xIlkogretim Online - Elementary Education Online,2021; Vol 20 (Issue 1): pp.1959-1969
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.01.212

Speech Emotion Recognition using Deep neural networks: Insight from Management perspective

Nabeel Sabir Khan, Department of Computer Science, School of Systems and Technology, University of Management and Technology, Lahore, Pakistan
Ahmad Hassan Butt, Department of Computer Science, School of Systems and Technology, University of Management and Technology, Lahore, Pakistan
Muhammad Ali, The Superior university, Lahore, Pakistan

ABSTRACT- Emotions play an important role in social interactions of humans and it is often said that emotions separate us from machines. Spoken words may have different interpretations depending on how they are uttered. Same sentences can have different meanings under different type of emotional states. Human brain understands different meanings by perceiving underlying emotions in speech. Finding the emotional content from speech signals is desirable because this enables us to teach emotional intelligence to computers.Speech emotion recognition is an important field of study with applications ranging from emotionally intelligent robot creation, audio surveillance, web-based E-learning, computer games, etc. The objective of this paper is to identify emotions in audio speech by using deep learning algorithms including Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to identify different emotional states of a person. In this regard, the RADVEES dataset, Ryerson Audio-Visual database of Emotional Speech and Song, is used to study speech emotion recognition. For experiments, we used approximately 1247 audio and song files containing eight different emotions for classification of audio data. The experimental results show that the best performing model was CNN based model with accuracy of 74.57% while RNN model only showed 55.47% accuracy which is far less in comparison. This work will be extended in future using different variants of RNNs and other DNNs like auto-encoders. Audio is a complex signal with linguistic and paralinguistic features and our future goal is to combine these features with different neural network architectures for developing improved SER systems.

Keywords: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), DNN

I. INTRODUCTION

Research in the area of cognition sciences has proved that emotions plays a vital role in communication, perception and social interaction [1]. Emotions can be designated as the psychological response of our interaction with environment and surrounding including human and animals [2]. Emotions influence the way we communicate with others, make decisions and our behavioral expression with others[2]. Machines are an important part of our daily lives. We are surrounded with machines. In near future it may happen that robots are found working with us in our social environment. Therefore, it is necessary for robots to understand, analyze and predict the human emotions and interact emotionally with them[3]. This can be achieved through speech, facial expression and gestures. There are many reasons suggesting development of technologies which enable robots to understand and respond to emotions. Firstly, it will help in human computer interaction. Secondly, it will provide machines a better understanding towards the humans and making decisions accordingly [3].

Last Decade have seen a number of studies on the objective to understand and comprehend the human brain and develop the systems that mimic the capacity of brain to make intelligent decisions. Brain is the most complex organ of human body and it has been a source of inspiration for exploration and investigation in the field of Artificial Intelligence (AI), Human Computer Interaction (HCI) and neural networks [4]. A social organism uses emotions in order to communicate with its species. Organisms use different means i.e. secretion of chemicals, body gestures or languages. If we take example of ants, they use secretion of chemicals to communicate with their species. Similarly, whales use body gestures (splashing their tales and making bubbles) for sending their messages to other members of its species.

The research in the field of emotions is not new, it started with Darwin, a famous scientist, who proposed a scientific method to recognize and analyze emotions [5]. He proposed his theory of evolution and explained with description the movement of various muscles during emotional states. He also supported

his theory with the help of images the movement of various muscles [5]. According to Darwin's theory of evolution, individuals living in a specific community adopt similar style of speaking which is common in that specific society however; each individual has different style of speaking and portraying emotions as compared to other individuals [5]. It also varies in gender (male or female) age, (child, adult or mature person) language, culture or social status of individual. As individual grows in age, their way of communicating becomes well defined and their way of portraying emotions changes. Therefore, we can say that a child of 8 year and a person of 50 year has a different way of expressing their emotions [5].

Emotions are usually defined as how people perceive them. An emotion is a complex behavioral sensation, which involves different intensities of neural and chemical incorporation. The term "emotion" refers to a complex state associated with a variety of physical, mental and physiological events [5]. Every person has different perception level of understanding for expressed emotions. Psychologists, based on human perception, did early work on emotion recognition. There are different applications of analyzing the emotions like in robots, audio surveillance [2], web-based E-learning [2], marketable applications, medical studies [6], entertainment, banking, call centers, car board systems [3], computer games and language learning [2]. Emotions are one of the most important aspect of human social interaction. They play a vital role in our interaction with others and help us to understand the feelings of the other, conveys our messages and feeling to others, and help us to understand the mental state of others. In addition, they are the central part of our interactions and conversation with others [7]. It is believed that the human being can recognize emotions from speech whether it is any type of language (foreign or local) that they understand or not. Similarly, machines should be capable to understand and learn emotions from any type of language. A typical Speech recognition system (SER) consist of three phases as shown in Figure 1. These phases are 1) Speech Signal Preprocessing, 2) Feature Extraction and Selection and 3) Classification of Signal into relevant class using Machine Learning algorithms. Figure 1. Illustrates the working of a traditional SER system. In a traditional SER, speech signals are used to extract and select useful features based on their ability to distinctively identify different emotions. The selected features are then utilized to create training set which is used to train an SER model using any of the available classification algorithms. Once trained, the model can be deployed to classify real-time audio data for recognition of emotions. Feature selection method aims to attain such information from speech signals, that are not redundant and to derive the feature values from it [28].

