IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 23, 2020, accepted March 9, 2020, date of publication March 30, 2020, date of current version April 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984368

Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network

NGOC-HUYNH HO, (Member, IEEE), HYUNG-JEONG YANG, (Member, IEEE), SOO-HYUNG KIM, (Member, IEEE), AND GUEESANG LEE, (Member, IEEE)
Department of Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, South Korea

Corresponding author: Hyung-Jeong Yang (hjyang@jnu.ac.kr)

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A2B5B01002085) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1A4A1015559).

ABSTRACT Speech emotion recognition is a challenging but important task in human computer interaction (HCI). As technology and understanding of emotion are progressing, it is necessary to design robust and reliable emotion recognition systems that are suitable for real-world applications both to enhance analytical abilities supporting human decision making and to design human–machine interfaces (HMI) that assist efficient communication. This paper presents a multimodal approach for speech emotion recognition based on Multi-Level Multi-Head Fusion Attention mechanism and recurrent neural network (RNN). The proposed structure has inputs of two modalities: audio and text. For audio features, we determine the mel-frequency cepstrum (MFCC) from raw signals using the OpenSMILE toolbox. Further, we use pre-trained model of bidirectional encoder representations from transformers (BERT) for embedding text information. These features are fed parallelly into the self-attention mechanism base RNNs to exploit the context for each timestamp, then we fuse all representatives using multi-head attention technique to predict emotional states. Our experimental results on the three databases: Interactive Emotional Motion Capture (IEMOCAP), Multimodal EmotionLines Dataset (MELD), and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI), reveal that the combination of the two modalities achieves better performance than using single models. Quantitative and qualitative evaluations on all introduced datasets demonstrate that the proposed algorithm performs favorably against state-of-the-art methods.

INDEX TERMS Speech emotion recognition, multi-level multi-head fusion attention, RNN, audio features, textual features.

I. INTRODUCTION

In speech enabled Human-Machine Interfaces (HMI), the context plays important role for improving the user interface. One of the important components of the context is the emotion in speaker's voice. Emotion recognition provide important priors in human decision handling, interaction and cognitive process [1], [2], making it possible to add human-like features to the HMI, such as empathy and responding with proper emotion in the text to speech engine. This fact has motivated researchers to think of speech as a fast and efficient method of interaction between human and machine. Speech emotion recognition (SER) is defined as extracting the emotional state of a speaker from his or her speech. The main objective of employing SER is to adapt the system response upon detecting frustration or annoyance in the speaker's voice.

However, the task of SER is still very challenging. First, approaching the automatic emotion recognition necessitates an appropriate emotion representation model. There are two models commonly used in recent, namely as categories and dimensions [3]. In this study, we mainly focus on emotion categories, including ''the four basic emotions'' with anger, happiness, sadness and neutral that appears to be the favored

The associate editor coordinating the review of this manuscript and approving it for publication was Jihwan P. Choi.

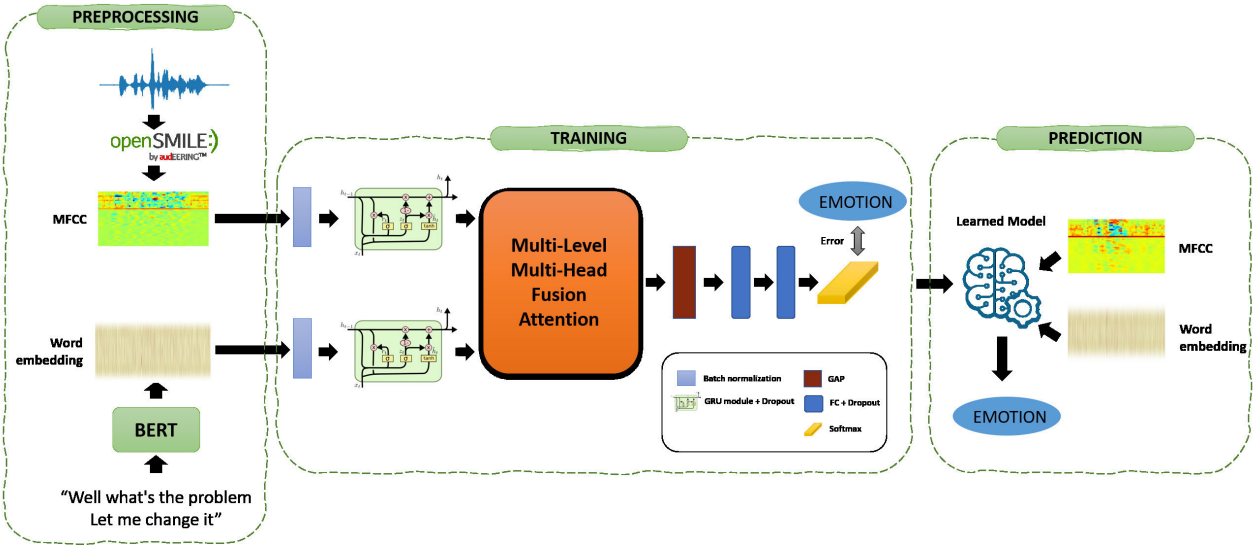

FIGURE 1. Overall architecture of the proposed MMFA-RNN model based SER.

approach in recent [3]. Moreover, it is an extremely gap to deal with SER in the wild. There may be more than one perceived emotion in the same utterance, and each emotion corresponds to a different portion of the spoken utterance. Consequently, emotion does not have a commonly agreed theoretical definition [4]. Meanwhile, capturing genuine emotion raises ethical issues, as well as difficulties in emotional labeling and control of recording situation. Furthermore, a complicated problem is the high dimensionality of the audio-textual feature space. Consequently, accurate modeling generally requires a reduction of the original input feature space. This reduction is commonly accomplished using feature selection, a method that identifies a subset of the initial features that provide enhanced classification accuracy [2]. However, it is not yet clear whether it is more advantageous to select a subset of emotionally relevant features or to capture the complex interactions across all features considered.

According to these facts, we proposed an effective, robust and reliable speech emotion recognition system using a mechanism called Multi-Level Multi-Head Fusion Attention (MMFA) based recurrent neural network (RNN) to combine the audio and textual features while capture their temporal-context information. The raw audio signals have been converted to a short-term power spectrum of sound, mel-frequency cepstrum coefficients (MFCC). Besides, we embed textual information to vector of works using pre-trained language model BERT [5], which is a deep bidirectional encoder representations from transformers achieving state-of-the-art on many language tasks. Then, we feed these representations of audio and text into two streams of self-attention based RNN to reclaim contextual information at each timestamp. Afterwards, the attention sequences of audio and text are combined using the adaptive multi-head attention mechanism [6], to learn the representatives. The whole process of extracting audio and textual features and fusing them is called ''Self-to-MultiHead'' attention. Next, global average pooling (GAP) is applied to minimize overfitting problem by reducing the temporal dimensions. Finally, its outputs are fed to two fully connected layers for predicting human emotional states. The overall process of the proposed model is illustrated in Fig. 1.

The main contributions of the proposed method are:

- Exploiting the temporal-contextual features of each single modality and effectively fusing two modalities with different temporal-length using proposed two-level attention mechanism, namely Multi-Level Multi-Head Fusion attention.
- Experimenting on classification of multi-class emotions on the three public datasets.
- Employing cross-validation to show the generalization of proposed model compared to conventional works and human performance.

The remainder of this paper is organized as follows. Section II presents the conventional works on speech emotion recognition. The proposed model is described in detail in Section III. Section IV gives experiments and analysis. A conclusion is drawn in Section V.

II. RELATED WORKS

Generally, feature extraction and emotion classification are two key steps in speech emotion recognition. In the following, we briefly review the classification strategies in literature, and then introduce the related works on IEMOCAP, MELD, and CMU-MOSEI datasets since they are more relevant to our work.

Recently, there are several classifiers have been utilized to distinguish the underlying emotion categories. One of early emotion classifiers is K-Nearest-Neighbor (KNN) [7]

or Artificial Neural Network (ANN) [8]. Other classifiers are GMM [9], HMM [10] and SVM [11], which are widely adopted for SER system. Furthermore, some advanced sparse representation-based classifiers [12], [13] have been published. Nevertheless, each classifier has its own advantages and disadvantages. To integrate the merits of different classifiers, ensembles of multiple classifiers have been investigated for speech emotion recognition [14], [15].

In 2011, Lee *et al.* [16] proposed a SER system using hierarchical binary decision tree approach, that maps an input speech utterance into one of the multiple emotion classes through subsequent layers of binary classifications. This method has made efforts to simplify multi-label to binary-label, however, there still occur loss at each stage of the hierarchical structure, that may lead to a large cumulative error for the whole system.

With the explosion of deep learning technique, Han *et al.* [17] utilized DNNs to extract high level features from raw data and shows that they are effective for speech emotion recognition. Nevertheless, this architecture is not enough to cover the long-time contextual effect in emotional speech. Consequently, the RNN-based emotion recognition framework from Microsoft Research has been proposed [18]. In this work, they consider the long-range context effect and the uncertainty of emotional label expressions. Neumann and Vu [19] proposed a model to predict the four-class emotion using an attentive convolutional neural network with multi-view learning objective function. Additionally, they compare the results on several representations like log Mel filter-banks (logMel), MFCC, extended Geneva minimalistic acoustic parameter set (eGeMAPS), and prosodic features.

