Emerging Technologies in Healthcare and Medicine, Vol. 116, 2023, 411–421
https://doi.org/10.54941/ahfe1004396

Neural Network Model for Visualization of Conversational Mood With Four Adjective Pairs

Koichi Yamagata[1,2], Koya Kawahara[1], Yuto Suzuki[1], Yuki Nakahodo[1], Shunsuke Ito[1], Haruka Matsukura[1], and Maki Sakamoto[1]

[1]The University of Electro-Communications, Chofu, Tokyo, Japan
[2]Kanazawa University, Kanazawa, Ishikawa, Japan

ABSTRACT

Several studies have examined the use of deep learning to estimate emotions during a conversation. However, when attempting to monitor or influence conversations conducted as part of a meeting or a chat, the mood of the conversation is more important than the emotion. In this study, we developed a deep neural network model that could read the "conversational mood" in real time, which is an important conversational feature in Japanese society. For this purpose, we constructed a new training data set containing 60 hours of conversations in Japanese. The data set was annotated to learn the mood of the conversations. For the annotation of mood, we selected four representative adjective pairs that could effectively describe the conversational mood. The evaluation results are shown to present the validity of our model. This model is expected to be applied to a system that can influence or control the mood of conversations in some ways, including presentation of ambient music and aromas, depending on the purpose of the discussion, such as during a conference, chatting, or business meeting.

Keywords: Mood, Conversation, Deep learning, Affective ambient intelligence

INTRODUCTION

In recent years, the accuracy of speech recognition has improved remarkably. This has been facilitated by the incorporation of transformer-based models and convolutional neural network-based models into speech recognition algorithms (Zhang et al., 2020), (Gulati et al., 2020). Speech recognition software can be used to obtain text information from conversational speech data. Although text can be treated as surface level information, several studies have indicated that speech recognition can also be used to estimate emotions, which represent higher level information in a conversation. (Ruusuvuori, 2013) describes the relationship between emotion/affect and conversation. In addition, several newly proposed models use long short-term memory (LSTM) or Gated Recurrent Unit (GRU) to estimate emotion in conversations (Majumder et al., 2019), (Hazarika et al., 2018), (Tzirakis et al., 2017), (Yoon et al., 2018), (Vrijkotte et al., 2000).

© 2023. Published by AHFE Open Access. All rights reserved.

In normal conversation, emotions such as anger and sadness are unlikely to be explicitly expressed for some purposes, including avoidance of getting into an unexpected argument and offending others. Thus, when attempting to control or monitor the state of a conversation during a meeting or casual discussion, it is often more important to estimate the mood than the emotion. Some researchers have examined the role of mood, as distinguished from emotion. According to (Jenkins et al., 1998), diffuse emotional states that persist over a long period of time are called "mood" and are usually distinguished based on duration and intensity of expression. However, these differences are rarely quantified, and no specific durations are fixed. (Lane et al., 2000) defines mood as "a series of emotions, ephemeral in nature, variable in intensity and duration, and usually accompanied by multiple emotions". Thus, emotion and mood are considered to be distinct. (Ekman, 1999) argued that mood can be estimated, at least in part, from the associated emotion signals. Based on previous psychological studies, (Katsimerou et al., 2015) concluded that emotion and mood are different but closely related concepts. Many studies have analysed the mood of music and images, and researchers in the English-speaking world appear to conceptualize mood as something that is influenced by the surrounding conditions and environment (Saari et al., 2016), (Tarvainen et al., 2020), (Thiparpakul et al., 2021).