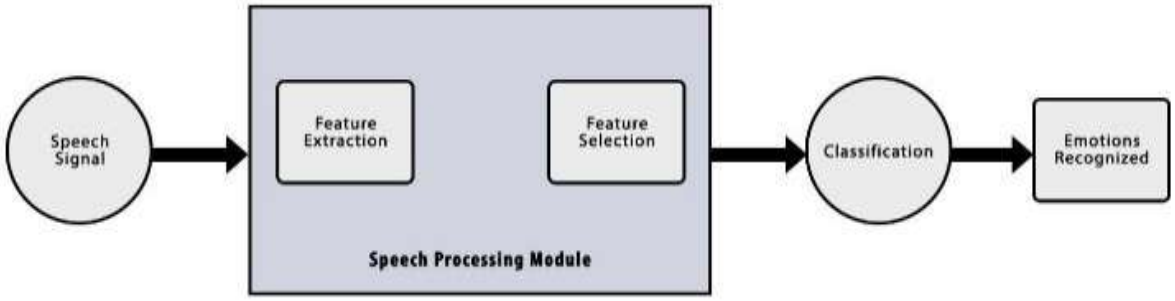

Figure 1: Architecture of Traditional Speech Emotion Recognition (SER) System

Existing technologies use feature engineering along with conventional machine learning models to develop speech recognition systems. These models are constrained by their ability to use high dimensional raw speech data and quality of engineered features. A better approach is to use deep neural networks (DNNs) for speech emotion recognition. DNNs are studied under Deep Learning. It is an emerging area of AI which uses simple mathematical structures as neurons for different learning tasks [8]. Using deep learning, contemporary data scientists have made great strides in developing solutions for problems involving computer vision, speech processing, and natural language processing and online-advertisements [9]. Deep learning models use multi-layered neural networks where first layer receives the input and last layer provides the output. Layers of DNNs non-linearly transform their input in a hierarchical manner, creating more abstract, task-specific representations, which are insensitive to unimportant variations, but sensitive to important features. With sufficient training of neural network on input/output data, the output of penultimate layer provides an optimal low-dimensional representation of input record which is used by output layer to predict the labels of input.

There are different ways through which neurons are combined to create DNNs, but mainly it is categorized in to two types. Cyclic graph in which information flows in one direction only. Feed forward

network shown in Figure.2 is an example of such type of network. In semi-cyclic graphs, information flows both in forward as well as backward direction. Recurrent network are examples of such type of graphs.

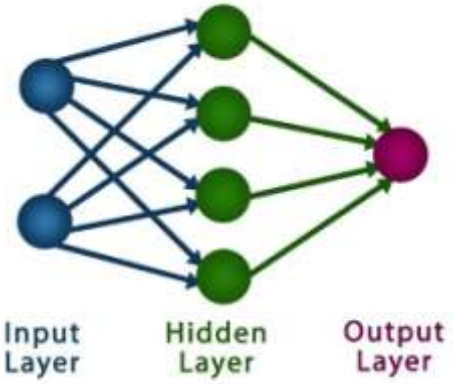

Figure 2: Standard neural network

As the intended outcome and impact of Speech Emotion Recognition (SER) has not been achieved yet, this study aims to utilize DNNs for developing SER system. Rest of this paper is organized as follows. Section II provides a review of literature relating to SER. Section III presents the materials and methods of this study. Section IV describes the results of proposed SER models and provide discussion on results. Conclusions drawn from this study are outlined in Section V of the study and paper is concluded by presenting the references on which this study is composed.

II. RELATED WORK

Speech Emotion Recognition can be inspected through only the processing of sound signals without involvement of any linguistic information. Emotions have a direct or indirect effect on human behaviors. Previously, emotion detection was considered one of the difficult problems for computers. But the research landscape has been constantly changing for better due to remarkable performance of deep neural networks in this area. Work has been done on emotion recognition from both visual and auditory (speech) data to make computers emotionally intelligent. There exists different types of features in speech including Paralinguistic features and Acoustic features [10]. Paralinguistic features include non-verbal cues and body gestures like hands movement during speech, body movement, facial expression, tone and pitch of sound [10]. Paralinguistic features convey implicit information from speech such as emotions in speech [10]. Acoustic features, on the other hand, include frequency and amplitude of speech. They also include qualitative, quantitate and spectral features [10]. In this article, paralinguistic features are focused by authors to develop SER system using different DNNs.

