sensors

MDPI

Article

Fusion-ConvBERT: Parallel Convolution and BERT Fusion for Speech Emotion Recognition

Sanghyun Lee [1], David K. Han [2] and Hanseok Ko [1,*]

[1] Department of Electronics and Electrical Engineering, Korea University, Seoul 136-713, Korea; shlee@ispl.korea.ac.kr

[2] Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104, USA; dkh42@drexel.edu

* Correspondence: hsko@korea.ac.kr

Received: 20 October 2020; Accepted: 20 November 2020; Published: 23 November 2020

Abstract: Speech emotion recognition predicts the emotional state of a speaker based on the person's speech. It brings an additional element for creating more natural human–computer interactions. Earlier studies on emotional recognition have been primarily based on handcrafted features and manual labels. With the advent of deep learning, there have been some efforts in applying the deep-network-based approach to the problem of emotion recognition. As deep learning automatically extracts salient features correlated to speaker emotion, it brings certain advantages over the handcrafted-feature-based methods. There are, however, some challenges in applying them to the emotion recognition problem, because data required for properly training deep networks are often lacking. Therefore, there is a need for a new deep-learning-based approach which can exploit available information from given speech signals to the maximum extent possible. Our proposed method, called "Fusion-ConvBERT", is a parallel fusion model consisting of bidirectional encoder representations from transformers and convolutional neural networks. Extensive experiments were conducted on the proposed model using the EMO-DB and Interactive Emotional Dyadic Motion Capture Database emotion corpus, and it was shown that the proposed method outperformed state-of-the-art techniques in most of the test configurations.

Keywords: speech emotion recognition; bidirectional encoder representations from transformers (BERT); convolutional neural networks (CNNs); transformer; representation; spatiotemporal representation; fusion model

1. Introduction

In general, emotions are often tightly coupled with social interaction, cognitive processes, and decision making. Human brain processes the multimodalities to extract the spatial and temporal semantic information, that are contextually meaningful to perceive and understand the emotional state of an individual. In particular, human speech contains a wealth of emotional content; therefore, speech emotion recognition (SER) can play an crucial role in communication between humans and computers, as it would aid machines in more accurately predicting speech itself or the speaker's intention [1]. As such, there is an anticipation that SER may become an important component in social media, and it attracted much interest and research efforts in the area [2–4]. For a speech-only system, automatic emotion recognition should understand the fundamental dynamics of emotional cues and be able to identify emotional states from utterances. As recognizing emotions in speech among humans is difficult when cultural differences exist among speakers, the same difficulty remains for SER [5]. Coupled with speech variations among speakers and dynamical features with low saliency, SER is a challenging problem. Previous speech emotion studies have used handcrafted

features consisting of low-level descriptors (LLDs), local features such as Mel-frequency cepstral coefficients (MFCCs), energy, or pitch, and have also considered global features by calculating local feature statistics [6,7]. However, extracting handcrafted features requires expensive manual labor and the data quality depends on expert knowledge of labelers. The recent introduction of deep learning models learns relevant features automatically from data without such expert knowledge [8–10]. Restricted-Boltzmann-machine (RBM)-based deep neural networks (DNNs) [11,12] or convolutional neural networks (CNNs) [13,14] have shown significant performance improvements over the handcrafted-feature-based methods in prediction. Trigeorgis et al. [9] proposed a hybrid architecture of CNNs extracting local features from a spectrogram and recurrent neural networks (RNNs) taking the output of the CNNs for considering sequentially relevant features from speech. While these deep-learning-based models are effective in the applied scenarios, there are still some significant challenges that these methodologies have yet to overcome. Those issues are as follows:

1. Most of the existing models assume fixed-size input length; however, speech signals are often of variable time durations.
2. Local features extracted by CNNs may not necessarily contain contextually important information on a global scale, as some emotional contents may reside over a lengthy utterance. Previous models of serial fusion that employed RNN structures following the local feature extraction process by CNNs may not be able to capture global scale contextual features [9,15].
3. A new architecture capable of independently capturing both local and global emotional features in lengthy utterances is needed. To address these challenges, we propose an end-to-end technique called "Fusion-ConvBERT", a parallel network and fusion architecture that automatically captures rich meaning in spatial and temporal features of two models.

Bidirectional encoder representations from transformers (BERT) [16], a recent transformer-based model [17] for natural language processing (NLP), is a pretrained model by unsupervised language models for various NLP tasks. Pretrained models using BERT dominate the NLP world as they have proved highly effective [18,19]. The pretrained models also learn robust phonetic expressions in speech processing tasks, such as speaker recognition and SER [20–23]. In Fusion-ConvBERT, log mel-spectrograms are extracted from acoustic signals first to be composed as inputs for BERT and CNNs. We employ Mockingjay [21], which is a speech recognition model by pretraining BERT with a large corpus speech data, for fine tuning it for emotion recognition. In the proposed architecture, BERT captures global features associated with lengthy sequential data, as it can maintain bidirectional dependencies, and CNNs are responsible for extracting salient SER features by perceiving local fields of data. In the proposed transformer fusion architecture, BERT and CNNs learn simultaneously. Our major contributions in this paper can be summarized as follows:

- We propose a novel framework to fuse both spatial and temporal representations for SER by leveraging transformer-based Fusion-ConvBERT with pretrained BERT and CNNs, an approach capable of automatically learning feature representations and modeling the temporal dependencies.
- Different from previous serial fusion methods, our method adopts input in multiple features in parallel and simultaneously fuses various emotion details in the transformer layer. The rich interaction that occurs in the fusion allows Fusion-ConvBERT to capture intermediate associations between the local and global patterns and also between different modalities at various representation depths.

We conduct extensive speaker-independent and speaker-dependent experiments, using labeled emotional speech data from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [24] and the Berlin Database of Emotional Speech (EMO-DB) [25] datasets. To the best of our knowledge, this is the first study in which the BERT model and CNNs are applied to a fusion model, to learn enhanced deep spectrum representations for SER. We demonstrate experimentally that our framework outperforms individual models. For the EMO-DB, our method achieves a weighted accuracy (WA) of

88.43% and an unweighted accuracy (UA) of 86.04% in the speaker-independent experiments. In the speaker-dependent experiments, it achieves a WA of 94.23% and UA of 92.1%. For the IEMOCAP, it achieves a WA of 66.47% and UA of 67.12% in the speaker-independent experiments and a WA of 69.51% and UA of 71.36% in the speaker-dependent ones. The remainder of this paper is organized as follows. In Section 2, existing works related to this research field are presented. In Section 3, we describe the proposed Fusion-ConvBERT framework for speech emotion classification. In Section 4, an experimental analysis comparing the performance of the proposed model on the IEMOCAP and EMO-DB corpora is detailed. Finally, in Section 5, a summary of the study and scope for future work is provided.

