IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 October 2022, accepted 26 October 2022, date of publication 3 November 2022, date of current version 8 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3219094

RESEARCH ARTICLE

Speech Emotion Recognition Based on Self-Attention Weight Correction for Acoustic and Text Features

JENNIFER SANTOSO[1], TAKESHI YAMADA[1], (Member, IEEE), KENKICHI ISHIZUKA[2], TAIICHI HASHIMOTO[2], AND SHOJI MAKINO[1,3], (Life Fellow, IEEE)

[1]Degree Programs in Systems and Information Engineering, University of Tsukuba, Ibaraki 305-8573, Japan
[2]RevComm, Inc., Tokyo 150-0002, Japan
[3]Graduate School of Information, Production and Systems, Waseda University, Fukuoka 808-0135, Japan

Corresponding author: Jennifer Santoso (j.santoso@mmlab.cs.tsukuba.ac.jp)

This work was supported by JST SPRING under Grant JPMJSP2124.

ABSTRACT Speech emotion recognition (SER) is essential for understanding a speaker's intention. Recently, some groups have attempted to improve SER performance using a bidirectional long short-term memory (BLSTM) to extract features from speech sequences and a self-attention mechanism to focus on the important parts of the speech sequences. SER also benefits from combining the information in speech with text, which can be accomplished automatically using an automatic speech recognizer (ASR), further improving its performance. However, ASR performance deteriorates in the presence of emotion in speech. Although there is a method to improve ASR performance in the presence of emotional speech, it requires the fine-tuning of ASR, which has a high computational cost and leads to the loss of cues important for determining the presence of emotion in speech segments, which can be helpful in SER. To solve these problems, we propose a BLSTM-and-self-attention-based SER method using self-attention weight correction (SAWC) with confidence measures. This method is applied to acoustic and text feature extractors in SER to adjust the importance weights of speech segments and words with a high possibility of ASR error. Our proposed SAWC reduces the importance of words with speech recognition error in the text feature while emphasizing the importance of speech segments containing these words in acoustic features. Our experimental results on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset reveal that our proposed method achieves a weighted average accuracy of 76.6%, outperforming other state-of-the-art methods. Furthermore, we investigated the behavior of our proposed SAWC in each of the feature extractors.

INDEX TERMS Speech emotion recognition, confidence measure, automatic speech recognition, self-attention mechanism.

I. INTRODUCTION

Human speech is the most basic and widely used communication modality. Human speech contains various types of information aside from the speech content, such as emotion and speaking styles. In particular, emotion is a cue important for understanding a speaker's intention. Therefore, speech emotion recognition (SER) is becoming a powerful technology to develop because of its enormous potential to achieve more natural human-computer and human-human interactions. SER is an essential component of many speech-based applications, such as call-center analysis [1], [2], [3], [4], spoken dialog systems [5], [6], and education [7].

To date, many groups have attempted to improve the performance of SER. The most common method is to extract statistical frame-based acoustic features such as Mel frequency cepstral coefficients (MFCCs), fundamental frequency (F0), and power, and then input them to classifiers such as a Gaussian mixture model (GMM) and a support vector machine (SVM) [8], [9]. In contrast, recently, deep neural network

The associate editor coordinating the review of this manuscript and approving it for publication was Fanbiao Li.

(DNN)-based approaches have been intensively applied to SER, resulting in significantly higher performance [10], [11], [12], [13], [14]. One possible reason for this higher performance is that DNN can automatically learn the representations of emotions from speech data. One of the DNN types that are particularly effective in extracting these acoustic features is bidirectional long short-term memory (BLSTM) [15], which has been widely applied to speech-related tasks because of its capability of handling and capturing sequential information. These DNN-based approaches can extract critical information without the need to hand-craft the acoustic features.

BLSTM-based networks have one demerit: information loss after going through long sequences. This issue causes some of the information in the earlier sequences to be considered not important in the end. To solve this issue, recent studies put focus and weight on the important parts and used the weights to calculate another sequence. The attention mechanism [16], which is a neural-network-based mechanism to capture the contextual information from a sequence, has been introduced. In particular, the self-attention mechanism [17] is widely used in classification tasks. The self-attention mechanism focuses on the important parts and applies weights in the same sequence. Together with LSTM-based networks, the use of the self-attention mechanism has improved the performance of many classification tasks, including SER [18], [19], [20], [21].

SER also benefits from combining the information in speech: acoustic features and textual content. One method to obtain the textual content is through manual transcriptions. Transcriptions provide accurate speech content and are reliable for SER tasks. Studies showed that the methods using acoustic features and transcriptions are effective in recognizing emotions [22], [23]. However, obtaining transcriptions is impractical because it requires numerous human annotators and transcriptions are not available in real time. In recent years, the automatic speech recognizer (ASR) has enabled us to automate the speech-to-text transcription.

State-of-the-art ASR has achieved excellent performance in recognizing speeches and is available in many languages. However, even for the most common languages, ASR is still prone to recognition errors, particularly when the speech contains emotion. Moreover, since the ASR text from emotional speeches is used as the text feature for SER, the attention mechanism might focus on incorrectly recognized words, decreasing the SER performance.

To solve this problem, we propose a BLSTM and self-attention-based SER method using self-attention weight correction (SAWC) with a confidence measure (CM) [24], which is an indicator of the reliability of ASR results. The idea is that the CM adjusts the importance weights in the acoustic features and text information following the possibility of a speech recognition error in each word and its corresponding speech segments. In terms of acoustic features, our proposed method improves the SER performance by using the emotional cues in speech recognition errors. SAWC emphasizes the importance of speech segments with low CMs; this is important for indicating the presence of emotion in a speech segment. On the other hand, our proposed method helps mitigate the speech recognition error effects on SER performance in terms of text features. SAWC reduces the importance of words that contain low CMs. Low CMs in words in ASR results indicate the high possibility of speech recognition error. The mechanism of our proposed method for text features helps mitigate the speech recognition error effects on SER performance.

In our previous work [25], we proposed using CM to mitigate the SER performance deterioration due to speech recognition errors. We showed that using CM to correct self-attention weights on the text feature extractor provides the highest SER improvement compared with using CM as a part of an input feature or an extracted feature. In this work, we generalize the previous work by including SAWC using CM as the key component in the acoustic and text feature extractors of our SER method. We evaluate the effectiveness of the SAWC in each feature extractor in improving the SER performance.

The remainder of this paper is organized as follows. In Section II, we discuss the related work on the SER classifier using acoustic and ASR results as input, as well as some recent methods to deal with ASR performance degradation in SER performance. In Section III, we describe the base SER classifier using acoustic and ASR text input. In Section IV, we describe our proposed method using the SAWC on acoustic and text feature extractors. The experimental setup, results, and discussion are presented in Section V. Finally, in Section VI, we conclude the study and suggest our future work.