SER can be explained using Brunswik's lens theory. According to this model, at one end, the speaker speaks a message, which is encoded and transmitted. On the other end, the message is decoded and perceived. Different types of emotions can be expressed including speech, verbal and nonverbal communication and music. In the field of SER, previous works have proved that gender difference affects the results of model due to pitch and quality of male and female [11].

Recent research contributions on SER [12], [13], [14], [15], [16], [17] used IEMOCAP database for recognition and classification of emotion from speech data by using different classifiers. Different datasets has been used in literature to create better SER systems. A list of important datasets and SER studies utilizing those datasets is shown in Table 1. Fayek et al. [12] used multiple variants of fully connected neural network (FCNN) and recurrent neural networks (RNN) and to devise effective SER models. Their experimental results show the qualitative and quantitative aspects of model performance. This contribution also describes the advantages and limitations of paralinguistic. The results presented by [12] describe how FCNN and their variants could be engaged to obtain paralinguistic emotion recognition from speech signals. According to results of Fayek et al. [12], convolutional neural network (CNN) shows better results as compared to other architectures of deep neural networks. In other studies comprising of [18], [19], the researchers used IEMOCAP and Emo-DB databases along with MFCC for emotion recognition by using state of the art deep learning algorithms (1D-CNN and LSTM) and attention based convolutional neural network(ACRNN) on speech data. The experiment was conducted by using 1D and 2D CNN LSTM.

Their deep neural networks based model have a benchmark accuracy of 91.6% and 92.9% as compared to previous models. But the study presented by [18] does not provide explicit training and testing accuracy to support their claim.

Table 1: Different Speech Corpus used in the Literature

Dataset Name	IEMOCAP	Berlin EMODB	eNTERFACE 05	INTER-SPEECH	Audio-visual emotional databases	Spanish dataset	MSP-IMPROV	RAVDEES
Relevant Literature	[2],[10], [13] [14],[15],[16] [20]	[2],[10],[13], [15],[21], [22] [20],[19],[23] [24],[25],[26]	[18],[27], [28],[29]	[21], [23]	[29],[30]	[2], [24]	[14]	[19], Current Study

Another experiment conducted by Guo et al. [13] used two state of the art datasets for SER, which include Emo-DB and IEMOCAP, to recognize emotional states by using audio data. In this research, kernel extreme learning machine (KELM) is used as classifier and to extract features from audio data CNN-BLSTM is used. The experimental results by [13] showed the accuracy of 57.99% on Emo-DB and 92.9% on IEMOCAP. IEMOCAP dataset, along with MSPIMPROV, was also used by Peng et al. [14] to investigate SER. They used attention-based sliding recurrent neural networks ASRNNs on Happy, Sad, Angry, and Neutral emotions. The percentage accuracy claimed by [14] for IEMOCAP and MSP-IMPROV is 62.6% and 55.7% respectively. The percentage accuracy of MSP-IMPROV is low as compared to IEMOCAP due to highly imbalanced nature of dataset.

EMO-DB Dataset was used by other research contributions [15], [17], [19], [20], [21], [22] and [23] to develop SER models using different classifiers and feature extractors. For example, Deng et al. [21], used semi supervised auto encoders to develop improved SER model. The major advantage of using auto encoder is their ability to learn useful nonlinear features. In this regard, researcher has also explored AEC, SUSAS, GeWEC and INTERSPEECH 2009 Emotion Challenge datasets. In another study, Demircan et al. [22], used fuzzy C mean clustering algorithm for classification of emotion. The Spectral features are obtained by using MFCC and LPC. After comparison from models, developed using different classifiers, the researcher claimed SVM showing highest success rate of 92.86%.

In another study, Sun et al. [31], used Berlin EmoDB database on universal emotions like anger, disgust, happiness, sadness, fear, boredom and neutrality. The experiment was recursively repeated using different configurations of convolutional neural network (CNN) to increase performance. Best accuracy was shown by model trained using a variant of Convolutional neural network (CNN) showing accuracy of 65.73%. According to this research, the percentage accuracy was not up to the mark but satisfactory for SER system and need of improvement exists in the model proposed by researcher.

