Speech Emotion Recognition with Fusion of Acoustic- and Linguistic-Feature-Based Decisions

Ryotaro Nagase*, Takahiro Fukumori* and Yoichi Yamashita*
*Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan

***Abstract*—In recent years, the advanced technique of deep learning has improved the performance of speech emotion recognition (SER) as well as speech synthesis and speech recognition. On the other hand, emotion recognition still has low accuracy and does not capture emotional features in detail. Multimodal processing for SER is one of the techniques that improve the performance of SER and can handle integrated emotional factors. Many researchers adopt various fusion methods to produce optimal methods for each case. However, it is insufficient to observe and analyze the respective fusion's synergistic effects in acoustic and linguistic features conveyed by speech. In this paper, we propose a method of SER with acoustic and linguistic features at the utterance level. Firstly, two emotion recognition systems using acoustic or linguistic features are trained with Japanese Twitter-based emotional speech (JTES). Then, we aim to improve accuracy by using early fusion, which fuses linguistic and acoustic features, and late fusion, which fuses the values predicted by each model. Consequently, proposed methods have about a 20% higher accuracy than the method that uses classifiers in only acoustic or linguistic information. Also, several methods improve the recognition rate for each emotion.**

I. Introduction

Speech emotion recognition (SER) is a technique of recognizing emotions conveyed in speech. This technique has recently attracted attention in the field of human–computer interaction (HCI). It is widely applicable to the development of more human-like robots [1], virtual agents of smart speakers or automobiles [2], and mental health analysis [3] and e-learning systems [4]. Many researchers have used deep learning to improve the performance of SER, as well as speech synthesis and speech recognition.

Although the performance of SER has improved over the past few years, SER is still a challenging task. One of the SER challenges is to improve the accuracy rate further. In previous studies, emotion recognition still has a low recognition rate. To expand the range of applications of emotion recognition, it is necessary to develop a more accurate recognition technology. Another challenge is to research integrated processing techniques for both acoustic and linguistic features. Since emotions are complex and ambiguous, it is difficult to understand emotions entirely using only acoustic features. The recognition of more human-like emotions requires multimodal processing that combines information such as speech, images, and language. The way of integrating some features is called fusion. Most researchers use late fusion, which fuses predictions from different recognizers. In this method, it is easy to combine values because the predicted scores are in the same format, but it does not consider the relationships among features. Hence, in this paper, we propose a SER method using two different fusion techniques. We attempt to improve the emotion recognition performance and consider different features by applying late fusion and early fusion. Furthermore, because SER with linguistic features has received less attention than SER with visual features in conventional research, in this study, we focus on acoustic and linguistic information obtained from speech. This process is intended to be incorporated and applied in automatic speech recognition. We use the Japanese Twitter-based emotional speech (JTES) [5] database for training and evaluation. This database includes Japanese speech and text data of four emotion categories: joy, sadness, anger, and neutral. For training with emotion recognition in acoustic features, we use a network comprising a convolutional neural network (CNN), a bidirectional long short-term memory (BLSTM), and an attention mechanism (AM). A bidirectional encoder representation from transformers (BERT) is used in training with emotion recognition in linguistic features.

The main contribution of this paper is the proposal of the novel SER method to fuse the features and predictions obtained from models of speech and text emotion recognition. We combine various fusion methods and analyze the performance of the proposed method.

This paper proceeds as follows. In Section II, we describe related works on emotion recognition in speech, text, and multimodality. In Section III, we present the proposed method for recognizing emotion with acoustic and linguistic features. In Section IV, we explain the setup for the experiment and show the results. Finally, Section V shows our conclusions and future work.

II. Related works

A. Speech emotion recognition

SER is a method of identifying emotions from acoustic information. Low-level descriptors (LLDs) and mel-spectrograms are mostly used and emotions are categorized or detected on the basis of deep learning. There is a variety of network architectures in this research, such as the CNN, recurrent neural network (RNN), long short-term memory (LSTM), and BLSTM. Badshah et al. [6] used a hierarchical CNN to extract detailed mel-spectrogram patterns and achieved high accuracy. Also, Xie et al. [7] observed that BLSTM can capture time-domain emotional features from speech. In addition to these network architectures, AM is a method of focusing on important information and is used in SER studies. AM, especially self-AM (SAM), can capture the

period of time-domain features related to emotions. In [8], [9], the network architecture comprising CNN, BLSTM, and SAM was reported to achieve high performance.

