IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 October 2022, accepted 16 November 2022, date of publication 21 November 2022,
date of current version 28 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3223705

RESEARCH ARTICLE

Attention-Based Multi-Learning Approach for Speech Emotion Recognition With Dilated Convolution

SAMUEL KAKUBA[1,2], ALWIN POULOSE[3], AND DONG SEOG HAN[4], (Senior Member, IEEE)

[1]Graduate School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, South Korea
[2]Faculty of Engineering, Technology, Applied Design and Fine Art, Kabale University, Kabale, Uganda
[3]Department of Electrical and Computer Engineering, University of Michigan, Dearborn, MI 48128, USA
[4]School of Electronics Engineering, Kyungpook National University, Daegu 41566, South Korea

Corresponding author: Dong Seog Han (dshan@knu.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2021R1A6A1A03043144.

ABSTRACT The success of deep learning in speech emotion recognition has led to its application in resource-constrained devices. It has been applied in human-to-machine interaction applications like social living assistance, authentication, health monitoring and alertness systems. In order to ensure a good user experience, robust, accurate and computationally efficient deep learning models are necessary. Recurrent neural networks (RNN) like long short-term memory (LSTM), gated recurrent units (GRU) and their variants that operate sequentially are often used to learn time series sequences of the signal, analyze long-term dependencies and the contexts of the utterances in the speech signal. However, due to their sequential operation, they encounter problems in convergence and sluggish training that uses a lot of memory resources and encounters the vanishing gradient problem. In addition, they do not consider spatial cues that may exist in the speech signal. Therefore, we propose an attention-based multi-learning model (ABMD) that uses residual dilated causal convolution (RDCC) blocks and dilated convolution (DC) layers with multi-head attention. The proposed ABMD model achieves comparable performance while taking global contextualized long-term dependencies between features in a parallel manner using a large receptive field with less increase in the number of parameters compared to the number of layers and considers spatial cues among the speech features. Spectral and voice quality features extracted from the raw speech signals are used as inputs. The proposed ABMD model obtained a recognition accuracy and F1 score of 93.75% and 92.50% on the SAVEE datasets, 85.89% and 85.34% on the RAVDESS datasets and 95.93% and 95.83% on the EMODB datasets. The model's robustness in terms of the confusion ratio of the individual discrete emotions especially happiness which is often confused with emotions that belong to the same dimensional plane with it also improved when validated on the same datasets.

INDEX TERMS Emotion recognition, multi-head attention, residual dilated causal convolution.

I. INTRODUCTION

The detection and classification of emotion states from the speech signal or the extracted features has been applied in a number of affective computing tasks that involve human-computer interactions. The affective computing task of this nature is called speech emotion recognition (SER). The advent of deep learning and the development of miniaturized technologies has seen the application of SER in social living assistance, health monitoring, authentication systems, smart homes and interactive robots. Resource-constrained devices that can be used ubiquitously are used to foster the SER

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

technology in these applications. Recently, SER systems were proposed for remote monitoring of senior citizens at smart homes [1], [2]. The progressive improvement of the performance of these systems using deep learning techniques is one of the major reasons for the enhancement of these systems. Moreover, endeavors have continuously been made by researchers to ensure that emotion recognition is transformed from laboratory experiments to real-life situations [3]. This has seen emotional artificial intelligence being incorporated in social robots [4] that are used in a number of human-machine-interaction tasks. They are also used in senior citizens' home control systems [1] to monitor their emotional states, health and wellness. This shows that SER systems are increasingly becoming important in the well-being of human life and hence the need for deep learning-based models that foster efficient, robust and computationally inexpensive emotion recognition.

Chatterjee et al. [2] asserted that the deep learning-based SER systems that use smart home assistants in an ubiquitous manner need to be computationally efficient and robust since they are resource constrained. Moreover, deep learning systems also face resource constraints during training and may slowly converge. The SER deep learning-based models also suffer from the vanishing gradient problem since the task often necessitates deep models which have to be trained for a long period of time with sluggish convergence. To alleviate this, the long short-term memory (LSTM) was proposed. However, the LSTM and its variants are computationally expensive and require a longer training time to arrive at convergence. This is due to the sequential operation that is characterized by moving from one cell state to another during training. Though it is very good at training and detecting temporal cues, LSTM does not consider spatial cues which may exist in speech features that may further help the model in learning how to infer emotions. The works in [5] and [6] proposed local and global feature learning strategies to improve the learning of spatial and temporal cues by the SER models. Mustaqeem et al. [6] which was inspired by the work in [5] considered learning both the spatial and temporal cues from raw signals using convolutional LSTM for local feature learning and the gated recurrent unit (GRU) for global feature learning. Though this system gives a good accuracy the confusion ratio of the high arousal emotions like happy with those of the same dimensional plane still needs improvement.

In this paper, we propose an attention-based multi-learning model (ABMD) that uses residual dilated causal convolutions blocks and dilated convolution layers with multi-head attention to achieve comparable performance while taking global contextualized dependencies between features in a parallel manner. The model uses spectral and voice quality features extracted from the raw speech signals as input since models that use raw signals tend to confuse happy and angry emotions or neutral and other emotions. It should be noted that this confusion happens because the two discrete emotions being confused by the model belong to the same emotional dimensional plane [7], [8]. This is also because it may require computationally efficient models that will aid efficient signal processing to enable them to overcome the said prediction error and be robust enough to be deployed in real-life scenarios. The contribution of this paper is threefold;

- We propose the ABMD model that uses residual dilated causal convolution blocks, dilated convolution layers and multi-head attention for speech emotion recognition in the North American and British English and German languages using spectral and voice quality features. We use the RAVDESS, SAVEE and EMODB datasets for the respective languages.
- We present a comparison between the proposed ABMD model and our previous model that uses residual bidirectional LSTM (RBLSTMA) [9] instead of the residual dilated causal convolution blocks used in the proposed model.
- The performance of the proposed ABMD model is also compared with existing approaches that are trained in an end-to-end approach.

The rest of the paper is organized as follows: related work is explained in Section II and the methods in Section III. We describe the experiments in Section IV. The results and discussion are presented in Section V. We finally conclude in Section VI.

II. RELATED WORK

The de facto deep learning method used in most speech recognition studies is the use of recurrent neural networks (RNNs) like LSTM [10]. This is due to the need of tracking sequential time steps and long-term dependencies between the features in time series studies which RNNs are good at. Recently, different attention mechanisms [11], [12], [13] have been deployed in combination with RNNs in order to take into consideration the context of the inputs and feature representations. However, some of these depend on the sequential operation of the RNNs which increases the need for resources yet the devices are resource constrained. In order to consider long-term dependencies and context in speech for emotional state prediction, the models proposed in [14], [15], [16], and [17] use attention mechanisms in combination with either LSTM or bidirectional LSTM (BiLSTM) [18]. In [19] the bidirectional gated recurrent unit (BiGRU) is used to model the long-term dependencies for SER. Particularly authors of [16] and [17] improve the performance of SER systems by considering spatial cues through the use of convolution layers in combination with recurrent neural networks. However, as suggested in [20] though RNNs achieve promising results, they encounter problems in convergence, sluggish training that uses a lot of memory resources due to the sequential manner in which they operate. Moreover, they do not consider spatial relationships between speech features.