2. Related Work

SER has attracted considerable attention in digital signal processing research. Recently, researchers have developed various approaches using digital speech signals to identify the emotional conditions of an individual speaker. The focus is on emotion classification by salient acoustic features of speech. Classical machine learning sentiment classifiers include the hidden Markov model (HMM) [26,27], Gaussian mixed model (GMM) [28], and support vector machine (SVM) [29,30]. In general, previous studies have considered both local features (LLDs or MFCCs and energy and pitch) and global features (statistical functionals) [6,31,32]. However, these studies have been based on handcrafted low-level features found to be effective in distinguishing speech emotions. Alternatively, much work has been conducted into applying deep learning to automatically learn useful features relevant to emotion from speech data. Stuhlsatz et al. [12] began with an RBM to preoptimize the network parameters and proceeded with supervised training of a DNN for recognizing speech emotions and improved upon the accuracy of other classifiers (e.g., SVMs). By extracting spectrograms from speech signals and learning hidden time-series information, CNNs can select high-level discriminatory features and recognize the emotional state of the speaker [33–36]. Traditional SER methods also include bidirectional long-term short-term memory (BLSTM) approaches [37]. Meng et al. [38] proposed a 3-D convolutional network for SER, by combining CNNs with BLSTM. In their model, the 3-D spectral features of segments were used as inputs. Mirsamadi et al. [39] adopted attention mechanisms and feature pooling; thus, the method automatically classified emotions from speech, focusing on the areas of speech signals that were more emotionally prominent. Extracting useful features is essential in good recognition performance, but a fusion method that combines different feature information may further improve classifier results. To this end, Jiang et al. [40] presented an SER fusion method using handcrafted and bottleneck features, whereas Guo et al. [41] and Lim et al. [42] proposed a hybrid CNN-BLSTM model without using any handcrafted features. Although these serial fusion models obtained good results for many speech processing tasks, numerous problems were left unaddressed. For instance, the serial fusion model may not be able to capture global scale emotion contextual features. Furthermore, the deeper the learning model, the easier it falls into overfitting when the database is small. To alleviate these difficulties, we propose a method based on combining a pretrained BERT model and CNN functions to identify salient local spatial features in the SER dataset. The pretrained BERT model prevents overfitting by fine-tuning features of temporal flow over the entire spectrogram frame. Additionally, the model learns by fusing local features extracted from the CNNs and BERT output. The proposed method is shown to be superior to an individual model based on our experiments.

3. Proposed Methodology

In this section, we describe Fusion-ConvBERT, which is formed of two parallel networks (the BERT model and CNNs) and uses log-mel spectrogram as its input (Figure 1). The BERT model extracts speech representations (temporal information) from a spectrogram, whilst the CNNs extract spectrotemporal information. Then, the speech features from these two networks are fused to form a combined spectrotemporal feature vector.

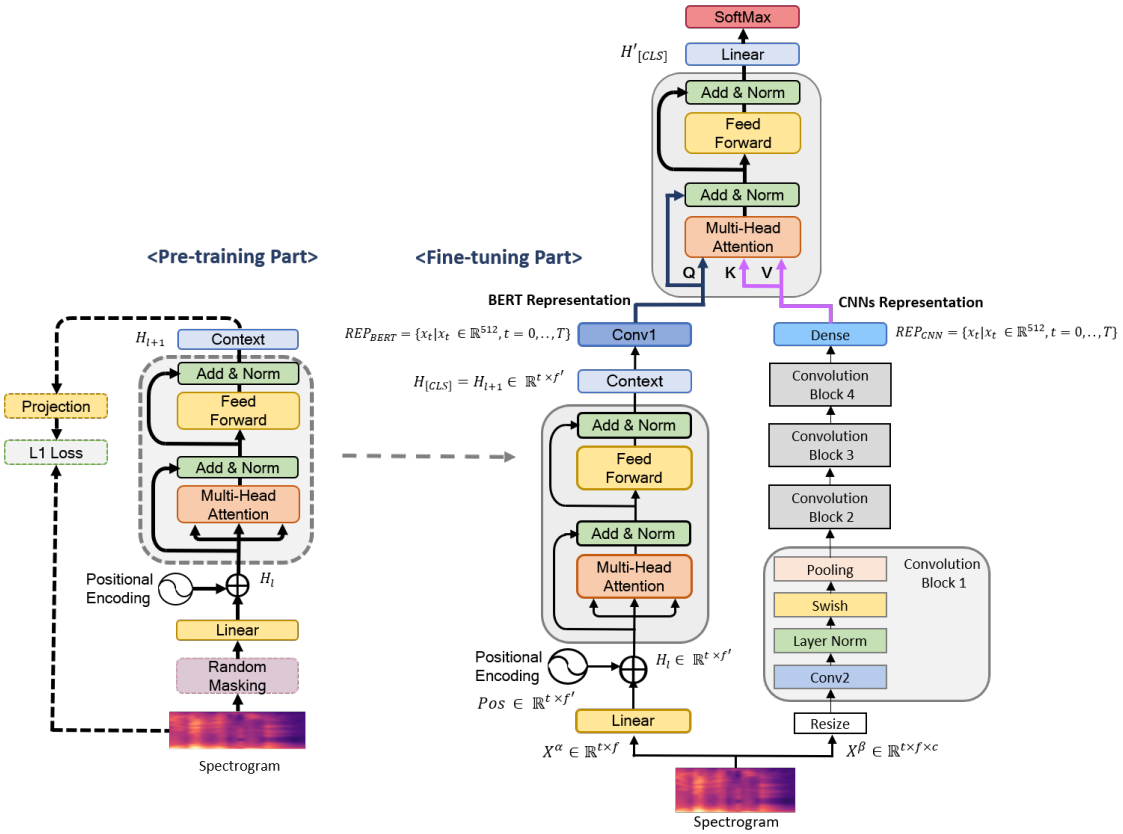


Figure 1. Overview of our proposed Fusion-ConvBERT method for speech emotion recognition. In the pretraining component, a random masking strategy is employed in the spectrogram frame using speech recognition data. Then, unsupervised learning is performed using an L1 loss between the masked and predicted frames. During the fine-tuning process of the BERT model, the spectrogram features are simultaneously input to the pretrained BERT model and CNNs. Meanwhile, Fusion-ConvBERT as a whole is trained by fusion at the transformer; it uses the final hidden-layer features obtained during the fine-tuning of both the CNN- and BERT-based learning components.

3.1. Log-Mel Spectrogram Generation

The first step in the process is to extract log-mel spectrograms as the input to our proposed network. We used the Librosa framework [43] to resample the one-dimensional audio signals received from a microphone to 16 Khz; then, we split them into short frames via a short-time Fourier transform (STFT) [44] using a Hamming window function with a frame length of 25 ms and a rate of 10 ms. Then, the power spectrum for each frame was calculated and transmitted through the mel-filter bank to generate the output $mel_{i,j} \in mel$, where i denotes the power spectrum components corresponding to the filter bank and j denotes the individual frequency components that span the filter bank region i. The relationship between the mel-spectrum (after scaling) and frequency f' is expressed via Equation (1):

$$mel = 2595 \log(1 + \frac{f'}{700\ \text{Hz}}). \tag{1}$$

The mel-frequency spacing approximates that of the human cochlea, and mel-spectrograms reflect the relative importance of different frequency bands [45]. Then, as shown in Equation (2), the features are normalized using the mean $\mu_{i,j}$ and variance $\sigma_{i,j}$ to obtain $x_{i,j} \in X$ such that $-1 \leq mel \leq 1$.

$$x_{i,j} = \frac{mel_{i,j} - \mu_{i,j}}{\sigma_{i,j}}. \tag{2}$$

Equation (3) shows how the log-mel features, calculated from the inputs of the CNNs and BERT model, are reconstructed for each model.

$$X^{\alpha} \in \mathbb{R}^{t \times f}, \quad X^{\beta} \in \mathbb{R}^{t \times f \times c}. \tag{3}$$

where t denotes the total number of time frames, f denotes the number of mel-filter banks ($f = 160$), and c denotes the number of feature channels. X^{α} denotes the single channel two-dimensional BERT model input while X^{β} represents the CNN input with its channel extended to three dimensions. As the number of frames in the frame-level features varies depending on speech, these features are zero-padded to match the predefined dimension.