B. Text emotion recognition

Text emotion recognition is a method of identifying emotions captured from linguistic information. Most researchers use word embedding such as Word2vec [10] and detect emotions using CNN or LSTM. Kim [11] combined pretrained word embedding and simple CNN and showed that his proposed method realizes the same performance as the conventional sentiment analysis method. Su et al. [12] reported that a proposed method using LSTM outperforms conventional methods. In recent years, most researchers have attempted to extract high-level semantic information from text by pretraining general tasks with large amounts of data using a transformer [13]. For example, one of the typical network configurations is BERT [14]. Moreover, a pretrained representation has been applied to various neural language processing tasks, including emotion recognition. For instance, Al-Omari et al. [15] used BERT for word embedding and trained networks including BLSTM. They achieved a high accuracy of multiple emotion prediction. Additionally, many studies have shown that using a network based on transformers makes it possible to improve the emotional recognizer's performance [16].

C. Multimodal emotion recognition

Multimodal emotion recognition is a method of identifying emotion from a variety of information. The processing of multimodalities is performed to recognize complex emotions. Emotions are highly ambiguous and complex. Therefore, to recognize emotion in more detail, it is necessary to detect feature-related emotions from among voices, texts, and images, as performed by a human. Recently, many researchers have tackled emotion recognition by multimodal processing using information obtained from multimedia. To deal with multimodalities, most researchers have used fusion, which is of two main types, namely, early fusion and late fusion.

Early fusion, in other words, feature-based fusion, is the method of integrating two or more features as input to the next network. Many researchers have used early fusion to improve the emotional recognizer's performance [17]. This method has the advantage that the correlation among multimodalities is considered, although there is "the curse of dimensionality".

Late fusion, that is, decision-based fusion, is the method of integrating two or more predictions into the final predicted result. Many researchers have also used late fusion and achieved good emotion recognition performance [18]. This method has the advantage of easier comparison than in the case of early fusion because the predicted values from each classifier have the same format. On the other hand, it is not possible to capture the correlation among features. In most cases, late fusion is used because it is easy to handle [19].

The introduction of multimodal features has improved SER because it provides various perspectives to make decisions.

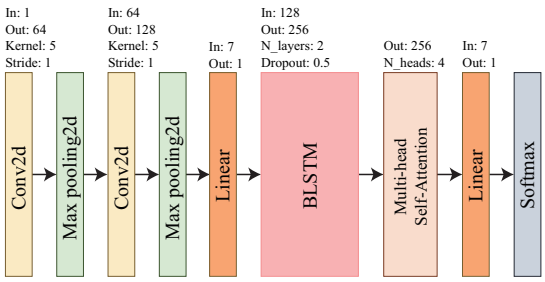

Fig. 1. CNN+BLSTM+Attention

However, although speech data contain acoustic and linguistic information, SER, in many cases, uses only acoustic features and cannot capture semantic information. If some transcriptions of the speech are available, we can extract the linguistic features. Then, by adding text information to the recognizer, the meanings of words or sentences can be considered. In some research, emotion has been recognized by considering two different features obtained from speech. Atmaja et al. [20] proposed emotion recognition with speech segments and text using early fusion. Cho et al. [21] integrated predictions from acoustic and text emotion recognition using late fusion and evaluated an annotated dataset and a call-center dataset. Although early fusion and late fusion were combined in some studies [22], the networks still consist of either early or late fusion in many studies. In this paper, we evaluate SER performance using the network combined with various early and late fusion methods.

III. Proposed Method

A. Emotion recognition with acoustic features

For emotion recognition using acoustic information, we build the network structure shown in Figure 1. This architecture is constructed on the basis of a model with which a high recognition rate was achieved in previous studies [8], [9]. In the input part of the network, the acoustic features for a whole utterance are divided into a sequence of segments consisting of a certain number of frames. Segment inputs make it possible to input time-domain features of different lengths and reduce the number of model parameters. In this work, we divide all input frames into segment units with a half-overlap. Inputs are the mel-cepstrum extracted from speech, and outputs are the predicted emotion category of joy, sadness, anger, or neutral.