Accurate identification of the mood of a conversation is especially important for Japanese people who are engaged in collaborative and democratic decision making. Accordingly, many Japanese studies have focused on techniques for assessing the mood in a conversation. These have included the use of linguistic information such as words and sentence topics as well as nonverbal information such as laughing, voice characteristics, facial expressions, and gestures. (Tokuhisa et al., 2006) clarified the types of utterances present in hu-man-human conversational dialogue, and showed that the inter-annotator agreement for specific tag schemes was relatively high. Subsequently, (Inaba et al., 2011) proposed a method for automatically determining the mood of a conversation between humans according to the cooccurrence of words in a text of the dialogue. Their goal was to design a dialogue agent to facilitate conversation. (Kondo et al., 2015) showed that the negativity of the mood in a conversation could be estimated according to the speakers' affective states (pleasantness, arousal, dominance, credibility, interest, positivity) using Poisson regression.

Nonverbal information such as facial expressions, postures, gestures, and heart rate may also be important for reading the mood of a conversation. Laughter is associated with positive emotions (Hofmann et al., 2017), a brighter and softer mood (Devillers et al., 2015), reduced stress and increased relaxation (Tanaka et al., 2018), (Cogan et al., 1987), (Akimbekov et al., 2021); it has been studied in various fields of research (Cosentino et al., 2016). The research reported in (Busso et al., 2008) used motion capture technology to obtain gesture information, while heart rate sensors have been used to estimate stress levels (Matsumoto et al., 2010), (TFES, 1996), (Pomeranz et al., 1985), (Takada et al., 2005). Furthermore, (Kunimasa et al., 2017) estimated task performance in intellectual workers via several physiological indices (pupil diameter and heart rate variability).

In this study, we propose a deep neural network model that could visualize the mood of conversations from only linguistic information obtained by generally used speech recognition technology. Obtaining of Nonverbal information mentioned above generally requires cameras and bio-signal monitors. The in-formation obtained by these devices are sensitive data in terms of individual privacy. Therefore, in this study, the data obtained only by microphones are used for estimation of conversational mood. Furthermore, in our proposed neural network model, the amount of laughter which is generally measured by capturing facial expression with camera is also estimated together with the conversational mood. Because laughter is considered to play an important role in creating a cheerful environment, it can be used to evaluate the conversational mood. For obvious and plain visualization of conversational mood, we propose to use only four adjective pairs as indica-tors of mood.

As linguistic information, we use text vectors generated by bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019) with input text obtained from the Google Cloud Speech-to-Text Application Programming Interface (API). BERT, developed by Google, is a transformer-based machine-learning model for natural language processing using masked language modelling and next sentence prediction. BERT text vectors are expected to contain semantic information about text. To capture the frequency of conversations, we also obtain the number of words via morphological analysis of text using MeCab. MeCab is an open-source text segmentation library for use with text written in the Japanese (Kudo et al., 2004). As non-verbal information, we used speech feature vectors using mel-frequency cepstral coefficients (MFCCs) (Mermelstein, 1976), (Davis et al., 1980). MFCCs represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCCs were originally widely used as features in speech recognition (Ganchev et al., 2005), and they have many other applications, including music classification (Müller, 2007).

Several open source data sets are available for performing various tasks, such as automatic speech recognition using machine learning, speech synthesis, and emotion recognition (Stappen et al., 2021), (Russell, 1980). In these data sets, emotional values are annotated with respect to video, sound, natural language, and biometric information. However, the size of the data sets varies according to the language, and while there are large English speech data sets, those for other languages, including Japanese, tend to be small (Alimuradov et al., 2020), (Ando et al., 2021).

At present, there are few databases of Japanese conversations between multiple people that can be used to estimate the mood of a conversation. Therefore, in this study, we record conversations between multiple people interacting with one another for several dozen hours and then annotated the data with four adjective pairs. We use the resulting data set to develop a model that can automatically estimate the mood in a Japanese conversation.

Thus, we developed a deep neural network-based model that can visualize the "conversational mood" in real time, which is an important feature for Japanese researchers. For this purpose, we constructed a new training data

set containing 60 hours of conversations in Japanese. This data set was annotated to learn the mood of the conversations. We expect that our research will also be applicable to non-Japanese contexts.

