Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks

Taiba Majid Wani
Electrical & Computer Eng. Department
International Islamic University Malaysia
Kuala Lumpur, Malaysia
wanitaiba1@gmail.com

Teddy Surya Gunawan
ECE Department, IIUM, Malaysia
FTIK, Universitas Potensi Utama, Indonesia
tsgunawan@iium.edu.my
tsgunawan@potensi-utama.ac.id

Syed Asif Ahmad Qadri
Electrical & Computer Eng. Department
International Islamic University Malaysia
Kuala Lumpur, Malaysia
syed17qadri@gmail.com

Hasmah Mansor
Electrical & Computer Eng. Department
International Islamic University Malaysia
Kuala Lumpur, Malaysia
hasmahm@iium.edu.my

Mira Kartiwi
Information Systems Department
International Islamic University Malaysia
Kuala Lumpur, Malaysia
mira@iium.edu.my

Nanang Ismail
Department of Electrical Engineering
UIN Sunan Gunung Djati
Bandung, Indonesia
nanang.is@uinsgd.ac.id

***Abstract*— An assortment of techniques has been presented in the area of Speech Emotion Recognition (SER), where the main focus is to recognize the silent discriminants and useful features of speech signals. These features undergo the process of classification to recognize the specific emotion of a speaker. In recent times, deep learning techniques have emerged as a breakthrough in speech emotion recognition to detect and classify emotions. In this paper, we have modified a recently developed different network architecture of convolutional neural networks, i.e., Deep Stride Convolutional Neural Networks (DSCNN), by taking a smaller number of convolutional layers to increase the computational speed while still maintaining accuracy. Besides, we trained the state-of-art model of CNN and proposed DSCNN on spectrograms generated from the SAVEE speech emotion dataset. For the evaluation process, four emotions angry, happy, neutral, and sad, were considered. Evaluation results show that the proposed architecture DSCNN, with the prediction accuracy of 87.8%, outperforms CNN with 79.4% accuracy.**

Keywords—speech emotion recognition; spectrogram; strides; convolutional neural network (CNN); deep stride convolutional neural network (DSCNN)

I. Introduction

An emerging area of research nowadays is Speech Emotion Recognition (SER), and thus several researchers have been producing various technologies in this area. Many are working to find some features ranging from effective, salient to discriminative features of speech signals. This process is essential for the classification of a speech signal to detect a particular emotion. We can already assume that this research area is very vast, owing to this feature to several social media users, low coast, and fast bandwidth of the internet. Semantic gaps have been occurring due to the usage of low-cost internet and social media. This semantic gap needs to be covered. Researchers are working tirelessly to introduce new methods that limit the extraction to just the most specific and salient features from any speech signal.

SER started as a minimal entity, a niche, and came up to a full-fledged component for Human-Computer Interaction (HCI) and that too a significant one [1]. Now, moving on these systems' work, it will be right to say that they act as facilitators during natural interaction with machines when using direct voice interaction instead of using traditional devices as input to understand verbal content and make it easy for human listeners to react [2]. SER's critical role is in the automatic translation systems and understanding human physical interaction in crowds, particularly for violent or destructive actions, as they are difficult to handle manually. Thus, we can now say that any SER system's central theme is to detect the characteristics of a speaker's voice in different emotional conditions.

A traditional SER system's functioning consists primarily of extracting features from the speech, which is then classified to predict the various classes of emotions in it [3]. The features can be extracted from a speech in two ways, the first being the hand-crafted features and the second being deep neural network-based features. Some generalized hand-crafted features are used more often than others, and these are spectral features, formant features, frequency features, energy-related features, and pitch. On the contrary, in deep learning, the features are absorbed learned hierarchically, learning higher-level abstractions of the low-level features [4].