Instead of using only audio features like previous studies, Jin *et al.* [20] generated feature representations from both acoustic and lexical levels for building an emotion recognition system. At the acoustic level, they first extract low-level features such as intensity, F0, jitter, shimmer, spectral contours, etc. At the lexical level, they use the traditional Bag-of-Words (BoW) feature and propose a new feature representation named emotion vector (eVector). They show fairly good performance on the four-class emotion recognition using late fusion technique. Eventually, the last review [21] in this section is about the use of deep recurrent neural network trained on a sequence of acoustic features calculated over small speech intervals. Furthermore, they use the Connectionist Temporal Classification (CTC) [22] approach classify emotional speakers by utilizing the additional NULL label which corresponds to the absence of any other label and extends the initial labels set. Conversely, they do not show a very high accuracy on emotion recognition.

Next, we describe the related works on MELD dataset. First, Zhang *et al.* [23] proposed a ConGCN architecture to model both context-sensitive and speaker-sensitive dependence for emotion detection. On the one hand, each utterance of the whole conversation corpus is represented as a node in a graph, with an edge between the two utterances in the same conversation to symbolize the contextual dependence. Each speaker of the whole corpus is represented as a node, and they bridge the specific-speaker dependence between each utterance and its speaker with an undirected edge. Jiao *et al.* [24] introduced an Attention Gated Hierarchical Memory Network (AGHMN) for real-time emotion recognition. Their work included a Hierarchical Memory Network (HMN) with a bidirectional GRU (BiGRU) as the utterance reader and a BiGRU fusion layer for the interaction between historical utterances For memory summarizing, they propose an Attention GRU (AGRU) to utilize the attention weights to update the internal state of GRU. Besides, Nadeem *et al.* [25] intended a novel confidence estimation method for predictions from a multi-class emotional classifier. The predicted confidence values by the proposed system are used to improve the accuracy of multi-modal emotion. The scores of different classes from the individual modalities are superposed on the basis of confidence values.

There are also several conventional works on CMU-MOSEI dataset. For instance, Sahay *et al.* [26] presented Relational Tensor Network architecture using the inter-modal interactions within a segment. They also generate rich representations of text and audio modalities by leveraging richer audio and linguistic context along with fusing fine-grained knowledge based polarity scores from text. Another work was introduced in [27] using a contextlevel inter-modal attention framework for simultaneously predicting the expressed emotions of an utterance. They hypothesize that the emotion of an utterance often has inter-dependence on other contextual utterances i.e. the knowledge of emotion for an utterance can assist in classifying its neighbor utterances. Particularly, they applied attention to contribute neighboring utterances and multimodal representations that may assist the network to learn in a better way.

While some previous works studied mainly on learning discriminant features or extracting temporal features from speech, others try to fuse the acoustic and linguistic features to archive the benefit of each single modality. Whereas, these features may not be discriminant enough to identify the subjective emotions, and the fusion models is still not powerful enough to capture complementary information. To tackle this issue, it may be feasible to employ temporal-contextual features in both audio and text information and combine them during training process for recognizing emotion state.

III. PROPOSED METHOD

In this section, we present our multimodal approach using RNN model with a mechanism of Multi-Level Multi-Head Fusion attention for SER. First, we introduce the preprocessing steps for audio signal and textual information. Next, we describe the process of attention models that can capture the contextual information. Finally, we report the proposed architecture for SER in details.

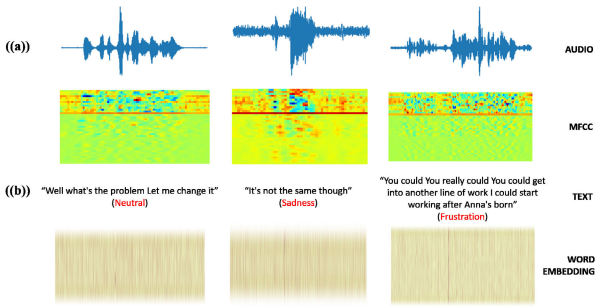

FIGURE 2. Example of audio conversion to MFCC and corresponding word embedding matrix.

A. PREPROCESSING

Unlike the works in [28], [29] that learned utterance features by utilizing convolutional neural networks (CNNs), we extract MFCC features for audio and BERT features for text. Through our exploration, we find that CNN feature is not a good representative for extracting temporal information such as raw audio and text. Generally, MFCC is an accurate representation of raw audio, which is widely used in automatic speech recognition. Meanwhile, BERT is the state-of-the-art model for extracting contextual features of text information even in different scenarios.

Especially, we use OpenSMILE toolkit [30] to extract MFCC features from audio signal. Unlike in speech signal processing, we set audio frames of 100 ms sampled at a rate of 50 ms using Hamming window to let the temporal length of the audio around ten times longer than the utterance length before re-sampling. This factor will reduce the influence of low signal-to-noise ratio (SNR) when estimating interpolation. For instance, if we choose small frame size and sampling rate (such as 20 ms of frame size and 5 ms of sampling rate), we must resample text features hundreds of times more than the original size. This leads to the generation of an various sequence of information and decreasing the SNR. In MFCC, the first feature is the log-energy of sound. Then, it computes 12 MFCC (1-12) from 26 Mel-frequency bands, and applies a cepstral liftering filter with a weight parameter of 22. The 13-delta and 13-acceleration coefficients are appended to the MFCC. These features are mean normalized with respect to the full input sequence. The frequency range of the Mel-spectrum is set from 0 to 8 kHz. Fig .2(a) shows examples of MFCC extraction from audio signals.

To utilize information from text data, we compute the vector representation of words using BERT. BERT is an open-sourced natural language processing (NLP) pre-trained model developed by researchers at Google in 2018. It is pre-trained on a large corpus of unlabelled text which includes the entire Wikipedia (about 2,500 million words) and a book corpus (800 million words). As opposed to directional models, which read the text input sequentially, BERT is considered bidirectional path. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word). In this paper, we use the baseline BERT model which includes 12 transformer blocks, 12 attention heads, and 110 million parameters. After embedding, each word in an utterance is represented by a 768-dimension vector. Fig. 2(b) presents example of work embedding using BERT corresponding to example in Fig 2(a).

B. MULTI-LEVEL MULTI-HEAD FUSION ATTENTION

Attention is motivated by how we pay visual attention to different regions of an image or correlate words in one sentence. Word attention allows us to focus on a contextual word with ''high attention'' while perceiving the surrounding words in ''low attention''. In literature, researchers experimented with Attention Mechanisms for machine translation tasks. Bahdanau *et al.* [31] proposed a neural machine translation based on jointly learning to execute translations concurrently. Then, attention mechanisms became common in NLP tasks based on neural networks such as RNN or convolutional neural network (CNN). In a nutshell, attention in the deep learning can be broadly interpreted as a vector of importance weights. In order to predict or infer one element, such as a pixel in an image or a word in a sentence, we estimate the attention vector how strongly it is correlated with other elements and take the sum of their values weighted by the attention vector as the approximation of the target.

In this section, we propose an attention mechanism that not only extracts contextual information of audio and text featutes but also combines these features from multimodality, as shown in Fig. 3. The proposed architecture is named Multi-Level Multi-Head Fusion Attention (MMFA), which included two levels of attention. The first-level compute a representation at different positions of a single sequence for each audio and text RNN-features. First of all, we perform a multiplicative operation $f_{att}\,(x_t, x_{t'})$ of current state, x_t at current timestamp t, over previous states $x_{t'}$ of previous timestamp t' to calculate the attention alignment. Next, we determine attention scores, a_t, by applying softmax function to $f_{att}\,(x_t, x_{t'})$. Then, we calculate the context vector, l_t, at position t as an average of the previous states weighted with the attention scores a_t. The first-level attention can be expressed as follows:

$$f_{att}\,(x_t, x_{t'}) = x_t^T W_a x_{t'} + b_a, \tag{1}$$

$$a_t = \text{softmax}(e_t), \tag{2}$$

$$l_t = \sum_{t'} a_{t,t'} x_{t'}. \tag{3}$$

where W_a and b_a are weight matrix and bias value to be learned in the attention model.

In the second-level attention, we modify the multi-head attention in [6] to fuse the attention features from audio and text. Rather than only computing the attention once, the multi-head mechanism runs through the scaled dot-product attention (SDPA) multiple times in parallel. However, the original SDPA [6] requires the same temporal lengths of inputs for computation. In fact, the temporal

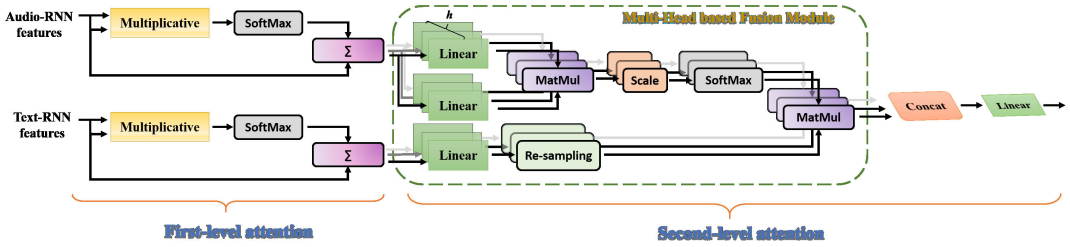

FIGURE 3. Multi-Level Multi-Head Fusion Attention (MMFA) consisting of two levels. **Multiplicative:** multiplicative operation; **Softmax:** softmax function; $\sum$: dot-product summation; **Concat:** feature concatenation; **Linear:** linear transformation; **MulMat:** dot-product operation; **Scale:** scaled operation; **Re-sampling:** interpolation technique.

dimension of audio and text are always different since the length of audio depends on the recording duration while the length of text is the number of words. To overcome this issue, we first fed the outputs of the first-level attention to multiple linear transformation modules, then we duplicate the audio branch and apply to them the sequential operations of dot-product, scaling, and softmax function. On the text branch, we add 'Re-sampling' block to interpolate the text features having the same temporal size with the audio features. Then, we perform another dot-product operation to extract the sympathetic representation between multiple models. The independent attention outputs are simply concatenated and linearly transformed into the expected dimensions. The second-level attention can be described as follows:

$$\begin{aligned}&\text{Multi-Head}\left(\mathbf{A_{1^{st}}}, \mathbf{A_{1^{st}}}, \mathbf{T_{1^{st}}}\right) = [\text{head}_1; \ldots; \text{head}_h]\,\mathbf{W}^{\text{O}}\\ &\text{head}_{\text{i}} = \text{softmax}\left(\frac{\left(\mathbf{A_{1^{st}}}\mathbf{W}_{\text{i}}^{1}\right)\left(\mathbf{A_{1^{st}}}\mathbf{W}_{\text{i}}^{2}\right)^{\text{T}}}{\sqrt{\text{d}_{\mathbf{A_{1^{st}}}}}}\right) \text{R}\left(\mathbf{T_{1^{st}}}\mathbf{W}_{\text{i}}^{3}\right)\\ &\text{where} \quad R(\bullet) = \text{Interpolation method}\end{aligned} \tag{4}$$

where $\mathbf{A_{1^{st}}}$ and $\mathbf{T_{1^{st}}}$ are the audio and text features from the first-level attention; $\mathbf{W}^{\text{O}}$, $\mathbf{W}^{1}$, $\mathbf{W}^{2}$, and $\mathbf{W}^{3}$ are parameter matrices to be learned.