MATERIALS AND METHODS

To collect training data for our model, we conducted a conversation experiment in a meeting room at the University of Electro-Communications library. In one conversation experiment, we asked three subjects to have a one-hour conversation, and we ran this experiment 35 times. The total duration of the conversation data was 35 hours, and there were 90 participants (17 women and 73 men, mean age 21.0 years, standard deviation 1.81). The participants were students at the same university. Because we wanted to obtain natural dialogue, no topic was specified. In addition, we conducted an online conversation experiment using Zoom online meeting software. Zoom Meetings is a videotelephony software program developed by Zoom Video Communications. This additional experiment was beneficial for the learning data because the demand for online conferencing has recently in-creased with the spread of infectious diseases, i.e., the COVID-19 pandemic. The total duration of the con-versation data was 25 hours, and there were 100 participants. The participants were paid to take part in the experiments, and all provided written informed consent.

To construct the teacher data for the model designed to estimate the conversational mood, we first selected representative adjective pairs that could describe the conversational mood. We utilized a system developed by Iiba et al. to estimate 21 affective scales of adjective pairs from input text (Iiba et al., 2013). Via cluster analysis, we further reduced the number of adjective scales by entering text corresponding to 30 theatrical scenarios into the system. The 21 adjective pairs were clustered into 4 groups based on the absolute values of the output scales using K-means++. As shown in Table 1, the 1st cluster was related to serious and easy, the 2nd cluster was related to aggressive and calm, the 3rd cluster was related to tidy and messy, and the 4th cluster was related to happy and gloomy. The 4 adjective pairs to be annotated were representative of the 4 clusters. We expected these 4 adjective pairs (gloomy-happy, easy-serious, calm-aggressive, tidy-messy) to capture the mood of a conversation.

Table 1. The clustering of 21 adjective pairs.

Representative adjective pairs	Adjective pairs that constitute the clusters
serious-easy	hard-soft, strong-weak, stable-unstable, heavy-light, serious-easy, pleasant - unpleasant
aggressive-calm	rational-passionate, quiet-noisy, aggressive-calm
tidy-messy	simple-complex, natural-unnatural, clean-dirty, Like-dislike, polite-impolite, Tidy-messy, young-old
happy-gloomy	flashy-plain, cheerful-gloomy, masculine-feminine, active-inactive, happy-sad

To construct a supervised learning data set, we conducted an annotation experiment in which we asked subjects to listen to recorded conversations and add annotations that described the mood of the con-versation using the 4 adjective scales on a 5-point se-mantic differential (SD) scale: −2, silent; −1, nearly silent; 0, neither; +1, slightly talkative; +2, talkative. Two subjects annotated each set of recorded data. The conversation data from the real-life conversations were annotated by seven university students, and that from the online meetings were annotated by five university students. Note that the annotated scale reflected the mood of the conversation, not the emotion. The purpose of this system was to enable visualization of the mood of a conversation.

To reinforce the estimation of the conversational mood, we asked the subjects to count the number of people laughing in parallel with the annotation of the 4 adjective pairs. We estimated not only the 4 adjective scales, but also the amount of laughter. The participants were paid to take part in the experiments, and all provided written informed consent.