Classification of speech emotion recognition can be carried out in two ways: (a) traditional classifiers like HMM (Hidden Markov Model), SVM (Support Vector Machines), GMM (Gaussian Mixture Model), and Bayesian Network model and (b) deep learning algorithms like Deep Belief Networks (DBN), Deep Neural Networks (DNN), Convolution Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM). These methods' success can be determined by the promising results that they have shown in various fields of speech recognition, including those of emotion recognition and other speech analysis applications. As a result of various new researches in the field, recently, deep learning has been confirmed with a research field's status

978-1-7281-7596-6/20/$31.00 ©2020 IEEE

having vast potential in machine learning and has gained immense attention in recent years [5].

SER's deep learning methods have several advantages over traditional methods, including their capability to detect the complex structure and features without the need for manual feature extraction and tuning. Other advantages include the tendency toward extraction of low-level features from the given raw data and the ability to deal with unlabeled data. Taking inspiration from the success of deep learning methods, we have modified a recently developed different network architecture of CNN, i.e., Deep Stride Convolution Neural Networks (DSCNN) [6]. The modification is carried out by taking five convolution layers with different filter sizes and kernels to increase computational efficiency. The detailed description of the methodology is explained in the subsequent section. Two experiments are carried out: Convolutional Neural Networks and Deep Stride Convolutional Neural Networks. Both are trained on the spectrograms generated from the SAVEE dataset.

II. Related Works

The distinct feature of deep learning methods that differentiate it from the traditional machine learning method is to extract high-level features. It has been shown to exceed human performance in visual tasks [7]. The most significant contributions of deep learning are making certain breakthroughs in multimedia, including speech emotion recognition [8]. Recently, CNN has become very popular for SER.

A congruent way to deal with executing an effective emotion recognition system dependent on deep convolution neural systems (DCNNs) utilizing categorized training audio data was proposed in [10]. An end-to-end model presented by Trigeorgis et al., in [11] encompassed CNN architecture used to extract features before feeding a Bi-directional LSTM (BLSTM) to model the temporal dynamics in the data. In [12], the authors proposed a model comprising three convolutional layers and three fully connected layers to extract the discriminative features from spectrogram images. They also used a pre-trained AlexNet model for scrutinizing the potency of transfer learning for emotion recognition.

An implementation of an end-to-end deep neural network for emotion recognition was presented in [13]. The authors implemented convolutional and recurrent networks extracted from paralinguistic data present in the speech. The network learned the data that represented emotions directly from spectrograms, which enabled a solution for noise reduction. Similarly, in [14], the authors investigated the deep spectrum features formed by feeding the spectrograms through AlexNet for speech emotion recognition and formed a feature vector from the activations of the last fully-connected hidden layer.

In [15], several investigations were carried out over various CNN and LSTM-RNN models. The CNN architectures achieved better performance as compared to the other LSTM-RNN architectures. Multiple models have been presented in [3] for the classification of emotion using phoneme and spectrogram.

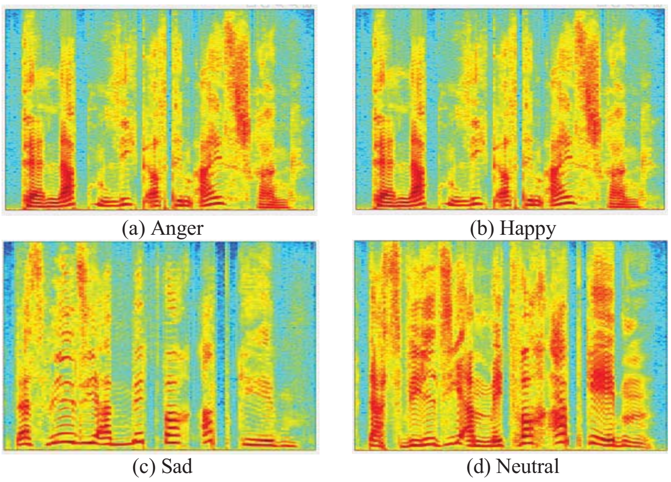
The phoneme sequence (embedded vector form) was taken as input, followed by convolution with multiple kernels in Model 1. Likewise, in Model 2, the spectrogram was used as input to the 2D CNN. Model 2 provided enhanced accuracy in the extraction of high-level features. In [4], the authors utilized a modified single frequency filtering (SFF) spectrogram as a substitute depiction of speech and CNN with batch normalization for each convolution layer for evaluating pitch-synchronous SFF. In a recent publication [6], the authors present the use of preprocessing of a speech signal to remove the unwanted noise to deliver a clean speech and used Deep Stride Convolution Neural Networks (DSCNN) for the classification process. We aim to design a high accuracy emotion recognition system for speech using a clean spectrogram and modified architecture of previous DSCNN.