C. MULTIMODAL SPEECH EMOTION RECOGNITION BASED ON MMFA AND RNN

In this section, we introduce a multimodal approach for SER using RNN and MMFA. After preprocessing step, we pass each input modality to a layer of batch normalization, as shown in Fig. 1, to normalize the input layer by adjusting and scaling the activation. Then, the outputs from batch normalization layers are fed to the gated recurrent unit (GRU) module [32], which is a variant of RNN model. The problem of the conventional RNN is suffering from vanishing or exploding gradients. Therefore, instead of having a simple neural network with four nodes as the RNN had previously, GRU has a cell containing multiple operations that allows it to carry forward information over many time periods in order to influence a future time period. Next, the returning sequences of GRU from both audio and text branches are applied to the MMFA module in the subsection III-B to extract the local features and combine the audio and text features into one matrix. In this work, we do not fix the temporal length of time-series data for all training samples as researchers usually did because the gap between the shortest and longest records is too large. Instead, we fix the temporal length on batch based on the maximum length of sample in that batch and pad zero-vector for the shorted ones. Therefore, this avoid to pad very long zero-vector to the most samples and reduce the size of input on each iteration. Since the temporal length of input is unknown, GAP layer is added to minimize overfitting problem by reducing the temporal dimensions. Finally, the reduced vector is passed to two fully connected (FC) layers for scaling and compressing feature-dimension to predict the probabilities of emotional states using 'Softmax' function. The detail of operational configuration is presented in Appendix A.

IV. EXPERIMENTS AND DISCUSSION

A. MATERIALS

IEMOCAP dataset [33] consists of approximately twelve hours of recordings. Audio, video and facial key points data was captured during the five sessions. Each session is a sequence of dialogues between man and woman. In total, ten people split into five pairs took part in the process. All involved people are professional actors and actresses from Drama Department of University of Southern California. The recording process took place at the professional cinema studio. Actors seated across each other at 'social' distance of meters. It enables more realistic communication. After recording of these conversations, authors divided them into utterances with speech. Note that audio was captured using two microphones. Therefore, the recordings contain two channels which correspond to male and female voices. Sometimes they interrupt each other. In these moments the utterances might intersect. This intersection takes about of all utterances time. It might lead to undesired results because microphones were place relatively near each other and thus inevitably captures both voices. The evaluation form contained ten options: *neutral*, *happiness*, *sadness*, *anger*, *surprise*, *fear*, *disgust*, *frustration*, *excitation*, and other. The total length is around 12 hours. To be comparable with related works, we divide two sets from the data: only *Improvised*, and *Mixed* (merging *Improvised* and *Scripted*) scenarios. In the *Improvised* scenario, subjects were asked to improvise based

TABLE 1. Number of utterances and duration per emotion class in the two scenarios of the IEMOCAP dataset: ***Improvised*** and ***Mixed***.

Emotion		Angry	Excited	Frustrated	Happy	Sad	Neurtal	Surprise	Fear	Disgust	Other
Improvised	# Utterance	289	663	971	284	608	1099	60	8	1	1
	Duration (min)	22.15	42.14	79.94	19.62	50.23	74.54	3.37	0.43	0.03	0.14
Mixed	# Utterance	1103	1041	1849	595	1084	1708	107	40	2	3
	Duration (min)	82.96	82.95	145.16	43.05	99.30	111.08	5.54	1.82	0.08	0.25

TABLE 2. Number of utterances per emotion in the MELD and CMU-MOSEI datasets.

Emotion	MELD			CMU-MOSEI		
	Training	Validation	Testing	Training	Validation	Testing
Neural	4513	450	1204	-	-	-
Joy	1661	157	380	-	-	-
Fear	264	36	50	257	28	55
Sadness	674	107	204	2669	279	502
Disgust	266	21	68	767	61	150
Anger	1065	148	328	1607	136	384
Surprise	1150	142	270	348	29	78
Happiness	-	-	-	7136	636	1438

on hypothetical scenarios while in the *Scripted* one, subjects were asked to memorize and rehearse with scripts. The *Mixed* scenario means we combine both improvised and scripted scenarios to cover a full range of situations. In summary, Table 1 presents the total number of utterances and duration of ten emotions in the two considerable scenarios.

MELD dataset [34] is an extended version of the EmotionLines dataset [35]. The data comes from the Friends TV series with multiple speakers involved in the conversations. It was split into training, validation, and testing sets with 9593, 1061, and 2504 utterances, respectively. Each utterance has been labeled led by one of the seven emotion types, namely, *anger*, *disgust*, *sadness*, *joy*, *neutral*, *surprise*, and *fear*.

CNU-MOSEI datasset [36] is the largest dataset of sentence level sentiment analysis and emotion recognition in online videos. CMU-MOSEI contains more than 65 hours of annotated video from more than 1000 speakers and 250 topics. The dataset was introduced in the 2018 Association for Computational Linguistics and used in the co-located First Grand Challenge and Workshop on Human Multimodal Language. Six emotions labeled in the MOSEI are *happy*, *sad*, *angry*, *fear*, *disgust*, and *surprise*. Similar to MELD dataset, the CMU-MOSEI was also divided into three sets: training, validation, and testing.

Table 2 summaries the total number of utterances of seven emotions in MELD dataset and of six emotions in CMU-MOSEI dataset. All three datasets are sampled at 16 kHz.

TABLE 3. Performance of the four emotional classification on the ***Improvised*** and ***Mixed*** presented by Mean ± STD in percentage. ***ACC***: accuracy; ACC_B: balanced accuracy; ***PRE***: precision; ***REC***: recall; F_β: ***F***-beta score.

Scenario	Modality	ACC	ACC_B	PRE	REC	F_β
Improvised	Audio	66.19±2.36	65.87±2.95	67.92±2.55	66.19±2.36	66.85±2.52
	Text	56.5±3.9	56.79±4.45	59.51±4.61	56.5±3.9	58.08±4.31
	Multimodal (Proposed)	**76±3.68**	**76.98±3.77**	**76.86±3.55**	**76±3.68**	**76.46±3.6**
Mixed	Audio	55.92±3.31	55.42±3.89	57.83±3.71	55.92±3.31	56.17±3.82
	Text	67.24±4.52	68.23±4.81	68.45±4.4	67.24±4.52	67.9±4.43
	Multimodal (Proposed)	**73.23±4.07**	**74.33±4.22**	**74.26±3.93**	**73.23±4.07**	**73.77±4**

B. PERFORMANCE ON SPEECH EMOTION RECOGNITION

Since the number of "Disgust" and "Other" samples in IEMOCAP dataset are too small, we drop them and decide two experiments corresponding to two classifiers: four emotions (neutral, angry, sad, and happy/excited), and eight emotions (neutral, frustrated, angry, sad, happy, surprise, excited, and fear). For each classifier, we conduct two scenarios which are *Improvised* and *Mixed*. To evaluate the IEMOCAP dataset, we configure 10-fold leave-one-speaker-out cross validation method, which assign a person in a session as validation while other person of the same session as test and vice verse, and the remaining sessions are used for training. Also, we assume that speaker identity information is not available in our study. Since the MELD and CMU-MOSEI datasets had been already divided in training, validation and testing sets, we learn the model's parameters using the training and validation sets, and evaluate the proposed model's performance using the testing set. Evaluation metrics used are *Accuracy*, *Balanced Accuracy*, *Precision*, *Recall*, and F_β, which are expressed in Appendix B. Especially for the CMU-MOSEI, we perform mean absolute error (*MAE*) metric for emotional score estimation to compare with conventional works. Besides, we implement experiments with single modality (using only audio or only text) to compare with our proposed multimodal approach.

Table 3 presents the performance of the four emotional classification on the *Improvised* and *Mixed* scenarios for the thee models: only audio, only text, and multimodal. From Table 3, we obtain a significant improvement of using multimodal by 9.81% of mean *Accuracy* compared

TABLE 4. Performance of the eight emotional classification on the *Improvised* and *Mixed* presented by Mean ± STD in percentage. *ACC*: accuracy; ACC_B: balanced accuracy; *PRE*: precision; *REC*: recall; F_β: *F*-beta score.