All data obtained in the conversation experiment were divided into 10-second intervals to treat them uniformly as time series data. The data recorded in the conversation experiment was converted into text using Google Speech API. The text files were then separated into 10-second segments and text data in 20-second sliding windows were converted into 768-dimensional text vectors using BERT (Devlin et al., 2019). The 768-dimensional vectors were then compressed into 20-dimensional vectors using principal component analysis. In addition, the recorded data were transformed into 20-dimensional speech feature vectors using MFCCs. We adopted LSTM as a time series deep learning model. LSTM is an artificial recurrent neural network (RNN) for deep learning. Unlike standard feedforward neural networks, LSTM and RNN have feedback connections and can process entire sequences of data. LSTMs were developed to deal with the vanishing gradient problem encountered when training traditional RNNs (Hochreiter et al., 1997). The input data for our model were the 20-dimensional compressed BERT text vectors, 20-dimensional MFCC vectors, and the number of words in the spoken text (Fig.1-(A)). The input data were concatenated and passed through 4 fully connected (FC) layers with exponential linear units (ELUs) (Fig.1-(B)). The sizes of the four FC layers were 41, 40, 30, and 20. In every FC layer, we used a dropout rate of 0.5 to avoid over-fitting. After passing through the 4 FC layers, the resulting 20-dimensional vector was entered into the LSTM layer, which had a 128-dimensional hidden vector (Fig.1-(C)). The hidden vector output from the LSTM was then input into the same LSTM. The hidden vector was expected to contain information about the mood of previously stored conversations. The 128-dimensional vector output from the LSTM layer was passed to the 4 FC layers (size: 128, 50, 30, and 5, respectively) using ELUs and dropout rates (Fig.1-(D)). For the 4 adjective scales, the hyperbolic tangent function (tanh) was used at the output layer because the values of the estimated adjective scales ranged from -1 to +1. For the laughter rate, the sigmoid function was used at the output layer because the estimated rate ranged from 0 to +1 (Fig.1-(E)). We used the mean squared error loss to train our model. We used

Adam as the optimizer with a learning coefficient of 0.01 and a weight decay of 0.00000001.

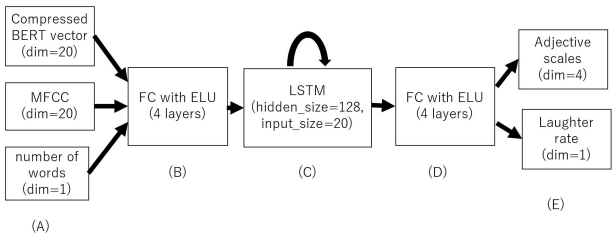

Figure 1: Overview of the proposed model.

RESULTS AND ACCURACY EVALUATION

We used a new cross-validation method that allowed us to visualize the effect of the size of the training data. All data acquired in the experiment were divided into 5 data sets: D0, D1, D2, D3, and D4. First, we performed cross-validation with D0 as the validation data and D1 as the training data. We then performed cross-validation with D0 as the validation data and D1 and D2 as the training data. In this way, training data sets were gradually added, and finally D1, D2, D3, and D4 were used as training data sets. We performed 4 cross-validations, each with a total of 1/5, 2/5, 3/5, and 4/5 of the training data sets.

Fig. 2 shows the progress of the loss function for learning with the 4 cross-validations. The left graph shows that the loss of the training data de-creased with the epochs. In contrast, the right graph shows that the loss of the test data decreased as the total number of training data sets increased. From this result, we expected to reduce the loss of test data by adding more training data in the future.

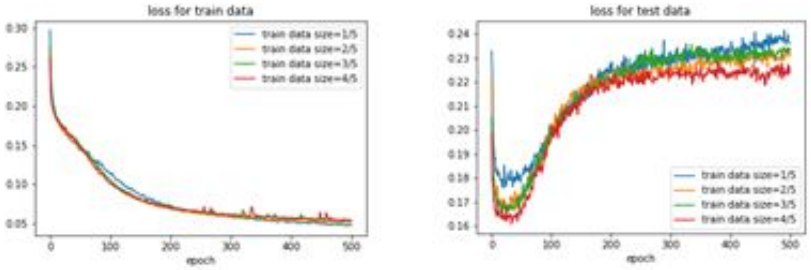

Figure 2: Learning curves for train data and test data.

Fig. 3 shows the correlation coefficients between the actual and estimated values for each sensitivity scale in the training data sets. The correlation coefficients increased as the epochs progressed in the training data sets. Fig. 4 shows the correlation coefficients between the actual values and the estimated

values for each sensitivity scale in the evaluation data. The graph shows that the correlation coefficient for the evaluation data tended to increase with the total number of training data sets.