III. Proposed Algorithms

The proposed framework endeavors to use a discriminative Convolution Neural Networks for feature learning schemes utilizing spectrograms produced from speech signals. The stride CNN architecture will have input layers, convolutional layers, a flatten layer, and fully connected layers followed by a SoftMax classifier. The spectrograms will be put into use as they tend to hold rich information. Also, extraction and application of such information when transforming the audio speech signal to text or phonemes are highly unlikely. This capability lets the spectrogram enhance the recognition of emotion. Therefore, the primary idea is to study high-level discriminative features from speech signals, making CNN architecture highly imperative.

A. Spectrograms

Spectrograms represent a signal quality across time at various frequencies present in the specific waveform. Its computation is based on the application of short-term Fourier transform (STFT) to the speech signal, which in turn forms the time-frequency representation.


(a) Anger (b) Happy

(c) Sad (d) Neutral

Fig. 1. Spectrogram Samples of Various Speech Emotions

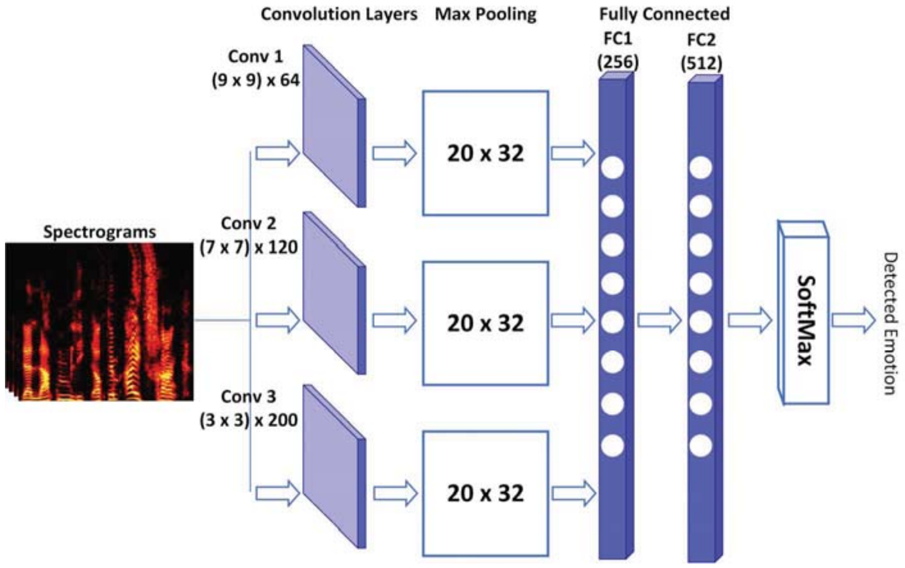


Fig. 2. Proposed CNN architecture for speech emotion recognition.

