A Research of Speech Emotion Recognition Based on CNN Network

Anurish Gangrade[1], Shalini Singhal[2]
Swami Keshvanand Institute of Technology, Management & Gramothan Rajasthan Department of Computer Science, Jaipur, India
Swami Keshvanand Institute of Technology, Management & Gramothan Rajasthan Department of IT, Jaipur, India
Email: anurish.gangrade@gmail.com, shalini.singhal@skit.ac.in
Received: 28.06.2022**, received in revised form** 26.07.2022, **accepted** 26.07.2022
DOI: 10.47904/IJSKIT.12.1.2022.24-31

Abstract- This paper proposed a novel method of feature extraction, using DBNs in DNN to automatically extract emotional options from speech signals. Speech emotion recognition relies heavily on feature extraction, which is why the paper focused on this aspect of the problem. Feature extraction is an essential component of the speech emotion recognition process. To extract speech emotion features, we used a 9-layer depth DBN, and we included numerous consecutive frames into the process to produce a high-dimensional feature. An improved CNN model is presented in this article. This model consists of a combination of convolution 1d layers and has been generalized to form a 9-layer architecture of CNN (convolutional neural network). The model accuracy has been checked with respect to emotion classes such as considering 5 emotions such as angry, calm, fearful, happy, and sad for both male and female speakers, and eventually a speech emotion recognition multiple classifier system was achieved. The voice emotion recognition rate of the system achieved 89.00 percent, which is around 14 percent more than the traditional approach could get.

Keywords – Convolution Neural Network

1. INTRODUCTION

Voice emotion identification is a technique that involves the use of a computer to extract emotional characteristics from speech signals, followed by the comparison and analysis of distinguishing parameters, in addition to the emotional modification that is not inheritable. In the end, the legislation that governed both speech and feelings was overturned, and from then on, speech and emotional states were assessed according to the law.

Speech emotion recognition is currently a popular research area in signal processing and pattern recognition, as well as a growing field of computing and artificial psychology. The analysis has been widely used in sectors such as humans & computer interaction, interactive teaching, enjoyment, and security. Speech feeling process and recognition system is mostly composed of 3 components that were speech input signal analysis a then, extracting any features from it, and finally the recognized emotion.

This prompts Speech Emotion Recognition (SER) developing exploration subject in which loads of progressions can prompt headways in different field like programmed interpretation frameworks, machine to human collaboration, utilized in combining discourse from text so on. Interestingly the paper center to overview and survey different speech extraction highlights, emotional speech information bases, classifier calculations, etc.

The field of technology known as speech recognition focuses on methods and strategies for extracting speech from input signals. Numerous technological advancements within the field of the artificial intelligence and signal process techniques, recognition of feeling created easier and attainable. It's conjointly referred to as Automatic Speech Recognition". It's found that voice may be next medium for human action with machines particularly once computer-based systems. Since there's a colossal development within the field of Voice Recognition. Numerous voice-activated products, such Google Home and Alexa from Amazon, Apple Home Pod have been created that functions principally on voice-based commands. It's evident that Voice are going to be the higher medium for human action to the machines.

1.1 CONVOLUTION NEURAL NETWORK

One of the most often used deep learning models is convolutional neural networks (CNNs), which have excelled in a variety of academic domains such as fourteen visual perception, face identification, handwriting recognition.

Convolutional neural networks, often known as CNNs, may be broken down into three distinct layers: the pooling layer, the totally connected layer, and the convolutional layer. After that, we will proceed to discuss these building parts in

conjunction with a selection of essential concepts like the SoftMax unit, rectified linear measurement, and dropout.

1.2 CONVOLUTIONAL LAYER

Convolutional layers in CNNs encrypt the output using convolution rather than multiplication. This design is distinctive due to the innate receptive field that neurons in the visual region have. In other words, the neurons are trained to react to inputs that are limited to a certain area and structure. The deep neural networks have been reduced as a result of parameter sharing and sparse connectivity in CNNs, which are a result of convolution's structure.