In order to solve the challenges of RNNs, researchers suggested solutions that could achieve similar or better performance with fewer challenges. Oord et al. [21] proposed

a generative model (WaveNet) for text-to-speech translation using dilated causal convolution layers which used fewer parameters in a deep neural network to achieve promising results. This architecture grows the receptive field exponentially with network depth covering many steps. The use of such light models improves the receptive field with less increase in the number of parameters compared to the number of layers involved yet taking into consideration the long-term dependencies, alleviating the vanishing gradient problem and hence eliminating the need for RNNs. This is because dilated causal convolutions process temporal sequential batches with no parametric increase. A number of researchers in the SER domain have used dilated convolution layers as an alternative instead of recurrent neural networks and achieved commendable performance. Pandey et al. [22] proposed a model that uses Oord et al.'s WaveNet with modifications for SER and compared its performance with the model that uses one-dimensional convolution neural networks and LSTM. Also, in [23], a dilated convolutional neural network was used in an end-to-end model for speech emotion recognition. Promising results were also obtained in both classification of discrete emotions and regression tasks for the dimensional emotions of arousal and valence. Based on the need for computationally efficient and robust SER systems, a model that uses dilated convolution with a multi-learning approach was suggested in [24]. This model takes into consideration the relationship between emotional cues, sequential learning, and long-term dependencies.

Most of these models use raw signals as input and learn emotional cues and their relationships during training. It is however, observed from the results in [22] and [24] that these models find difficulties in discriminating high arousal emotions especially happy and angry during the classification tasks. They sometimes also exhibit prediction errors in discerning between discrete emotions that belong to the low arousal plane. This affects their performance especially in discerning these particular discrete emotions. Moreover, these models don't consider global contextualized dependencies in a parallel manner which can further reduce the computational cost while improving the performance of the model.

As mentioned earlier, researchers implement SER models with attention mechanisms to improve their performance. Attention mechanisms enable the model to consider the context and significance of a given input feature in relation to others in the sequence. If attention mechanisms are implemented, they improve the performance of deep learning-based speech emotion recognition systems [25]. Among the attention mechanisms mentioned earlier, multi-head attention used in [13] considers global contextual relationships among features in a parallel manner. This improves the robustness of the models with less parametric increase as compared to the other kinds of attention mechanisms. In the bid to alleviate the loss of temporal cues in SER systems [26] used multi-head attention in combination with a dilated residual network. However, they also used the LSTM layers which increased the number of parameters and therefore introduced high complexity of the network. In [27] multi-head attention is employed in combination with focus and calibration attention mechanisms in order to align the signal amplitudes with the relevant emotion regions. In our previous work in which we proposed a residual BiLSTM with multi-head attention (RBLSTMA) model [9], we also used the residual BiLSTM which operates sequentially in combination with multi-head attention mechanism for SER. Multi-head and self-attention mechanisms which are similar in operation have also been applied in multi-task SER studies that involve auxiliary tasks like gender classification [17], [28]. Due to its powerful contribution multi-head attention has also been applied in multimodal SER studies that involve facial, physiological and lexical unaligned multimodality emotional cues [29]. This attention mechanism has also been applied in studies that involve lexical and acoustic multimodality features [30], [31], [32] and acoustic and video features [33].

Over the years, progressive research in SER has provided models that exhibit a good performance with some robust and convergence issues that we seek to solve in this paper. The existing deep learning models are mostly end-to-end that involve both feature extraction and classification. Sun et al. [34] proposed an end-to-end model that auxiliary learns gender information in addition to emotion cues from raw signals for SER. In recent studies, transformer-based SER models have been proposed. Andayani et al. [35] proposed a hybrid model that replaced the position encodings of the transformer encoder with LSTM in order to learn contextualized long-term dependencies for emotion recognition. Pre-trained models and data augmentation techniques have also been used to improve SER performance in recent research. In [36], a CNN-based model that employs data augmentation of a combination of zero crossing rate (ZCR), mel spectrograms, mel frequency cepstral coefficients (MFCCs), and chroma grams was proposed. Padi et al. [37] proposed to augment a combination of spectral and time domain features using multi-window data augmentation instead of single window approach with a CNN model for SER. Al-onazi et al. [38] also proposed to augment a combination of first and second delta MFCCs, chroma grams, tonnetz, and spectral contrast which were used as input to a transformer-based SER model. In [39], harmonic and percussive components were extracted from the mel spectrograms and later concatenated with the mel spectrograms to form augmented input to a pre-trained VGG16 model for SER. They also experimented with a combination of MFCCs, mel spectrograms and chroma grams. The pre-trained VGG16 was also used for feature extraction in models proposed in [40] for SER. The pre-trained AlexNet model is also used for feature extraction in combination with correlation-based feature selection (CFS) and dense layers for speaker-dependent and independent emotion recognition in [41]. Nevertheless, it should be noted that there is a scarcity of SER pre-trained models. The available pre-trained models were primarily trained on general deep learning tasks that involve a wide range of object classification among other

tasks. Data augmentation involves changes in pitch, loudness, amplitude, and addition of noise in the audio signal or extracted acoustic features which tampers with the natural emotion clues that would otherwise give an even better performance. Therefore, with these challenges, there is a need for models that use the small datasets available for training without distortion, and solve the overfitting and vanishing gradient problems through fast convergence.

In this paper, we used residual dilated causal convolution blocks and dilated convolutions in a multi-learning approach in order to achieve what the LSTM and/or its variants have been used for in previous studies. We boost the fast training and convergence in both branches of learning with multi-head attention blocks to benefit from the parallel global contextualization of feature representations.

III. METHODS

A. THE SER SYSTEM OVERVIEW

The general end-to-end deep learning-based SER system consists of three major processes as shown in Fig. 1. It consists of data processing, feature extraction, feature learning and classification. Data processing depends on the nature of the speech signal. Generally, it involves ensuring equal sequence length of all the speech signals in the dataset which is the input to the model. The speech signals that are of shorter length than the required are padded and the longer ones are truncated. The data processing stage also involves removal of the silent regions since they do not carry any emotional cues that are useful to the model. Preemphasis and bandpass filters can also be used to allow only the frequencies of the speech signal that are considered to have pertinent cues for emotion recognition. To speed up the fast fourier transform process and avoid spectral leakage framing and windowing are normally done. We particularly used the hamming window function for windowing. The frames are overlapped after windowing to avoid loss of signal information. After data processing, the efficiency and robustness of SER models depend on two stages; feature extraction and emotion classification. Robust feature extraction enables the model to learn characteristic features that exist in the speech signal and can help the model distinguish between emotions like pitch, loudness, vocal tract and formant frequencies. It involves the extraction of low-level descriptors of the speech signal that can depict the emotional cues. The classification process involves the analysis of the global low-level features to discern between discrete emotions. In deep learning-based SER models, local and global features are learned and classified in an end-to-end approach.

The proposed SER system follows the above described overview to enhance robust and efficient emotion recognition. The proposed model exhibits fast convergence and training that improves local and global feature learning with minimal resources. The model learns the spatial and temporal cues of emotions with a small number of layers that eventually reduces its number of parameters making it possible to be deployed in resource constrained applications. The proposed model consists of two branches that are dedicated to extraction of spatial and temporal cues simultaneously. One branch is responsible for learning the long-term dependencies of especially temporal cues using the residual dilated causal convolution blocks (RDCC). The second branch uses a block of two dilated convolution (DC) layers to extract spatial cues that may exist among the speech features. The learned latent representations from both branches are separately passed through multi-head attention mechanism blocks before being concatenated for emotion classification using a dense layer and softmax activation function.