3.2. CNN Architecture

In this session, we describe the CNN-based architecture for extracting log-mel features for SER. The network is primarily composed of four convolution blocks with indices 1–4.

Convolutional blocks are composed of a two-dimensional convolution layer, a layer normalization [46], and a Swish activation [47] in Figure 2. The advantage of the two-dimensional convolution layer is that it can extract local features using the connectivity and shared weights of the spatial information [48,49]. Performing layer normalizations of the convolution layer's activations at each batch improves the performance and stability of the network. The recently proposed Swish activation is defined as follows:

$$x\varphi(\theta x) = x\varphi(1 + exp(-\theta x))^{-1}. \tag{4}$$

where $\varphi(\cdot)$ denotes the sigmoid function and θ is either a constant or a trainable parameter. The convolution and pooling kernels in each convolution block are two-dimensional. The convolution kernels are of the same size (3×3), stride (1×1), and shape. Convolutional layers can be composed of multiple feature maps. Layers 1–4 contain 64, 128, 256, and 512 convolution kernels, respectively. The parameters of the CNNs are shown in Table 1. Local connections, weight sharing, and downsampling in CNNs can effectively reduce the complexity of the network model and the number of training parameters. The process abstracts input data into high-level feature representations via algorithmic operations between the layers. When only the CNN network is used, the dense output from the fourth layer as shown in Equation (5) is mapped to K emotional categories via Softmax.

$$REP_{CNN} = \left\{ x_t | x_t \in \mathbb{R}^{512}, t = 0, \ldots, T \right\}. \tag{5}$$

Table 1. Layer parameters of the CNNs. The output shape is given by the time steps (T), mel bins (F), and feature maps (C). The output K denotes the number of emotion targets.

Layers	Output Shape ($Time\ Steps: T \times Mel\ Bins: F \times Feature\ Maps: C$)	Kernel Size	Stride
Input	$T \times F \times 1$	–	–
Convolution (1)	$T \times F \times 64$	3×3	1×1
Pooling (1)	$T/4 \times F/4 \times 64$	4×4	4×4
Convolution (2)	$T/4 \times F/4 \times 128$	3×3	1×1
Pooling (2)	$T/16 \times F/16 \times 128$	4×4	4×4
Convolution (3)	$T/16 \times F/16 \times 256$	3×3	1×1
Pooling (3)	$T/64 \times F/64 \times 256$	4×4	4×4
Convolution (4)	$T/64 \times F/64 \times 512$	3×3	1×1
Pooling (4)	$1 \times 1 \times 512$	2×2	2×2
Dense	512	–	–
Output	K	–	–

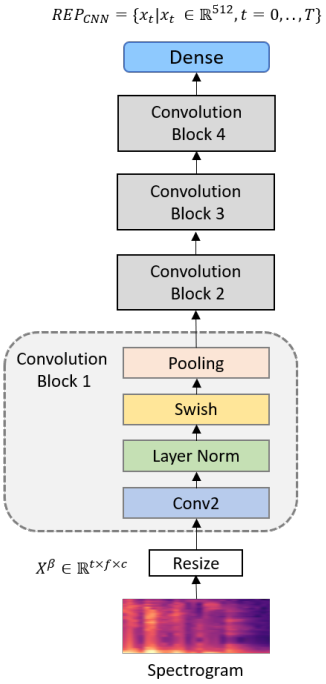


Figure 2. Proposed CNN architecture.

3.3. Architecture of BERT Model

This section introduces the BERT model, which consists of a transformer architecture for extracting temporal features via the self-attention process [16]. Unlike recent language processing models, the BERT model is designed to pretrain deep bidirectional representations from the unlabeled text. Once pretrained on a large language corpus, the BERT model can effectively deliver transfer learning for multiple NLP tasks, such as extracting expressive global features of sentences. The BERT model is commonly trained in a two-step strategy: (1) pretraining and (2) fine-tuning. In our architecture, we begin with the Mockingjay [21] model, which is essentially a BERT pretrained on acoustic data, and proceed to fine-tuning for the emotion classification task. The process starts by taking sequences of mel-feature frames $X = (x_0, \ldots, x_T) \in \mathbb{R}^{t \times f}$ and passing them through the transformer to produce the learned encoding of BERT representations $H = (h_0, \ldots, h_T) \in \mathbb{R}^{t \times f}$. In 15% of the frames selected during pretraining, we mask 80% of the time frame, replacing the remainder with 10% of the frames randomly shuffled and 10% unaltered. Given the observed set, the BERT model is trained to minimize reconstruction errors between the prediction and ground-truth frames selected using L1 loss (15%). The BERT structure is presented in Figure 3.

The flow path denoted by dotted lines and the random masking process shown in Figure 3 are used only in the pretraining process while the other elements in the figure were used for both pretraining and fine-tuning. In the BERT architecture, H_l denotes an input representation combining a linear layer and position embedding, and it is supplied as the transformer input. Specifically, for the i_{th} head attention in Equation (6), the input layer is based on the dot-product attention mechanism [17] as follows:

$$\begin{aligned} att_i &= \varepsilon(X), \\ &= softmax\left(\frac{QK^T}{\sqrt{m}}\right)V, \\ &= softmax\left(\frac{XW_{Q_i}W_{K_i}^T X^T}{\sqrt{m}}\right)W_{V_i}X. \end{aligned} \tag{6}$$

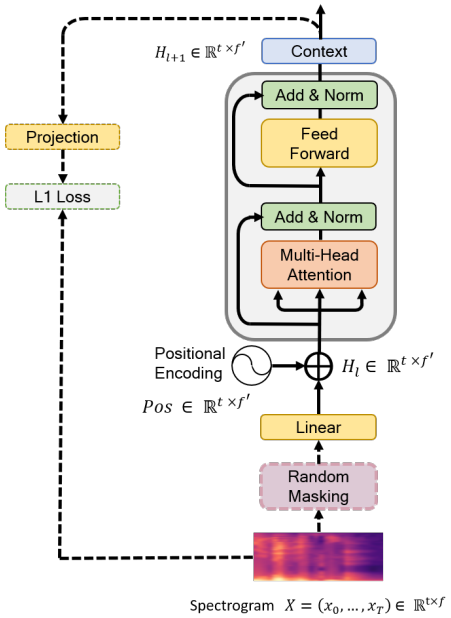


Figure 3. Proposed BERT architecture.