Table 2: Popular speech corpus and emothins contained therein

Corpus name	***Language***	***Emotion***
Berlin emotional DB	German	Neutral, Anger, Fear, Disgust, Boredom, Joy, Sadness.
RAVDESS DB	English	Anger, Fear, Disgust, Surprise, Happy, Sad, Neutral
SAVEE [47],[49]	English	Anger, Disgust, Fear, Happiness, Sadness, and surprise
IEMOCAP	English	Anger, Happiness, neutral Sadness, Excitement, Surprise, Frustration, Fear, Disgust
eNTERFACE '05	English	sadness, Happiness, anger surprise, disgust, fear

INTERSPEECH,	English	Paralinguistic And Acoustic Features (Pitch, durations and intensity)

The raw speech signals are complex and contain linguistic information as well as additional information such as intensity, pitch, loudness and other features [10]. This additional information makes the speech signals highly complex. Feature extraction method aims to attain such information from speech signals, that are not redundant and to derive the feature values from it [32]. Some contributions of literature used feature engineering with conventional Machine learning models to develop SER systems. MFCC (Mel frequency Cepstral Coefficients) are most widely used in Automatic speech recognition (ASR) [2]. It is considered a standard method for extracting features from speech [33]. MFCCs are considered to be very accurate in certain recognition problems such as human voice recognition and speaker identification [2]. The drawback of MFCCs is that these coefficients are sensitive to noise due to its dependence on the spectral form [33]. This study also uses MFCC for feature extraction because it is considered a standard feature extraction method from speech signals in multiple SER research contributions.

III. MATERIALS AND METHODS

In this section we provides a discussion on dataset used and methodology adopted for research.

a.Dataset introduction and preprocessing

In literature, different studies used different datasets. Some popular datasets used by different authors for SER research are presented in Table 1. Each dataset contained different range of human emotions as outlined in Table 2. There is no single well-defined standard to decide which type of dataset is more useful for different type of speech signals. In this article, Ryerson Audio-Visual database of Emotional Speech and Song (RAVDESS) dataset is used to study SER problem. RAVDESS contains total 7356 audio and files having size of 24.8 GB [34]. The dataset contains 24 professional actors vocalizing two lexically matched statements in a neutral North American accent. The statements are, “Kids are walking by the door” and “Dogs are sitting by the door”. The average duration of audio files is 3s. And the voice segments are recorded in .wave file. It contains two category of sound signals i.e. speech and song. Emotions contained in RAVDESS include but not limited to calm, happy, sad, angry, and fearful expressions. Each expression is produced at two levels of emotional intensity e.g. normal and strong [34]. Although RAVDESS also contains visual data, for this research, we only used audio data comprising of approximately 1247 files containing song and speech to run our experiments for training SER models. The methodology of proposed research is shown in Figure 3.

For preprocessing, the amplitude of audio files is normalized in the range of [-1, 1]. Preprocessed audio files were stored in h5 format to conserve computational resources. Data for experiments was generated by extracting MFC coefficients from preprocessed audio files. MFC coefficients captures variation in such dimension which are considered to be very useful in speech recognition problems such as human voice recognition and speaker identification. The number of MFC coefficients was fixed at 80 to generate 80 features representing each audio sample.

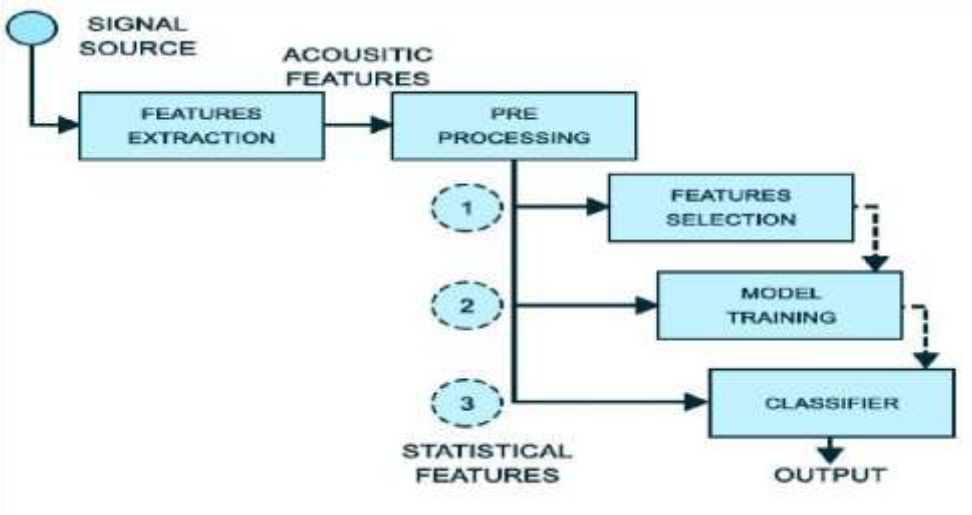

Figure 3: Methodology of Research

b.Deep Neural Networks for Model Training

This section presents the overview of DNNs used in this along with discussion on architecture and optimization of model.