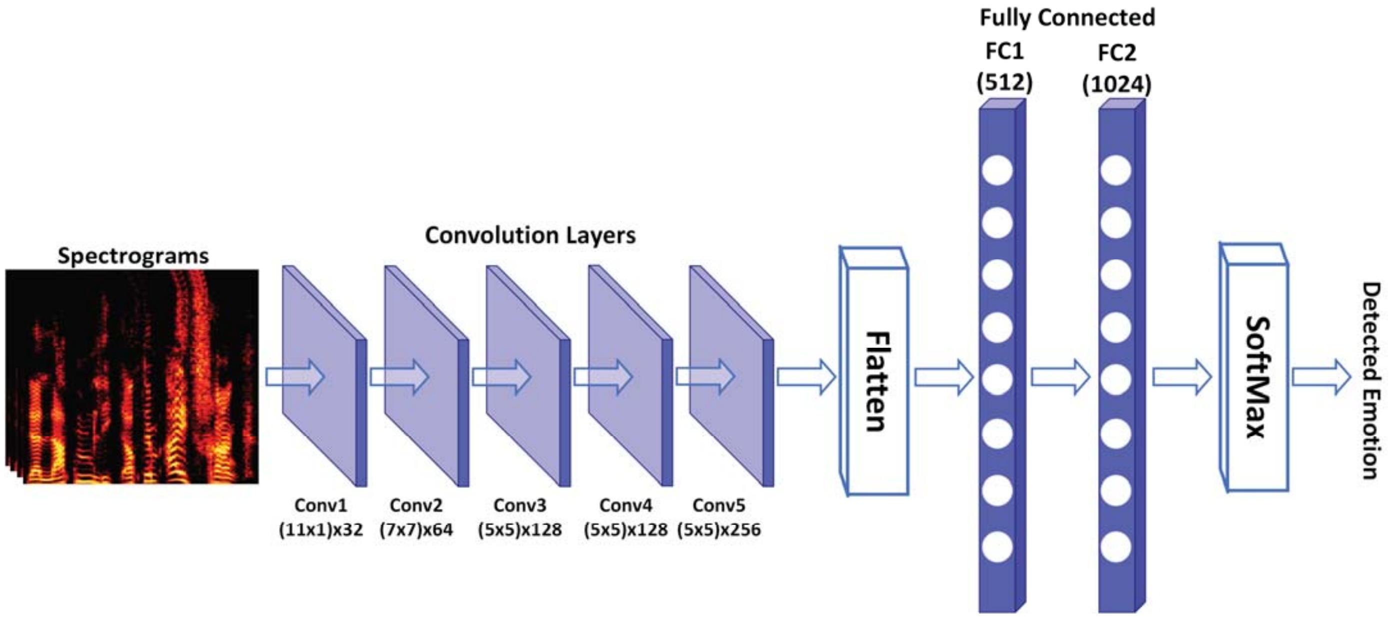


Fig. 3. Proposed Deep Stride CNN architecture for speech emotion recognition

One of the difficult tasks in SER is the dimensioning of the signal using 2D CNN. Therefore, 1D representation of speech signal is modified into a suitable 2D representation for 2D CNN, since we intend to learn high-level features from speech signals using the CNN architecture. Spectrograms are utilized to represent the audios files present in the SAVEE speech emotion database. Sample of the extracted spectrograms of audio files from the SAVEE database depicting the angry, happy, sad, and neutral emotion by applying STFT is illustrated in Fig. 1.

B. Convolutional Neural Networks (CNN)

CNN is an organized neural network consisting of various layers in a sequential pattern. The model generally consists of various convolution layers, pooling layers, fully connected layers, and a SoftMax unit. This sequential network forms a feature extraction pipeline modeling the input in the form of an abstract.

Initially, the input spectrograms are convolved with different types of filters that are learned during the training phase, and feature maps are obtained. The polling layers accumulate the maximum activation functions from the feature maps, thus, reducing their dimensionality. In fully connected layers, all the neurons of the input layer related to every other neuron in the layer. Finally, the task of classification is performed by the SoftMax unit. The proposed CNN architecture is shown in Fig. 2.

A spectrogram of size 256×256 is used as an input generated from the speech SAVEE database. The CNN model consists of 3 convolution layers: Conv1, Conv2, and Conv3 with 64 kernels of size (9×9), 120 kernels size (7×7), and 200 kernels of size (3×3) respectively. The generated features are fed to the top pooling layer. For all the three convolutional blocks, the size of the maximum pooling is the same 20x32. The top pooling layer is followed by two fully connected layers FC1 and FC2, 256, and 512 neurons. Only FC1 is followed by the dropout layer with a 25% dropout ratio to avoid overfitting. The ReLU activation function follows all the convolutional layers. This activation is used because of its higher convergence rates. Finally, the classification task is carried out by the SoftMax unit.