We define the query as $Q = XW_Q$, the key as $K = XW_{K_i}$, and the value as $V = XW_{V_i}$, where, $W_{Q_i} \in \mathbb{R}^{f \times f}$, $W_{K_i} \in \mathbb{R}^{f \times f}$ and $W_{V_i} \in \mathbb{R}^{f \times f}$ are weights. Note that att_i has the same dimensions as Q, K, and V. The Softmax function in Equation (6) measures the attention given for specific parts of the mel spectrogram, thus att_i is features of V weighted by attention calculated from Q and K. The outputs of the m attention heads are concatenated together and followed by a linear layer as

$$Linear(X) = W_m[att_1(X), \ldots, att_m(X)]^T + b_m. \tag{7}$$

where $W_m \in \mathbb{R}^{f \times f}$ and b_m are learning parameters. The BERT model adds the residual connection from the input to output and then adds the layer normalization [46] as follows:

$$Z = LayerNorm(X + Linear(X)). \tag{8}$$

$$H = LayerNorm(X + Feedforwrad(Z)). \tag{9}$$

The entire model stacks L layers, and the final representation H_{l+1} of the frame is used for the fusion networks.

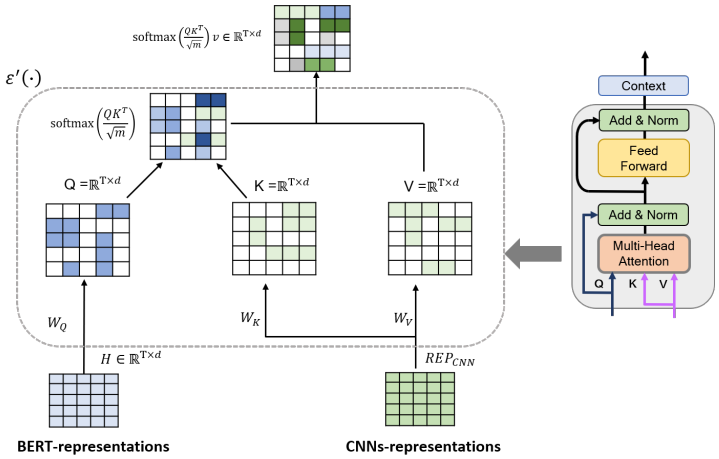
3.4. Fusion-ConvBERT

While following the basic layout of NLP encoder transformers, we propose a hybrid transformer model by incorporating a fusion of taking two separate input features in different combinations of Q, K, and V. BERT representations and CNN features are the two input types considered, and one of these was taken as input for Q and K while V was provided from the other representation. As illustrated in Figure 4, an intuitive but general solution that incorporates the two features into the BERT architecture is to directly concatenate the two representations with the final hidden states of the input sequence and then stack additional BERT layers on top, to capture the intermodel interactions between the CNNs and BERT representations. In Equation (5), the output features of the CNNs and H_{l+1} pass through a one-dimensional convolution, to project them into the same feature space. Furthermore,

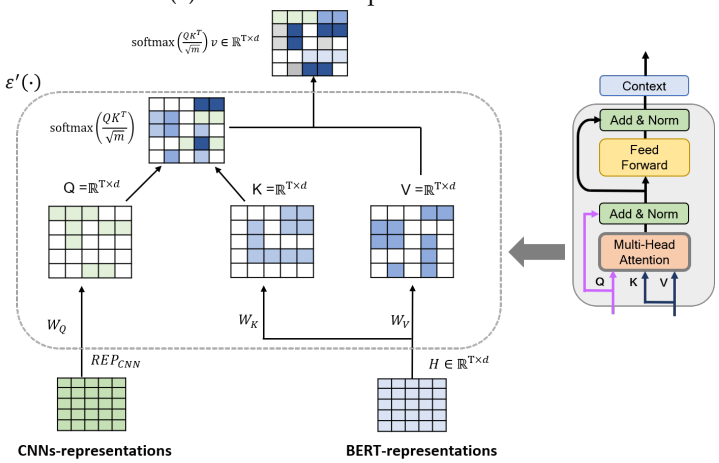
Fusion-ConvBERT applies a transformer to automatically model the rich interactions between the CNNs and BERT model. The i_{th} head attention takes the form of Equation (10) as

$$\begin{aligned} att'_i &= \varepsilon'\left(REP_{CNN}, H_{l+1}\right) \\ &= softmax\left(\frac{H_{l+1}W'_{Q_i}W'^{T}_{K_i}REP^{T}_{CNN}}{\sqrt{m}}\right)W'_{V_i}H_{l+1}. \end{aligned} \tag{10}$$

where W'_{Q_i}, W'_{K_i}, and $W'_{V_i} \in \mathbb{R}^{f\times f}$ are the parameters. Similarly to the BERT model, the L layers are stacked to obtain the final representation by connecting the feedforward layer and residual connections. Finally, we direct the final hidden state of H'_{l+1}, which captures salient features of different emotions of interest, to the linear function for classification. All modules (i.e., the CNNs and fine-tuned BERT model) are trained simultaneously, to ensure that the model can learn the emotional content of each utterance. As shown in Figure 4, we experimented with two different design configurations of fusion. The first configuration, shown as Figure 4a, applies the CNN-extracted representation for K and V computations, thus greater attention weight was given to the features obtained by the CNN. Figure 4b depicts the other design option that applies the BERT-extracted representation toward K and V calculations for giving a larger emphasis to the BERT delivered features. The results of both experiments are discussed in detail in Section 4.2.4.



(**a**) Fusion with emphasis on CNNs



(**b**) Fusion with emphasis on BERT

Figure 4. Proposed Fusion-ConvBERT strategy. (**a**), CNN representation is given more emphasis in the attention process; (**b**), BERT representation is given more emphasis in the attention process.

4. Experiments and Results

In this section, we evaluate the proposed system's effectiveness for SER and compare it against other baseline methods on a publicly available benchmark speech emotion dataset.

4.1. Dataset

The Fusion-ConvBERT network was evaluated on two public speech datasets: EMO-DB [25] and IEMOCAP [24]. These two databases contain predetermined sentences spoken by invited actors according to the emotions required. A detailed description of the datasets is given in the following subsections.

4.1.1. EMO-DB Emotion Database

The Berlin EMO-DB was recorded in 2005 and is a German-language SER database. The dataset contains 353 sentences with an average length of approximately 2.7 s recorded at 16 Khz sampling rate. We used seven categories of emotions and they are listed in Table 2.

Table 2. Emotion class distribution of EMO-DB dataset.

Emotion	Total Utterances	Proportion in (%)
Anger	127	23.74
Fear	69	12.90
Boredom	81	15.14
Disgust	46	8.60
Happiness	71	13.27
Neutral	79	14.77
Sadness	62	11.59
Total	535	100

4.1.2. IEMOCAP Emotion Database

The IEMOCAP database is widely used by researchers in the field of SER. It features two types of dialogue: scripted and improvised. It is an acted English-language SER dataset consisting of five sessions; each session includes two actors (one male and one female), and the database contains 12 h of audio-visual data from all 10 actors recording different emotions, including anger, disgust, fear, sadness, neutral, happiness, and excitement. It is also worth noting that the data distribution of each emotion class is heavily imbalanced. Therefore, following the approach of [50,51], we merged the happiness and excitement utterances into the happiness class. We used four categories of emotions—namely neutral, happiness, sadness, and anger—for training and evaluation. Details are given in Table 3.

Table 3. Emotion class distribution of IEMOCAP dataset.