Scenario	Modality	ACC	ACC_B	PRE	REC	F_β
Improvised	Audio	49.88±4.04	40.42±5.15	54.36±2.48	49.88±4.04	50.46±3.31
	Text	41.56±3.72	30.86±3.33	43.32±3.78	41.56±3.72	41.26±3.5
	Multimodal (Proposed)	**60.71±3.76**	**53.41±5.64**	**62.13±3.32**	**60.72±3.77**	**61.09±3.38**
Mixed	Audio	39.91±5.01	27.81±3.87	38.59±5.26	39.91±5.01	37.4±5.81
	Text	49.88±4.18	37.1±3.68	49.87±3.87	49.88±4.18	48.3±3.9
	Multimodal (Proposed)	**56.7±4.1**	**48.71±6.35**	**57.93±3.66**	**56.7±4.1**	**56.82±3.91**

TABLE 5. Comparison of the proposed model and conventional studies on IEMOCAP dataset for the four emotional classification. Metric used is mean *Accuracy* in percentage. CV : cross-validation; LOSO: leave-one-speaker-out.

Model	CV Scheme	Modality	Mixed	Improvised
V. Chernykh et al. [21]	Grouped CV	Audio	54	-
K. Han et al. [17]	5-fold CV	Audio	54.3	-
C. Lee et al. [16]	LOSO	Audio	56.83	-
J. Lee et al. [18]	5-fold CV	Audio	-	62.85
M. Neumann et al. [19]	5-fold CV	Audio	56.1	62.11
Mirsamamdi et al. [37]	5-fold CV	Audio	63.5	-
Etienne et al. [38]	10-fold CV	Audio	-	64.5
Tzinis et al. [39]	5-fold CV	Audio	-	64.2
Huang et al. [40]	5-fold CV	Audio	-	59.4
Y. Zhang et al. [41]	5-fold CV	Audio	-	70.2
G. Ramet et al. [42]	5-fold CV	Audio	62.5	68.8
Q. Jin et al. [20]	LOSO	Audio	57.4	-
		Text	53.5	-
		Multimodal	69.2	-
S. Tripathi et al. [43]	Holdout	Audio	55.65	62.72
		Text	64.78	-
Proposed Model	LOSO	Audio	55.92	66.19
		Text	67.24	56.5
		Multimodal	**73.23**	**76.98**
Human Performance [33]			74.6	83.1

to audio model and 19.5% compared to text model for the *Improvised* case. Furthermore, the use of multiple model brings expressive enhancement on *Precision*, *Recall*, and F_β compared to the use of single model. Since people only express their emotion through voice intensity rather than the semantics of words in the *Improvised* case, the audio model obtain better performance than the text model. Besides, Table 3 shows that, for the *Mixed* scenario, our proposed model is healed by 5.99−17.31% on *Accuracy*, 6.1−18.88% on *Balanced Accuracy*, 5.81−16.43% on *Precision*, 5.99−17.31% on *Recall*, and 5.27−17.6% on F_β compared to using text or audio model only.

Moreover, the average performance of eight emotional classification for the *Improvised* and *Mixed* scenarios is introduced in Table 4. For the *Improvised* case, we achieve the *Accuracy* of 49.88±4.04% for audio model, 41.56±3.72% for text model, and 60.71±3.76% for multimodal, which is improved from 10.83% to 19.15%. For the *Mixed* case, the *Accuracy* of audio, text and multimodal are 39.91±5.01%, 49.88±4.18%, and 56.7±4.1%, respectively. However, the *Balanced Accuracy* scores of eight emotional classification are low because the dataset is imbalanced on "Surprise" and "Fear" classes. This challenge can encourage researchers preferring topics on the advanced emotion recognition to the basic one.

The evaluation of our proposed model compared to the conventional studies is presented in Table 5. Most of studies focus on emotion recognition using audio signal. The results prove that we can heal a big gap of accuracy using multiple models as in [20] and proposed model. Even the evaluation settings are different from most studies who used 5-fold cross validation, our results are still presented objectively since we conduct experiment with speaker-independence.

TABLE 6. Performance of the emotional classification on the MELD and CMU-MOSEI presented by Mean value in percentage. *ACC*: accuracy; ACC_B: balanced accuracy; *PRE*: precision; *REC*: recall; F_β: *F*-beta score.

Dataset	Modality	ACC	ACC_B	PRE	REC	F_β
MELD	Audio	48.84	43.25	43.77	48.84	42.93
	Text	61.66	57.18	58.06	61.66	58.05
	Multimodal (Proposed)	**63.26**	**59.94**	**60.11**	**63.26**	**59.66**
CMU-MOSEI	Audio	89.91	70.28	89.69	89.91	89.27
	Text	93.86	93.87	93.88	93.86	93.74
	Multimodal (Proposed)	**99.19**	**98.23**	**99.2**	**99.19**	**99.2**

TABLE 7. Comparison of the proposed model and conventional studies on MELD and CMU-MOSEI datasets. *ACC*: accuracy; F_1: *F*-1 score; *MAE*: mean absolute error.

Dataset	Model	Modality	ACC	F_1	MAE
MELD	W. Jiao et al. [24]	Text	60.3	58.1	-
	P. Zhong et al. [44]	Text	-	58.18	-
	I. Nadeem et al. [25]	Multimodal	61.15	59.47	-
	D. Zhang et al. [23]	Audio	-	57.4	-
		Text	-	42.2	-
		Multimodal	-	59.4	-
	Proposed Model	Audio	48.84	45.34	-
		Text	61.66	58.98	-
		Multimodal	**63.26**	**60.59**	-
CMU-MOSEI	S. Sahay et al. [26]	Multimodal	-	-	0.1551
	M.S. Akhtar et al. [27]	Audio	56.2	74.6	-
		Text	60.2	76.9	-
		Multimodal	60.5	77.6	-
	J. Williams et al. [45]	Audio	-	-	0.146
		Text	-	-	0.156
		Multimodal	-	-	0.14
	S. Sangwan et al. [46]	Audio	57.75	76.26	-
		Text	61.19	77.88	-
		Multimodal	61.98	78.21	-
	C.W. Lee et al. [47]	Multimodal	88.89	-	-
	Proposed Model	Audio	89.91	89.2	0.1622
		Text	93.86	93.67	0.1647
		Multimodal	**99.19**	**99.19**	**0.1321**

TABLE 8. Per-class performance of the four emotional classification on the *Improvised* and *Mixed* of IEMOCAP dataset presented by Mean ± STD in percentage. ***PRE***: precision; ***REC***: recall; F_1: *F*-1 score.

Scenario	Emotion	Audio				Text				**Multimodal (Proposed)**			
		PRE	*REC*	F_1	Support	*PRE*	*REC*	F_1	Support	*PRE*	*REC*	F_1	Support
Improvised	Neutral	60.06±6.56	68.91±8.99	63.90±6.41	172	53.74±8.68	52.10±9.83	52.11±5.87	171	69.32±7.08	70.22±9.36	**69.57±7.18**	171
	Angry	77.45±9.02	62.82±9.56	68.42±6.05	110	60.82±16.96	60.56±8.37	58.87±7.75	110	81.43±9.99	77.14±6.41	**78.87±6.34**	110
	Sad	63.45±9.59	74.13±10.42	67.99±8.52	108	51.52±11.09	55.68±11.49	52.76±8.7	108	73.95±7.21	83.86±4.33	**78.49±5.5**	108
	Happy/ Excited	70.30±8	57.62±9.12	62.68±5.22	164	65.12±10.27	58.81±8.66	60.87±4.55	164	80.13±8.7	76.35±3.73	**77.87±4.16**	164
	Weighted Average	67.92±2.55	66.19±2.36	66.02±2.56	553	59.51±4.61	56.50±3.9	56.76±4.08	553	76.86±3.55	76.00±3.68	**76.08±3.67**	553
Mixed	Neutral	49.69±8.20	60.64±11.81	54.28±9.04	171	62.20±8.25	61.70±7.75	61.60±6.21	171	65.91±7.29	66.69±10.01	**66.06±7.56**	171
	Angry	67.62±8.64	51.48±11.31	56.59±11.42	110	71.34±10.42	71.47±8.3	70.86±6.7	110	79.55±10.79	75.24±6.9	**76.92±6.97**	110
	Sad	54.63±10.32	65.39±12.79	59.04±10.03	108	61.90±10.68	68.46±6.08	64.53±7.12	108	71.10±8.14	82.06±5.08	**76.02±6.21**	108
	Happy/ Excited	58.77±8.89	44.18±10.32	49.39±8.3	164	72.22±7.59	71.30±5.12	71.48±4.51	164	77.89±9.15	73.34±4.33	**75.20±4.62**	164
	Weighted Average	57.83±3.71	55.92±3.31	55.24±3.83	553	68.45±4.4	67.24±4.52	67.38±4.49	553	74.25±3.93	73.23±4.07	**73.32±4.07**	553

TABLE 9. Per-class performance of the eight emotional classification on the *Improvised* and *Mixed* of IEMOCAP dataset presented by Mean ± STD in percentage. ***PRE***: precision; ***REC***: recall; F_1: *F*-1 score.