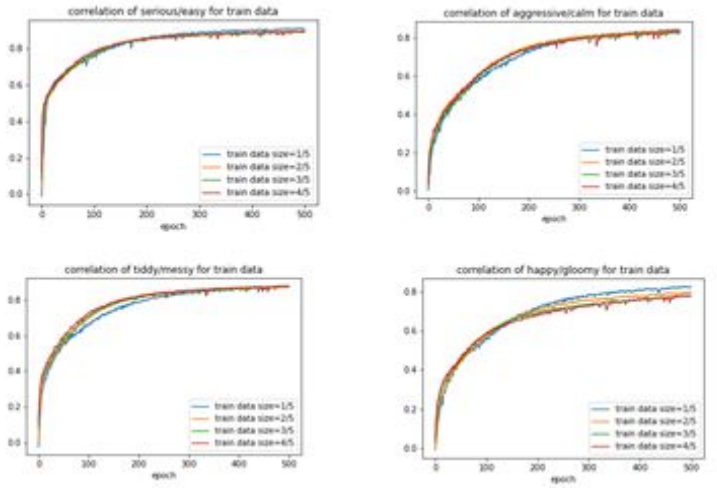

Figure 3: Curves showing the change in the correlation coefficient between the real and estimated values for train data.

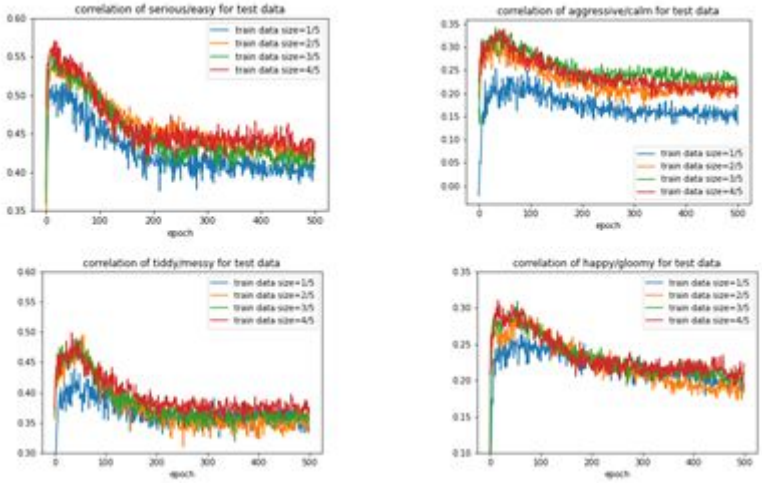

Figure 4: Curves showing the change in the correlation coefficient between the real and estimated values for test data.

CONCLUSION

In this study, we developed an LSTM-based model that automatically visualizes the mood of a conversation in real time from speech information alone. Reading the mood of a conversation is very important in Japanese culture. To achieve this goal, we selected 4 adjective pairs that could describe the moods

of conversations, and conducted conversation experiments to collect training data. Our system can also accurately detect the amount of laughter, which is important for estimating the mood of a conversation, from speech information alone. To evaluate the accuracy of the learned models, we employed a novel cross-validation method. Our cross-validation method showed that the model accuracy improved as the amount of training data increased. Further improvements in accuracy are expected with accumulated training data.

Our system could be used in a variety of situations. Mental stress monitoring has been used in education-al, workplace, medical, and everyday life settings (Singh et al., 2022), (Seo et al., 2022), (Torkamani-Azar et al., 2022), (Jiang et al., 2022). However, existing devices generally require participants to wear a sensor and undergo video recording. Our proposed system could detect stress levels without sensors or video images. The relationships between the wellbeing index and our adjectival scales are as follows. Serious-easy reflects the tension of the mood and is expected to measure stress. Aggressive-calm is expected to reflect the degree of heated discussion and to indicate productivity. Happy-gloomy is expected to reflect the participants' empathy toward happy or sad stories. Tidy-messy is expected to reflect the level of engagement in the conversation.