Emotion	Total Utterances	Proportion in (%)
Neutral	1708	30.88
Happiness	1636	29.58
Sadness	1084	19.60
Anger	1103	19.94
Total	5531	100

4.2. Experimental Setup and Evaluation Metrics

We conducted several experiments to evaluate the performance of the proposed framework. These experiments are divided mainly into two categories: speaker-dependent and speaker-independent. Each of these experiments was conducted in two phases: (1) pretraining the BERT setup, and (2) the Fusion-ConvBERT setup. As stated earlier, we applied Mockingjay [21], a speech recognition version

of BERT, by pretraining it with the LibriSpeech [52] corpus train-clean-360 containing 1000 h of data. While we followed the main structure of Mockingjay, we found the effect of its downsampling and upsampling parts to be minimum. Thus, for expediency, they were excluded from our architecture. We used the mel-features as inputs to be converted into a high-level representation. In the pretraining phase, we used the same hidden dimension size of $H = 768$, an attention head of $A = 12$, a layer number of $L(-layer) = 3$, and a consecutive masking number of $C = 7$. A total of 500 K epochs were used in the pretraining. In the fusion phase, Fusion-ConvBERT implemented an attention dropout [53] of 0.3 inside the transformer and 20 K training epochs total. The fusion transformer layer was applied with a hidden dimension size of $H = 120$, an attention head of $A = 6$, and a layer number of $L'(-layer) = 3$. We used a learning rate of $2e - 3$ for all models and applied the Adam [54] optimizer and Swish [47] as activation function. We also applied SpecAugment [55] to prevent overfitting. The overall model performance was quantitatively measured using the F1 score, WA, and UA. WA and UA better reflect class-to-class imbalance class averages. We determined WA as the ratio between the correctly classified emotions and the total emotions of the same class. Similarly, UA is the ratio between the correctly predicted emotions and all emotions in the dataset. The metrics are expressed as follows:

$$F1 = \frac{2 * precision * recall}{precission + recall}. \tag{11}$$

$$WA = \sum_i \frac{correct\ utterances}{utterances}. \tag{12}$$

$$UA = \sum_i \frac{correct\ utterances\ for\ emotion\ i}{utterances\ of\ emotion\ i}. \tag{13}$$

Furthermore, we performed a detailed ablation study to justify our design choices. First, we built CNN, BERT, and transformer models, to investigate the impacts of spatial and temporal information. The model comparison is summarized as follows: (1) CNNs: we used the proposed CNNs only (see Section 3.2); (2) BERT: we fine-tuned the pretraining BERT model by adding a transformer downstream (see Section 3.3); (3) Transformer: the BERT strategy of masking pretraining was not used and only the transformer encoder architecture was applied; (4) Fusion-ConvBERT: during the fine-tuning process of the BERT model, the pretrained BERT model and the features extracted from CNNs were trained simultaneously in transformer networks. We also compared the fusion performances obtained when using the BERT model for fine-tuning and as a feature extractor in Fusion-ConvBERT. Finally, for Fusion-ConvBERT, we compared the transformer fusion configurations of placing greater emphasis on CNN features versus greater emphasis on BERT features. In other words, the value compares how weights were assigned to spatial features (CNN features) or the temporal features (BERT features). All models were implemented using the Pytorch framework and the suggested models were trained and evaluated on two NVIDIA GeForce GTX 2080TIs.

4.2.1. Speaker-Independent Performance Experiments

We conducted extensive speaker-independent experiments on all the labeled emotional speech data of the EMO-DB and IEMOCAP datasets. The EMO-DB and IEMOCAP corpora contain 10 speakers, and we performed 10-fold cross-validation using a leave-one-out strategy. In each training process, eight speakers from four sessions were used as training data, and the remaining sessions were separated into two parts; thus, eight, one, and one folds were used for the training, development, and testing sets, respectively. Therefore, speaker independence is strictly enforced by testing the model performance by an unseen speaker. As such, the arrangement may also shed some light on how generalizable the models are. We evaluated the proposed system using these datasets and compared several models to verify our design decisions. The comparisons are shown in Tables 4 and 5.

Table 4. Performance of the proposed model for speaker-independent emotion recognition on EMO-DB.

Model	F1-Score	WA (%)	UA (%)
CNNs	0.77	79.42	78.12
BERT	0.82	84.77	83.01
Transformer	0.78	80.25	79.1
Fusion-ConvBERT	**0.84**	**88.43**	**86.04**

Table 5. Performance of the proposed model for speaker-independent emotion recognition on IEMOCAP.

Model	F1-Score	WA (%)	UA (%)
CNNs	0.63	62.12	64.75
BERT	0.64	64.3	65.11
Transformer	0.63	61.41	64
Fusion-ConvBERT	**0.66**	**66.47**	**67.12**

It is apparent from the tables that Fusion-ConvBERT outperformed all the other models. The highest WA and UA achieved were 88.43% and 86.04%, respectively. For IEMOCAP, a WA and UA of 66.47% and 67.12% were achieved, respectively. Since IEMOCAP data contains more realistic expressions of speaker emotions rather than the EMO-DB data collected from speeches by actors instructed to express content emotions, the model performance is overall lower as can be expected. Models containing the pretrained BERT (BERT or Fusion-BERT) showed higher performances than the other models. From the results, it can be surmised that CNN and the pretrained BERT are extracting information that is mutually exclusive, as the fusion of the two delivers improved performances. It is also observed that the pretraining of BERT delivers better results compared to simply applying the transformer. The proposed architecture of employing a transformer as a fusion structure for combining features taken from CNN and pretrained BERT clearly delivers improved overall performances. The confusion matrix for the proposed Fusion-ConvBERT shows the actually predicted emotions and the model confusion results of each class (Figure 5).

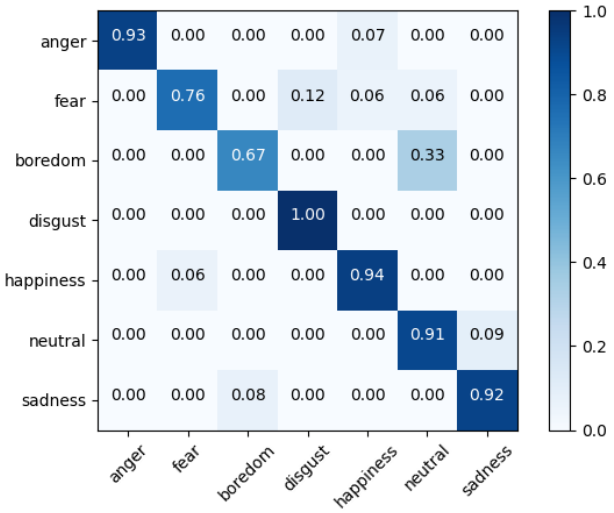


Figure 5. Confusion matrix of speaker-independent emotion prediction on EMO-DB with 86.04% unweighted accuracy overall; the confusion between actual and predicted emotions is shown in the corresponding row.

The figure shows improved overall emotion recognition performance for the EMO-DB dataset, however, both “fear” and ”boredom” were identified with accuracies below 80%. While recognition rates for the other five emotions (i.e., “anger”, “disgust”, “happiness”, “neutral”, and “sadness”)

exceed 90%; "boredom" was confused with "neutral" due to diversities in how these emotions are expressed among people.