Convolution Neural Network

Convolution Neural Network (CNN) is a deep neural network which learns the filters, when applied on input using convolution operation, allows the Neural Network to learn representation of data which is useful for predicting the correct outputlabel. CNN is an inspired biological architecture and is described with its features like shared weights, receptive structure, sparse connectivity and different layers. CNNs are very successful in the field of pattern recognition, image processing and emotion recognition and natural language processing [25], [35]. CNN is composed of layers consisting of number of filters which, when convolved with input, produce a representation comprising of multiple feature maps of same dimensions arranged depth-wise. Purpose of subsampling layer, the other core component of a CNN, is twofold. Not only, it controls over-fitting by decreasing the size of representation with a fixed down sampling technique (e.g. max-pooling or mean pooling), it also performs dimensionality reduction by merging generalized features learned from upper layers to form more fine-grained and abstract representation at its output [8].

Like any other DNN, last layer of CNN is called output layer and it can be composed of one of many different activation functions including Softmax, Sigmoid, Hyperbolic Tangent and ReLU depending on the problem being solved. Output layer can be considered as a simple classifier function which uses the representations of penultimate layer of CNN to predict the output class. A sample illustration of 1-D CNN is shown in Figure 4.

Recurent Neural Network

Traditional Neural Networks are unable to share features learned across different positions of the sequence. Recurrent Neural Networks (RNNs) are designed to address this issue by using loops in their structure. An RNN is a connectivity pattern that performs computations on a sequence of vectors $\{x_1, \ldots, x_n\}$ using a recurrence formula of the form $a_t = f_\theta(a_{t-1}, x_t)$, allowing us to process sequences with arbitrary lengths where f an activation function is and θ is a parameter and both are used at every timestamp. The hidden vector a_t intuitively provides a temporal summary of sequence at previous steps and it is known as the state of the RNN. In a Simple RNN cell, the parameters governing

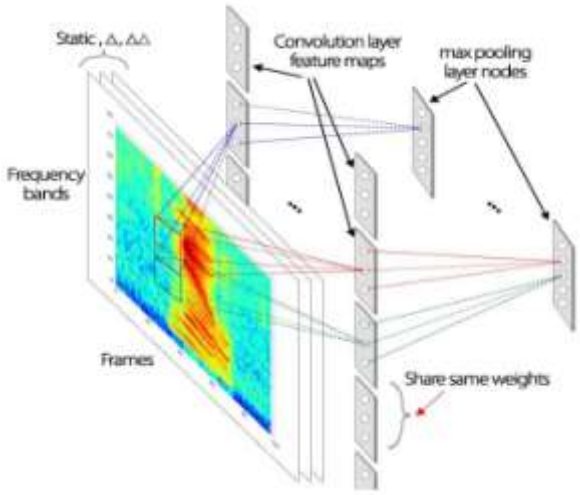

Figure 4: An illustration of CNN with 1D convolution

the connections from input to the hidden layer, horizontal connections between activations and connections from hidden to output layers are shared. A forward pass in simple RNN cell is governed by following a set of equations:

$$a^{<t>} = g(W_a[a^{<t-1>}, X^{<t>}] + b_a)$$

$$y^{<t>} = f(W_y . a^{<t>} + b_y)$$

where <t> describes the timestamp, g represents the activation function which is usually tanh, $X^{<t>}$ represents the input at time step t, b_a represents the bias, W_a represents the shared weights and $a^{<t>}$ represents the activation at timestamp t. Simple RNN can use this activations $a^{<t>}$ to calculate predictions $y^{<t>}$ at timestamp, if required. Let T_x be the length of the input sequence and T_y be the length of the output sequence, different architectures of RNNs are possible. We used a many to one architecture where Tx ≠Ty for multiclass classification of the audio sequences. c.Experimental Work

We used both CNN and RNN for developing SER models in this study in a manner shown in Figure 7. In order to obtain better results, CNN-1D and RNN with vanilla cell architecture were implemented. The number of epochs used was 500 and utilized number of MFC coefficients as features was fixed at 80. Initially the number of features were increased to different values but the accuracy of the models were compromised so in order to obtain good results to improve the accuracy of DNNs, the number of MFC coefficient features were fixed at 80 for all experiments.