B. Emotion recognition with linguistic features

In emotion recognition using text information, we use the BERT model pretrained with parameters using the Japanese dataset and constructed a recognizer by transfer learning. In this work, inputs are text and outputs are the predicted emotional label. We focus on improving the accuracy of SER with linguistic features. Therefore, we used complete sentences in our experiments, not transcriptions obtained by automatic speech recognition.

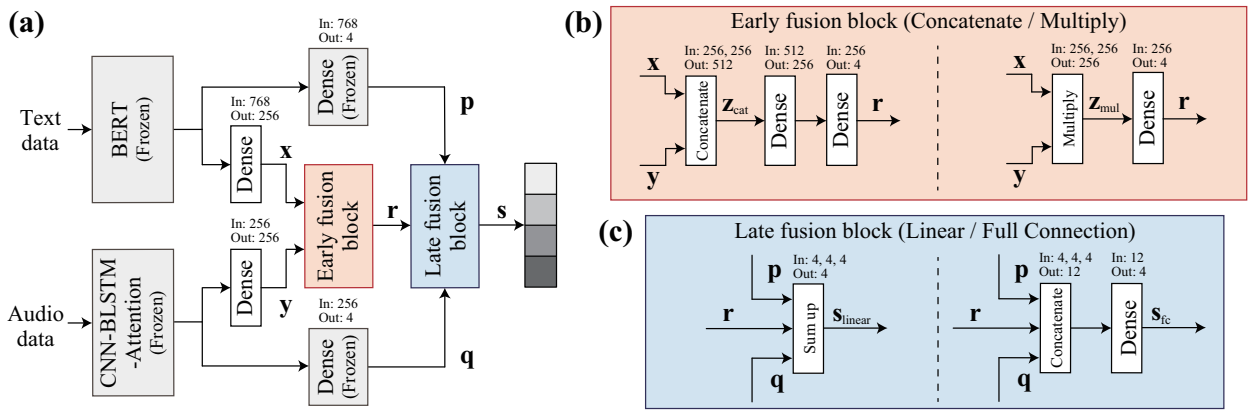

Fig. 2. The model for processing acoustic and linguistic features using (1) or (2) as the early fusion and (3) or (4) as the late fusion

C. Emotion recognition with acoustic and linguistic features

Figure 2 (a) shows the network structure of the proposed method. We expect that the proposed method, which considers each acoustic or linguistic feature or prediction, improves the performance of SER. The proposed method involves two components: the pretraining-based part and the fusion part. In the pretraining-based part, each of the models described in sections III-A and III-B is trained for emotion recognition. The pretraining-based part consists of two SER classifiers with speech or text input, as described in sections III-A and III-B. The fusion part is implemented as a network combined with early fusion and late fusion. Early fusion uses the outputs of the pretraining-based part, which are transformed to the same format. Late fusion uses the predicted results of the two pretrained classifiers as in early fusion. In our method, we aim to use the advantages of early fusion and late fusion. To recognize emotion in detail, the different fusion methods must be handled well. We attempt to align two different features by considering the classified results from each pretrained model.

The equations in the proposed method are as follows. Equations (1) and (2) show the fused features $\mathbf{z}_{\mathrm{cat}}$ and $\mathbf{z}_{\mathrm{mul}}$. The calculation process in early fusion is shown in Figure 2 (b). Let $\mathbf{x} = [x_1, \cdots, x_i, \cdots, x_N]$ and $\mathbf{y} = [y_1, \cdots, y_i, \cdots, y_M]$ denote outputs of the middle layer from speech and text recognizers, respectively. Also, x_i and y_i represent the i-th output, N and M indicate the number of dimensions of the output, and $\circ$ is the hadamard product.

$$\mathbf{z}_{\mathrm{cat}} = [\mathbf{x}, \mathbf{y}] \tag{1}$$

$$\mathbf{z}_{\mathrm{mul}} = \mathbf{x} \circ \mathbf{y} \tag{2}$$

Equation (1) is a simple concatenation and Equation (2) decreases the dimension of the fused feature and takes the alignment between networks directly. Equations (3) and (4) show the fused features $\mathbf{s}_{\mathrm{liner}}$ and $\mathbf{s}_{\mathrm{fc}}$, respectively. The calculation process in late fusion is shown in Figure 2 (c). Let $\mathbf{p} = [p_{\mathrm{ang}}, p_{\mathrm{joy}}, p_{\mathrm{sad}}, p_{\mathrm{neu}}]$ and $\mathbf{q} = [q_{\mathrm{ang}}, q_{\mathrm{joy}}, q_{\mathrm{sad}}, q_{\mathrm{neu}}]$ denote the predictions of speech and text recognizers, respectively. Also, $\mathbf{r} = [r_{\mathrm{ang}}, r_{\mathrm{joy}}, r_{\mathrm{sad}}, r_{\mathrm{neu}}]$ represents the prediction using the feature from early fusion, where FC is a fully connected layer. Each prediction contains the scores of four emotions: joy, sadness, anger, and neutral.