II. RELATED WORK

SER has been an increasingly important field of study. In this section, we introduce some of the SER methods using ASR features and explore ways to overcome the drawbacks of ASR errors on SER performance.

SER methods using ASR features have been recently studied, and many of them have achieved state-of-the-art SER performances. In some studies, ways to fuse the acoustic and text features from ASR results were investigated [23], [26], [27], [28], [29]. However, it might be challenging to fuse the acoustic and text features optimally. Other existing methods adopt transfer learning from a pre-trained ASR model [30], [31] or the intermediate layer output of ASR as a feature for SER [32]. However, the information provided by these methods does not include text information itself, which contains essential cues for emotion recognition. Moreover, in one study [33], it is shown that the methods with the text feature from ASR results, in addition to other input features, still provide better performance for SER than those without, indicating that text information is still essential to improve SER performance.

In some of the previous studies [23], [26], [27], the effectiveness of the SER methods on both transcriptions and ASR results was tested. The experimental results showed that using ASR results degrades the overall SER performance compared with using transcriptions. Emotions cause ASR performance degradation, an ongoing issue that many studies need to address.

One of the most intuitive methods to alleviate this issue is to improve the ASR performance for emotional speeches. The lower word error rate has been shown in some studies [34], [35] to correlate with higher SER performance. Therefore, it would be reasonable to reduce speech recognition errors for emotional speeches to improve SER performance. In one study [36], it was proposed to conduct SER based on BLSTM and self-attention mechanisms using ASR results while fine-tuning ASR simultaneously, which results in SER performance improvement and ASR robustness to emotions. The main drawback of this method is that the fine-tuning of ASR performance alongside SER is computationally costly. Moreover, it is difficult to collect data with emotional labels and transcriptions that are essential to perform fine-tuning properly. Combining the ASR results from different recognizers and text embeddings has reduced the impact of ASR performance degradation on SER [35]. In this method, a late fusion is performed in the form of majority voting based on SVM predictions on the individual feature sets combining acoustic features and text features obtained from different text encodings and pre-trained speech recognizers. These combinations can reduce the incorrect emotion recognition caused by incorrect recognition results without adapting the ASR to emotional speeches. However, this method is complex and requires access to many different speech recognizers, text encodings, and acoustic features. Another approach to indirectly reducing the effects of ASR errors on SER performance is to employ a joint self-supervised training approach using text, audio, and visual information, as shown in some studies [37], [38]. The experiment using ASR results as text information showed that adding visual information to the text and audio as the main information can improve the SER performance. Although the performance is improved, visual information is unavailable in many situations, making it challenging to apply this approach practically.

In our study, we regard ASR performance deterioration in emotional speeches as an essential cue for the presence of emotion in different segments of speeches. For this reason, we keep ASR and its results as is and instead utilize CM in ASR results, which might hold essential emotional cues in the speech segments, in our proposed SAWC to improve SER performance. Our proposed method improves SER performance without the need for the fine-tuning of ASR to be robust to emotional speeches. The fine-tuning of ASR helps reduce speech recognition errors in emotional speeches, although it removes many essential emotional cues in the speech segments. In addition, our proposed method assumes the accessibility to only one speech recognizer and focuses on utilizing helpful information found in the ASR results, in contrast to the approach that combines multiple speech recognizers and the one that utilizes other types of information outside of speech and text.

III. BASIC SER METHOD

A. OVERVIEW

Figure 1 shows the general structure of a basic SER method, which consists of the acoustic feature extractor, the text feature extractor, and the emotion classifier. First, the SER method receives inputs of speech and its ASR text. These inputs are then fed to the acoustic and text feature extractors. Then, the outputs of the acoustic feature extractor $\mathbf{z}_{acoustic}$ and text feature extractor $\mathbf{z}_{text}$ are concatenated as the output of the intermediate layer $\mathbf{z} = \mathbf{z}_{acoustic} \oplus \mathbf{z}_{text}$. Finally, these $\mathbf{z}$ data are fed to the emotion classifier consisting of a dense network, which outputs the probability of each emotion class.

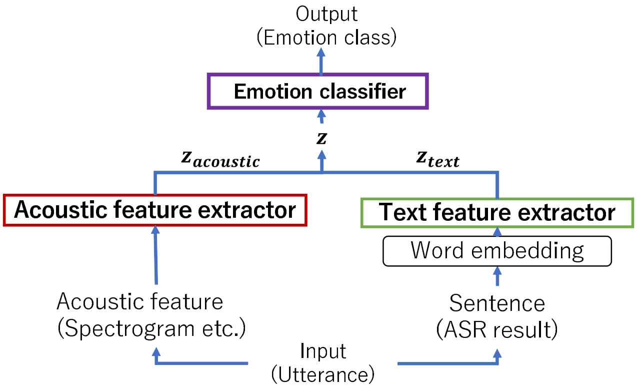

FIGURE 1. General structure of a basic SER method.

In this study, the acoustic and text feature extractors are based on the self-attention-based BLSTM. BLSTM is one of the most commonly used networks to handle sequential data, and the self-attention mechanism helps focus on the important parts of the output of the BLSTM.

B. ACOUSTIC FEATURE EXTRACTION

Figure 2(a) illustrates the acoustic feature extractors in the basic SER method. The input of our basic SER method consists of MFCC, constant Q-transform (CQT), and F0. These features with sequence length T, defined as $\mathbf{x}_1, \ldots, \mathbf{x}_T$, are then fed to the BLSTM to obtain $\mathbf{e}_i$, which is defined for each frame index i as

$$\mathbf{e}_i = \mathbf{g}_i \oplus \mathbf{h}_i, \tag{1}$$

where $\mathbf{g}$, $\mathbf{h}$, and $\oplus$ represent the forward hidden states of BLSTM, backward hidden states of BLSTM, and concatenation, respectively.

As not all the segments in an acoustic feature contribute to the emotion class in a speech, we employ the self-attention mechanism [17] that processes the output of the BLSTM. The self-attention mechanism helps focus on the specific words or segments from the input, which contains essential information on emotion. In the acoustic feature extraction,

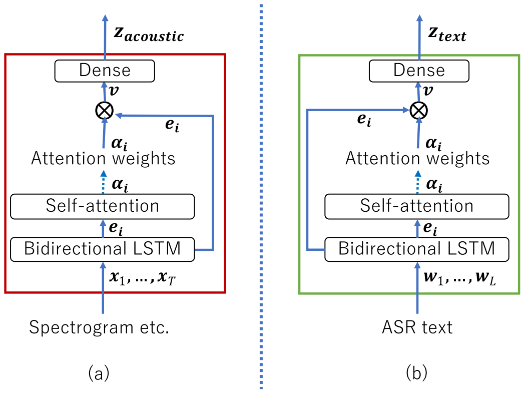

FIGURE 2. Feature extractors in the basic method: (a) acoustic feature extractor, (b) text feature extractor.