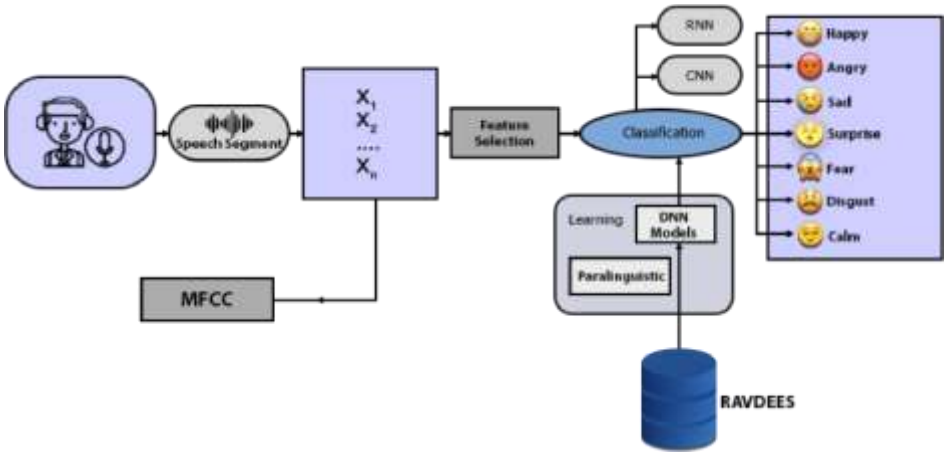

Figure 5: Architecture of Experiments for Creating SER models in this study

Hyper-parameters were optimized using RandomizedSearchCV [36]. For RNN, we used a many to one architecture where $T_x \neq T_y$ and $T_y = 1$ as shown in Table 3. For CNN-1D, selected hyper-parameters include learning rate of 0.005, MaxPooling-1D with filter-size of 8 and rmsprop [37]as optimizer. The architecture of CNN-1D is shown in Table 4. Two convolution layers with Relu activations and three pooling layer with a dropout layer is used. Global-Average-Pooling function was added in the model to

converts the pooled feature map to a single vector that is passed to the output layer for making label predictions.

Table 3: Architecture of implemented RNN for SER

Layer Type	Output Shape	Trainable Parameters
R1: Recurrent layer with 16 Simple RNN units	(Batch_size, 80, 20)	560
GlobalAveragePooling ID	(Batch_size, 16)	0
Output: Output Layer	(Batch_size, 8)	(16+1) * 8 = 136

Table 4: Architecture of Implmented 1-D CNN for SER

Layer Type	Output Shape	Weight Parameters
C1: Conv1D with 20 filters of size 3	(Batch_size, 78, 20)	(9+1)*20 = 200
S1: Maxpooling 1D	(Batch_size, 26, 20)	Not Applicable
D1: Dropout with 25% of probability		Not Applicable
C2: Conv1D with 10 filters of size 3	(Batch_size, 24 , 10)	((3 * 20) + 1)* 10)= 610
S2: Maxpooling 1D	(Batch_size, 8, 10)	Not Applicable
S3: GlobalAveragePooling1D	(Batch_size, 10)	Not Applicable
Output: Output layer with one sigmoid	(Batch_size, 8)	(10+1)* 8 = 88

IV. RESULT AND DISCUSSION

In this section, result of this study for SER model are presented along with discussion thereon. Unfortunately RNN model did not show promising results for SER while CNN based SER showed promising results. CNN based SER model is evaluated using Confusion matrix. For a binary classification problem, Confusion matrix is comprised of following four values:

- True Positive (TP): If the sample is from positive class and model under consideration also predicts it as a positive class then result is considered True Positive
- False Positive (FP): if a sample is from negative class and model under consideration predicts it as a positive class then result is accepted as False positive
- False Negative (FN): If the sample belongs to positive class but model under consideration predicts it as a negative class then result is accepted as False Negative
- True Negative (TN): If the sample belongs to negative class and model under evaluation also predicts it as a negative class then the result is considered True Negative.

For multiclass problem, the confusion matrix is calculated in one-versus rest (OVR) manner and the main diagonal of confusion matrix represents TP for each class. Values above the main diagonal represent collective False positives and values below the main diagonal represent collective false negatives.

Table 5: Confusion Matrix of CNN-1D based SER Model

	Neutral	**Calm**	**Happy**	**Sad**	**Anger**	**Fear**	**Disgust**	**Surprise**
Neutral	**18**	4	1	1	1	0	1	0
Calm	3	**50**	2	2	0	1	2	1
Happy	1	7	**46**	1	1	2	2	2
Sad	4	6	3	**42**	4	8	6	3
Anger	0	2	5	1	**53**	1	2	0

Fear	0	1	6	1	4	**37**	4	2
Disgust	0	0	1	1	2	1	**31**	1
Surprise	0	0	4	0	5	3	4	**16**

Table 5 shows the confusion matrix of CNN-1D based SER model. The True Positives for each class are shown on the main diagonal of confusion matrix. As main diagonal entries are large integers as compared to other entries, this shows that the proposed model is performing reasonablywell. To quantify the performance of Model, we used accuracy which is a popular summary model evaluation metric.

Accuracy provides the fraction of results which were correctly classified by a model. The train accuracy (accuracy of model on training data) and test accuracy (accuracy of model at test data) is shown in Figure 6. The model achieved training accuracy of 94.63% and test accuracy of 74.57%. The gradual increase of both train and test accuracy show that model is neither under-fitted nor over-fitted on training data. Over-fitting in CNN-1D based model was mitigated by use of dropout layer proposed by Srivastav et al. [38].