C. Deep Stride Convolutional Neural Network (DSCNN)

The architecture of the Deep Stride Convolutional neural network is depicted in Fig. 3. Unlike the CNN model, this model excludes the pooling layer instead uses special strides for dimension reduction or downsampling. Strides simply mean the number of pixels shifted over the input matrix and minimizes the considerable number of computations that are to be done by the subsequent layers in the network. The DSCNN utilizes the concept of direct networks for recognizing the emotions from the speech. The proposed DSCNN consists of convolutional layers, a flatten layer, fully connected layers, and a SoftMax unit. The network has mostly used a similar filter size (5×5) for learning the deep features in the convolutional layers. Further, to reduce the resolution of feature maps sizes, the stride setting of 2×2 pixels is used.

The spectrograms generated from the SAVEE database are taken as input. There are five convolutional layers in the proposed architecture: one flatten layer, two fully connected layers, and a SoftMax unit. The first convolutional layer consists of 32 kernels of size (11×11) with zero paddings and stride (2×2) pixels to effectively adjust the data's dimensionality. Similarly, the second convolutional layer has 64 kernels of size (7×7) with stride (2×2). The convolutional layers 3 and 4 have a similar number of kernels 128 and size (5×5). The last convolutional layer has 256 (5×5) kernels with a stride setting of (2×2) pixels. The activation function used is a rectified linear unit (ReLU). Also, the batch normalization is carried out to improve the stability and performance of the model.

After the last convolutional layer, a flatten layer is used, converting the data shape into a vector form. The output of the flatten layer is fed to fully connected layers. The two fully connected layers have 512 and 1024 neurons, respectively. Both fully connected layers are followed by 50% of the dropout ratio to overcome the overfitting. Finally, for the classification, the features are fed to the SoftMax classifier.

IV. Experimental Results and Discussion

In this study, we have performed two experiments. In the first experiment, convolutional neural networks (CNN) were trained over the SAVEE database, and the performance is evaluated. For the second experiment, Deep Stride Convolutional Neural Network (DSCNN) was trained over the same database. A comparative analysis of both the experiments is shown in the subsequent section. This section accords the description of the dataset, experimental setup, and the analysis of the results.

A. Speech Emotion Dataset

Surrey Audio-Visual Expressed Emotion (SAVEE) is an acted English dataset. The SAVEE dataset consists of 480 British English utterances. The recordings were done by four male speakers DC, JE, JK, and KL. It consists of 15 sentences for each of the seven different emotions, anger, fear, happiness, disgust, sadness, surprise, and neutral. Each emotion has 60 utterances except neutral with 120 utterances. In this study, only four emotions, including anger, sadness, happy and neutral, were considered. Moreover, 240 utterances were taken for the evaluation of the experiment. The dataset was divided into 70% for training and 30% for testing.

B. Experimental Setup

Matlab software was used to generate spectrograms from the SAVEE database. The considered number of emotions was 4. For each emotion, 60 images were collected from the dataset. Approximately 240 spectrograms were generated. The data was split in the ratio of 70 to30, in which 70% of data was used for the training process and the remaining 30% for the testing. The training process was run for three different numbers of epochs 500, 1200, and 1500.

The training was performed on a single i5-8250 CPU with 8 GB RAM. We performed two experiments. The first experiment involved the CNN model's training on spectrograms and evaluated the prediction performance for accuracy. In the second experiment, DSCNN was trained on the spectrograms and evaluated the model's accuracy performance. Moreover, the proposed architectures are shown in Fig. 2 and Fig. 3 were trained on the SAVEE database with three different epochs 500, 1200, and 1500. The recognition or prediction rate was recorded for all three epochs.