Scenario	Emotion	Audio				Text				**Multimodal (Proposed)**			
		PRE	*REC*	F_1	Support	*PRE*	*REC*	F_1	Support	*PRE*	*REC*	F_1	Support
Improvised	Neutral	46.05±9.03	60.53±7.61	51.83±6.83	171	41.53±8.52	40.91±6.88	40.67±6.41	171	55.63±9.25	60.34±11.91	**57.17±8.19**	171
	Frustrated	49.75±9.29	48.38±2.93	48.58±4.99	185	42.65±8.39	48.04±5.42	44.54±5.06	185	60.58±8.17	59.79±4.36	**59.94±5.21**	185
	Angry	60.80±8.34	42.30±10.54	49.07±8.12	110	48.03±12.67	43.54±13.02	43.23±6.03	110	72.15±8.66	61.41±9.37	**65.62±5.55**	110
	Sad	49.57±10.26	75.89±6.16	59.15±8.14	108	45.56±12.42	51.60±9.01	46.99±6.67	108	63.85±11.23	79.49±4.95	**70.07±6.08**	108
	Happy	70.55±22.43	21.10±9.1	30.04±7.1	60	29.36±33.82	5.00±5.68	7.94±8.9	60	54.06±8.81	41.52±11.36	**45.66±5.97**	60
	Surprise	65.00±47.43	16.48±15.22	25.62±22	11	15.00±33.75	4.58±10.84	6.22±13.77	11	50.20±25.17	31.09±22.59	**34.68±18.15**	11
	Excited	50.25±14	35.15±7.8	40.07±6.95	104	41.92±11.27	50.20±6.24	44.46±7.21	104	59.28±10.52	63.27±9.31	**60.14±6.41**	104
	Fear	60.00±51.64	19.18±21.83	27.62±28.07	4	0±0	0±0	0±0	4	50.00±47.14	24.94±28.91	**31.45±33.26**	4
	Weighted Average	54.36±2.48	49.88±4.04	48.74±4.33	753	43.32±3.78	41.56±3.72	40.57±3.53	753	62.13±3.32	60.72±3.77	**60.35±3.73**	753
Mixed	Neutral	38.07±8.4	52.36±9.26	43.62±7.28	171	47.34±9.49	50.13±9.57	47.94±7.45	171	51.66±9.26	56.63±12.75	**53.26±8.78**	171
	Frustrated	40.039.44	38.71%	38.89%	185	49.81±8.05	55.83±3.47	52.22±4.48	185	56.17±8.14	55.80±4.74	**55.71±5.35**	185
	Angry	48.86%	32.16±4.11	37.81±6.11	110	56.07±10.64	55.12±12.91	53.92±6.31	110	67.88±9.85	56.04±9.76	**60.53±6.09**	110
	Sad	43.51±6.52	71.25±11.97	53.19±10.34	108	53.09±12.21	59.82±8.27	55.38±7.87	108	61.30±11.38	77.64±5.39	**67.66±6.03**	108
	Happy	11.52±12.41	3.58±3.88	5.23±5.62	60	33.86±33.13	6.98±7.08	10.73±10.49	60	47.21±10.44	33.80±11.98	**38.00±7.94**	60
	Surprise	0±0	0±0	0±0	11	15.00±33.75	4.58±10.84	6.22±13.77	11	45.39±26.2	25.52±21.92	**28.33±17.86**	11
	Excited	32.18±13.29	21.36±10.65	24.37±10.48	104	47.93±12.74	60.50±5.43	52.31±8.69	104	56.27±10.61	60.44±30.19	**57.18±6.38**	104
	Fear	0±0	0±0	0±0	4	0±0	0±0	0±0	4	27.00±39.31	18.83±30.19	**21.58±32.86**	4
	Weighted Average	38.59±5.26	39.91±5.01	37.28±5.91	753	49.87±3.87	49.88±4.18	48.19±4.07	753	57.93±3.66	56.7±4.1	**56.12±4.16**	753

The performance on single models and proposed multi-model for the MELD and CMU-MOSEI datasets is displayed in Table 6. As discussed above, the proposed multi-model outperforms both single models (audio and text) in term of emotion recognition. Particularly, we obtain an accuracy of 63.26% using multimodal on the MELD data while they are 48.84% and 61.66% for audio and text, respectively. For the CMU-MOSEI data, we achieve an accuracy of 99.19%, which improves a bit gap compared to audio by 9.28% and text by 5.33%.

Table 7 presents comparison of the proposed model and conventional works on the MELD and CMU-MOSEI datasets in term of *ACC*, F_1, and *MAE* (for CMU-MOSEI data only). For both scenarios, we achieve great effectiveness of the proposed approach. Precisely, we obtain 63.226% and 60.59% in term of accuracy and *F*-1 score, respectively, on the MELD dataset. Also, we achieve almost perfect performance on the test set of CMU-MOSEI data, which are 99.19% of accuracy, 99.19% of *F*-1 score, and 0.1312 of *MAE*.

Further results can be found in the Appendix C.

C. DISCUSSION

Our proposed "Self-to-Multihead" attention based RNN serves many advantages over the other published studies introduced in Section II. The main contributions of our paper are 1) the uses of "Self-to-Multihead" to exploit important information from each modality and fuse multiple temporal-feature from audio and text, and 2) the combination of multiple modalities (audio and text) to enhance the overall performance compared to unimodal. In details, we first preprocess raw audio signal and text information to produce good representatives. MFCC is a very compressible representation for audio signal since it is more "biologically inspired" and have been proven to be more successful in automatic speech recognition or speech segregation. Meanwhile, BERT is state-of-the-art in language modeling which embed word into vector depended on the context in which it occurs. It means BERT can construct the same word in different representation and provide meaningful information. Table 5 proves that the BERT-based text model results higher accuracy than in [20] and [43].

TABLE 10. Per-class performance of the emotional classification on the MELD and CMU-MOSEI datasets presented by Mean value in percentage. *PRE*: precision; *REC*: recall; F_1: *F*-1 score.

Dataset	Emotion	Audio				Text				Multimodal (Proposed)			
		PRE	*REC*	F_1	Support	*PRE*	*REC*	F_1	Support	*PRE*	*REC*	F_1	Support
MELD	Neural	49.98	97.18	66.01	1204	71.70	84.39	77.53	1204	74.46	83.31	**78.64**	1204
	Surprise	29.78	5.93	9.73	270	54.90	58.15	56.47	270	49.32	67.41	**56.96**	270
	Fear	0.00	0.00	0.00	50	14.29	2.00	3.51	50	50.00	20.00	**38.46**	50
	Sadness	0.00	0.00	0.00	204	35.85	18.63	24.52	204	40.37	21.57	**28.12**	204
	Joy	34.44	4.21	7.69	380	49.76	55.26	52.37	380	53.72	60.79	**57.04**	380
	Disgust	0.00	0.00	0.00	68	28.57	5.88	9.76	68	30.00	14.00	**25.00**	68
	Anger	40.88	6.40	10.61	328	46.83	35.98	40.69	328	50.00	37.50	**42.86**	328
	Weighted Average	43.77	48.84	42.34	2504	58.06	61.66	58.98	2504	60.11	63.26	**60.59**	2504
CMU-MOSEI	Happy	93.79	98.68	96.17	1438	95.92	99.72	97.78	1438	99.72	99.72	**99.72**	1438
	Sad	83.42	94.22	88.49	502	89.60	94.42	91.95	502	98.04	99.60	**98.81**	502
	Angry	87.86	79.17	83.29	384	96.90	81.51	88.54	384	99.47	98.44	**98.95**	384
	Fear	77.27	30.91	44.16	55	86.11	56.36	68.13	55	98.18	98.18	**98.18**	55
	Disgust	88.89	53.33	66.67	150	84.21	85.33	84.77	150	99.31	96.00	**97.63**	150
	Surprise	73.91	65.38	69.39	78	93.06	85.90	89.33	78	96.20	97.44	**96.82**	78
	Weighted Average	89.69	89.91	89.20	2607	93.88	93.86	93.67	2607	99.20	99.19	**99.19**	2607

Recognizing human emotion from speech needs characteristic audio and textual features before feeding data into a suited deep learning algorithm. The synchronization of audio and text in temporal space is always a controversial problem. Traditional studies usually resize the temporal dimension to a fix length such as zero-padding technique. However, this can produce very largely non-informative samples if there is a big gap of temporal length between trials. Therefore, in this study, we proposed a dynamic-size inputs that only resize the observations' temporal lengths in a batch rather than for all samples. Besides, we apply self-attention on each timestep to take advantage of its importance in the whole time-series data and avoid non-informative features being learned during training. The modified multi-head attention allows us to join two different-length inputs flexibly while extracting the discriminant features of both. That is the reason why the multimodal system can predict hard-class emotions as shown in Table 9 and Table 10. Generally, the combination of multiple models achieves closing prediction to human performance, as shown in Table 5.

Nevertheless, Out proposed model still has a few limitations. The audio-based emotion recognition model from this study is worse than conventional models. The proposed model has recently focused on individual dialog, while the dataset provides several conservation. It is necessary to figure out the relationship between dialogs in a conservation.

V. CONCLUSION

Accurate emotion recognition systems are essential for the advancement of human behavioral informatics and in the design of effective human–machine interaction systems. Such systems can help promote the efficient and robust processing of human behavioral data as well as in the facilitation of natural communication. In this paper, we proposed a multimodal-based speech emotion recognition using recurrent neural network and Self-to-MultiHead attention mechanism. The proposed framework is developed based on two types of speech representations, which is the MFCC of audio signal and word embedding from text data. By training these features on temporal space parallelly, we obtain the state-of-the-art performance on the IEMOCAP, MELD, and CMU-MOSEI datasets.

However, there are still many future modifications integrated within this framework. Instead of fusing the two modalities at late layers, a synchronization between audio signal and text data can be used at low-level representations to determine the relationship between these data. Additionally, we will consider using other features of audio such as perceptual linear prediction (PLP), chroma, prosody, etc. For text data, we will select the emotional words in an utterance rather than using all to eliminate non-related emotional information in the speech. Finally, another trend is to apply domain adaptation techniques and transfer the knowledge from the speech recognition methods to the emotion detection using pretraining and fine-tuning.