B. OUR PREVIOUS RBLSTMA MODEL

As shown in Fig. 2, our previous RBLSTMA model [9] uses the residual BiLSTM and the multi-head attention mechanism to take into consideration the long-term dependencies and pays attention to the most important features in the sequence while preserving the context in which they were uttered. In order to solve the vanishing gradient problem, this model uses two residual blocks of 3 stacked bidirectional layers of 128 units each, with a skip connection that provides an identity mapping of the previous layer to the next layer. The second part of the model is a multi-head attention mechanism block of eight heads with 2 layers each. Similar to the proposed ABMD model, the RBLSTMA uses the multi-head attention network to consider the context of each utterance feature in relation to others in the sequence which helps the model to understand the semantics cues together with the paralinguistic cues of speech. The inclusion of the multi-head attention together with residual BiLSTM in this models ensures long and short-term dependency among features or their latent representations coupled with control of vanishing gradient with skip connections which improves the SER system performance.

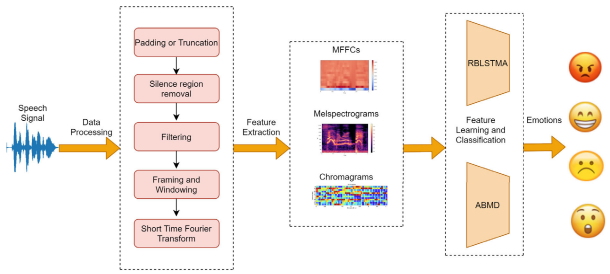

FIGURE 1. The Speech Emotion Recognition (SER) system overview.

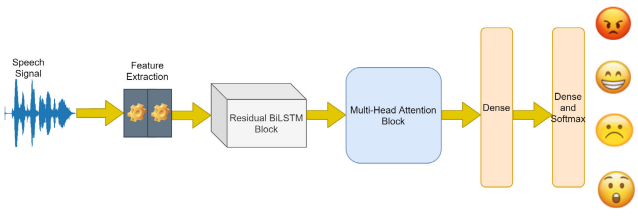

FIGURE 2. A framework of the Residual Bidirectional LSTM with Multi-Head Attention model (RBLSTMA).

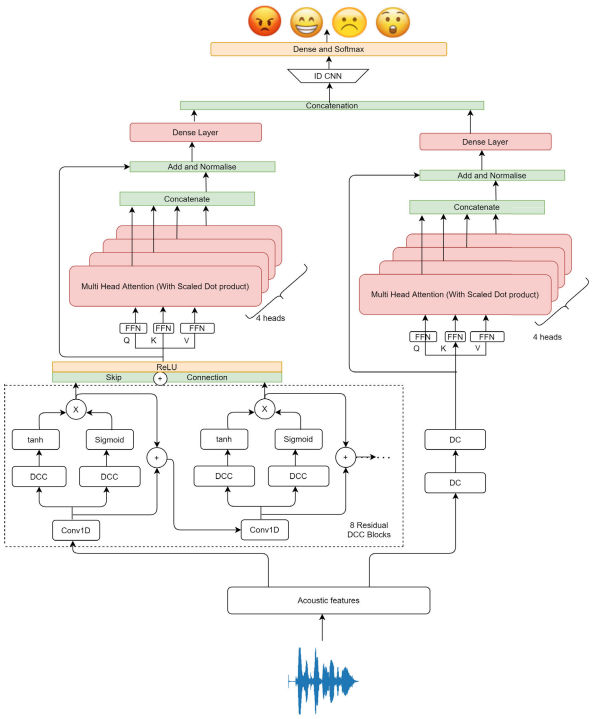

FIGURE 3. A framework of the proposed Attention-based Multi-Learning model (ABMD) that uses dilated causal convolution and multi-head attention.

In the previous work, the RBLSTMA was validated on a slightly different set of emotions however, to favorably compare its performance with the proposed ABMD model, both of them were subjected to the same emotion categories of happy, sad, angry and neutral. Though the RBLSTMA model obtains a good performance it uses the BiLSTM whose operation is sequential in a deep network with many layers that increase the number of parameters and takes a longer training time to arrive at convergence. In addition, the RBLSTMA has a weakness of confusing the emotions that belong the to the same emotional dimensional plane especially happy and angry. Moreover, for the datasets where it performs well in classifying the happy emotion, there is an uneven distribution of the model's performance on the other emotions in terms of the confusion ratio.

C. THE PROPOSED ABMD MODEL

As shown in Fig. 3, the proposed ABMD model for speech emotion recognition uses residual dilated causal convolution blocks, dilated convolution layers and multi-head attention mechanism. The main objective of this model is to learn contextualized speech emotion cues with long-term dependencies from spectral and voice quality features of a speech signal using residual dilated causal convolution blocks with gated activation, dilated convolution layers and multi-head attention in a multi-learning approach without the use of LSTM or its variants. The proposed model consists of two branches that learn emotional cues from the features. Besides Fig. 3, Algorithm 1 provides further information about the multi-learning procedure of the proposed model.

Algorithm 1 Attention-Based Multi-Learning Model (ABMD) for SER

Input: Speech signal
Extract MFCCs, Mel spectrograms, Chroma grams
Compute the mean of each feature
Concatenate all the features
Branch 1:
a) RDCC, number of blocks = 8
b) Aggregate with skip connections
c) Apply multi head attention, number of heads = 4
Branch 2:
a) Dilated layer, dilation rate = 1
b) Dilated layer, dilation rate = 2
c) Apply multi head attention, number of heads = 4
Concatenate the feature representations
One-dimensional Conv1D layer
Fully connected layer
Output: Softmax function

One multi-learning branch of the proposed model considers long-term dependencies using eight residual dilated causal convolution blocks that are aggregated with skip connections before using four multi-head attention blocks that allow the model to pay global attention to the most relevant feature representations in the sequence while preserving the utterance level context. The residual dilated causal convolution block consists of gated activation units with dilated causal convolution layers. In these blocks, the hyperbolic tangent function is referred to as a filter and the sigmoid as a gate. This output of the residual dilated causal convolution blocks is fed into the multi-head attention mechanism (MHA) after being aggregated by a skip connection.

The other branch consists of a block of two dilated convolution layers of dilation rates 1 and 2 and four multi-head attention heads. The two dilated convolution layers were for extracting spatial cues from the features in this branch of the model.

The contextualized feature representations with long-term dependencies that are extracted using these two branches are concatenated and passed through a one-dimensional convolution layer, a fully connected layer to further learn the relationship between the feature representations before finally feeding them into a dense layer that uses the softmax as the activation function to compute the likelihood of the input sequence belonging to a given discrete emotion class. However, it should be noted that we used a variable number of dilation rates, number of heads, and blocks in the experiments. However, an increase in the blocks, heads, and dilation rates increases the complexity of the model which leads to overfitting and vanishing gradient problems. As an example, an increase of the number of heads from 4 to 8 alone, with a similar number of blocks and dilation rates as reported in this paper increases the number of learnable parameters from 5,600,544 to 6,641,644 with no significant improvement in

the performance. On the other hand, a decrease in the number of heads, blocks, and dilation rates results in a model that is robust on the angry and sad emotions but not the happy and neutral emotions which are often confused because of their dimensional positions in the emotion plane.

The main components of the proposed model are discussed in detail in this section.