C. SER Experiment using the proposed CNN architecture

For the proposed CNN architecture, Table I shows the numerical confusion matrix with 500 training epochs. The numbers shown diagonally are the percentage of each emotion class that was identified clearly; the rest of the numbers show incorrectly identified percentage of each emotion class. It is clearly depicted that sad and neutral prediction performance was above 50% but lower than 50% for angry and happy. However, the overall accuracy of the model is 65.5%.

TABLE I. Performance of SER System using CNN Architecture with 500 Epochs

		Predicted Class			
		Anger	**Sad**	**Neutral**	**Happy**
Actual Class	**Anger**	**43.3**	12.6	33.7	10.4
	Sad	9.6	**78.3**	0	12.1
	Neutral	3.6	0.3	**93.3**	2.8
	Happy	25.9	0	27	**47.1**

TABLE II. PERFORMANCE OF SER SYSTEM USING CNN ARCHITECTURE WITH 1200 EPOCHS

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	**64.8**	13.2	15.8	6.2
	Sad	5.8	**83.3**	0	10.9
	Neutral	2.8	1.2	**93.3**	2.7
	Happy	0	32.1	0	**67.9**

TABLE III. PERFORMANCE OF SER SYSTEM USING CNN ARCHITECTURE WITH 1500 EPOCHS

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	**71.2**	0	12.8	16
	Sad	5	**83.3**	0	11.7
	Neutral	0.8	0	**98.6**	0.6
	Happy	0	23.8	11.7	**64.5**

Similarly, Table II and Table III show the numerical confusion matrix with 1200 and 1500 epochs. Both the tables show better predictions for all the emotions. However, angry and happy have fraternized with each other. Overall, prediction accuracy for 1200 and 1500 epochs is 77.3% and 79.4%, respectively. This result shows that with the increase in the training epochs, an improvement in the prediction accuracy is indicated.

D. SER Experiment using the proposed CNN architecture

For the proposed DSCNN architecture, Table IV shows the numerical confusion matrix with 500 training epochs. It is depicted, the prediction performance of sad, happy, and neutral were above 50% but lower than 50% for angry. However, the overall accuracy of the model is 68.3%%.

TABLE IV. PERFORMANCE OF SER SYSTEM USING DSCNN ARCHITECTURE WITH 500 EPOCHS

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	**30**	10.2	59.8	0
	Sad	6.8	**71.2**	5.9	16.1
	Neutral	0	0	**100**	0
	Happy	0	16.2	11.8	**72**

TABLE V. PERFORMANCE OF SER SYSTEM USING DSCNN ARCHITECTURE WITH 1200 EPOCHS

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	**80**	0	0	20
	Sad	2.8	**94.4**	0	2.8
	Neutral	3.6	0.7	**95.7**	0
	Happy	12.6	0	10.3	**77.1**

TABLE VI. PERFORMANCE OF SER SYSTEM USING DSCNN ARCHITECTURE WITH 1500 EPOCHS

		Predicted Class			
		Anger	Sad	Neutral	Happy
Actual Class	Anger	**83.3**	0	15	1.7
	Sad	1	**95**	1.4	2.6
	Neutral	0	0	**96.6**	3.4
	Happy	2.6	18	2.8	**76.6**

In addition, Table V and VI show the numerical confusion matrix with 1200 and 1500 epochs, respectively. Both the tables show better predictions for all the emotions. However, sad and neutral have fraternized with each other. Overall, prediction accuracy for 1200 and 1500 epochs is 86.8% and 87.8%, respectively.

E. Discussion

Two architectures CNN and DSCNN were employed for emotion recognition using speech dataset SAVEE. Only four emotions were taken for both the experiments, Anger, Sad, Neutral, and Happy. CNN model consisted of 3 convolutional layers, two fully connected layers. For reduction in resolution, the pooling layer was used, and for the classification, a SoftMax unit was deployed. While in DSCNN, five convolutional layers were used in which small blocks of size (5×5) were used. For a decrease in dimensionality, particular strides (2×2) pixels were used. A flatten layer, following that, were two fully connected layers and a SoftMax unit. Both models were subjected to three different sets of training epochs 500. 1200 and 1500. With training epochs 500, both the models did not perform well. However, DSCNN was responded better than CNN. For 1200 training epochs, DSCNN showed quite good results with 86.8% and CNN with 77.3%. Finally, with 1500 epochs, there was not much hike as compared to previous training epochs. The proposed DSCNN algorithm resulted in 87.8% and CNN with 79.4% of accuracy. As the number of training epochs was increased, there was an enhancement in the prediction accuracy.