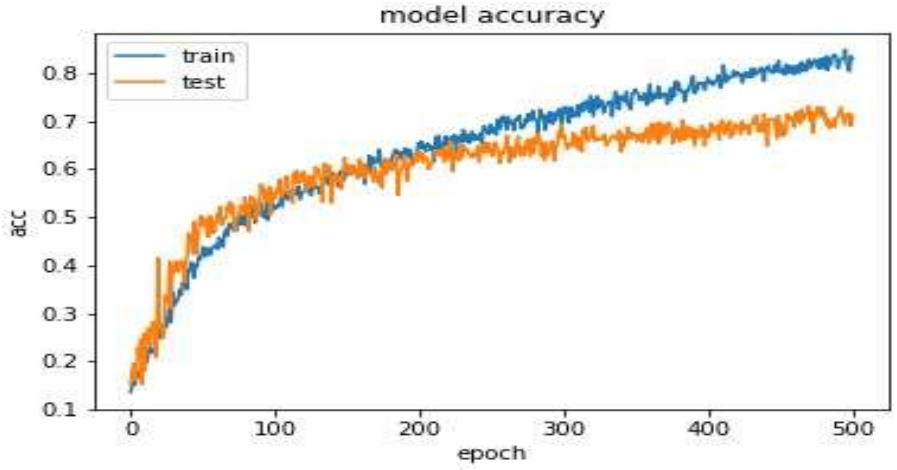

Figure 6: Accuracy plot of CNN-1D based SER Model

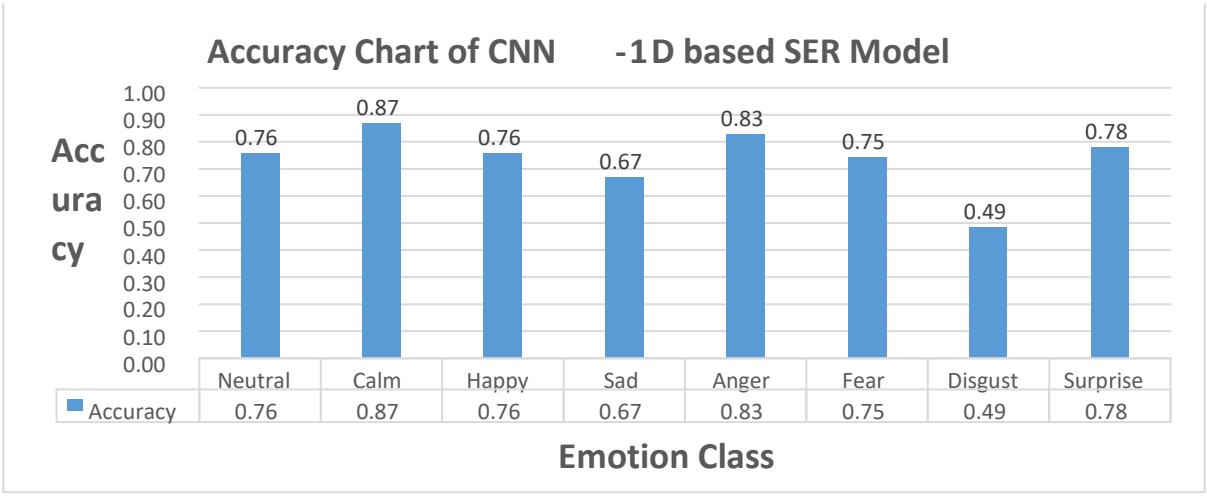

Figure 7: Accuracy Chart of individual emotions in CNN-1D based SER Model

Table 6: Individual Emotion Accuracy Scores for CNN-1D based SER

Emotion Class	**Neutral**	**Calm**	**Happy**	**Sad**	**Anger**	**Fear**	**Disgust**	**Surprise**
Percentage Accuracy	76.00%	86.89%	75.81%	67.11%	82.82%	74.55%	48.65%	78.13%

Table 6 shows the percentage accuracy of different emotion classes using CNN-1D. The most difficult emotion class to predict for CNN-1D based SER model turned out to be "Disgust" which is shown by less than 50% accuracy score for the class. "Calm" emotion class, on the other hand, proved to be highest accuracy emotion with accuracy score of 86.89%. We believe this is due to minimal variation in audio features captured by MFC coefficients which enabled this class to stand out from other emotion classes and achieve accuracy score higher than any other emotion class. Accuracy chart of CNN-1D based SER model is shown in Figure 7.

V. CONCLUSION

This article aims to study the problem of recognizing human emotions from sound signals i.e speech and songs. In this study, we proposed a new speech emotion recognition (SER) system by combining MFC coefficients and deep neural networks. The dataset used in this study is RAVDESS. This dataset was chosen due to its notoriety in classifying different emotions contained in two type of audio signals consisting of speech and song. We used 1247 files of RAVDESS containing audio and song files containing eight different emotions i.e. calm, happy, sad, angry, fearful, disgust, surprise, neutral for classification of audio data. Audio data was preprocessed to extract MFC coefficients as features which were then used to train speech emotion predictors using Convolutional neural network and recurrent neural network with vanilla cells to train SER models. Experimental results showed better performance of convolutional neural network with test accuracy of 74.57%. This work will be extended in future using different variants of RNNs and other DNNs like auto-encoders. Audio is a complex signal with linguistic and paralinguistic features. Our ultimate target is to use other available audio features such as acoustic and lexical features and combine them with different DNNs to develop improved SER systems.