Figure 6 shows that on the IEMOCAP dataset, "neutral" was identified with the highest accuracy (68%), "anger" and "happiness" were identified with the lowest accuracy (64%), and both were affected by "neutral."

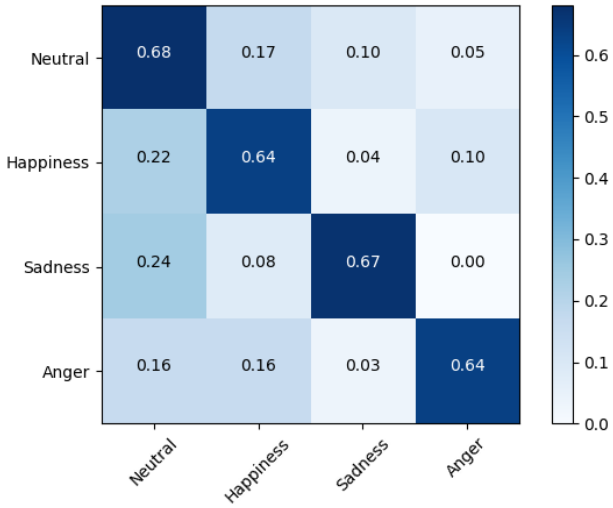


Figure 6. Confusion matrix of speaker-independent emotion prediction on IEMOCAP with 67.12% unweighted accuracy overall; the confusion between actual and predicted emotions is shown in the corresponding row.

4.2.2. Speaker-Dependent Performance Experiments

In the speaker-dependent experiments, the data were shuffled and the entire set was divided by randomly selecting an 80:20 split ratio for model training and testing, respectively. As before, we investigated the speaker-dependent model using F1-score, WA, and UA. The detailed numerical results for the EMO-DB and IEMOCAP datasets are given in Tables 6 and 7.

Table 6. Performance of the proposed model for speaker-dependent emotion recognition on EMO-DB.

Model	F1-Score	WA (%)	UA (%)
CNNs	0.84	87.56	85.44
BERT	0.91	91.42	91.12
Transformer	0.86	88.4	87.34
Fusion-ConvBERT	**0.91**	**94.23**	**92.1**

Table 7. Performance of the proposed model for speaker-dependent emotion recognition on IEMOCAP.

Model	F1-Score	WA (%)	UA (%)
CNNs	0.62	64.11	66.41
BERT	0.68	70.45	71.12
Transformer	0.66	63.16	65.75
Fusion-ConvBERT	**0.69**	**69.51**	**71.36**

Fusion-ConvBERT again was found superior in this set of experiments. The EMO-DB dataset shows 94.23% WA and 92.1% UA, and the IMEOCPA dataset shows 69.51% WA and 71.36% UA. The confusion matrices for Fusion-ConvBERT in the speaker-dependent experiments are shown in Figures 7 and 8.

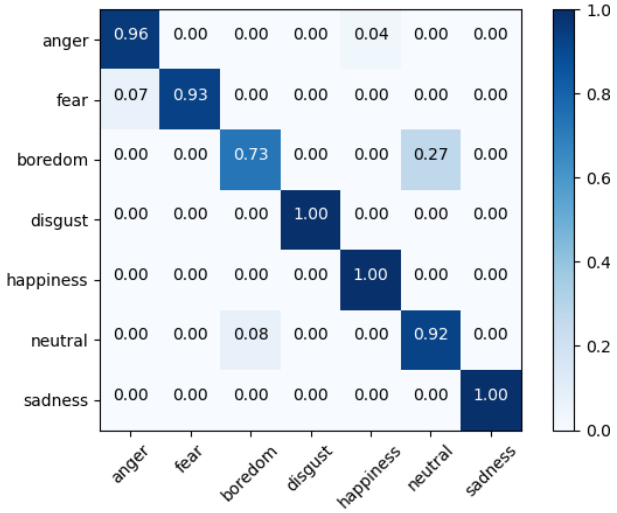


Figure 7. Confusion matrix of speaker-dependent emotion prediction on EMO-DB with 92.1% unweighted accuracy overall; the confusion between actual and predicted emotions in the corresponding row.

In this experiment, the model recognized "disgust", "happiness, "' and "sadness" with high accuracy while "boredom" was recognized with a 73% ratio. As in the speaker-independent experiments, "boredom" was confused with "neutral."

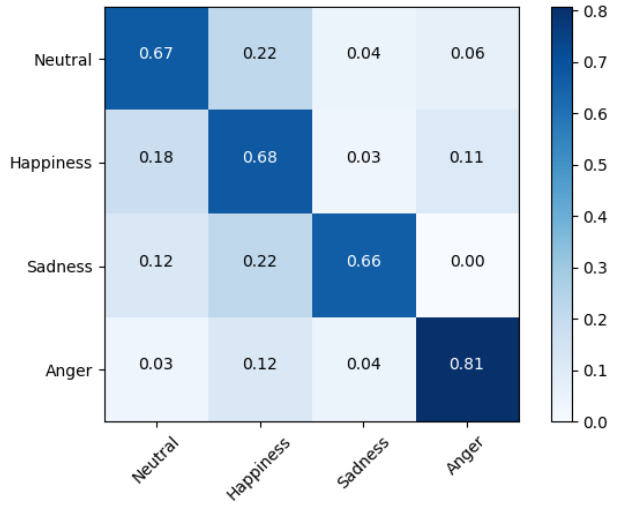


Figure 8. Confusion matrix of speaker-dependent emotion prediction on IEMOCAP with 71.36% unweighted accuracy overall; the confusion between actual and predicted emotions is shown in the corresponding row.

In the speaker-dependent, IEMOCAP dataset experiments on our network architecture, we found a notable superiority for the "anger" emotion state, with an 81% UA. For "neutral", "happiness", and "sadness", we found almost equal performance results of 67%, 68%, and 66%, respectively. The overall accuracy of the system for speaker-dependent emotion recognition exceeded that of the speaker-independent emotion recognition experiment.

4.2.3. Analysis of Fine-Tuning Effects of Fusion-ConvBERT

As described earlier, Fusion-ConvBERT was first pretrained on the BERT model to compensate for the sparseness of the emotional recognition data. A fusion transformer encoder was added on top of the BERT model, and the overall model is fine-tuned (simultaneously with CNN representation training) as shown in Figure 1. Here, we demonstrate the effectiveness of Fusion-ConvBERT when fused with

a pretrained BERT model by comparing performances of two configurations, namely "Fusion-A" and "Fusion-B". In Fusion-A, learning is applied to the whole network in the fine-tuning process including the pretrained parameters while in Fusion-B the pretrained parameters are frozen in the fine-tuning process.

We compared these strategies using both the EMO-DB and IMEOCAP emotion datasets as shown in Figure 9. From the figure, we can see that the Fusion-A outperforms Fusion-B in general. Thus, the value of the fine-tuning method in our architecture is clear.

