counseling: A study of an interactive TV system," *Jpn. J. Psychol.*, vol. 68, no. 1, pp. 9–16, 1997, doi: 10.4992/jjpsy.68.9.