$$\mathbf{s}_{\mathrm{linear}} = \mathbf{p} + \mathbf{q} + \mathbf{r} \tag{3}$$

$$\mathbf{s}_{\mathrm{fc}} = \mathrm{FC}([\mathbf{p}, \mathbf{q}, \mathbf{r}]) \tag{4}$$

Equation (3) is a simple linear addition. Equation (4) is the concatenation for outputs from different networks and makes it possible to train all proposed networks. The emotion with the maximum value of $\mathbf{s}$ is the final prediction in this model. Eventually, the proposed method has four forms of combinations of two early fusion and two late fusion processes.

IV. Experiment

A. Dataset

In our experiment, we used JTES, a database that contains Japanese emotional speech. It is one of the most commonly used emotional speech data corpora for Japanese SER [23]. JTES was created with a balance of phonemes and prosodies. Sentences for recording were taken from Twitter. The method of selecting sentences is as follows. Takeishi et al. [5] collected 34 million tweets from 20,000 Twitter users and selected 124,000 tweets from among them. Extracted tweets were thinned out using the emotional polarity recognition. Moreover, these were selected with an entropy-based balancing algorithm to balance phonemes and prosodies. Eventually, they determined 50 different sentences for each emotion by subjective evaluation. This dataset consists of 20,000 emotional utterances by 50 male and 50 female speakers. Each speaker utters 50 sentences with one of four emotions: joy, sadness, anger, or neutral. In evaluating emotion recognition with acoustic features, we divided the dataset into five parts without allowing duplication of speakers and evaluated the method with speaker-open cross-validation. In evaluating emotion recognition with linguistic features, we configurated five folds from the dataset

TABLE I
RESULTS OF EMOTION RECOGNITION WITH ACOUSTIC AND LINGUISTIC FEATURES

		WA (%)	UA (%)	Accuracy of each emotion (%)			
				Ang	Joy	Sad	Neu
Acoustic feature	CNN + BLSTM + Attention	68.47	71.31	63.59	74.51	72.47	74.68
Linguistic feature	BERT	65.95	66.34	67.67	65.52	64.87	67.30
Acoustic + Linguistic features*	Early (cat) [20]	80.18	80.60	84.03	79.58	71.26	87.53
	Early (cat) + Late (linear) [22]	80.38	80.56	82.48	80.90	77.55	81.31
	Late (linear) [21]	90.70	91.22	86.15	92.07	91.48	95.20
	Early (mul) [Ours]	85.67	86.14	88.85	84.65	80.76	90.30
	Early (mul) + Late (linear) [Ours]	**86.33**	**86.92**	**90.14**	82.41	**82.98**	92.13
	Early (cat) + Late (fc) [Ours]	80.08	80.27	81.48	81.40	73.44	84.75
	Early (mul) + Late (fc) [Ours]	85.95	86.52	87.86	**86.94**	78.80	**92.47**

*Early (cat/mul): Early fusion using Eq. (1) or (2), Late (linear/fc): Late fusion using Eq. (3) or (4)

with no text duplication and analyzed the performance by text-open cross-validation. In evaluating emotion recognition with acoustic and linguistic features, the performance evaluation is the same as above.

B. Setup for emotion recognition with acoustic features

In the configuration of emotion recognition with acoustic features, we adopted the combination of CNN, BLSTM, and Attention, which is shown in Figure 1. The input was a segment of fixed frames with a half-overlap of the segment. The sampling rate was 16,000 Hz. We extracted the mel-cepstrum of 36 dimensions from speech signals. The analysis frame length was 5 ms using the WORLD analyzer [24]. The network was trained in 100 epochs, eight per batch, with early stopping. The optimizer was Adam with a warm-up, the learning rate was 0.0001, and the objective function was Cross Entropy.