The models' overall prediction performance is shown in Table VII, and it is observed that DSCNN outperforms CNN for all the three sets of training epochs.

TABLE VII. PERFORMANCE COMPARISION BETWEEN CNN AND DSCNN

Epochs	CNN Accuracy (%)	DSCNN Accuracy (%)	CNN Training Time (second)	DSCNN Training Time (second)
500	65.5	68.3	31	30
1200	77.3	86.8	79	86
1500	79.4	87.8	80	184

V. CONCLUSIONS AND FUTURE WORKS

The focus of Speech Emotion Recognition research is to design proficient and robust methods to recognize emotions. In this paper, we have modified the recently proposed algorithm Deep Stride Convolutional Neural Networks (DSCNN) by decreasing the number of convolutional layers with different sizes. This network completely excludes the pooling layers instead makes use of special strides for decreasing the dimensionality of feature maps. Two experiments were carried out to check the effectiveness of the state-of-art model of CNN and DSCNN. Both models were trained on the SAVEE database considering only four emotions: angry, happiness, sadness, and neutral. The input to the models was spectrograms generated from the speech database. Both the models' performance was aggrandized as the number of training epochs was incremented from 500 to 1200 and 1500. However, the presented model DSCNN outperforms the state-of-art model CNN by a considerable margin. 87.8% of accuracy was obtained for DSCNN and 79.4% for CNN. Still, more work is needed to improve the given architecture for persuasively recognizing the emotions.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the Malaysian Ministry of Education (MOE), which has provided research funding through the Fundamental Research Grant, FRGS19-076-0684. The authors would like to acknowledge and support from International Islamic University, Universitas Potensi Utama, and UIN Sunan Gunung Djati.

REFERENCES

[1] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," Inf. Fusion, 2019, doi: 10.1016/j.inffus.2018.09.008.

[2] J. G., D. Sundgren, R. Rahmani, A. Larsson, A. Moran, and I. Bonet, "Speech emotion recognition in emotional feedback for Human-Robot Interaction," Int. J. Adv. Res. Artif. Intell., 2015, doi: 10.14569/ijarai.2015.040204.

[3] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," 2018, doi: 10.21437/Interspeech.2018-1811.

[4] S. Gupta, M. S. Fahad, and A. Deepak, "Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition," Multimed. Tools Appl., 2020, doi: 10.1007/s11042-020-09068-1.

[5] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," IEEE Access, 2019, doi: 10.1109/access.2019.2936124.

[6] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," Sensors (Switzerland), 2020, doi: 10.3390/s20010183.

[7] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," Nature, 2016, doi: 10.1038/nature16961.

[8] Deng and Y. Dong, "Foundations and Trends in Signal Processing, Deep Learning: Methods and Applications," Signal Processing, 2014.

[9] A. M. Badshah et al., "Deep features-based speech emotion recognition for smart affective services," Multimed. Tools Appl., 2019, doi: 10.1007/s11042-017-5292-7.

[10] W. Q. Zheng, J. S. Yu, and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," 2015, doi: 10.1109/ACII.2015.7344669.

[11] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," 2016, doi: 10.1109/ICASSP.2016.7472669.

[12] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," 2017, doi: 10.1109/PlatCon.2017.7883728.

[13] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," 2017, doi: 10.21437/Interspeech.2017-200.

[14] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," 2017, doi: 10.1145/3123266.3123371.

[15] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," Neural Networks, 2017, doi:10.1016/j.neunet.2017.02.013.