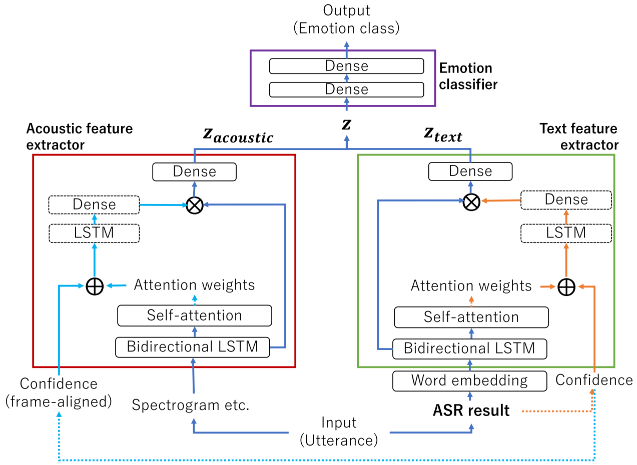

FIGURE 3. Proposed SER method.

the self-attention mechanism weighs the importance of $\mathbf{e}_i$, defined as

$$y_i = \mathbf{m}tanh(\mathbf{N}\mathbf{e}_i^T), \tag{2}$$

$$\alpha_1, \ldots, \alpha_T = softmax(y_1, \ldots, y_T). \tag{3}$$

α_i is the attention weight at frame i, and $\mathbf{m}$ and $\mathbf{N}$ are trainable parameters that can be represented as a layer of a dense neural network. The weighted sum $\mathbf{v}$ from BLSTM and attention weights is defined as

$$\mathbf{v} = \sum_{i=1}^{T} \alpha_i \mathbf{e}_i. \tag{4}$$

After the weighted sum $\mathbf{v}$ is calculated, it is fed to a single fully connected layer to obtain a fixed-length intermediate layer representation, $\mathbf{z}_{acoustic}$, of acoustic features.

C. TEXT FEATURE EXTRACTION

Information from texts consists of word sequences and therefore must be converted to numeric vectors before being fed to any deep-learning-based methods. The most common way to process the texts is through word embeddings, which are vector representations of words. The word embeddings enable the deep learning-based methods to process the text data.

Figure 2(b) illustrates the text feature extractor in the basic SER method. The text feature extractor in the SER method uses ASR text of the input utterance. ASR text is first encoded by bidirectional encoder representations from transformers (BERT) word embedding [39]. The resulting embeddings with length L, defined as $\mathbf{w}_1, \ldots, \mathbf{w}_L$, are then fed to the text feature extractor by the same process as that of the acoustic feature extractor. Here, we obtain the fixed-length intermediate layer representation $\mathbf{z}_{text}$ of text features.

IV. PROPOSED METHOD

A. PROBLEMS OF THE BASIC SER METHOD

In this section, we discuss the problems of the basic SER method and then explain the details of our proposed method. One main issue with the basic SER method is that the SER performance deteriorates owing to ASR errors. One of the most prominent causes of these errors is the presence of emotion in speech, because emotion changes the intonation and pronunciation of the intended speech content.

As a reference, we investigated the word error rate of ASR for neutral speeches and emotional speeches. In this study, we use a pretrained ASR based on the Kaldi speech recognition toolkit [40], using the speech data from Librispeech [41], which consists of English speech from audiobooks. The word error rate for this dataset is 3.8% under the clean condition. On the other hand, the word error rate for the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [42], the English emotional speech dataset used to evaluate our proposed method, using the same pretrained ASR is 43.5%. The high word error rate for the IEMOCAP dataset indicates that the presence of emotion considerably deteriorates ASR performance.

Since ASR is not robust to the presence of emotion in speech, the ASR text would contain many speech recognition errors upon recognizing emotional speech. As explained in the previous section, the acoustic and text feature extractors of the basic SER method use BLSTM and a self-attention mechanism. The self-attention mechanism focuses on the important words and speech segments to determine emotion. Focusing on the words incorrectly recognized by ASR results in many incorrectly recognized emotions, thus deteriorating the SER performance. One of the solutions to this problem is to improve ASR performance to be robust to emotions through retraining or fine-tuning as shown in Section II. However, this solution requires a high computational cost and might not be effective in improving SER performance. The essential information regarding the presence of emotion in segments containing speech recognition errors, which can be focused on by a self-attention mechanism in acoustic feature extraction, might be lost owing to ASR being more robust to emotions.

We propose a method to improve the basic SER method by adjusting the self-attention weights using CM and named this method self-attention weight correction (SAWC). It is a critical component in the acoustic and text feature extractors. SAWC resolves the issue without retraining or fine-tuning ASR to be robust to emotions. The proposed SER method is illustrated in Figure 3. The details of CM and the proposed SAWC are explained in Sections IV-B and IV-C, respectively.

B. CM

In the field of speech recognition, one of the most prominently used metrics for ASR reliability is CM [24]. CM indicates how reliable ASR results is. CM falls in the range of 0 to 1; 0 indicates an unreliable result and 1 indicates a reliable result. CM has long been used in ASR to evaluate word-level and sentence-level recognition results, accurately discriminating parts that contain possible speech recognition errors. We employ CM in the Kaldi speech recognition toolkit, which is based on the lattice posterior estimation. An example of ASR results and CM aligned for each speech segment and its corresponding spectrogram are illustrated in Table 1 and Figure 4, respectively.

TABLE 1. Example of ASR results including the speech segments and their CMs.

start (s)	duration (s)	result (1st candidate)	CM
0.70	0.65	CLEARLY	1.00
2.27	0.14	YOU	0.97
2.41	0.11	KNOW	0.86
2.52	0.42	D‘AVRIL	0.38
2.95	0.19	AGO	0.39
3.15	0.78	SUPERVISOR	1.00
3.93	0.13	OR	0.78
4.06	0.41	SOMETHING	1.00
4.47	0.33	YAH	0.44

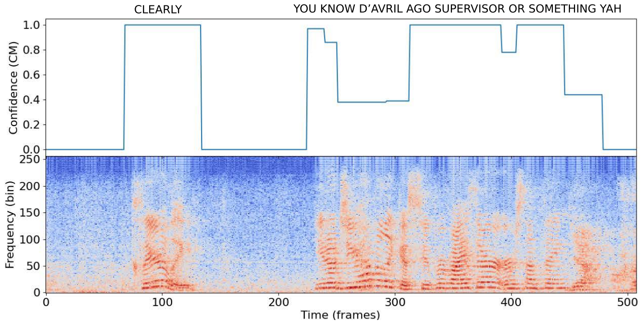

FIGURE 4. CM aligned for each speech segment and its corresponding spectrogram. For display purposes, the CM of the silent segment is set to 0.