Furthermore, it may soon be possible to provide therapeutic treatments via combined spatial presentation techniques such as aroma, video, and music. In particular, aromas help people to connect with their environment (Flavián et al., 2021). During the COVID-19 pandemic, telecommuting and working in virtual offices became more common (Graves et al., 2020). Casual conversations in these new settings may have characteristics that differ from those in typical work environments. Our system could be used as a marketing tool, for instance, to estimate a customer's willingness to purchase and level of interest based on their interactions with the salesperson.

ACKNOWLEDGMENT

This work was supported by JST-Mirai Program Grant Number JPMJMI17DB and JSPS KAKENHI Grant Number JSPS22H0367. All the experiments including human subjects in this paper were conducted under the approval of Ethics Committee of the University of Electro-Communications (register number 21028).

REFERENCES

Akimbekov, N. S.; and Razzaque, M. S. (2021). "Laughter Therapy: A Humor-Induced Hormonal Intervention to Reduce Stress and Anxiety". Curr. Res. Physiol., vol. 4, pp. 135–138.

Alimuradov, A. K.; Tychkov, A. Y.; Mezhina, V. A.; Fokina, E. A.; Zhurina, A. E.; Ageykin, A. V.; Gorbunov, V. N.; Reva, E. K. (2020). "Development of Natural Emotional Speech Database for Training Automatic Recognition Systems of Stressful Emotions in Human-Robot Interaction". Proc. 4th Scientific School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR), pp. 11–16.

Ando, S.; Fujihara, H. (2021). "Construction of a Large-Scale Japanese ASR Corpus on TV Recordings," Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 6948–6952.

Busso, C. et al. (2008). "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," Lang. Resour. Eval., vol. 42, no. 4, pp. 335–359.

Cogan, R.; Cogan, D.; Waltz, W.; McCue, M. (1987). "Effects of Laughter and Relaxation on Discomfort Thresholds". J. Behav. Med., vol. 10, no. 2, pp. 139–144.

Cosentino, S.; Sessa, S.; Takanishi, A. (2016). "Quantitative Laughter Detection, Measurement, and Classification: A Critical Survey". IEEE Rev. Biomed. Eng., vol. 9, pp. 148–162.

Davis, S. B.; Mermelstein, P. (1980). "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". IEEE Trans. Acoust., Speech, Signal Process., vol. 28, no. 4, pp. 357–366.

Devillers, L. et al. (2015). "Multimodal Data Collection of Human-Robot Humorous Interactions in the Joker Project". Proc. Int. Conf. Affective Computing and Intelligent Interaction (ACII), pp. 348–354.

Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. 17th Annu. Conf. the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.

Ekman, P. (1999). "Basic emotions". Handbook of Cognition and Emotion, T. Dalgleish and M. J. Power, eds., Oxford, UK: John Wiley & Sons, pp. 45–60.

Flavián, C.; Ibáñez-Sánchez, S.; Orús, C. (2021). "The Influence of Scent on Virtual Reality Experiences: The Role of Aroma-Content Congruence". J. Bus. Res., vol. 123, pp. 289–301.

Ganchev, T.; Fakotakis, N.; Kokkinakis, G. (2005). "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification T". Proc. 10th Int. Conf. Speech and Computer (SPECOM 2005), pp. 191–194.

Graves, L. M.; Karabayeva, A. (2020). "Managing Virtual Workers: Strategies for Success". IEEE Eng. Manag. Rev., vol. 48, no. 2, pp. 166–172.