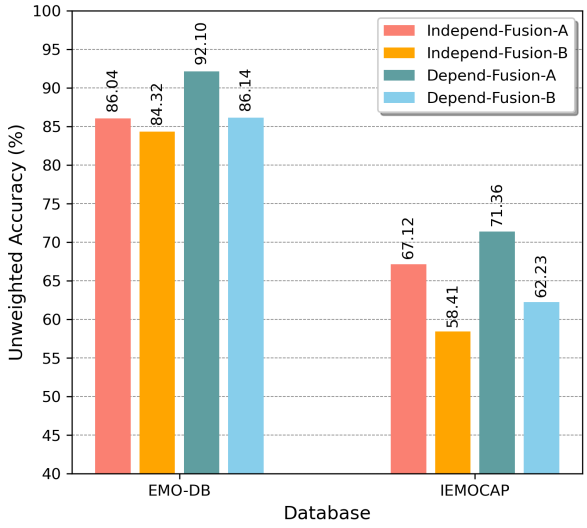


Figure 9. Performance comparison of fine-tuning and feature extractor strategies when training Fusion-ConvBERT with pretrained BERT model, evaluated using the EMO-DB and IEMOCAP datasets with unweighted accuracy overall.

4.2.4. Analysis of Fusion Strategies

We investigated fusion strategies for Fusion-ConvBERT. As shown in Figure 4, in Fusion-ConvBERT, two alternative fusion strategies (BERT-focused and CNN-focused) were applied to the attention mechanism by taking different value combinations of the query, key, and value. Figure 10 shows their experimental performances on the EMO-DB and IEMOCAP datasets.

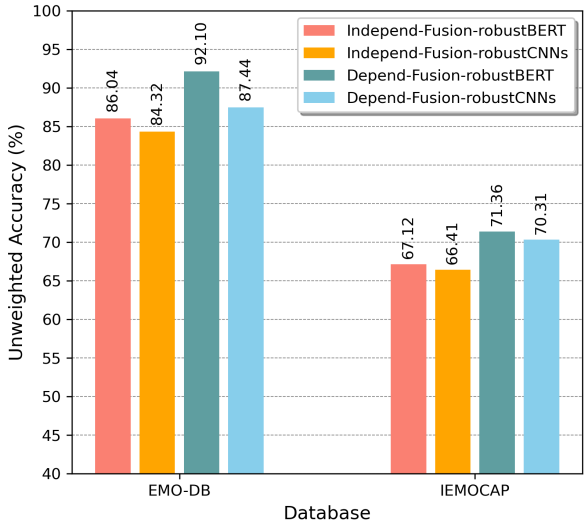


Figure 10. Unweighted overall accuracy comparison of EMO-DB and IEMOCAP according to the two fusion strategies in Fusion-ConvBERT.

"Fusion-robustBERT" in Figure 10 indicates the experiments with a query from CNN representation, and the key and value taken from BERT representation, thus greater weight, in essence, is given to the BERT-extracted features. "Fusion-robustCNNs" represents the experiments with query taking the value from BERT representation, and the key and value are from the CNN representation, therefore the CNN features are given greater emphasis. Both the speaker-dependent and speaker-independent tests were conducted using the two fusion configurations and the results are labeled accordingly in the figure. As shown from the figure, the BERT-focused fusion methodology consistently outperforms the CNN-focused fusion. It seems that in the case of emotion recognition, considering the temporal context in large scales, as BERT is capable of, would yield better results compared to placing more emphasis on spectral-temporal correlations and context.

4.2.5. Discussion

SER performance of Fusion-ConvBERT was compared to the existing methods over both the speaker-independent and the speaker-dependent configurations and the results are summarized in Tables 8 and 9. As most of the existing methods were evaluated in terms of UA scores, only limited performance comparison was possible via WA scores. It can be seen that the approach proposed outperforms existing methods on the EMO-DB and IEMOCAP datasets in all cases except one in the speaker-independent test of EMO-DB dataset.

Table 8. Speaker-independent and speaker-dependent comparison of the proposed model against baseline methods for the EMO-DB dataset. The optimal results are highlighted in bold. Our model outperforms the current state-of-the-art methods across most evaluation metrics.

Model	Speaker-Indep (WA%)	Speaker-Indep (UA%)	Speaker-Dep (WA%)	Speaker-Dep (UA%)
Guo, L. et al. [41]	87.85	**87.49**	–	87.85
Meng, H. et al. [38]	–	84.99	–	90.37
Chen, M. et al. [15]	–	82.82	–	–
Badshah, A.M. et al. [56]	–	80.79	–	89.46
Jiang, P. et al. [57]	–	84.53	–	86.44
Our model	**88.43**	86.04	**94.23**	**92.1**

Table 9. Speaker-independent and speaker-dependent comparison of the proposed model against baseline methods for the IEMOCAP dataset. Optimal results are highlighted in bold. Our model outperforms the current state-of-the-art methods across most evaluation metrics.

Model	Speaker-Indep (WA%)	Speaker-Indep (UA%)	Speaker-Dep (WA%)	Speaker-Dep (UA%)
Guo, L. et al. [41]	56.55	57.99	–	–
Zheng, W. et al. [58]	–	40.02	–	–
Behnke, S. et al. [49]	–	51.24	–	–
Luo, D. et al. [59]	60.35	63.98	–	–
Chen, M. et al. [15]	–	64.74	–	–
Our model	**66.47**	**67.12**	**69.51**	**71.36**

The reason that the speaker-independent UA score of our model is lower than the one by Guo, L. et al. [41] is that Guo's model delivered better classification in the Boredom category. It must be noted the dataset on the Boredom class is sparse compared to the other categories, thus the WA score of the same experiment shows our model delivering a better result. It is not clear whether this disparity is due to the sparseness of the data, and warrants future investigation.

5. Conclusions

This paper presented Fusion-ConvBERT, a novel fusion network model for SER. Through a series of experiments, the proposed method demonstrated its effectiveness in correctly recognizing the emotional state of a speaker based on the utterance using spectrogram-based features. Two types of features employed here were CNN-based and the BERT-extracted features to consider both spectral-temporal correlations as well as extended temporal contextual information. In the proposed model, these two features were processed simultaneously using a parallel internal structure within the transformer architecture to extract high-level features containing different emotional details. From the ablation study, it was clearly shown that the proposed method delivers improved performance over the individual CNNs, BERT, and transformer models. After verifying the effectiveness of the fusion structure, the model was finalized for performance evaluation. The experiment showed that Fusion-ConvBERT can effectively mine emotional information from spectral and temporal features via an end-to-end technique, and its performance on the EMO-DB and IMEOCAP datasets are superior compared to the state-of-the-art techniques. Fusion-ConvBERT has shown high performance, but one of the areas to be improved for future work is the model complexity resulting in a large number of weight that requires additional computational resources.We will explore simpler structures for combining two different features while retaining the effectiveness of the proposed approach. Additionally, we will investigate improving the model performance in the Boredom class. To improve the Boredom class, we will introduce binary classification to further refine the emotion class and experiment. We expect to achieve real-time speech emotion recognition for human-machine interaction.