1) RESIDUAL DILATED CAUSAL CONVOLUTION (RDCC)

In this part of the model, the dilated causal convolution (DCC) layers in residual blocks are used to provide temporal sequence modeling while covering a large receptive field. Dilated convolutions spread causal filters by skipping values in the input sequence in specified predetermined steps. Since a stack of dilated convolution layers provides a large receptive field with a few layers, we use eight residual blocks, exponentially increasing their receptive fields by increasing the dilation rates from 2^1 to 2^4 twice. In so doing the filters are applied over a large receptive area without an increase in the kernel inputs and number of parameters. This, therefore, ensures temporal dependencies and therefore lowers the need for LSTM. In order to ensure a similar gated operation of the hidden states in RNNs, we use the gated activation units (GAU) [21]. The GAU consists of dilated convolutions with the hyperbolic tangent activation function (tanh) and the sigmoid. The tanh is used as a filter (f) that chooses the outputs of the unit while the sigmoid is a gate (g) that emphasizes the significance of the output of the tanh filter. Mathematically GAU is given by [22]

$$z = tanh(W_{f,s} * u) \odot \sigma(W_{g,s} * u) \tag{1}$$

where * denotes element wise multiplication, s is the layer index and $W_{f,s}$ and $W_{g,s}$ denote the learnable parameters for convolution filter (f) and gate (g). u denotes the input sequence, σ is the sigmoid activation function.

The outputs of the residual dilated causal convolution blocks are added together using skip connections. The RDCC and the skip connections solve the convergence and training speed problem as well as the avoiding the vanishing gradient problem. It should be noted that as one branch concentrates on long-term dependencies of especially temporal cues using the residual dilated causal convolution blocks, the second branch uses a block of two dilated convolution (DC) layers to extract spatial cues that may exist among the speech features.

2) MULTI-HEAD ATTENTION MECHANISM (MHA)

The other part of the model that is used in both branches to aid contextualized multi-learning is the multi-head attention mechanism block of four heads. Multi-head attention mechanism operates the four self-attention heads in parallel which improves its performance, reduces complexity and training time. The position encoding is handled by the use of the triangle positional embedding used in [13]. The sinusoidal function is used for the even positions while the cosine function is used for the odd positions. The encoding result is passed through linear models to obtain the query Q, key K, and value V that would later be used in the scaled dot product computation to obtain similarities between features. For linear projection, we used feed-forward neural networks (FFN). The multi-head attention block in both branches helps the model to consider the global context of each utterance feature in relation to the others. This further helps the model to learn the relevant cues that are most important for a particular emotion class and in so doing solve the problem of discriminating between emotions that belong to the same arousal or valence plane.

Multi-head attention mechanism derives attention weights of current inputs in relation to all other inputs in the sequence as 2 [13]

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

where query, Q is $W^q x_i$, key, K is $W^k x_i$ and value, V is $W^v x_i$. x_i is a word or feature at position i. d_k denotes the feature dimension which ensures that the dot product of the query and key do not grow too large.

IV. EXPERIMENTS

In this section, we present the experiments carried out to validate the model on three datasets. To carry out the experiments, we used Keras 2.8.0 API, TensorFlow 2.6 as the back-end with python programming, and Nvidia GeForce RTX 2080 super graphics processing unit (GPU). We used an initial learning rate of 0.0005 and varying batch sizes depending on the datasets with the Adam optimizer. The range of values used for tuning the hyper-parameters and other parameters in the experiments are shown in Table 1. For each corpus, the available number of samples was split into train, validation, and test portions with a ratio of 80:10:10. For all experiments drop out, regularization and layer normalization are configured to overcome overfitting. In all the experiments, the model was trained for a minimum of 30 epochs and a maximum of 100 epochs. We also used similar hyperparameters where possible to carry out experiments on residual BiLSTM with multi-head attention model in order to compare the proposed model with others that use LSTM or its variants. The following sections describe the datasets and features used in the experiments to validate the model. In datasets where there was an imbalance in the number of samples we we configured appropriate class weights.

A. DATASETS

To study the performance of the proposed model and comparison with other models, we used the Berlin database of emotional speech (EMODB) [42] for the German language, Ryerson audio-visual database of emotional speech and song (RAVDESS) [43] and surrey audio-visual expressed emotion (SAVEE) [44] for the English language. Since each of these datasets is of a particular language with different culture and accent, the proposed model's performance in a variety of languages, accents, and cultures is evaluated. In addition, the utterances in EMODB and RAVDESS datasets were recorded

from 10 and 24 participants respectively. Among these are an equal number of male and female participants which is substantial enough to provide the proposed model with a variety of utterances that exist in the real world. Though the utterances in the SAVEE dataset were recorded from only four male actors, they contain balanced phonemes which help the model to learn the differences among the different emotions. The EMODB dataset is a German language dataset that consists of 493 utterances from 10 actors that speak German language sentences in anger, boredom, sadness, happiness, anxiety and neutral. The RAVDESS dataset consists of speech and song files. The speech files depict calmness, disgust, happiness, sadness, fear, anger, surprise and neutral. The song files depict all except neutral, disgust and surprise. The dataset consists of 1440 files for speech and 1012 for songs. We however only use the speech files in this paper. The SAVEE dataset contains 480 utterances in British English depicting the seven basic emotions in addition to neutral from four native British English speakers. For all the datasets we consider happiness, sadness, neutral and anger as emotional states.

B. FEATURE EXTRACTION

For each of the datasets, after preprocessing the signal data as explained in Section III, we extracted spectral and voice quality features using Librosa 0.9.2 and used them as input to the proposed model. We considered features that can depict loudness, pitch and quality of sound. The spectral low-level descriptors of sound extracted from the speech signal were mel frequency cepstral coefficients (MFCCs) and chroma grams. In terms of voice quality, mel spectrograms were extracted. MFCCs are low-level descriptors of sound that describe changes per time interval of different sound spectrum bands. They depict the vocal tract frequency response of sound. They are obtained by creating triangular filters on an already constructed log mel spectrum and decorrelating the obtained filter banks using the discrete cosine transform (DCT). Mel spectrograms depict the sound quality according to the human auditory system through the multiplication of the frequency domain values on the mel scale with filter banks. They provide the model with perceptual frequency representations that are relevant for emotion recognition in speech. To present the model with changes in pitch as one speaks we extract the Chroma grams. In order to present the model with one-dimensional inputs we compute the mean of the extracted features from each frame. We carried out experiments to find out the best features to use for the proposed model on the EMODB dataset. In the first experiments we subjected the proposed model to each of the three extracted features. In the other two experiments, a combination of either MFCCs and mel spectrograms (mel) or MFCCs, chroma grams (chroma) and mel spectrograms were used as input to the model after concatenation. As observed in Table 1, the propopsed model obtains the best performance when a combination of MFCCs, chroma grams and mel spectrograms is used.

TABLE 1. Ranges of parameters used in the experiments.

Parameter	Range of Values
Optimizer	Adam
Learning rate	0.0001 - 0.0025
Batch size	8 - 16
Epochs	30-100
Number of heads	2 - 8
Number of residual blocks	2 - 10
Embedding dimension	180
FFN units	32
Dilation rate	1 - 32

V. RESULTS AND DISCUSSION

In this section, we present the results and discussion of the impact of the proposed model on the speech emotion recognition system's performance. We also discuss the significance of the residual dilated causal convolution blocks in combination with multi-head attention mechanism in a context-aware multi-learning approach for speech emotion recognition in comparison with existing approaches.