APPENDIXES

APPENDIX A

CONFIGURATION OF MULTI-LEVEL MULTI-HEAD FUSION ATTENTION BASED RNN ARCHITECTURE

The architecture of the MMFA based RNN includes two branches as shown in Fig. 1. The first branch inputs the MFCC features, which is converted from audio signals, of $t_a \times 39$ dimensions, where t_a is timesteps and 39 is number of features extracted from each timestep by the OpenSMILE. The second branch inputs the word embedding matrix determined from the pre-trained BERT model that provides a $t_w \times 768$ dimensions, where t_w is the number of word for

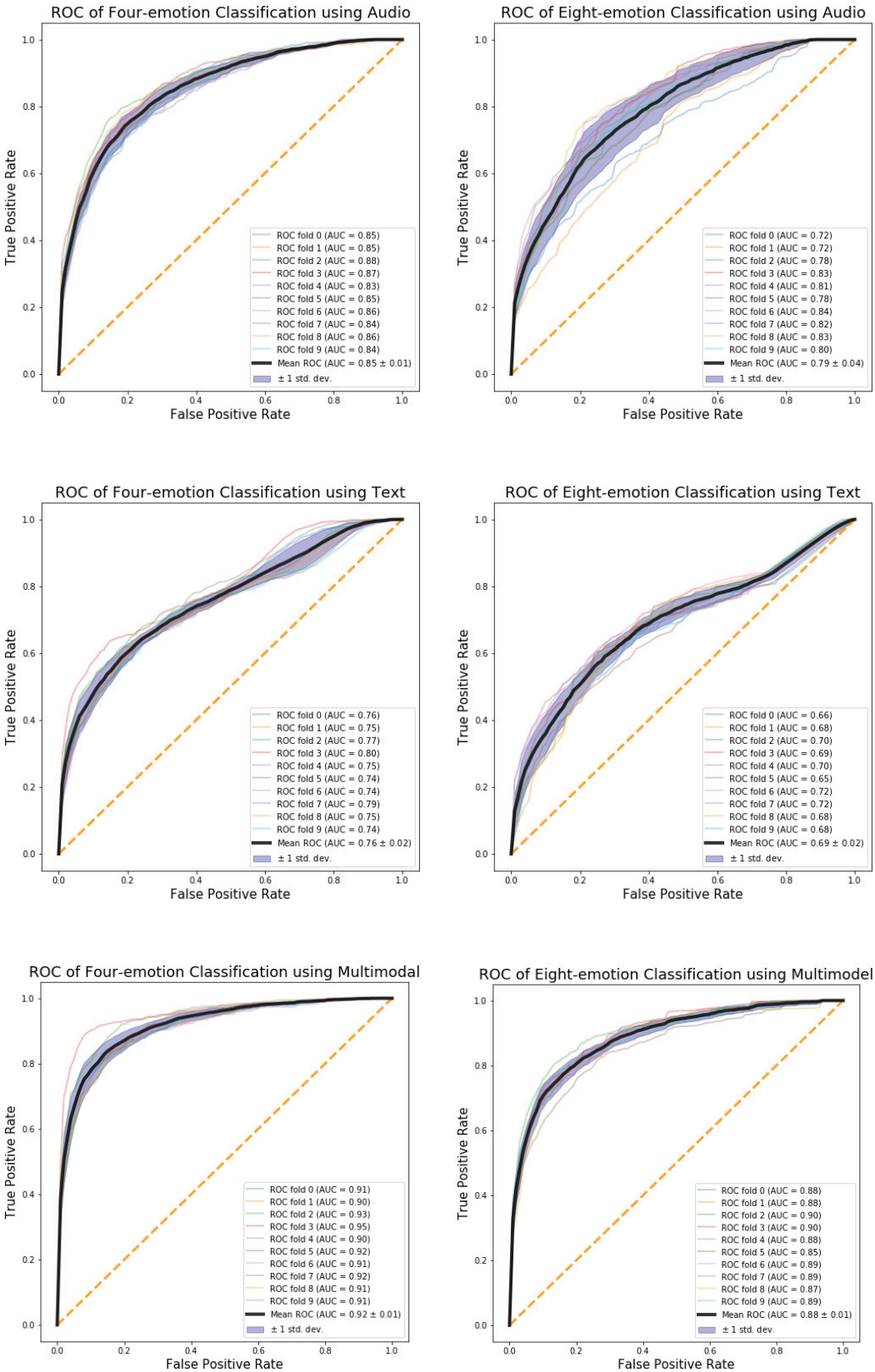

FIGURE 4. ROC-AUC of the four-emotion and eight-emotion classifiers for the *Improvised* scenario.

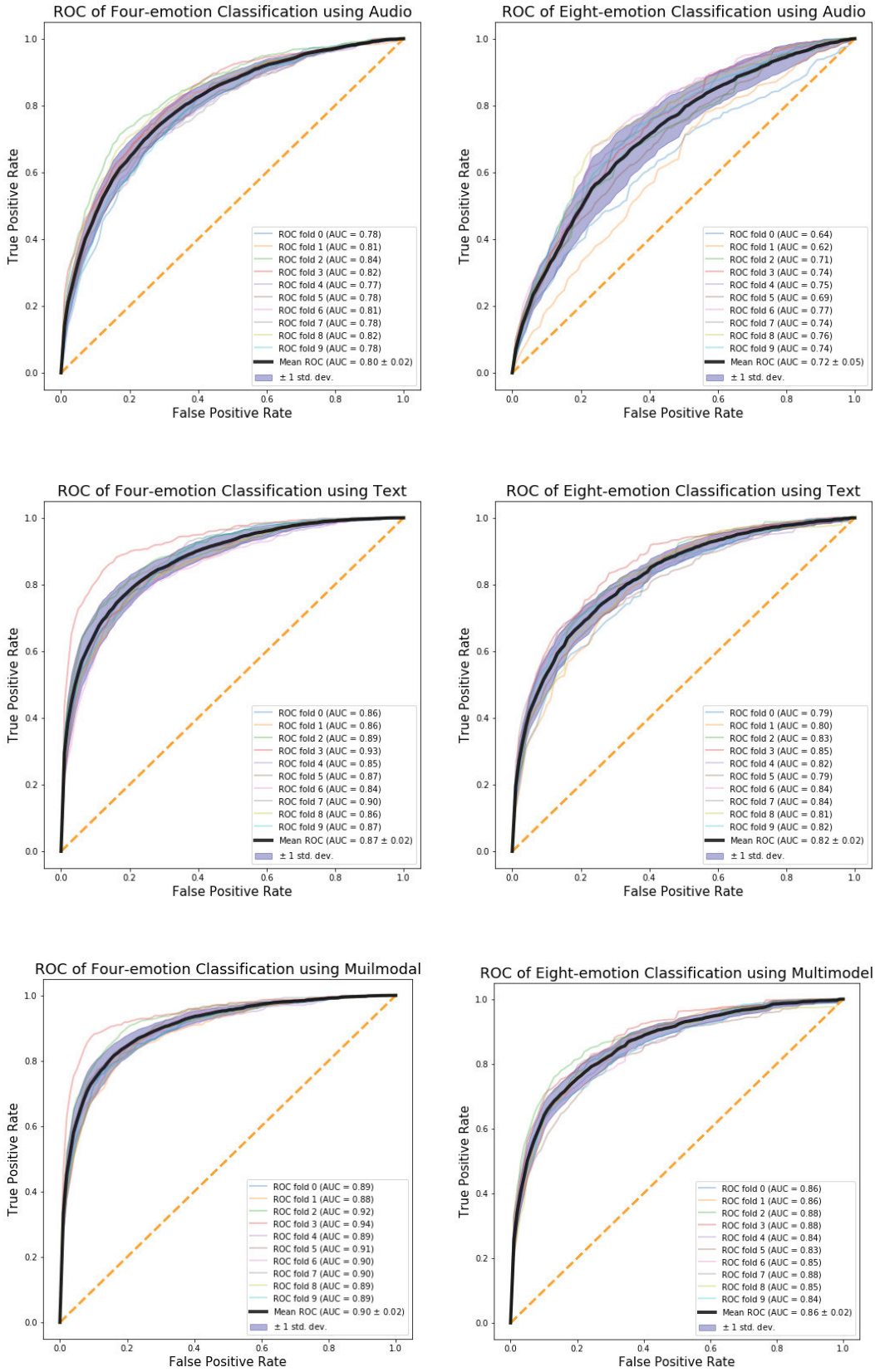

FIGURE 5. **ROC-AUC of the four-emotion and eight-emotion classifiers for the *Mixed* scenario.**