C. Setup for emotion recognition with linguistic features

In the configuration of emotion recognition with linguistic features, we adopted BERT [25] and used pretrained parameters. This model had been trained with extensive data from Wikipedia in Japanese [26]. When BERT was retrained, pretrained parameters were frozen.

Since text data in JTES are insufficient for the learning of the deep neural network, we implemented data augmentation for text when emotion recognition was pretrained with text. To extend text data, a Japanese sentence from JTES was translated to a foreign language and retranslated back to Japanese. The set of foreign languages included the top 14 popular languages. The category of anger had the smallest amount of augmented text data. Each of the categories of emotion had 434 sentences selected randomly from extended data. Eventually, text data were augmented to include 1736 sentences. The network was trained in 300 epochs, eight per batch, with early stopping. The optimizer was Adam with a warm-up, the learning rate was 0.0001, and the objective function was Cross Entropy.

D. Setup for emotion recognition with acoustic and linguistic features

In the configuration of emotion recognition with acoustic and linguistic features, we proposed the network model using pretrained speech and text emotional recognizers for JTES. In this paper, we selected single modal SER, Early (cat), Early (cat) + Late (linear), and Late (linear) as conventional methods; these were often used in previous studies. We compared the proposed method with them. The network architecture of pretrained emotion recognition with acoustic features is shown in Figure 1. The network architecture of pretrained emotion recognition in the text is BERT. These model parameters were frozen. Both of them were adapted to early fusion (cat/mul) and late fusion (linear/fc). The network architectures of the fusion part are shown in Figure 2. When only late fusion was used, we adopted late fusion (linear) and did not train the fusion model. In training, we calculated loss as the distance between the final output in the fused network and the correct labels. Training and testing data comprised 200 texts and 20,000 speeches. The network was trained in five epochs, eight per batch, with early stopping. The optimizer was Adam with a warm-up, the learning rate was 0.0001, and the objective function was Cross Entropy.

E. Result

Table I shows the experimental results and weighted and unweighted accuracies, including the accuracy of each emotion. Weighted accuracy (WA) is the average of overall accuracy, and unweighted accuracy (UA) is the average accuracy of each emotion. According to the recognition results of the experiment with only the acoustic or linguistic feature, WA and UA are about 66% to 71%. According to the results of experiments with acoustic and linguistic information, all

proposed methods have an approximately 14% to 21% higher accuracy than conventional methods, except Late (linear).

Comparing only Early (cat) with Early (mul), we find that multiplication has approximately 5% higher WA and UA than concatenation in early fusion. This result shows the possibility that the proposed method of multiplication has a synergistic effect on middle features, especially in the accuracy of sadness.

Comparing Early (cat) + Late (linear) with Early (mul) + Late (linear) reveals that the method with multiplication has higher WA and UA than that with concatenation in early fusion. Moreover, comparing these methods with the methods of early fusion only shows that the methods combined with early fusion and late fusion with simple linear combination have higher WA and UA. In the result of Early (mul) + Late (linear), the accuracies of anger and neutral are higher than 90%. These results show the possibility that the method with the multiplication of middle features and the summing of detections has a synergistic effect, especially in the accuracy of anger and neutral.

Comparing Early (cat) + Late (fc) with Early (mul) + Late (fc), the method with multiplication is seen to have higher WA and UA than that with concatenation in early fusion. Comparing these methods with the methods of early fusion only shows that the methods combined with early fusion and late fusion with full connection have no significant improvement of WA and UA. These results show the possibility that the method with full connection in late fusion has slight overfitting.

Comparing Early (cat) + Late (linear) with Early (cat) + Late (fc) reveals that there are some different accuracies in each emotion, but these two methods exhibit almost equal accuracies. Comparing Early (mul) + Late (linear) with Early (mul) + Late (fc), we find that the methods combined with multiplication and linear have higher accuracies for anger and sadness than that with multiplication and full connection. These results show no significant difference between linear and full connection in late fusion in improving accuracy.

Comparing all proposed methods, Late (linear) was found to have the highest WA and UA. For the accuracy of each emotion, Early (mul) + Late (linear) has the highest accuracy for anger. These results show the possibility that detection-based fusion easily relates multimodalities, although it may not be possible to show associations among different features sufficiently depending on the emotion. Hence, it is possible that the proposed method with early fusion and late fusion is effective in improving accuracy for specific emotions.