C. SAWC USING CM

1) TEXT ATTENTION WEIGHT CORRECTION

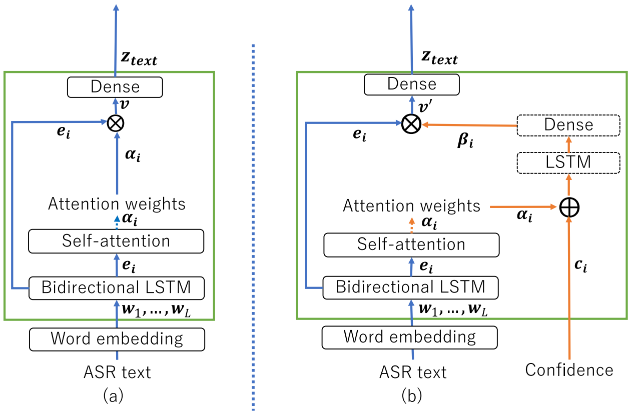

FIGURE 5. Structures of the text feature extractors of (a) the basic SER method and (b) our proposed method with SAWC.

In text features, SAWC aims to mitigate the effects of ASR error on SER performance. SAWC here uses CM to suppress incorrectly recognized words or emphasize the more correctly recognized words. Figure 5 illustrates the structures of the text feature extractors of the basic SER method and our proposed method with SAWC. In both text feature extraction structures, the flow begins with inputting the text feature consisting of BERT word embeddings. These features are fed to the text feature extractor using BLSTM and the self-attention mechanism. CM c_i is concatenated with self-attention weights α_i and will be fed to an LSTM network and dense network. The output from the network is then normalized using the softmax function to obtain new attention weights. SAWC is defined as

$$s_i = Dense(LSTM(\alpha_i \oplus c_i)), \tag{5}$$

$$\beta_1, \ldots, \beta_T = softmax(s_1, \ldots, s_T), \tag{6}$$

where $\beta_1, \ldots, \beta_T$ indicate the resulting self-attention weights. Here, the LSTM layer learns and adjusts the attention weights by also considering the CM sequence. $\beta_1, \ldots, \beta_T$ are then used to calculate the weighted sum of the BLSTM outputs defined as

$$\mathbf{v}' = \sum_{i=1}^{T} \beta_i \mathbf{e}_i, \tag{7}$$

where $\mathbf{v}'$ represents the new weighted-sum feature, now used as the updated $\mathbf{z}_{text}$.

In a previous study [25], we investigated three applications of CM in text feature extraction: early fusion, late fusion, and SAWC. The early fusion concatenates CM with the input text feature before inputting it into BLSTM, whereas the late fusion concatenates CM with the features extracted by BLSTM. As indicated in the previous study, the result shows that the SAWC mentioned above achieves the best performance among the various applications.

2) ACOUSTIC ATTENTION WEIGHT CORRECTION

Figure 6 illustrates the structures of acoustic feature extractors of the basic SER method and our proposed method with SAWC. In both acoustic feature extractors, the flow begins with inputting the acoustic features MFCC, CQT, and F0. These features are fed to the acoustic feature extractor consisting of BLSTM and the self-attention mechanism.

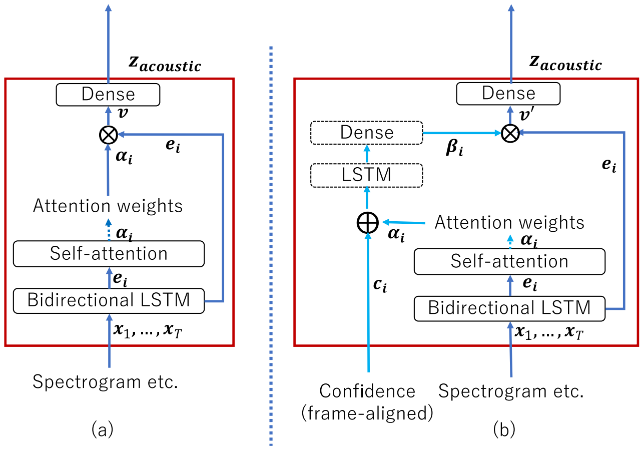

FIGURE 6. Structures of acoustic feature extractors of (a) the basic SER method and (b) our proposed method with SAWC.

On the basis of the application of CM in the text features, we apply SAWC to acoustic feature extraction. The aim of SAWC in the acoustic features is to utilize the information contained in the speech segments having a high probability of ASR errors and focus on these segments. The idea is that emotions cause pronunciation changes in specific speech segments resulting in ASR errors; therefore, the speech segments with a high probability of ASR errors contain information helpful in determining emotions. Here, we align the CM from each word to the corresponding speech segments in the acoustic features, as illustrated in Figure 4. Since SAWC needs to have the same sequence length of CM as that of self-attention, for each of the words in the recognition result, the CMs on the start and end times are aligned to the corresponding speech segments. The silent segments are assumed to be correctly recognized, and CM on those segments is set to 1. The aligned CM, which has the same length as the acoustic feature and the calculated self-attention weights, is used for SAWC in the acoustic features. The self-attention weights of the acoustic features are concatenated with the aligned CM and updated, similarly to the calculations in Eqs. (5) and (6). The updated self-attention weights of the acoustic feature are then multiplied by the output of the BLSTM in the acoustic feature extractor similarly to the calculation in Eq. (7), and the updated $\mathbf{z}_{acoustic}$ is produced.

V. EXPERIMENT AND DISCUSSIONS

A. EXPERIMENT OVERVIEW

To evaluate the effectiveness of our proposed method, we conducted three experiments using the same dataset and evaluation metrics explained in Section V-B. First, we compare the performance of the basic SER method with our proposed method with different input feature combinations. The input features used for each of the SER methods include the combination of acoustic and text features, where the text features can be either transcriptions or ASR results. We evaluate the performance of the proposed method with SAWC applied to only acoustic features, text features, and both features. We used the same input features and classifier specifications throughout this experiment. In addition, we also analyze the SAWC mechanism in correcting the attention weights through visualization.

Second, we compare the performance of our proposed method with the state-of-the-art SER methods. All the state-of-the-art methods compared are deep-neural-network-based SER methods using acoustic and text information. Three methods use transcriptions as text information, while the rest use ASR results as text information. As our proposed method uses ASR results as text information, the methods that use transcriptions as text information will only be used as a reference in ideal situations and will not be used for a direct performance comparison with our proposed method.

Finally, following the visualization analysis conducted in the first experiment, we investigated whether SAWC can be replaced by CM by comparing our proposed method with a method using CM as attention weight.