A. RESULTS

The experimental results in Table 2 show the efficiency and robustness of the proposed ABMD model when subjected to the extracted features of the EMODB dataset separately and in combination. The performance of the model is presented in terms of accuracy (A) and F1 score (F1) obtained by the model. To realize the robustness of the proposed model when subjected to the features we also present results in terms of the confusion ratio of the different classes of emotions for each input. CH, CS, CA and CN are confusion ratios for happy, sad, angry and neutral respectively. The results show that for all the inputs, the accuracy and F1 score are commendable. However, for the best robustness results in terms of the confusion ratio of high arousal dimension emotions especially happiness which is often confused with anger, a combination of all the features is the best option. This option also gives the best evenly distributed performance results across all the individual emotions.

In Table 3, the performance of the proposed ABMD model is compared with that of the RBLSTMA model in terms of accuracy (A), precision (P), recall (R), F1 score (F1) and loss. We also compare the confusion ratio (CR) of happiness as a highly confused arousal discrete emotion class in previous studies including our previous RBLSTMA model. We also present the confusion matrices obtained after testing the proposed ABMD model on unseen data in Figs. 4 (a), 4 (b) and 4 (c). The performance of the proposed ABMD model on individual classes of all the datasets used is also presented in Table 4.

The results portray a tremendous significance of the residual dilated causal convolution blocks aggregated by skip connections on one learning branch and dilated convolutions on the other, in combination with multi-head attention mechanisms for contextualized speech emotion recognition using a multi-learning approach. The results also show that the model gains spatial knowledge from the dilated con-

TABLE 2. Performance of the proposed ABMD model on extracted features.

Input	A(%)	F1(%)	CH(%)	CS(%)	CA(%)	CN(%)
Mel	85.42±1.20	84.17±1.00	**36±5.00**	94±6.00	82±4.00	86±2.00
MFCCs	84.42±2.00	85.42±1.50	**68±6.00**	100±4.00	77±3.00	92±3.00
Mel & MFCCs	81.25±3.00	81.25±2.50	**52±5.00**	94±4.00	91±2.00	92±2.00
Chroma, Mel & MFCCs	95.93±1.18	95.83±2.00	**80±5.00**	100±3.00	84±4.00	82±5.00

TABLE 3. Performance comparison of the proposed ABMD Model with the RBLSTMA model.

Model	Datasets	A(%)	P(%)	R(%)	F1(%)	CR(%)	Loss
RBLSTMA [9]	EMODB	92.73±0.50	92.73±0.20	92.73±0.20	91.73±0.10	47±9.00	0.6614±0.0923
	RAVDESS	86.50±0.20	89.40±0.02	82.82±1.00	84.47±1.10	84±4.00	0.6632±0.1000
	SAVEE	85.42±2.00	85.42±1.00	85.42±1.00	85.42±1.20	100±6.00	0.3854±0.0612
Proposed ABMD	EMODB	95.93±1.18	95.83±0.10	95.83±0.11	95.83±2.00	80±5.00	0.3402±0.0863
	RAVDESS	85.89±3.00	89.40±2.00	82.82±2.00	85.34±1.20	86±3.00	0.4041±0.1020
	SAVEE	93.75±1.50	93.62±1.00	91.67±2.00	92.50±1.20	93±2.00	0.2132±0.0774

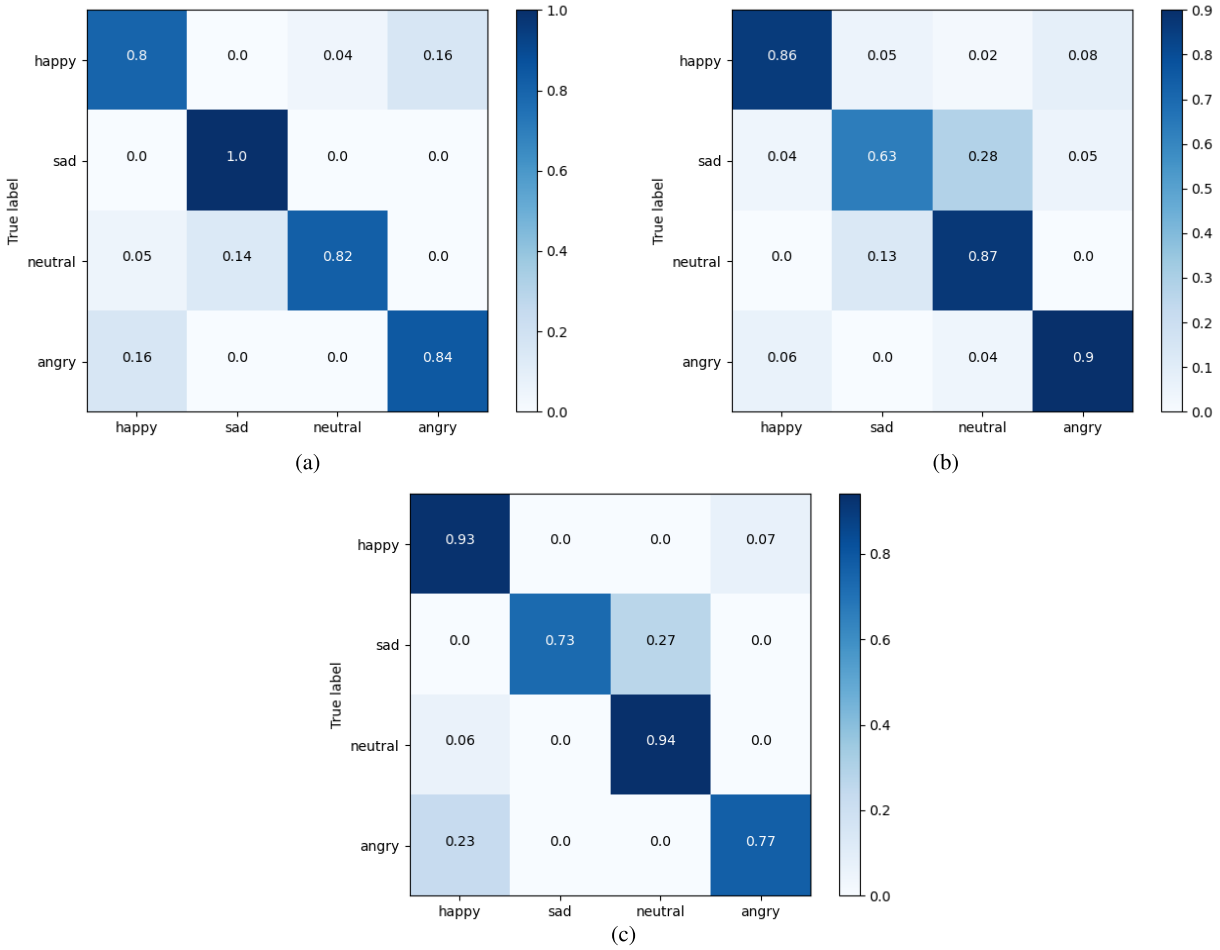

FIGURE 4. The confusion matrix results of the proposed ABMD when tested on test datasets. (a) EMODB. (b) RAVDESS. (c) SAVEE.

volution layers used in its second branch. The proposed model achieves comparable performance in both English and non-English language datasets and exhibits a reduction in the prediction error rate (confusion ratio) as observed in the confusion matrices. We observe from the confusion matrices that the model is good at discriminating emotions that belong to the same arousal dimension plane of emotions. This robustness is observed especially for anger and happiness which were our main concerns in this paper for all datasets.