The following is a summary of the above. The method with late fusion only is more effective for improving the recognition performance than the integration of early fusion and late fusion in this experiment. On the contrary, the proposed method with multiplied early fusion and the summing of late fusion makes it possible to improve the accuracy for anger.

V. Conclusions

In this paper, to improve SER using the Japanese speech database, we proposed a method of integrated early fusion and late fusion with acoustic and linguistic features. As a result, all proposed methods have an almost 14% to 21% higher accuracy than classifiers for only acoustic or linguistic features. Furthermore, using early fusion with multiplication and late fusion with linear addition was useful for recognizing the category of anger. It is considered that these techniques can capture the part of complementary relationships between acoustic and linguistic features in specific emotions. In our future work, we will analyze the proposed method using automatic speech recognition and reconsider the fusion method or training parameters. We will also try to use another speech dataset that includes four or more emotional categories or foreign languages and evaluate the effectiveness of the proposed method.

VI. Acknowledgements

This work is supported by Grant-in-Aid for Scientific Research (C) No. 20K11898.

References

[1] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," Information Sciences, vol. 509, pp. 150–163, 2020.

[2] Y. Gao, Z. Pan, H. Wang, and G. Chen, "Alexa, my love: Analyzing reviews of amazon echo," 2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 372–380, 2018.

[3] K. Y. Huang, C. H. Wu, M. H. Su, and Y. T. Kuo, "Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model," IEEE Transactions on Affective Computing, vol. 11, no. 3, pp. 393–404, 2020.

[4] W. Li, Y. Zhang, and Y. Fu, "Speech emotion recognition in e-learning system based on affective computing," Third International Conference on Natural Computation (ICNC 2007), vol. 5, pp. 809–813, 2007.

[5] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 16–21, 2016.

[6] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik,"Speech emotion recognition from spectrograms with deep convolutional neural network," 2017 International Conference on Platform Technology and Service (Plat-Con), pp. 1–5, 2017.

[7] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, et al., "Speech emotion classification using attention-based lstm," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1675–1685, 2019.

[8] Y. Li, T. Zhao, and T. Kawahara, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning," in Proc. Interspeech 2019, pp. 2803–2807, 2019.

[9] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440–1444, 2018.

[10] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, pp. 1–12, 2013.

[11] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751, 2014.

[12] M. Su, C. Wu, K. Huang, and Q. Hong, "Lstm-based text emotion recognition using semantic and emotional word vectors," 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), pp. 1–6, 2018.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, et al.,"Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, 2017.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova,"BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, 2019.
[15] H. Al-Omari, M. A. Abdullah, and S. Shaikh, "Emodet2: Emotion detection in english textual dialogue using bert and bilstm models," 2020 11th International Conference on Information and Communication Systems (ICICS), pp. 226–232, 2020.
[16] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal,"SemEval-2019 task 3: EmoContext contextual emotion detection in text," in Proc. the 13th International Workshop on Semantic Evaluation, pp. 39–48, 2019.
[17] A. Khare, S. Parthasarathy, and S. Sundaram, "Multimodal embeddings using multi-task learning for emotion recognition," in Proc. Interspeech 2020, pp. 384–388, 2020.
[18] B. T. Atmaja and M. Akagi, "Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm," Speech Communication, vol. 126, pp. 9–21, 2021.
[19] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98–125, 2017.
[20] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech emotion recognition using speech feature and word embedding," 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 519–523, 2019.
[21] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, et al., "Deep neural networks for emotion recognition combining audio and transcripts," in Proc. Interspeech 2018, pp. 247–251, 2018.
[22] M. Chen, X. Zhao, "A Multi-scale Fusion Framework for Bimodal Speech Emotion Recognition," in Proc. Interspeech 2020, pp. 374-378, 2020.
[23] Y. Chiba, T. Nose, A. Ito, "Multi-stream Attention-based BLSTM with Feature Segmentation for Speech Emotion Recognition," in Proc. Interspeech 2020, pp. 3301-3305, 2020.
[24] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," IEICE Transactions on Information and Systems, vol. E99.D, no. 7, pp. 1877–1884, 2016.
[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al., "Transformers: State-of-the-art natural language processing," in Proc. the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45, 2020.
[26] Pretrained Japanese BERT models. https://github.com/cl-tohoku/bert-japanese.