2. LITERATURE REVIEW

Complete review on the speech feeling recognition is explained during which reviews properties of dataset, speech feeling recognition study classifier selection. Varied acoustic options of speech area unit investigated and a few of the classifier strategies are analyzed during which is useful within the additional investigation of contemporary strategies of feeling recognition. This study examined how emotional voice cues may be used to anticipate future reactions, which helped to justify the widespread use of emotions, using totally different classes of classifiers. In order to categorize speech feelings, a few classification algorithms, such as K-NN and Random Forest, are used. In the discipline of information science, recurrent neural networks are incredibly prevalent and attempt to address many problems. Utilized are deep RNN models of LSTM and bi-directional LSTM trained for acoustic alternatives. Various CNN units that are being used and taught to recognize speech emotions are assessed. Filter banks and Deep CNN are used to infer emotion from voice signals, and their high accuracy rate raises the possibility that deep learning might potentially be used to identify feelings. Speech emotion recognition is frequently accomplished in conjunction with image spectrograms and enforced deep convolutional networks.

3. IMPLEMENTATION

In this study, we used convolutional neural networks (CNNs) to identify voice utterances based on their emotional components. In addition to three widely known standards for emotion recognition from voice utterances, we leverage RAVDESS data to teach and evaluate our models. To implement our CNN and LSTM models, we often employ Tensor Flow (an ASCII text file library written in Python and C++). This chapter outlines the current work's experimental setting. The initial section describes the databases used in this research. The second section goes over the preprocessing steps. The third section discusses the training and validation configuration that we utilized to train and validate our models.

4. DATASET RAVDESS: BRITISH ENGLISH DATABASE

Dataset here is a dependable source of emotional speech and music across media. With twenty-four skilled actors, vocal music, and lexically matched phrases given in a very objective North American dialect, the material is gender balanced. Unlike song, which also incorporates expressions of peace, happiness, grief, wrath, fear, surprise, and disgust, speech can also display these emotions. With a neutral expression in the middle, every expression has two levels of emotional expression. There are voice-only, face-only, and face-and-voice variants of every condition. Each of the 7356 recordings was given a score based on how emotionally authentic, intense, and valid it was. 247 participants from North America who were representative of untrained analysis participants provided ratings. 72 people in a separate group took a knowledge test and retest. Measures of corrected accuracy and composite "goodness" are supplied to aid investigators in the selection of stimuli.

5. PROPOSED METHOD

A typical pattern recognition workflow, which includes vocal emotion recognition, consists mostly of feature extraction and classification. The proposed approach uses the CNN Classifier attention model, which serves as a bidirectional long short-term memory network (LSTM) for classification, silent elimination before feature extraction, and adds both to the feature extraction process. Figure 1 depicts the suggested system.

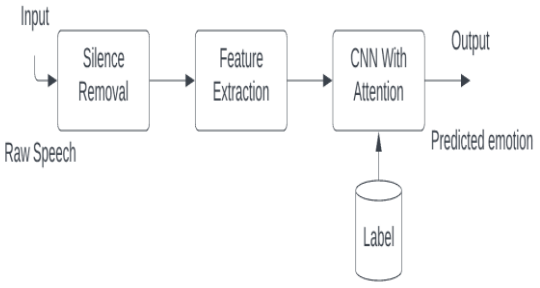

Figure 1: Silence removal and classification system

5.1 SILENCE REMOVAL

The first step in dealing with audio knowledge would be to view the voice recording as a vector or matrix. We prefer to execute silence removal on every file supported by the threshold and also the least variety of samples whenever every sound file on the dataset

is browsed as a vector. These two parameters are applied to each vocalization in the speech dataset. Filtered speech is the result of this silence reduction approach. Formula one contains the whole formula for silence removal.

6. FEATURE EXTRACTION

The methods in [4] are followed for feature extraction. To begin, each spoken auditory communication is divided into window frames that are affected by overlapped one another. The feature extraction steps conducted over every frame at intervals every auditory communication. At initially we fed input shape of 259 x_train shape parameter and 256 as conv 1d filter with kernel size as 8, Hence, the dimensions of the feature vector for every auditory communication is (None,259,256). These options area unit fed to CNN networks.