Figs. 5 (a), 5 (b) and 5 (c), present the confusion matrices obtained by the RBLSTMA model on the test dataset of the EMODB, RAVDESS and SAVEE datasets. The confusion matrices show poor robustness of the RBLSTMA model for some emotions yet good for others. The discriminative robustness results of the RBLSTMA model in terms of the confusion ratio of the different individual classes of emotions for each dataset clearly show its weakness if deployed in real-life scenarios. It should be noted that this model was trained

TABLE 4. Performance of the proposed ABMD model on individual emotional classes.

Dataset	EMODB				RAVDESS				SAVEE			
Emotion	P(%)	R(%)	F1(%)	CR(%)	P(%)	R(%)	F1(%)	CR(%)	P(%)	R(%)	F1(%)	CR(%)
Happy	74±6	80±2	77±3	80±5	92±2	86±3	89±2	86±3	76±4	93±3	84±3	93±2
Sad	86±2	100±4	92±2	100±3	84±3	63±7	72±5	63±8	100±2	73±7	85±6	73±4
Neutral	95±2	82±3	88±4	82±5	59±10	87±2	70±6	87±4	81±4	94±2	87±4	94±3
Angry	89±4	84±4	86±2	84±4	85±2	90±3	88±3	90±3	91±4	77±4	83±3	77±4

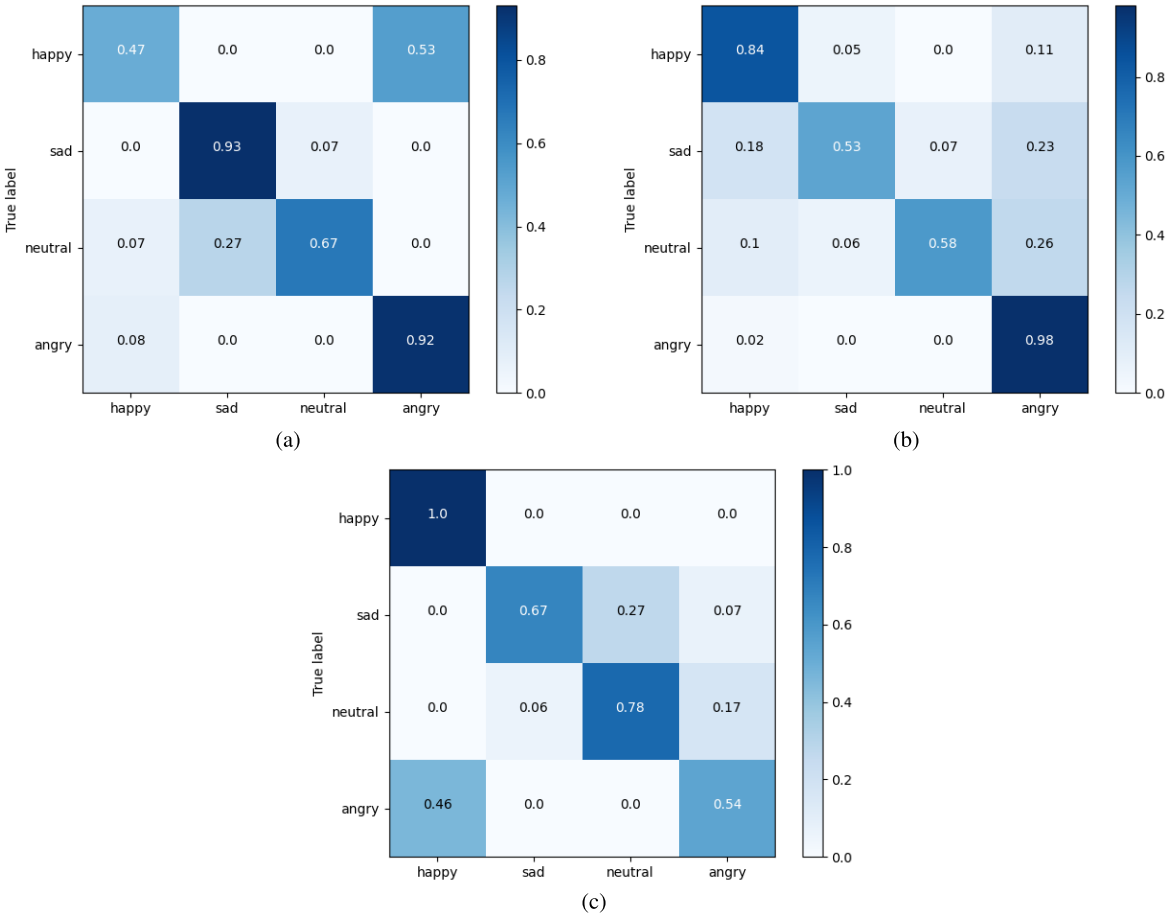

FIGURE 5. The confusion matrix results of our previous RBLSTMA model when tested on test datasets. (a) EMODB. (b) RAVDESS. (c) SAVEE.

for a minimum of 500 epochs to obtain the results that would be favorably compared with those obtained by the proposed ABMD model which was trained for a minimum of 30 and a maximum of 100 epochs faster than the training of the RBLSTMA model.

B. DISCUSSION

As observed in Figs. 4 (a), 4 (b) and 4 (c), the proposed ABMD model achieved commendably improved confusion ratio results. This is in comparison with the confusion ratio results shown in Fig. 5 (a), 5 (b) and 5 (c) for the RBLSTMA model for individual discrete emotion classes. The commendable performance of the proposed ABMD model was obtained after training for a maximum of 100 epochs compared to a minimum of 500 epochs for the RBLSTMA model which means that the proposed ABMD model exhibits faster training and convergence. These results indicate that the use of residual dilated causal convolution blocks in one branch and dilated convolutions in another with utterance contexts consideration using the multi-head attention mechanism gives accurate and robust results for SER models as compared to the use of LSTM or its variants with multi-head attention. The robustness of the proposed model is analyzed using Figs. 6 (a) and 6 (b) which show a comparison of the confusion ratio and loss of the proposed ABMD model with our previous RBLSTMA model. Particularly, Fig. 6 (a) shows that the proposed model commendably improves the prediction rate of the emotions that had a poor performance in the existing approaches while evenly predicting the others without being very robust at predicting one emotion and very poor at others. The results in Fig. 6 (b) present the reduction in loss values when the proposed ABMD model is used for

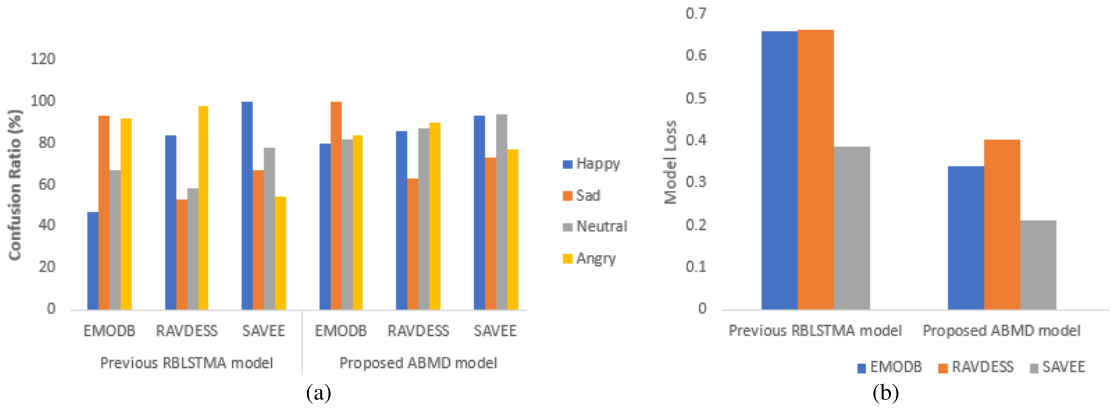

FIGURE 6. Comparative results of model loss and individual emotion class confusion ratio obtained by the RBLSTMA model and the proposed ABMD model. (a) Individual class confusion ratio. (b) Model loss.