B. DATASET

In this study, we use the IEMOCAP dataset [42], which is one of the benchmark datasets for emotion recognition, to evaluate the effectiveness of the proposed method. The IEMOCAP dataset was developed when conventional machine learning methods, such as support vector machines, decision trees, logistic regression, and early neural networks, were the most commonly used methods to conduct SER. It is also used to evaluate the recent deep-learning-based methods. The IEMOCAP dataset consists of approximately 12 h of speech. The IEMOCAP dataset consists of scripted and improvised emotional speeches divided into five sessions, each containing one male and one female speaker. There are ten speakers (five males and five females) in the IEMOCAP dataset. We used the data from four emotion classes (happy, sad, neutral, and angry). To make the dataset conditions similar to those in previous works, we grouped the utterances labeled as excited with the utterances labeled as happy. The experiments were performed with five-fold cross-validation. The training set consists of speech data from four sessions, and the test set consists of the speech data from the remaining one session, ensuring speaker independence. The evaluation metrics for this study are the average unweighted accuracy (UA) and the average weighted accuracy (WA), as in the previous studies. In addition, we also use the F-score as the evaluation metric for the performance of each emotion class. The UA, WA, and F-score are respectively defined as

$$\mathrm{UA} = \frac{\sum_{i=1}^{N} t_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N} t_{ij}}, \tag{8}$$

$$\mathrm{WA} = \frac{1}{N}\sum_{i=1}^{N}\frac{t_{ii}}{\sum_{j=1}^{N} t_{ij}}, \tag{9}$$

$$\text{F-score} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}, \tag{10}$$

where N is the number of classes and t_{ij} is the number of data labeled as class i and predicted as class j. The details of the dataset used in this study are shown in Table 2, and the number of speech data for each emotion class and recording session is shown in Table 3.

TABLE 2. Dataset specifications.

Dataset	IEMOCAP	
Speakers	5 males and 5 females	
Utterance length	$1-19$ s	
# of utterances	Happy	1689
	Sad	1084
	Neutral	1708
	Angry	1103

TABLE 3. Number of speech data for each emotion class and recording session.

Session	Happy	Sad	Neutral	Angry	Total
1	286	194	384	229	1093
2	335	197	362	137	1031
3	322	305	320	240	1093
4	303	143	258	327	1031
5	443	245	384	170	1242
Total	1689	1084	1708	1103	5584

C. INPUT FEATURES

The features used as input to the basic and the proposed SER method were divided into acoustic and text features. For the acoustic features, we extracted a 33-dimensional feature consisting of 20-dimensional MFCCs, 12-dimensional CQT, and one-dimensional F0. All of the acoustic features are extracted using Librosa [43]. For the text features, first, we conducted ASR on the input speeches using a recognizer based on the Kaldi acoustic recognition toolkit pretrained with the Librispeech dataset. The Librispeech dataset consists of approximately 1000 h of speeches sampled at 16 kHz. Next, we encoded ASR texts using BERT pretrained using lower-case English texts. The pretrained BERT consists of 12 layers and 110 M parameters, resulting in 768-dimensional text features. The pretrained BERT is named bert-based-uncased, a public BERT model available in [44].

D. CLASSIFIER SPECIFICATIONS

The SER consists of an acoustic feature extractor, a text feature extractor, and an emotion classifier. The basic SER method comprises a two-layer BLSTM with 128 units and a self-attention mechanism. Each of the feature extractors consists of BLSTM with 128 units and a self-attention mechanism, resulting in a 128-dimensional vector representation for $\mathbf{z}_{acoustic}$ and $\mathbf{z}_{text}$ for the acoustic and text feature extractors, respectively. SAWC in acoustic and text feature extractors uses BLSTM that receives two-dimensional inputs, providing a two-dimensional output and a dense layer that outputs a one-dimensional output. The resulting intermediate layer representation $\mathbf{z}$ is a 256-dimensional vector consisting of a 128-dimensional vector from each of the acoustic and text features. The intermediate layer representation is fed to the emotion classifier, consisting of two dense layers with (256–64–4) units. In this experiment, we used Adam [45] as the optimizer with a learning rate of 0.0001 and a weight decay of 0.00001. The dropout was set to 0.3. The batch size was set to 40. The results were taken from the highest WA out of 100 epochs.

Figure 7 shows the graphical representation of loss and WA during the training on one of the folds in our proposed method using SAWC on both acoustic and text feature extractors. Here, the x-axis represents the number of model training epochs, and the y-axis respectively represents loss and WA in the left and right graphs. In the loss representation, the loss in the training phase decreases until the last epoch, whereas that in the testing phase decreases up to epoch 80 and starts to overfit afterward. On the other hand, in the WA representation, the accuracy in the training phase improves close to 100% until the last epoch, whereas that in the testing phase increases to epoch 40 and tends to plateau afterward.

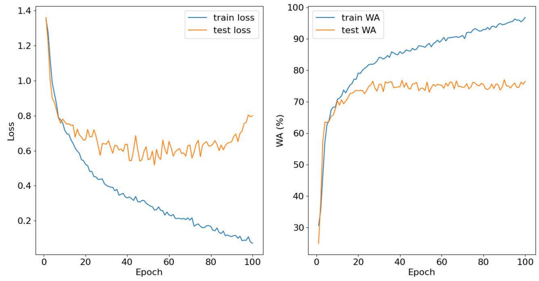

FIGURE 7. Graphical representation of loss and WA during the training of the proposed method.

E. COMPUTATION ENVIRONMENT

In the experiments, we conducted both the training and testing phases using the GPU. Owing to the large amount of calculation needed during the training phase of our proposed method, it is recommended to use GPU for training. On the other hand, the testing phase can be conducted using either GPU or CPU. The computer used for the experiment has NVIDIA Quadro RTX 6000 GPU with 24 GB of RAM and Intel Core i9-10940X 3.3 GHz CPU with 32 GB of RAM. All the programs were run using the operating system Ubuntu 18.04. All the programs were written in the Python 3 programming language using PyTorch 1.4.0 [46] as the library. The evaluation metrics are calculated using the Scikit-learn [47] toolbox.

We measured the average computational time in our proposed method of SER using SAWC on both acoustic and text features. Our proposed method ran for 41.798 s on GPU for each epoch in the training phase. On the other hand, our proposed method ran for 0.011 s on GPU and 0.375 s on CPU for each utterance in the test phase.

TABLE 4. Performance comparison of the basic SER method with different input features and our proposed method with different SAWC combinations.

Method	UA (%)	WA(%)	F-score (%)			
			Happy	Sad	Neutral	Angry
(Basic) Acoustic	61.1	64.3	47.6	70.5	56.4	69.7
(Basic) Transcriptions	75.5	75.6	82.8	77.8	65.3	75.4
(Basic) ASR text	71.8	71.9	80.4	74.6	58.6	70.9
(Basic) Acoustic + Transcriptions	78.6	78.4	83.4	81.1	68.5	82.2
(Basic) Acoustic + ASR text	73.9	74.2	79.7	71.8	64.7	81.1
(Proposed) Acoustic w/o SAWC + ASR text with SAWC [25]	75.5	75.3	81.4	76.3	65.8	79.7
(Proposed) Acoustic with SAWC + ASR text w/o SAWC	76.1	76.1	82.5	77.2	64.4	80.9
(Proposed) Acoustic with SAWC + ASR text with SAWC	76.8	76.6	83.0	77.6	66.2	80.6

F. EXPERIMENTAL RESULTS

1) COMPARISON OF THE APPLICATION OF SAWC

Table 4 shows the performance of the basic SER method with different combinations of input features and SAWCs using CM. Here, we conducted the experiment of the basic SER method using only acoustic features, only text features, and both features. The text features were divided into two types, one using human-based transcriptions provided in the dataset and the other using ASR text. The experiments were run with the same classifier specifications for each input feature.