7. CNN AS BIDIRECTIONAL LSTM

The concept of employing an LSTM network stems from the belief that humans have the ability to retain memories for lengthy periods of time [9]. Every second, Humans don't begin to think in a new way. We have a tendency to see every word in this article as supporting our understanding of prior words as we browse through it. We have a habit of not throwing everything away and starting over. Tenacity is in our thinking. [8].

Only data from the previous time period is saved by LSTM, but it keeps data across layers. That is, we are able to reserve historical facts but not future ones. This problem is addressed bidirectional through victimization. Each data point, to enhance network performance.

Bidirectional LSTMs outperform unidirectional LSTMs in the majority of cases ([3], [6]). We have a propensity to develop a technique for extracting features to recognize speech emotions from the speech section using bifacial LSTM as shown in [11]. The speech segments are then processed to produce alternatives trained with bifacial LTSMs.

8. ATTENTION MODEL

We usually add an attention model as a step after the bidirectional LSTMs. As previously stated, BLSTMs will pass data from long run to past, but which data is the most significant is unknown. The attention model is intended to handle this issue in this scenario. The attention model will be used to select relevant facts (e.g. avoid noises). Provided speech segments, for example, the components of these speech segments contribute significantly to feeling recognition. As said attention model comprises of these components: encoder, attention manipulation, and decoder. On the encoder aspect, A BLSTM network is employed integrating the input features x = (x1, x2. . . xt) and provide output like encoded sequence of h = (h1, h2, . . . , hT)

For each sequence, T may represent a range of input features. Some processing will occur before h is delivered to the decoder. In this scenario, we choose the final encryption purely because it reflects the overall state outline. For any input feature at point t on the decoder encoded sequence is accepted by the output decoder of h = (h1, h2,..., hT) as well as the previous state st1 that is shared within the decoder cell and a feeling label yt1 that represents one of the four emotions. The ultimate result is now one of four feelings y = (y1, y2, y3, y4) in binary type, one for anticipated emotions and zero for the remaining emotions.

The context vector was formed by the attention mechanism. First, attention probability chances = (α1, α2. . . αT) are calculated using the encoded sequence (Eq. (1)) of the decoder cell's inner hidden state, st1.This probability computation is illustrated in equivalent (2). It's possible that a softmax will perform. The context vector ct (Eq. (3)) is then computed as the weighted sum of the encoded sequences with attention probability.

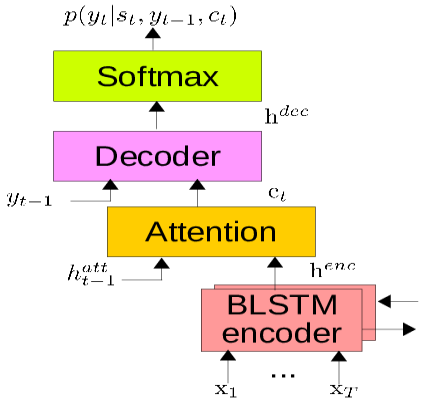

Figure 2: Bidirectional LSTM encoder attention model schematic diagram

Depending on how many classes are utilized to categorize the emotions, different SoftMax will be employed. For running such model, it generally took around 10hrs to 24 hrs. to be trained. CPU's uses more of time to train the model, so instead of that GPU's can be preferred. The GPU's several cores increase speed and cut down on waiting time. The Figure 1 shows the Convolutional Neural Network architecture used in this paper. Additionally, the CNN architecture is employed to categorize among more classes, with successful results.