REFERENCES

[1] L. Tian, J. D. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with acoustic and lexical cues," in *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*, 2017, pp. 45–46.

[2] D. Jeanmonod, K. Suzuki, and others, "We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists TOP 1% Control of a Proportional Hydraulic System," *Intech Open*, vol. 2, p. 64, 2018.

[3] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*, 2017, pp. 1–5.

[4] S. Shahsavarani, "Speech Emotion Recognition using Convolutional Neural Networks," 2018.

[5] P. Ekman, "Darwin's contributions to our understanding of emotional expressions," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 364, no. 1535, pp. 3449–3451, 2009.

[6] T. Rajisha, A. Sunija, and K. Riyas, "Performance analysis of Malayalam language speech emotion recognition system using ANN/SVM," *Procedia Technol.*, vol. 24, pp. 1097–1104, 2016.

[7] Y. Kim, "Automatic Emotion Recognition: Quantifying Dynamics and Structure in Human Behavior.," PhD Thesis, 2016.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436, May 2015.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[10] 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, 2019.

[11] B. Zhang, "Improving the generalizability of emotion recognition systems: towards emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 582– 586.

[12] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Netw.*, vol. 92, pp. 60–68, 2017.

[13] L. Guo, L. Wang, J. Dang, Z. Liu, and H. Guan, "Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine," *IEEE Access*, vol. 7, pp. 75798–75809, 2019.

[14] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech Emotion Recognition Using 3D Convolutions and Attention-Based Sliding Recurrent Networks With Auditory Front-Ends," *IEEE Access*, vol. 8, pp. 16560–16572, 2020.

[15] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.
[16] F. Chenchah and Z. Lachiri, "Speech Emotion Recognition in Acted and Spontaneous Context.," in *IHCI*, 2014, pp. 139–145.
[17] C. Busso*et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, p. 335, 2008.
[18] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5858–5869, 2014.
[19] Y. Huang, K. Tian, A. Wu, and G. Zhang, "Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 5, pp. 1787–1798, 2019.
[20] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, 2018.
[21] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 31–43, 2017.
[22] S. Demircan and H. Kahramanli, "Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech," *Neural Comput. Appl.*, vol. 29, no. 8, pp. 59–66, 2018.
[23] Q. Mao, G. Xu, W. Xue, J. Gou, and Y. Zhan, "Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition," *Speech Commun.*, vol. 93, pp. 1–10, 2017.
[24] S. Kuchibhotla, H. D. Vankayalapati, and K. R. Anne, "An optimal two stage feature selection for speech emotion recognition using acoustic features," *Int. J. Speech Technol.*, vol. 19, no. 4, pp. 657–667, 2016.
[25] J. Jia *et al.*, "Inferring Emotions From Large-Scale Internet Voice Data," *IEEE Trans. Multimed.*, vol. 21, no. 7, pp. 1853–1866, 2018.
[26] M. Deriche and others, "A Two-Stage Hierarchical Bilingual Emotion Recognition System Using a Hidden Markov Model and Neural Networks," *Arab. J. Sci. Eng.*, vol. 42, no. 12, pp. 5231–5249, 2017.
[27] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domainadaptive least-squares regression," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 585–589, 2016.
[28] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, no. 5, pp. 768–785, 2011.
[29] C. Huang, B. Song, and L. Zhao, "Emotional speech feature normalization and recognition based on speakersensitive feature clustering," *Int. J. Speech Technol.*, vol. 19, no. 4, pp. 805–816, 2016.
[30] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep neural networks," in *International Conference on Neural Information Processing*, 2017, pp. 811–819.
[31] L. Sun, J. Chen, K. Xie, and T. Gu, "Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition," *Int. J. Speech Technol.*, vol. 21, no. 4, pp. 931–940, 2018.
[32] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, 2006, pp. 8–8.
[33] U. Shrawankar and V. M. Thakare, "Techniques for feature extraction in speech recognition system: A comparative study," *ArXivPrepr. ArXiv13051145*, 2013.
[34] 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, 2018.
[35] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha, "Recognition of emotion in music based on deep convolutional neural network," *Multimed. Tools Appl.*, vol. 79, no. 1, pp. 765–783, 2020.
[36] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *JMLR*, p. 305, 2012.
[37] G. Hinton, N. Srivastava, and K. Swersky, "Rmsprop: Divide the gradient by a running average of its recent magnitude," *Neural Netw. Mach. Learn. Coursera Lect. 6e*, 2012.
[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.