Gulati, A. et al. (2020). "Conformer: Convolution-augmented Transformer for Speech Recognition". Website: https://arxiv.org/abs/2005.08100

Hazarika, D. et al. (2018). "ICON: Interactive Conversational Memory Net-work for Multimodal Emotion Detection". Proc. 2018 Conf. Empirical Methods in Natural Language Processing, pp. 2594–2604.

Hochreiter, S.; Schmidhuber, J. (1997). "Long Short-Term Memory". Neural Comput., vol. 9, no. 8 pp. 1735–1780.

Hofmann, J.; Platt, T.; Ruch, W. (2017). "Laughter and Smiling in 16 Positive Emotions". IEEE Trans. Affect. Comput., vol. 8, no. 4, pp. 495–507.

Iiba, S.; Doizaki, R.; Sakamot, M. (2013). "Color and Font Recommendations based on Mental Images of Text". (in Japanese), Trans. Virtual Real. Soc. Jpn., vol. 18, no. 3, pp. 217–226.

Inaba, M.; Toriumi, F.; Ishii, K. (2011). "Automatic Determination of "Enthusiasm" in Dialogues Using Word co-occurrence". (in Japanese), IEICE Trans. Inf. and Syst., vol. J94-D, no. 1, pp. 59–67.

Jenkins, J. M.; Oatley, K.; Stein, N. L. (1998). Human Emotions: A Reader. Malden, MA, USA: Blackwell.

Jiang, S.; Firouzi, F.; Chakrabarty, K.; Elbogen, E. B. (2022). "A Resilient and Hierarchical IoT-Based Solution for Stress Monitoring in Everyday Settings". IEEE Internet Things J., vol. 9, no. 12, pp. 10224–10243.

Katsimerou, C.; Heynderickx, I.; Redi, J. A. (2015). "Predicting Mood from Punctual Emotion Annotations on Videos". IEEE Trans. Affect. Comput., vol. 6, no. 2, pp. 179–192.

Kondo, T.; Kamashima, M. (2015). "Mood Estimation in a Conversation by Utilizing the Speakers' Affective States: Examination with Spoken Dialogue Corpus" (in Japanese), Trans. Jpn. Soc. Kansei Eng., vol. 15, no. 2, pp. 279–285, 2015.

Kudo, T.; Yamamoto, K.; Matsumoto, Y. (2004). "Applying Conditional Random Fields to Japanese Morphological Analysis". Proc. 2004 Conference on Empirical Methods in Natural Language Processing, pp. 230–237.

Kunimasa, S.; Seo, K.; Shimoda, H.; Ishii, H. (2017). "A Trial of Intellectual Work Performance Estimation by Using Physiological Indices". Advances in Neuroergonomics and Cognitive Engineering, C. Baldwin, eds., vol. 586, Cham, Switzerland: Springer, Cham, pp. 305–315.

Lane, A. M.; Terry, P. C. (2000). "The Nature of Mood: Development of a Conceptual Model with a Focus on Depression". J. Appl. Sport Psychol., vol. 12, no. 1, pp. 16–33.

Majumder, N. et al. (2019). "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations". AAAI, vol. 33, no. 1, pp. 6818–6825.

Matsumoto, Y. et al. (2010). "Study of Mental Stress Evaluation based on analysis of Heart Rate Variability," (in Japanese), J. Life Support Eng., vol. 22, no. 3, pp. 105–111.

Mermelstein, P. (1976). "Distance Measures for Speech Recognition: Psychological and Instrumental". Pattern Recognition and Artificial Intelligence, C. H. Chen, ed., NY, USA: Academic Press, pp. 374–388.

Müller, M. (2007). Information Retrieval for Music and Motio, Heidelberg, Germany: Springer.

Pomeranz, B.; Macaulay, R. J.; Caudill, M. A.; Kutz, I.; Adam, D.; Gordon, D.; Kilborn, K. M.; Barger, A. C.; Shannon, D. C.; Cohen, R. J.; and Benson, H. (1985). "Assessment of Autonomic Function in Humans by Heart Rate Spectral Analysis". Am. J. Physiol. Heart and Circ. Physiol., vol. 248, no. 1, pp. H151–153.