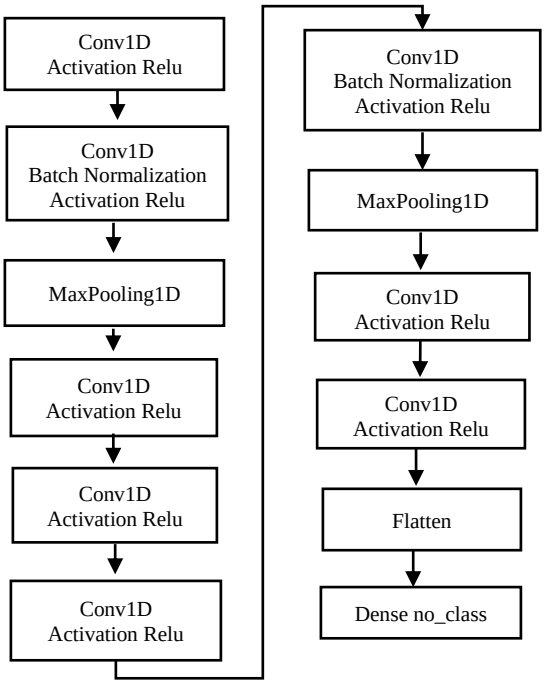

Figure 3: The CNN baseline architecture utilized to identify voice utterances according on their emotional states.

9. TRAINING AND TEST SETS

A 9-fold cross-validation was used throughout the training and testing of the models. In a different way of putting it, the data were divided into nine folds. The first fold was employed as a control set, and the subsequent folds were use in the process of training the models. The remaining folds were used for training, while the additional fold was used to test the models. This process was repeated with the remaining folds. The data sets were enlarged by adding white Gaussian noise to each audio signal ten or twenty times with a signal-to-noise magnitude ratio (SNR) of +15. This was done to avoid overfitting and the detrimental effect of limited database sizes. The signal-to-noise ratio, abbreviated as SNR, is calculated using the formula ten log10 (Pspeech/Pnoise), where P is the average signal power.

Knowledge sets with ten times augmentation (10x) and data sets with twenty times augmentation (20x) were created as a result of the data augmentation (20x). We frequently employ the first knowledge without noise to validate our models. The enhanced knowledge was only used for training purposes. Finally, both the training and test knowledge labels were encoded as vectors. The table displays the category labels for each piece of information. The number of training epochs was varied from 100 to 4000.Due to computation and time costs, the beneficial training epoch was chosen to 700.

10. ARCHITECTURE

The deep neural network that was used in this investigation had a baseline design that consisted of 9 convolutional layers in a convolutional neural network, and one completely linked layer with 1024 hidden neurons. This design was enforced within the context of the present study. In consideration of the number of categories, either a 5-way or a 7-way SoftMax unit was used to make an estimate of the probability distribution of the categories. After each convolutional layer came either a maxpooling or average-pooling layer, depending on which one was chosen. Rectified In order to include nonlinearity into the model, the convolutional and totally connected layers used linear units, also known as ReLU, as activation functions. The kernel size of the pooling layers was initially fixed to eight, and the initial range of kernels was determined to be the kernel size. The training of the networks developed in this investigation took anywhere from eight to twenty-four hours, depending on the Graphics Process Units used (GPUs). As a rule, graphics processing units (GPUs) are used rather than central processing units (CPUs) in order to increase the computing performance. This is due to the fact that GPUs have several cores and are able to handle a broad variety of synchronized threads.

11. RESULT ANALYSIS

The information model predicts emotions with an accuracy rating of around 89.00 percent. On the dataset, we did numerous language-dependent gender-independent studies. We usually begin the research by implementing the baseline CNN design. Later on, depending on the performance of the models, we tend to adjust the hyper parameters such as the scale of convolution kernels and hence the deletion chance of the dropout formula.

This chapter aims to present the results of those experiments and to debate the outcomes. a number of the explanation for fewer accuracy rate are, Transfer Learning is employed to train the model, there could’ve been less spectrograms used for coaching, which results in the less accuracy. There are additionally less knowledge set used for the coaching method that also results in the case.

12. EXPERIMENT RESULTS

To assess the effectiveness of the proposed model, we frequently run speaker-independent SER tests on the RAVDESS.