Author Contributions: Conceptualization, S.L.; methodology, S.L.; software, S.L.; validation, S.L.; formal analysis, S.L.; investigation, S.L.; resources, S.L.; data curation, S.L.; writing, original draft preparation, S.L. and D.K.H.; writing, review, and editing, S.L. and H.K.; visualization, S.L.; supervision, H.K.; project administration, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a National Research Foundation (NRF) grant funded by the MSIP of Korea (2019R1A2C2009480).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gupta, R.; Malandrakis, N.; Xiao, B.; Guha, T.; Van Segbroeck, M.; Black, M.; Potamianos, A.; Narayanan, S. Multimodal prediction of affective dimensions and depression in human-computer interactions. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Orlando, FL, USA, 3–7 November 2014; pp. 33–40.
2. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **2018**, *61*, 90–99.
3. Wu, C.H.; Lin, J.C.; Wei, W.L. Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course. *IEEE Trans. Multimed.* **2013**, *15*, 1880–1895.
4. Lin, J.C.; Wu, C.H.; Wei, W.L. Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition. *IEEE Trans. Multimed.* **2011**, *14*, 142–156.
5. Altrov, R.; Pajupuu, H. The influence of language and culture on the understanding of vocal emotions. *J. Est. Finno Ugric Linguist.* **2015**, *6*, 11–48.
6. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587.
7. Tzinis, E.; Potamianos, A. Segment-based speech emotion recognition using recurrent neural networks. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; pp. 190–195.
8. Anand, N.; Verma, P. Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data. In *Technical Report*; Stanford University: Stanford, CA, USA, 2015.

9. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
10. Zhu, L.; Chen, L.; Zhao, D.; Zhou, J.; Zhang, W. Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. *Sensors* **2017**, *17*, 1694.
11. Kim, Y.; Lee, H.; Provost, E.M. Deep learning for robust feature generation in audiovisual emotion recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 3687–3691.
12. Stuhlsatz, A.; Meyer, C.; Eyben, F.; Zielke, T.; Meier, G.; Schuller, B. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 5688–5691.
13. Huang, Z.; Dong, M.; Mao, Q.; Zhan, Y. Speech emotion recognition using CNN. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 801–804.
14. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213.
15. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444.
16. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
18. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
19. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. Available online: https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/languageunsupervised/languageunderstandingpaper.pdf (accessed on 23 November 2020).
20. Ling, S.; Liu, Y.; Salazar, J.; Kirchhoff, K. Deep contextualized acoustic representations for semi-supervised speech recognition. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6429–6433.
21. Liu, A.T.; Yang, S.W.; Chi, P.H.; Hsu, P.C.; Lee, H.Y. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6419–6423.
22. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. Wav2vec: Unsupervised pre-training for speech recognition. *arXiv* **2019**, arXiv:1904.05862.
23. Gideon, J.; Khorram, S.; Aldeneh, Z.; Dimitriadis, D.; Provost, E.M. Progressive neural networks for transfer learning in emotion recognition. *arXiv* **2017**, arXiv:1706.03256.
24. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335.
25. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
26. Nwe, T.L.; Foo, S.W.; De Silva, L.C. Speech emotion recognition using hidden Markov models. *Speech Commun.* **2003**, *41*, 603–623.
27. Schuller, B.; Rigoll, G.; Lang, M. Hidden Markov model-based speech emotion recognition. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), Hong Kong, China, 6–10 April 2003; Volume 2, pp. 1–4.
28. Ververidis, D.; Kotropoulos, C. Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005; pp. 1500–1503.

29. Schuller, B.; Rigoll, G.; Lang, M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; Volume 1, pp. 577–580.
30. Seehapoch, T.; Wongthanavasu, S. Speech emotion recognition using support vector machines. In Proceedings of the 2013 5th international conference on Knowledge and smart technology (KST), Chonburi, Thailand, 31 January–1 February 2013; pp. 86–91.
31. Nalini, N.; Palanivel, S. Music emotion recognition: The combined evidence of MFCC and residual phase. *Egypt. Inform. J.* **2016**, *17*, 1–10.
32. Wen, G.; Li, H.; Huang, J.; Li, D.; Xun, E. Random deep belief networks for recognizing emotions from speech signals. *Comput. Intell. Neurosci.* **2017**, *2017*.
33. Khamparia, A.; Gupta, D.; Nguyen, N.G.; Khanna, A.; Pandey, B.; Tiwari, P. Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access* **2019**, *7*, 7717–7727.
34. Huang, J.; Chen, B.; Yao, B.; He, W. ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network. *IEEE Access* **2019**, *7*, 92871–92880.
35. Cummins, N.; Amiriparian, S.; Hagerer, G.; Batliner, A.; Steidl, S.; Schuller, B.W. An image-based deep spectrum feature representation for the recognition of emotional speech. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 478–484.
36. Ocquaye, E.N.N.; Mao, Q.; Song, H.; Xu, G.; Xue, Y. Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition. *IEEE Access* **2019**, *7*, 93847–93857.
37. Keren, G.; Schuller, B. Convolutional RNN: An enhanced model for extracting features from sequential data. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 3412–3419.
38. Meng, H.; Yan, T.; Yuan, F.; Wei, H. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access* **2019**, *7*, 125868–125881.
39. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
40. Jiang, W.; Wang, Z.; Jin, J.S.; Han, X.; Li, C. Speech emotion recognition with heterogeneous feature unification of deep neural network. *Sensors* **2019**, *19*, 2730.
41. Guo, L.; Wang, L.; Dang, J.; Liu, Z.; Guan, H. Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine. *IEEE Access* **2019**, *7*, 75798–75809.
42. Lim, W.; Jang, D.; Lee, T. Speech emotion recognition using convolutional and recurrent neural networks. In Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea, 13–15 December 2016; pp. 1–4.
43. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. Librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; Volume 8, pp. 18–25.
44. Sejdić, E.; Djurović, I.; Jiang, J. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digit. Signal Process.* **2009**, *19*, 153–183.
45. Douglas, O.; Shaughnessy, O. *Speech Communications: Human and Machine*; IEEE Press: New York, NY, USA, 2000; pp. 367–433.
46. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
47. Xie, C.; Tan, M.; Gong, B.; Yuille, A.; Le, Q.V. Smooth adversarial training. *arXiv* **2020**, arXiv:2006.14536.
48. Palaz, D.; Collobert, R.; Doss, M.M. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *arXiv* **2013**, arXiv:1304.1018.
49. Behnke, S. Discovering hierarchical speech features using convolutional non-negative matrix factorization. In Proceedings of the International Joint Conference on Neural Networks, Portland, OR, USA, 20–24 July 2003; Volume 4, pp. 2758–2763.
50. Xia, R.; Liu, Y. DBN-ivector Framework for Acoustic Emotion Recognition. In Proceedings of the INTERSPEECH, San Francisco, CA, USA, 8–12 September 2016; pp. 480–484.

51. Zhao, Z.; Bao, Z.; Zhang, Z.; Cummins, N.; Wang, H.; Schuller, B.W. Attention-Enhanced Connectionist Temporal Classification for Discrete Speech Emotion Recognition. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 206–210.
52. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.
53. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
54. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
55. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779.
56. Badshah, A.M.; Rahim, N.; Ullah, N.; Ahmad, J.; Muhammad, K.; Lee, M.Y.; Kwon, S.; Baik, S.W. Deep features-based speech emotion recognition for smart affective services. *Multimed. Tools Appl.* **2019**, *78*, 5571–5589.
57. Jiang, P.; Fu, H.; Tao, H.; Lei, P.; Zhao, L. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE Access* **2019**, *7*, 90368–90377.
58. Zheng, W.; Yu, J.; Zou, Y. An experimental study of speech emotion recognition based on deep convolutional neural networks. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 827–831.
59. Luo, D.; Zou, Y.; Huang, D. Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018; pp. 152–156.