From the comparison, the performance of the basic SER method using only acoustic features yields UA and WA of 61.1% and 64.3%, respectively. The SER method using only transcriptions yields the UA and WA of 75.5% and 75.6%, whereas that using ASR text as the input yields a lower performance of 71.8% and 71.9% in UA and WA, respectively. The method combining acoustic features and transcriptions achieved the UA and WA of 78.6% and 78.4%, respectively, which are significantly higher than those obtained by the method using only acoustic features or transcriptions. The same increase can also be observed by combining acoustic features and ASR text, achieving the UA and WA of 73.9% and 74.2%, respectively. The decline in the performance of the basic SER method using ASR text as the input text features compared with that using transcriptions is due to the performance deterioration of ASR caused by the presence of emotions in speeches, resulting in incorrect recognition results being used.

Now, we compare the SER performance of the proposed SER method using SAWC with CM on the acoustic and text features with the performance of the basic SER method using acoustic features and ASR text. First, applying SAWC with CM to the text feature only (Acoustic without SAWC + ASR text with SAWC) improved both the UA and WA compared with the basic SER method by 1.6% and 1.1% to 75.5% and 75.3%, respectively. One explanation is that SAWC considers the words with low CMs as speech recognition errors, thereby reducing the attention weights on these words and adjusting the attention weights to focus more on the correctly recognized words.

On the other hand, applying SAWC to the acoustic feature only (Acoustic with SAWC + ASR text without SAWC) also improved both the UA and WA compared with the basic method by 2.2% and 1.9% to 76.1% and 76.1%, respectively. The results show that SAWC can improve the acoustic feature extraction by adjusting the importance weights of the speech segments in accordance with CM. One possible explanation is that the speech recognition errors in the speech segments contain information essential to determining the emotion; therefore, emphasizing these parts results in their being considered more in the decision of the emotional output label.

Furthermore, the proposed SER method combining SAWC with CM on acoustic and text features yields the UA and WA of 76.8% and 76.6%, respectively, which is a further performance improvement compared with the method applying SAWC on either acoustic or text feature extractors. The result implies that combining the two types of input enhanced with SAWC can improve the overall SER performance. The performance of our proposed method is close to that of the basic SER method using acoustic features and transcription.

We also evaluated the performance using F-score for each emotion class, as shown in Table 4. Overall, the trend of improvement of the F-score for each emotion on different input features is similar to those in UA and WA. The neutral class has the lowest F-score among the four emotion classes; we will discuss this in Section V-F-3.

2) COMPARISON WITH STATE-OF-THE-ART METHODS

Next, we compare the performance of our proposed method with those of state-of-the-art methods, as shown in Table 5. Most reports did not show the results in other metrics such as F-score for each emotion class. Therefore, we compare the performance of our proposed method with those of state-of-the-art methods only in terms of UA and WA. The state-of-the-art methods used for comparison are SER methods using acoustic and text features. The text feature is further separated into transcriptions and ASR text, where ASR text has a word error rate of 43.5%, indicating that many of the speech data contain incorrect text information, whereas the transcriptions can be assumed to have no such errors.

In terms of UA and WA, our proposed method outperforms the state-of-the-art SER methods using acoustic and ASR results as input information. Although our proposed method has yet to achieve the performance of the state-of-the-art SER methods using acoustic and transcriptions, the differences in UA and WA from those of the best state-of-the-art method are 1.6% and 0.9%, respectively, which means their accuracies are similar.

TABLE 5. Proposed and state-of-the-art methods.

Method	Input	UA (%)	WA (%)
Yoon et al. [23]	Acoustic + Transcriptions	77.6	76.5
Wang et al. [49]	Acoustic + Transcriptions	77.1	76.8
Wu et al. [27]	Acoustic + Transcriptions	78.4	77.5
Kim and Shin [26]	Acoustic + ASR text	68.7	66.6
Xu et al. [28]	Acoustic + ASR text	69.5	70.4
Yoon et al. [23]	Acoustic + ASR text	73.9	73.0
Feng et al. [36]	Acoustic + ASR text	69.7	68.6
Heusser et al. [48]	Acoustic + ASR text	71.0	73.5
Wu et al. [27]	Acoustic + ASR text	75.6	74.7
Proposed method	Acoustic + ASR text + CM	76.8	76.6

3) CONFUSION MATRIX OF THE SER CLASSIFIER

Figure 8 shows the confusion matrices of the basic SER method and the proposed method, which combines both acoustic and text feature extraction with SAWC using CM. Compared with the basic SER method, most emotions except for neutral emotions have gains in the number of correctly classified speeches. The F-score for neutral emotions is only about 66.2%, whereas those for other classes are above 75%. Neutral speeches are mistaken for all classes, especially happy speeches. One possible explanation is that SAWC might have emphasized the speech segments that contain incorrect recognition results, indicating the possible presence of emotion despite the speech being neutral.

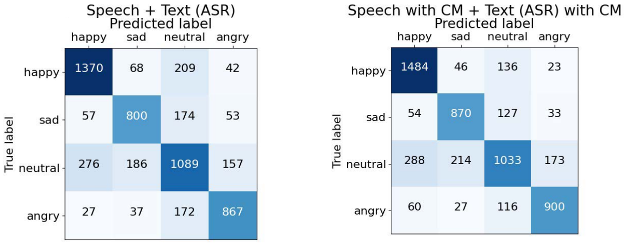

FIGURE 8. Confusion matrices of the basic SER method and our proposed method showing the best result.

4) VISUALIZATION OF SAWC

We discuss SAWC with CM in the proposed method through visualization. Here, we take an example of an utterance labeled as angry that had been incorrectly recognized as neutral by the basic SER method; it is the same utterance as the example shown in Table 1 and Figure 4.

Figure 9 shows SAWC with CM in the proposed method applied to the text feature extractor. The graphs from top to bottom respectively show the text attention weight before the update, CM aligned to each word, and the updated text attention weights. Here, some of the words with low CMs, which are more likely to be incorrectly recognized, were weighted more, whereas the words with high CM or the correctly recognized words were weighted less. By applying our proposed SAWC, we can reduce the text attention weights on the words with low CMs, whereas the words with high CMs are slightly emphasized. The visualization of the updated self-attention weights showed that applying the proposed method to the text features successfully improved the performance by suppressing the effects of ASR errors.