SER as compared to our previous RBLSTMA model. It is worthy mentioning that the reduction in losses is due to the reduced number of layers and fast arrival at convergence that the proposed model exhibits.

Further more, though trained for many epochs, the RBLSTMA model is robust on the prediction of the happy emotion for the English datasets but not the German language dataset. The results show that the model is not robust enough for the emotions of sad, angry and neutral for the SAVEE dataset. It also poorly predicts sad and neutral emotion classes for the RAVDESS dataset. In addition, the previous RBLSTMA model poorly predicts happy and neutral for the EMODB dataset as well. This generally could be because of the dimension planes that the happy and neutral emotions share with the other emotions. Because of the discrepancies in the confusion ratio of individual discrete emotions achieved by the model on different datasets, it may not be able to discriminate well some emotions if deployed in real-life applications. We were therefore motivated to improve the robustness of SER systems using the proposed ABMD model.

The proposed ABMD model achieves an accuracy of 95.93% and an F1 score of 95.83% on the EMODB dataset. The minimum accuracy of 85.89% and F1 score of 85.34% are registered when the model is tested on the RAVDESS dataset. There is also an improvement in the confusion ratio for the high arousal dimensional discrete emotions especially for happiness which was highly confused with anger in the previous works. We observe the best confusion ratio for happiness in the English datasets of SAVEE and RAVDESS at 93% and 86% respectively. It is however observed in Table 3 that though the confusion ratio of happy of 80% for EMODB dataset is not as good as the English datasets, it tremendously improved from that observed in the previous RBLSTMA model by 33%. In addition, Table 4 shows that the proposed ABMD model is generally robust on the individual classes. This is shown in terms of precision, recall F1 score and individual confusion ratio.

The improved performance is due to the multi-learning nature of the proposed model with one branch extracting spatial emotion cues using dilated convolution layers and multi-head attention and the other extracting temporal cues and considering their long-term dependencies using residual dilated causal convolution blocks in combination with multi-head attention aids performance improvement. The multi-head attention mechanism also plays a very important role in this model since it operates in parallel to learn the global attention context of the feature representations from both branches. The residual dilated causal convolution blocks used in this model ensure track of time steps in the input sequence thereby learning the long-term dependencies that further improve local feature learning without a very high increase in the number of learnable parameters that would otherwise make the model more complex and resource intensive. The performance of the proposed model also shows that it is able to contextualize the tone, loudness, and pitch in spoken North American English, British English and German languages. Compared with the existing models, the proposed ABMD model enables parallelism through the use of dilated causal convolution layers, dilated convolution layers and multi-head attention to improve the receptive field with a small increase in the number of parameters compared to the number of layers there by improving the robustness of the model. On the whole, these results indicate that models that use LSTM and its variants in a single learning approach may achieve commendable accuracies but may not be as robust as necessary for the real-life scenarios or applications in which SER systems are deployed.

1) COMPARISON WITH EXISTING APPROACHES

We present the a performance evaluation of the proposed ABMD model in comparison with other recent SER approaches in terms of accuracy (A) and F1 score (F1) where available in Table 5. In comparison with the existing models, we observe that the proposed model improves the accuracy when validated on the EMODB dataset in terms of accuracy in the range of 2.53% to 12.11% and 5.83% of F1 score. A range of 8.65% to 26.85% of accuracy and 21.50% of F1 score is observed for the SAVEE dataset. A comparable performance in the range of 3.95%

TABLE 5. Performance comparison of the proposed ABMD model with the existing models.

Reference	Dataset	Input Features	Method	A(%)	F1(%)
[22]	EMODB	Raw signal	WaveNet	83.82	-
[6]	RAVDESS	Raw signal	Hierarchical ConvLSTM	80.0	80.0
[35]	EMODB	Raw signal	Residual CNN	90.3	-
[42]	EMODB	Spectrograms	AlexNet + CFS + FC	90.50	-
	RAVDESS	Spectrograms	AlexNet + CFS + FC	73.50	-
	SAVEE	Spectrograms	AlexNet + CFS + FC	66.90	-
[24]	EMODB	Raw signal	Multi-learning 1D Dilated CNN	90.0	90.0
[37]	RAVDESS	ZCR, MFCCs, Mel spectrograms, Chroma grams	CNN with data augmentation	89.0	-
[41]	RAVDESS	Mel spectrograms	VGG16	81.94	-
[36]	EMODB	MFCCs	Hybrid LSTM-Transformer encoder	85.55	-
	RAVDESS	MFCCs	Hybrid LSTM-Transformer encoder	75.62	-
[38]	RAVDESS	Time domain and spectral features	Multi-window augmentation + CNN	88	88
	SAVEE	Time domain and spectral features	Multi-window augmentation + CNN	70	71
[40]	EMODB	MFCCs, Mel spectrograms, Chroma grams	VGG16 + FC	88.39	-
	EMODB	Mel spectrograms, Harmonic and Percussive cues	VGG16 + FC	92.79	-
[39]	EMODB	Spectral Features	Transformer Encoder with augmentation	93.4	-
	SAVEE	Spectral Features	Transformer Encoder with augmentation	85.1	-
Proposed ABMD model	EMODB	MFCCs, Mel spectrograms, Chroma grams	RDCC + DC + Multi-head attention	**95.93**	**95.83**
	RAVDESS	MFCCs, Mel spectrograms, Chroma grams	RDCC + DC + Multi-head attention	**85.89**	**85.34**
	SAVEE	MFCCs, Mel spectrograms, Chroma grams	RDCC + DC + Multi-head attention	**93.75**	**92.50**

to 12.39% and 5.34% of accuracy and F1 score respectively is also observed for the RAVDESS dataset. The performance comparison with the existing approaches further shows that the parallelism operation through the use of dilated causal convolution layers, dilated convolution layers and multi-head attention for long term dependencies and context computation improves the performance of SER models significantly.

VI. CONCLUSION

In this paper, we proposed an attention-based multi-learning approach for speech emotion recognition that uses residual dilated causal convolution blocks, dilated convolution layers and multi-head attention. The proposed ABMD model consists of two branches that learn emotional cues from the extracted features in a multi-learning approach. It learns contextualized speech emotion cues with long-term dependencies from spectral and voice quality features of a speech signal using residual dilated causal convolution blocks with gated activation and multi-head attention. The model also learns spatial cues using the dilated convolution layers. The model trains for a maximum of 100 epochs to arrive at convergence with minimum model loss which enables it to be deployed in resource-constrained devices. It is also observed that the model discriminates well the discrete emotions in the same arousal plane. However, it is important to investigate possible methods of improving the robustness in multi-modal SER models and cross-lingual learning.