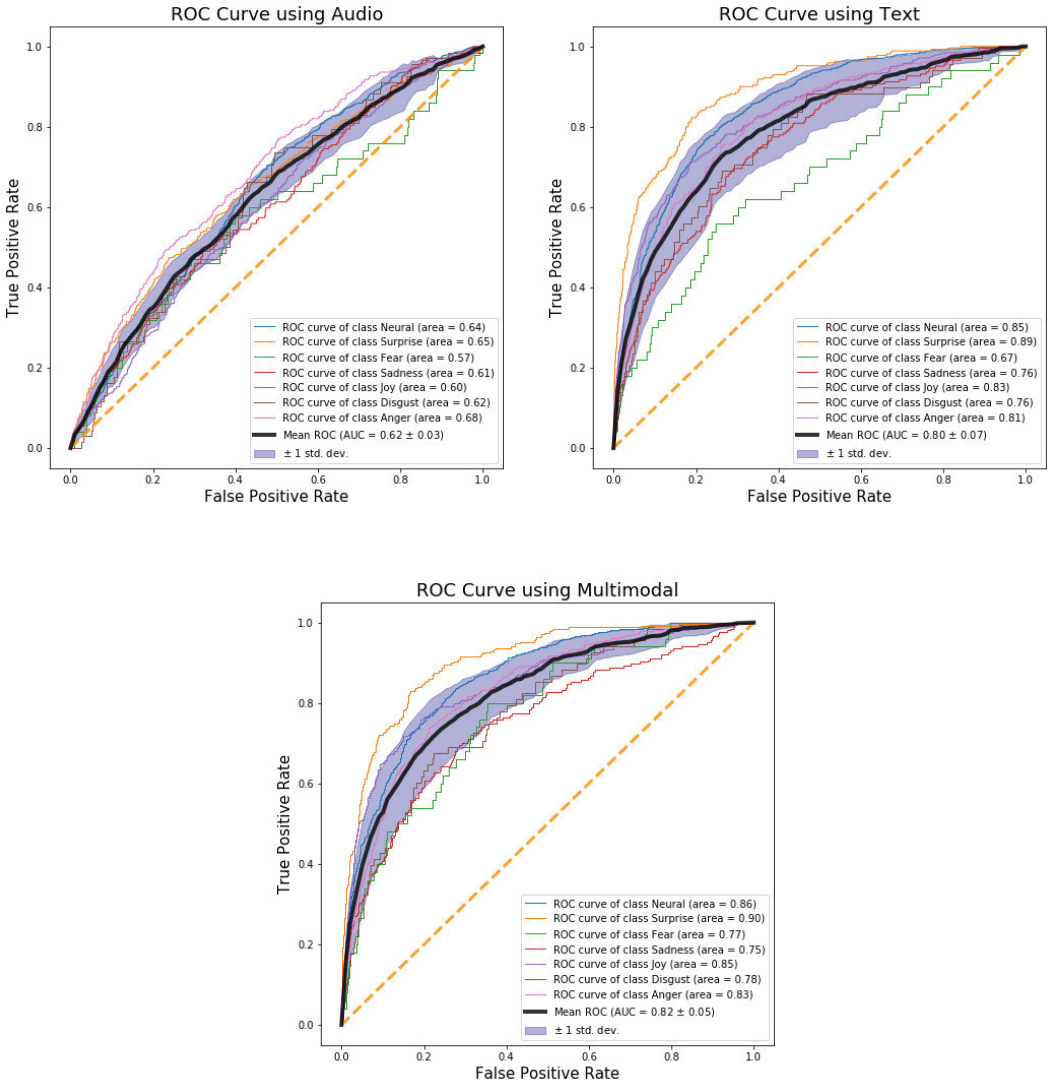

FIGURE 6. ROC-AUC of the seven-emotion classifier for the MELD dataset.

each utterance and 768 is the output size of the BERT model. Each input is connected to a batch normalization layer to prevent bias from different modalities before it is fed to one GRU layer. In the GRU, the number of recurrent units, n_{GRU}, is adapted as follows:

$$\text{Audio} : n_{GRU} = f_a * 39 \tag{5}$$

$$\text{Text} : n_{GRU} = 768/f_w \tag{6}$$

where f_a and f_w are multiplicative factors for audio and text models. In this work, we set $f_a = 2$ and $f_w = 4$ because they are large enough to avoid overfitting. We use $L2$ as regularization function with parameter of $1e^{-4}$. Recurrent dropout is set to 0.3. For the MMFA, the number of heads is set to 2. Moreover, we add $L2$ function as regularizer of $1e^{-3}$ for the Multi-Head module.

Then, the GAP is applied to scale the temporal space into feature vector, and its output is linked to two fully connected (FC) layers. Each FC layer contains 80 perceptive units followed by a batch normalization layer and a dropout layer of 0.3. The activation is fast Gaussian error linear units (fastGELU), which is utilized from [48], and can be expressed as follows:

$$\text{fastGELU}\,(x) = \max\left(0, \min\left(1, \frac{1.702 * x + 1}{2}\right)\right) * x \tag{7}$$

Finally, we use 'Softmax' function to predict the emotional probabilities, and decide the closing emotion that gets the highest score. Loss function and optimization strategy used in this study is cross entropy and stochastic gradient descent (SGD), respectively. Learning rate is set to $1e^{-3}$ with decay of $1e^{-6}$.

APPENDIX B
EXPERIMENTAL ENVIRONMENT AND EVALUATION CRITERIA

The environments of the experiments is the Linux operating system (Ubuntu 16.04 LTS), along with the

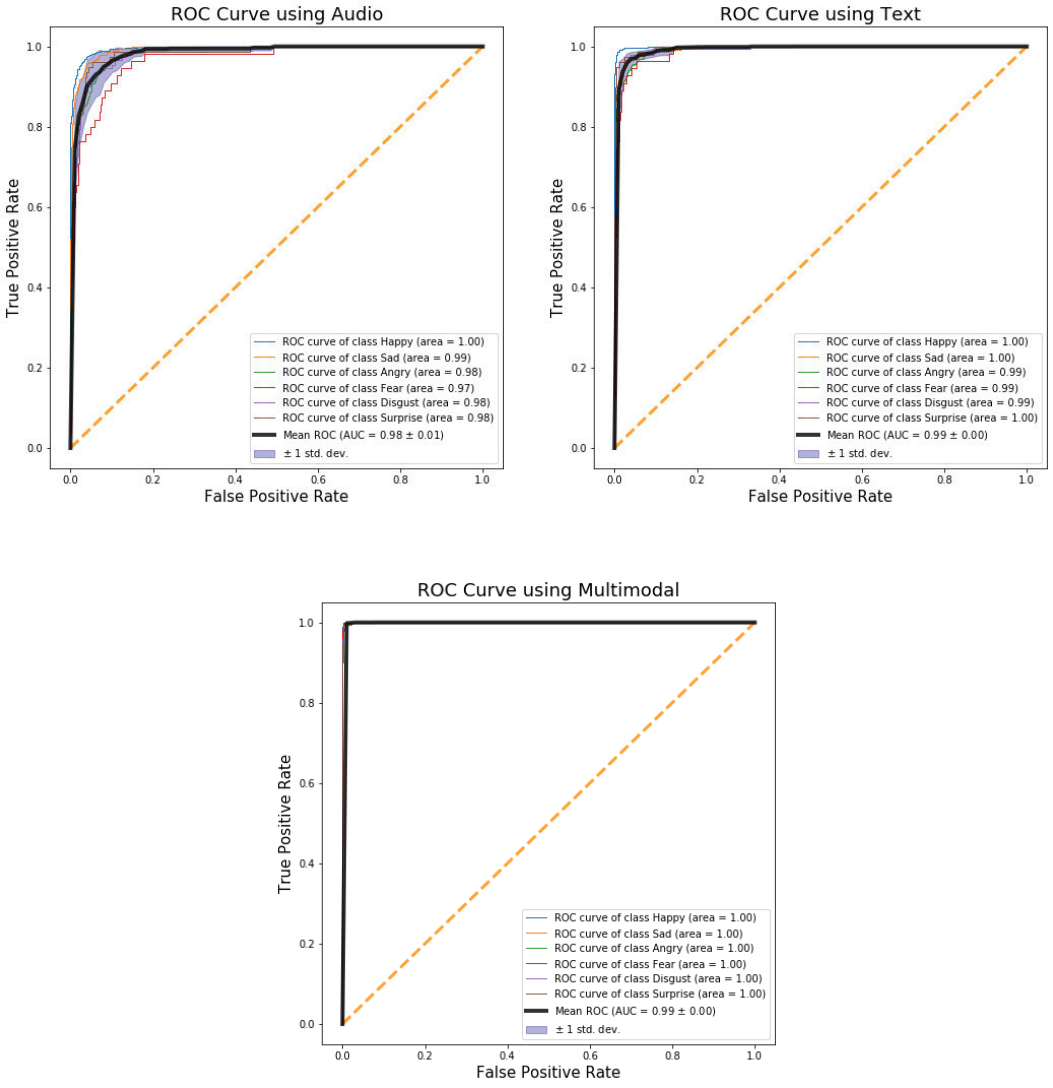

FIGURE 7. ROC-AUC of the six-emotion classifier for the CMU-MOSEI dataset.

Tensorflow[1] framework version 2.0, and CUDA 10.0 and cuDNN 7.6 dependencies (Nvidia Corporation, Santa Clara, CA, USA) for graphics processing unit acceleration. The computing systems have a GPU of 11GB GTX 1080 Ti and two GPUs of 16GB Tesla P100 with 64GB RAM.

We evaluate the proposed model for SER using 10-fold leave-one-speaker-out cross validation, which assign a person in a session as validation while other person of the same session as test and vice verse, and the remaining sessions are used for training. Also, we assume that speaker identity information is not available in our study. The quantitative measurement of speech emotion recognition is per-sample *Accuracy*. Additionally, other terms such as *Balanced Accuracy*, *Precision*, *Recall*, F_1 and F_β scores are also calculated for comparison. *Balanced Accuracy* is used to deal with imbalanced dataset and defined as the average of recall obtained on each class. *Precision* basically tells us how many positive samples classified by model are actually positive. *Recall* provides how many true positives are found by model. F_1 score takes into account both *Precision* and Recall as we can't always evaluate them and then take the higher one for our model. It is the harmonic mean of *Precision* and *Recall*. F_β score is used as a evaluation metric to assign different weights to *Precision* and *Recall*. Using the notation of true positive (TP), true negative (TN), false positive (TP), and false negative (FN), these metrics are presented in eqs. (8) to (13) as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$Accuracy_{Balanced} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

[1]https://www.tensorflow.org/

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{12}$$

$$F_\beta = \left(1+\beta^2\right)\frac{Precision \cdot Recall}{\left(\beta^2 \cdot Precision\right)+Recall} \tag{13}$$

In this study, $\beta = 0.5$ was chosen to lend more weight to precision. The cross-entropy loss for emotion classification is formed as eq. 14:

$$\mathscr{L}_{CE} = -\sum_{i=1}^{C} y_i \log \hat{y}_i \tag{14}$$

All metrics are computed using Scikit-learn[2] toolbox.