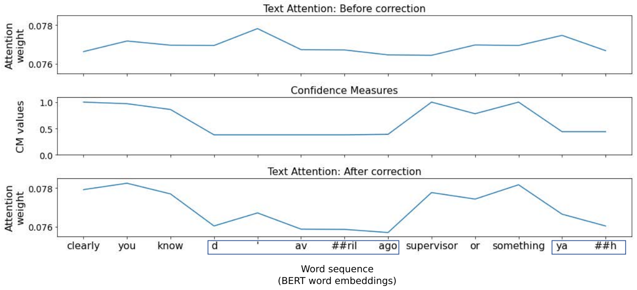

FIGURE 9. SAWC applied to the text feature extractor.

Figure 10 shows SAWC with CM in the proposed method applied to the acoustic feature extractor. The graphs from top to bottom respectively show the acoustic attention weight before the update, CM aligned to the acoustic frames and their corresponding ASR text, and the updated acoustic attention weights. Similarly to Figure 4, we set the silent speech segments to 0 for display purposes, in contrast to the experiment where the silent speech segments are set to 1. Here, the plot of updated attention weights resembles the inverted shape of the plot of CM aligned to the acoustic frame, where some parts contain the peak values from the attention weight before applying SAWC. SAWC works differently on the acoustic features from that on the text features. In the acoustic features, self-attention is adjusted to focus on the speech segments containing the speech recognition errors, which would likely have low CM. SAWC still considers the part previously focused on by self-attention, although not as much as CM. The visualization of SAWC confirms that the speech segment emphasized might be affected by emotion, thus containing information essential for SER.

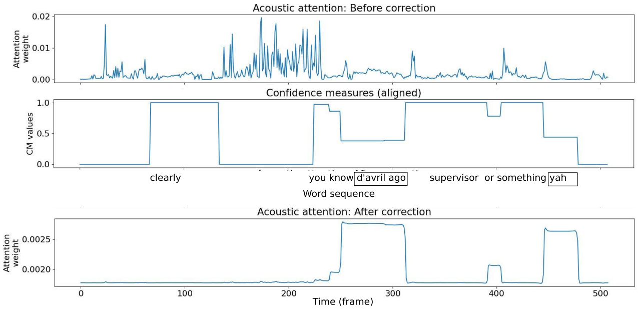

FIGURE 10. SAWC applied to the acoustic feature extractor.

5) COMPARISON WITH THE METHOD USING CM AS ATTENTION WEIGHT

To extend our results and discussion, we also investigate whether replacing the self-attention mechanism in the acoustic feature extractor with the inversely aligned CM, which is obtained by replacing CM $c_1, \ldots, c_T$ by $1 - c_1, \ldots, 1 - c_T$, yields results similar to those obtained by SAWC with CM.

TABLE 6. UA and WA of proposed method and basic SER method using CM as attention weight.

Method	UA (%)	WA (%)	F-score (%)			
			Happy	Sad	Neutral	Angry
(Proposed) Acoustic with SAWC + ASR text with SAWC	76.8	76.6	83.0	77.6	66.2	80.6
Acoustic with CM as attention + ASR text w/o SAWC	76.3	76.3	81.7	77.2	64.0	80.0
Acoustic with CM as attention + ASR text with SAWC	76.3	76.2	82.8	76.8	63.6	80.0

The inversely aligned CM is considered to be due to the updated attention weights in the bottom part of Figure 10 showing a similar pattern. Here, we substitute the attention weights with the inversely aligned CM instead of applying the correction to the attention weights.

$$r_1, \dots, r_T = softmax(1 - c_1, \dots, 1 - c_T) \tag{11}$$

$$\mathbf{c}' = \sum_{i=1}^{T} r_i \mathbf{e}_i \tag{12}$$

In this evaluation, we applied softmax to the inversely aligned CM weights and use them as the attention weights in the acoustic feature extractor.

Table 6 shows the UA and WA of the proposed method and the method with inversely aligned CM used as attention weights. The result shows that the inversely aligned CM used as attention weights yields a slightly lower UA, WA, and overall F-score for each emotion class. Despite the similarity of the updated attention weights to the inversely aligned CM, the attention weights in the proposed method before applying SAWC still hold some significance in the weights of the acoustic feature. It can be inferred that both the weights from the attention mechanism and the CM aligned to the acoustic frames are still essential in determining the important segment in the acoustic features.

VI. CONCLUSION

We proposed a BLSTM- and self-attention-based SER method using SAWC with CM. The idea is to mitigate the effects of ASR error on text feature extraction by reducing the weight of the words with low CM, which are likely to be a speech recognition error, and to emphasize the speech segments with low CM as segments with a higher probability of containing emotion in the acoustic feature. By utilizing the information from CM in ASR results and SAWC, our method can improve the SER performance. Our method does not require fine-tuning of ASR to be robust to emotion; this fine tuning incurs a high computational cost and might lose the important emotional cues in the segments with speech recognition errors. The experimental results demonstrated that our proposed method using SAWC in acoustic and text feature extractors improved the classification performance parameters UA and WA by 2.9% and 2.4%, respectively, compared with those of the basic SER method. In addition, our proposed method outperformed the state-of-the-art SER methods.

For future directions, we would like to investigate the effectiveness of our proposed method in other emotional speech corpora, including those in other languages. Other directions include applying our proposed method in practical situations such as business conversations.

REFERENCES

[1] V. Petrushin, "Emotion in speech: Recognition and application to call centers," *Proc. ANNIE*, vol. 710, p. 22, Nov. 1999.

[2] P. Gupta and N. Rajput, "Two-stream emotion recognition for call center monitoring," in *Proc. Interspeech*, Aug. 2007, pp. 2241–2244.

[3] B. Li, D. Dimitriadis, and A. Stolcke, "Acoustic and lexical sentiment analysis for customer service calls," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5876–5880.

[4] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, Y. Aono, and T. Toda, "Customer satisfaction estimation in contact center calls based on a hierarchical multi-task model," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 715–728, 2020.

[5] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion detection in task-oriented spoken dialogues," in *Proc. Int. Conf. Multimedia Expo. (ICME)*, 2003, pp. 549–552.

[6] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "A dialogical emotion decoder for speech emotion recognition in spoken dialog," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6479–6483.

[7] W. Li, Y. Zhang, and Y. Fu, "Speech emotion recognition in E-learning system based on affective computing," in *Proc. 3rd Int. Conf. Natural Comput. (ICNC)*, Aug. 2007, pp. 809–813.

[8] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic emotion recognition using prosodic parameters," in *Proc. Interspeech*, Sep. 2005, pp. 493–496.

[9] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2009, pp. 552–557.

[10] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Interspeech*, Sep. 2014, pp. 223–226.

[11] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[12] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *Proc. 9th Int. Conf. Signal Process. Commun. Syst. (ICSPCS)*, Dec. 2015, pp. 1–5.