REFERENCES

[1] X. Wu and Q. Zhang, "Intelligent aging home control method and system for Internet of Things emotion recognition," *Frontiers Psychol.*, vol. 13, May 2022, Art. no. 882699.
[2] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-time speech emotion analysis for smart home assistants," *IEEE Trans. Consum. Electron.*, vol. 67, no. 1, pp. 68–76, Feb. 2021.
[3] S. Saganowski, "Bringing emotion recognition out of the lab into real life: Recent advances in sensors and machine learning," *Electronics*, vol. 11, no. 3, p. 496, Feb. 2022.
[4] H. Abdollahi, M. Mahoor, R. Zandie, J. Sewierski, and S. Qualls, "Artificial emotional intelligence in socially assistive robots for older adults: A pilot study," *IEEE Trans. Affect. Comput.*, early access, Jan. 18, 2022, doi: 10.1109/TAFFC.2022.3143803.
[5] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019.
[6] S. Kwon, "CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network," *Mathematics*, vol. 8, no. 12, p. 2133, Nov. 2020.
[7] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots," *Int. J. Social Robot.*, vol. 11, no. 4, pp. 555–573, 2019.
[8] G. K. Verma and U. S. Tiwary, "Affect representation and recognition in 3D continuous valence–arousal–dominance space," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2159–2183, Jan. 2017.
[9] S. Kakuba and H. Dong, "Residual bidirectional LSTM with multi-head attention for speech emotion recognition," in *Proc. Korea Commun. Assoc. Summer Gen. Academic Conf.*, 2022, pp. 1419–1421.
[10] S. Hochreiter, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[12] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[14] Y. Yu and Y.-J. Kim, "Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database," *Electronics*, vol. 9, no. 5, p. 713, Apr. 2020.
[15] Y. Xie, R. Liang, Z. Liang, and L. Zhao, "Attention-based dense LSTM for speech emotion recognition," *IEICE Trans. Inf. Syst.*, vol. 102, no. 7, pp. 1426–1429, 2019.
[16] H. Zhang, H. Huang, and H. Han, "Attention-based convolution skip bidirectional long short-term memory network for speech emotion recognition," *IEEE Access*, vol. 9, pp. 5332–5342, 2021.
[17] M. Xu, F. Zhang, and S. U. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *Proc. 10th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2020, pp. 1058–1064.
[18] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[19] Z. Zhu, W. Dai, Y. Hu, and J. Li, "Speech emotion recognition model based on bi-GRU and focal loss," *Pattern Recognit. Lett.*, vol. 140, pp. 358–365, Dec. 2020.
[20] R. A. Hamad, M. Kimura, L. Yang, W. L. Woo, and B. Wei, "Dilated causal convolution with multi-head self attention for sensor human activity recognition," *Neural Comput. Appl.*, vol. 33, no. 20, pp. 13705–13722, Oct. 2021.

[21] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
[22] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Emotion recognition from raw speech using wavenet," in *Proc. IEEE Region 10 Conf. (TENCON)*, Oct. 2019, pp. 1292–1297.
[23] D. Tang, P. Kuppens, L. Geurts, and T. van Waterschoot, "End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, pp. 1–16, Dec. 2021.
[24] S. Kwon, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114177.
[25] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, p. 7530, Nov. 2021.
[26] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6675–6679.
[27] J. Kim, Y. An, and J. Kim, "Improving speech emotion recognition through focus and calibration attention mechanisms," 2022, *arXiv:2208.10491*.
[28] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7179–7183.
[29] J. Zhang, L. Xing, Z. Tan, H. Wang, and K. Wang, "Multi-head attention fusion networks for multi-modal speech emotion recognition," *Comput. Ind. Eng.*, vol. 168, Jun. 2022, Art. no. 108078.
[30] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross- and self-attention network for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4275–4279.
[31] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 985–1000, 2021.
[32] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.
[33] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multi-modal multi-head attention for continuous affect recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 4171–4183, 2021.
[34] T.-W. Sun, "End-to-end speech emotion recognition with gender information," *IEEE Access*, vol. 8, pp. 152423–152438, 2020.
[35] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-transformer model for emotion recognition from speech audio files," *IEEE Access*, vol. 10, pp. 36018–36027, 2022.
[36] U. A. Asiya and V. K. Kiran, "Speech emotion recognition—A deep learning approach," in *Proc. 5th Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, Nov. 2021, pp. 867–871.
[37] S. Padi, D. Manocha, and R. D. Sriram, "Multi-window data augmentation approach for speech emotion recognition," 2020, *arXiv:2010.09895*.
[38] B. B. Al-onazi, M. A. Nauman, R. Jahangir, M. M. Malik, E. H. Alkhammash, and A. M. Elshewey, "Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion," *Appl. Sci.*, vol. 12, no. 18, p. 9188, Sep. 2022.
[39] D. H. Rudd, H. Huo, and G. Xu, "Leveraged mel spectrograms using harmonic and percussive components in speech emotion recognition," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2022, pp. 392–404.
[40] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. A. Alnuaim, A. Alhadlaq, and H.-N. Lee, "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, no. 6, p. 2378, Mar. 2022.
[41] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. B. Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors*, vol. 20, no. 21, p. 6008, Oct. 2020.
[42] F. Burkhardt, A. Paeschke, M. Rolfes, and W. F. Sendlmeier, "A database of German emotional speech," in *Proc. Interspeech*, vol. 5, 2005, pp. 1517–1520.
[43] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.
[44] P. Jackson and S. Haq, *Surrey Audio-Visual Expressed Emotion (Savee) Database*. Guildford, U.K.: Univ. Surrey, 2014.

SAMUEL KAKUBA received the B.Sc. degree in computer engineering from Busitema University Tororo, Uganda, in 2011, and the M.Sc. degree in data communication and software engineering from Makerere University, Kampala, Uganda, in 2018. He is currently pursuing the Ph.D. degree with the Graduate School of Electronic and Electrical Engineering, College of IT Engineering, Kyungpook National University (KNU), Republic of Korea. He is also an Assistant Lecturer at Kabale University, Uganda. He has worked as a Research Assistant on projects in the fields of data communication systems, embedded systems engineering, the Internet of Things, emotion recognition, computer vision, affective computing, and other machine and deep learning systems.

ALWIN POULOSE received the B.Sc. degree in computer maintenance and electronics from the Union Christian College (affiliated to Mahatma Gandhi University), Aluva, India, in 2012, the M.Sc. degree in electronics from the MES College (affiliated to Mahatma Gandhi University), Marampally, India, in 2014, the M.Tech. degree in communication systems from Christ University, Bengaluru, India, in 2017, and the Ph.D. degree in electronics and electrical engineering from Kyungpook National University, Daegu, South Korea, in 2021. From 2021 to 2022, he was a Researcher at the Center for ICT & Automobile Convergence (CITAC), Kyungpook National University, where he developed a multi-intelligence-based human-centric autonomous driving core technology. He is currently a Research Fellow at the Department of Electrical and Computer Engineering, University of Michigan, Dearborn, USA. His research interests include localization, human activity recognition, facial emotion recognition, and human behavior prediction. He is a reviewer of prominent engineering and science international journals and has served as a technical program committee member/the session chairing at several international conferences.

DONG SEOG HAN (Senior Member, IEEE) received the B.S. degree in electronic engineering from Kyungpook National University (KNU), Daegu, South Korea, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1989 and 1993, respectively. From 1987 to 1996, he was with Samsung Electronics Company Ltd., where he developed the transmission systems for QAM HDTV and Grand Alliance HDTV receivers. Since 1996, he has been a Professor with the School of Electronics Engineering, KNU. He was a Courtesy Associate Professor with the Department of Electrical and Computer Engineering, University of Florida, in 2004. He was the Director of the Center of Digital TV and Broadcasting, Institute for Information Technology Advancement (IITA), from 2006 to 2008. He is currently the Director of the Center for ICT and Automotive Convergence, KNU, where he is also the Dean of the IT College. His research interests include intelligent signal processing and autonomous vehicles.