APPENDIX C
FURTHER RESULTS

This section presents further experimental results on IEMOCAP, MELD, and CMU-MOSEI datasets using our proposed model. Tables 8 and 9 provide per-class performance of the four and eight emotional classification on the *Improvised* and *Mixed* scenarios of the IEMOCAP data. For four-emotion classifier, Table 8 shows the highest scores on ''Angry'' while obtaining the lowest score on ''Neutral''. For eight-emotion classifier, we get high performance on ''Sad'' and ''Angry'' and low performance on ''Surprise'' and ''Fear'', as shown in Table9.

Table 10 details the per-class performance of the speech emotion recognition for both MELD and CMU-MOSEI datasets. Similar as discussed above, multimodal approach shows outperforming performance compared to using audio or text model only, and is able to predict hard cases such as *Fear*, *Sad*, and *Disgust* in the MELD dataset.

Fig. 4 displays the mean and standard deviation of the area under the receiver operating characteristic curve, usually called AUC, in *Improvised* cases for both four-emotion and eight-emotion classifiers over 10 folds. Similarly, Fig. 5 illustrates the AUC result for the *Mixed* case.

Fig. 6 illustrates the mean and standard deviation of per-class ROC-AUC of the seven-emotion classifier on the MELD dataset. Furthermore, Fig. 7 presents the per-class ROC-AUC on the CMU-MOSEI dataset.

All above results prove that the use of attention mechanism and combination of multiple models help to gain a significant success compared to using unimodal. Especially, multimodal system can predict hard-class emotion, which accounts for small quantities in the imbalanced dataset, while unimodal system is almost impossible to recognize.

REFERENCES

[1] M. Sreeshakthy and J. Preethi, ''Classification of human emotion from Deap EEG signal using hybrid improved neural networks with cuckoo search,'' *Broad Res. Artif. Intell. Neurosci.*, vol. 6, nos. 3–4, pp. 60–73, 2016.

[2] M. El Ayadi, M. S. Kamel, and F. Karray, ''Survey on speech emotion recognition: Features, classification schemes, and databases,'' *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[3] H. Gunes and B. Schuller, ''Categorical and dimensional affect analysis in continuous input: Current trends and future directions,'' *Image Vis. Comput.*, vol. 31, no. 2, pp. 120–136, Feb. 2013.

[4] P. R. Kleinginna and A. M. Kleinginna, ''A categorized list of emotion definitions, with suggestions for a consensual definition,'' *Motivat. Emotion*, vol. 5, no. 4, pp. 345–379, Dec. 1981.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[7] F. Dellaert, T. Polzin, and A. Waibel, ''Recognizing emotion in speech,'' in *Proc. 4th Int. Conf. Spoken Lang. Process. (ICSLP)*, 1996, pp. 1970–1973.

[8] J. Nicholson, K. Takahashi, and R. Nakatsu, ''Emotion recognition in speech using neural networks,'' *Neural Comput. Appl.*, vol. 9, no. 4, pp. 290–296, Dec. 2000.

[9] D. Neiberg, K. Elenius, and K. Laskowski, ''Emotion recognition in spontaneous speech using GMMs,'' in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2006, pp. 809–812.

[10] A. Nogueiras, A. Moreno, A. Bonafonte, and J. Mariño, ''Speech emotion recognition using hidden Markov models,'' in *Proc. 7th Eur. Conf. Speech Commun. Technol.*, 2001, pp. 746–749.

[11] A. Milton, S. Sharmy Roy, and S. Tamil Selvi, ''SVM scheme for speech emotion recognition using MFCC feature,'' *Int. J. Comput. Appl.*, vol. 69, no. 9, pp. 34–39, 2013.

[12] X. Zhao, S. Zhang, and B. Lei, ''Robust emotion recognition in noisy speech via sparse representation,'' *Neural Comput. Appl.*, vol. 24, nos. 7–8, pp. 1539–1553, Jun. 2014.

[13] X. Zhao and S. Zhang, ''Spoken emotion recognition via locality-constrained kernel sparse representation,'' *Neural Comput. Appl.*, vol. 26, no. 3, pp. 735–744, Apr. 2015.

[14] D. Morrison, R. Wang, and L. C. De Silva, ''Ensemble methods for spoken emotion recognition in call-centres,'' *Speech Commun.*, vol. 49, no. 2, pp. 98–112, Feb. 2007.

[15] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, ''Spoken emotion recognition using hierarchical classifiers,'' *Comput. Speech Lang.*, vol. 25, no. 3, pp. 556–570, Jul. 2011.

[16] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, ''Emotion recognition using a hierarchical binary decision tree approach,'' *Speech Commun.*, vol. 53, nos. 9–10, pp. 1162–1171, Nov. 2011.

[17] K. Han, D. Yu, and I. Tashev, ''Speech emotion recognition using deep neural network and extreme learning machine,'' in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 223–227.

[18] J. Lee and I. Tashev, ''High-level feature representation using recurrent neural network for speech emotion recognition,'' in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1–4.

[19] M. Neumann and N. Thang Vu, ''Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,'' 2017, *arXiv:1706.00612*. [Online]. Available: http://arxiv.org/abs/1706.00612

[20] Q. Jin, C. Li, S. Chen, and H. Wu, ''Speech emotion recognition with acoustic and lexical features,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4749–4753.

[21] V. Chernykh and P. Prikhodko, ''Emotion recognition from speech with recurrent neural networks,'' 2017, *arXiv:1701.08071*. [Online]. Available: http://arxiv.org/abs/1701.08071

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, ''Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,'' in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376.

[23] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, ''Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations,'' in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 10–16.

[24] W. Jiao, M. R. Lyu, and I. King, ''Real-time emotion recognition via attention gated hierarchical memory network,'' 2019, *arXiv:1911.09075*. [Online]. Available: http://arxiv.org/abs/1911.09075

[25] U. Nadeem, M. Bennamoun, F. Sohel, and R. Togneri, ''Learning-based confidence estimation for multi-modal classifier fusion,'' in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 299–312.

[2]https://scikit-learn.org/stable/index.html

[26] S. Sahay, S. H. Kumar, R. Xia, J. Huang, and L. Nachman, "Multimodal relational tensor network for sentiment and emotion classification," 2018, *arXiv:1806.02923*. [Online]. Available: http://arxiv.org/abs/1806.02923
[27] M. Shad Akhtar, D. Singh Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," 2019, *arXiv:1905.05812*. [Online]. Available: http://arxiv.org/abs/1905.05812
[28] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2018, pp. 2122–2132.
[29] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, pp. 6818–6825.
[30] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, 2013, pp. 835–838.
[31] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473
[32] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078
[33] 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, Dec. 2008.
[34] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2018, *arXiv:1810.02508*. [Online]. Available: http://arxiv.org/abs/1810.02508
[35] S. Chen, C. Hsu, C. Kuo, and L. Ku, "EmotionLines: An emotion corpus of multi-party conversations," 2018, *arXiv:1802.08379*. [Online]. Available: https://arxiv.org/abs/1802.08379
[36] A. Zadeh, P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.
[37] 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.
[38] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "CNN+LSTM architecture for speech emotion recognition with data augmentation," 2018, *arXiv:1802.05630*. [Online]. Available: http://arxiv.org/abs/1802.05630
[39] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 190–195.
[40] C.-W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *Proc. Interspeech*, Sep. 2016, pp. 1387–1391.
[41] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 1771–1775.
[42] 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.
[43] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," 2018, *arXiv:1804.05788*. [Online]. Available: http://arxiv.org/abs/1804.05788
[44] P. Zhong, D. Wang, and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," 2019, *arXiv:1909.10681*. [Online]. Available: http://arxiv.org/abs/1909.10681
[45] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, "Recognizing emotions in video using multimodal DNN feature fusion," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang. (Challenge-HML)*, 2018, pp. 11–19.
[46] S. Sangwan, D. S. Chauhan, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis," in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 662–669.
[47] W. Y. Choi, K. Y. Song, and C. W. Lee, "Convolutional attention networks for multimodal emotion recognition from speech and text data," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang. (Challenge-HML)*, 2018, p. 28.
[48] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*. [Online]. Available: http://arxiv.org/abs/1606.08415

NGOC-HUYNH HO (Member, IEEE) received the B.S. degree from the Department of Telecommunication Engineering, Ho Chi Minh City University of Technology, Vietnam, in 2015, and the M.S. degree from the School of Electronics and Computer Science, Kookmin University, South Korea, in 2017. He is currently pursuing the Ph.D. degree with the School of Electronics and Computer Engineering, Chonnam National University, South Korea. His research interests include the multimodal-based emotion recognition, machine learning, deep learning and its applications, and bioinformatics.

HYUNG-JEONG YANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Chonbuk National University, South Korea. She is currently a Professor with the Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea. Her main research interests include multimedia data mining, medical data analysis, social network service data mining, and video data understanding.

SOO-HYUNG KIM (Member, IEEE) received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, in 1988 and 1993, respectively. Since 1997, he has been a Professor with the School of Electronics and Computer Engineering, Chonnam National University, South Korea. His research interests include pattern recognition, document image processing, medical image processing, and ubiquitous computing.

GUEESANG LEE (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree in computer engineering from Seoul National University, South Korea, in 1980 and 1982, respectively, and the Ph.D. degree in computer science from Pennsylvania State University, in 1991. He is currently a Professor with the Department of Electronics and Computer Engineering, Chonnam National University, South Korea. His primary research interests include image processing, computer vision, and video technology.