[13] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2016, pp. 1–4.

[14] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5089–5093.

[15] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. ICANN*, vol. 2, 2005, pp. 799–804.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[17] Z. Lin, M. Feng, C. N. D. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *Proc. ICLR*, 2017, pp. 1–15.

[18] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2227–2231.

[19] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 126–131.

[20] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proc. Interspeech*, Sep. 2018, pp. 152–156.

[21] Y. Li, T. Zhao, and T. Kawahara, "Improved End-to-End speech emotion recognition using self attention mechanism and multitask learning," in *Proc. Interspeech*, Sep. 2019, pp. 2803–2807.

[22] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition," in *Proc. Interspeech*, Oct. 2020, pp. 374–378.

[23] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2822–2826.

[24] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Commun.*, vol. 45, no. 4, pp. 455–470, Apr. 2005.

[25] J. Santoso, T. Yamada, S. Makino, K. Ishizuka, and T. Hiramura, "Speech emotion recognition based on attention weight correction using word-level confidence measure," in *Proc. Interspeech*, Aug. 2021, pp. 1947–1951.

[26] E. Kim and J. W. Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6720–6724.

[27] W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6269–6273.

[28] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *Proc. Interspeech*, Sep. 2019, pp. 3569–3573.

[29] Y. Lee, S. Yoon, and K. Jung, "Multimodal speech emotion recognition using cross attention with aligned audio and text," in *Proc. Interspeech*, Oct. 2020, pp. 2717–2721.

[30] S. Zhou and H. Beigi, "A transfer learning method for speech emotion recognition from automatic speech recognition," 2020, *arXiv:2008.02863*.

[31] N. Tits, K. El Haddad, and T. Dutoit, "ASR-based features for emotion recognition: A transfer learning approach," in *Proc. Grand Challenge Workshop Human Multimodal Lang. (Challenge-HML)*, 2018, pp. 48–52.

[32] C. Chen and P. Zhang, "CTA-RNN: Channel and temporal-wise attention RNN leveraging pre-trained ASR embeddings for speech emotion recognition," in *Proc. Interspeech*, Sep. 2022, pp. 4730–4734.

[33] Y. Li, P. Bell, and C. Lai, "Fusing ASR outputs in joint training for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7362–7366.

[34] S. Sahu, V. Mitra, N. Seneviratne, and C. Espy-Wilson, "Multi-modal learning for speech emotion recognition: An analysis and comparison of ASR outputs with ground truth transcription," in *Proc. Interspeech*, Sep. 2019, pp. 3302–3306.

[35] S. Amiriparian, A. Sokolov, I. Aslan, L. Christ, M. Gerczuk, T. Hübner, D. Lamanov, M. Milling, S. Ottl, I. Poduremennykh, E. Shuranov, and B. W. Schuller, "On the impact of word error rate on acoustic-linguistic speech emotion recognition: An update for the deep learning era," 2021, *arXiv:2104.10121*.

[36] H. Feng, S. Ueno, and T. Kawahara, "End-to-end speech emotion recognition combined with Acoustic-to-Word ASR model," in *Proc. Interspeech*, Oct. 2020, pp. 501–505.

[37] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-supervised learning with cross-modal transformers for emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 381–388.

[38] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, vol. 1, 2019, pp. 4171–4186.

[40] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and J. Silovsky, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011, pp. 1–4.

[41] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[42] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[43] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.

[44] Hugging Face. *BERT Base Model (Uncased)*. Accessed: May 23, 2022. [Online]. Available: https://huggingface.co/bert-base-uncased

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NIPS*, 2019, pp. 8024–8035.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[48] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," 2019, *arXiv:1912.02610*.

[49] Y. Wang, G. Shen, Y. Xu, J. Li, and Z. Zhao, "Learning mutual correlation in multimodal transformer for speech emotion recognition," in *Proc. Interspeech*, Aug. 2021, pp. 4518–4522.

JENNIFER SANTOSO received the B.Eng. degree from Binus University, Indonesia, in 2016, and the M.Eng. degree from the University of Tsukuba, Japan, in 2020, where she is currently pursuing the Dr.Eng. degree. Her research interests include speech analysis and emotion recognition. She is a Student Member of the ASJ.

TAKESHI YAMADA (Member, IEEE) received the B.Eng. degree from Osaka City University, Japan, in 1994, and the M.Eng. and Dr.Eng. degrees from the Nara Institute of Science and Technology, Japan, in 1996 and 1999, respectively. He is currently an Associate Professor with the Faculty of Engineering, Information and Systems, University of Tsukuba, Japan. His research interests include speech recognition, sound scene understanding, multichannel signal processing, media quality assessment, and e-learning. He is a member of the IEICE, IPSJ, and ASJ.

KENKICHI ISHIZUKA received the B.Eng., M.Eng., and Dr.Eng. degrees from the University of Tsukuba, Japan, in 2005, 2007, and 2013, respectively. He is currently a Senior Research Engineer at RevComm Inc., Japan. His work focuses on the development of business communication analysis systems. He is a member of the ASJ and the JSKE.

TAIICHI HASHIMOTO received the B.Eng., M.Eng., and Dr.Eng. degrees from the Tokyo Institute of Technology, Japan, in 1997, 1999, and 2002, respectively. He is currently a Research Director at RevComm Inc., Japan. His work focuses specifically on speech and dialogue analytics in business communication. He is a member of the IPSJ, the ASJ, the ANLP, and the JSAI.

SHOJI MAKINO (Life Fellow, IEEE) received the B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1979, 1981, and 1993, respectively. He joined NTT, in 1981, and the University of Tsukuba, Japan, in 2009. He is currently a Professor at Waseda University, Japan. He has authored or coauthored more than 400 papers in journals and conference proceedings and is responsible for more than 200 patents. His research interests include adaptive filtering technologies, the realization of acoustic echo cancellation, blind source separation of convolutive speech mixtures, and acoustic signal processing for speech and audio applications. He was a recipient of 30 Awards, including the IEEE SPS Leo L. Beranek Meritorious Service Award, in 2022, the IEEE SPS Best Paper Award, in 2014, the IEEE MLSP Competition Award, in 2007, and the ICA Unsupervised Learning Pioneer Award, in 2006. He was on the IEEE SPS Board of Governors (2018–2020), Technical Directions Board (2013–2014), Awards Board (2006–2008), Conference Board (2002–2004), Fellow Evaluation Committee (2018–2020), and the Chair of the Technical Committee on Audio and Acoustic Signal Processing. He was a member of the IEEE Jack S. Kilby Signal Processing Medal Committee (2015–2018) and the James L. Flanagan Speech & Audio Processing Award Committee (2008–2011). He was an IEEE SPS Distinguished Lecturer (2009–2010), an IEICE Fellow, a Board Member of the ASJ, and a member of EURASIP.