Russell, J. A. (1980). "A Circumplex Model of Affect". J. Pers. Soc. Psychol., vol. 39, no. 6, pp. 1161–1178.

Ruusuvuori, J. (2013). "Emotion, affect and conversation". The Hand-book of Conversation Analysis, J. Sidnell and T. Stivers, eds., Hoboken, NJ, USA: Wiley-Blackwell pp. 330–349.

Saari, P. et al. (2016). "Genre-Adaptive Semantic Computing and Audio-Based Modelling for Music Mood Annotation," IEEE Trans. Affect. Comput., vol. 7, no. 2, pp. 122–135.

Seo, W.et al. (2022). "Deep Learning Approach for Detecting Work-Related Stress Using Multimodal Signals". IEEE Sensors J., vol. 22, no. 12, pp. 11892–11902.

Singh, M. et al. (2022). "A Facial and Vocal Expression Based Comprehensive Framework for Real-Time Student Stress Monitoring in an IoT-Fog-Cloud Environment". IEEE Access, vol. 10, pp. 63177–63188.

Stappen, L. et al. (2021). "The Multi-modal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements". IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1334–1350.

Takada, H.; Takada, M.; Kanayama, A. (2005). "The Significance of "LF-component and HF-component which Resulted from Frequency Analysis of Heart Rate" and "the Coefficient of the Heart Rate Variability": Evaluation of Autonomic Nerve Function by Acceleration Plethysmography", (in Japanese), Health Evaluation and Promotion, vol. 32, no. 6, pp. 504–512.

Tanaka, A.; Tokuda, N.; Ichihara, K. (2018). "Psychological and Physiological Effects of Laughter Yoga Sessions in Japan: A Pilot Study". Nurs. Health Sci., vol. 20, pp. 304–312.

Tarvainen, J.; Laaksonen, J.; Takala, T. (2020). "Film Mood and Its Quantitative Determinants in Different Types of Scenes". IEEE Trans. Affect. Comput., vol. 11, no. 2, pp. 313–326.

Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology (TFES). (1996). "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use". Circulation, vol. 93, no. 5, pp. 1043–1065.

Thiparpakul, P.; Mokekhaow, S.; Supabanpot, K. (2021). "How Can Video Game Atmosphere Affect Audience Emotion with Sound". Proc. 9th Int. Conf. Information and Education Technology (ICIET), pp. 480–484.

Tokuhisa, R.; Terashima, R. (2006). "The Relationship between Utterances and "Involvement" in Conversational Dialogue," (in Japanese), Trans. Jpn. Soc. Artif. Intell., vol. 21, no. 2, pp. 133–142.

Torkamani-Azar, M.; Lee, A.; Bednarik, R. (2022). "Methods and Measures for Mental Stress Assessment in Surgery: A Systematic Review of 20 Years of Literature". IEEE J. Biomed. Health Inform., vol. 26, no. 9, pp. 4436–4449.

Tzirakis, P.; Trigeorgis, G.; Nicolaou, M. A.; Schuller, B. W.; Zafeiriou, S. (2017). "End-to-End Multi-modal Emotion Recognition Using Deep Neural Networks". IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1301–1309.

Vrijkotte, T. G.; Van Doornen, L. J.; De Geus, E. J. (2000). "Effects of Work Stress on Ambulatory Blood Pressure, Heart Rate, and Heart Rate Variability". Hypertension, vol.35, no. 4, pp. 880–886.

Yoon, S.; Byun, S.; Jung, K. (2018). "Multimodal Speech Emotion Recognition Using Audio and Text". Proc. IEEE Spoken Language Technology Workshop, pp. 112–118.

Zhang, Y. et al. (2020). "Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition". Website: https://arxiv.org/abs/2010.10504.