For RAVDESS, we tend to consider the temporary information with 5 emotional categories: happy, angry, sad, fearful and neutral, and use all seven emotions. during this work, we tend to be had probe for an acceptable dataset that's offered and includes

the basic emotions, conjointly we tend to need the dataset to incorporate several actors instead of counting on one or 2 performers. We'd like the dataset to have a large number of ages for each gender so that we may test the models (Male, Female). As a result, we select the RAVDESS dataset. The dataset RAVDESS includes video and audio files for twenty-four artists who were required to sing and say two sentences ("Kids are talking by the door," "Dogs are sitting by the door") while displaying radically different emotional states.

For the verbal information, eight emotional states are covered. ("Neutral, calm, happy, sad, angry, fearful, disgust, surprised") and 6 emotions for sing recording (neutral, calm, happy, sad, angry, and fearful). There are 7356 files altogether in the dataset. The recording is made up by twelve male and twelve feminine performers. In our work, we tend to use the speaking recording solely.

With completely different parameter initializations, we are able to acquire big selection of results. To ensure that our analyses produce a large number of trustworthy results, we run each analysis five times using different random seeds and give the average and variance. Since the take a look at category distributions are unbalanced. On the test set, we typically provide the unweighted average recall (UAR). Note that, all model architectures, together with the quantity of epochs are elect by increasing the UAR on the validation set.

After adding gender to the label and defining the truth label, we came with below truth labels:

Index (['male_happy', 'male_fearful', 'male_angry', 'male_calm', 'male_sad',
'female_angry', 'female_fearful', 'female_sad', 'female_happy',
'female_calm', 'male_neutral', 'male_surprised', 'female_surprised',
'female_disgust', 'male_disgust', 'female_neutral'], dtype='object')

Then plotting the emotion distribution for the above truth labels:

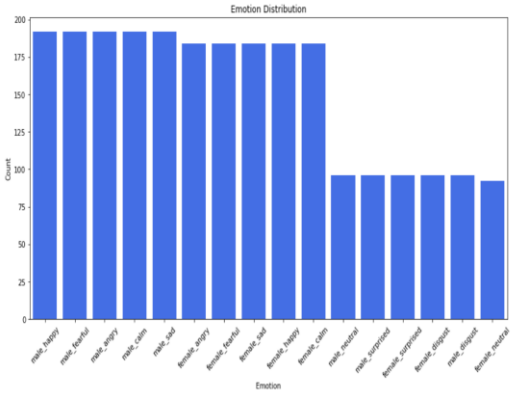

Figure 4: Emotion Distribution plot for ten different classes

We break the voice signal into equal-length 3 second segments for improved parallel acceleration, and zero-padding and then applied to utterances with durations less than 3 seconds. Each sub-segment predicts one emotion during the training phase, and during the testing phase, we evaluate the phrase as a whole by applying max pooling to the posterior probabilities of each sub-sentence.

Validation was done on 528 samples, Training on 2112 samples for the train valid loss graph on 3 class emotions and Training on 1920 samples, validate on 480 samples for 2 class emotions

1–20 actors are used for Training / Validation sets with 8:2 splitting ratio.21–24 Actors are excluded for testing usage. In this to minimize model complexity to analyze male emotions first isolated two actors to be test set with 8:2 stratified shuffle split, Neutral, Disgust, Surprised are excluded in 10 class recognition from dataset. The model is trained with batch size of 16 and 700 epoch's parameter.

Model is trained on the training data and then it is tuned with the results of metrics (accuracy, loss etc.) that is achieved from validation set.

Data splitting is done for male and female labels.

Table 1: Data Splitting from Data frame

	path	source	actor	gender	intensity	statement	repetition	emotion	label
0	C:/Users/Anurish Gangrade/Desktop/speech-proje...	1	1	male	0	0	0	1	male_neutral
1	C:/Users/Anurish Gangrade/Desktop/speech-proje...	1	1	male	0	0	1	1	male_neutral
2	C:/Users/Anurish Gangrade/Desktop/speech-proje...	1	1	male	0	1	0	1	male_neutral
3	C:/Users/Anurish Gangrade/Desktop/speech-proje...	1	1	male	0	1	1	1	male_neutral
4	C:/Users/Anurish Gangrade/Desktop/speech-proje...	1	1	male	0	0	0	2	male_calm

After Data splitting, we get the features of audio files using librosa, then data augmentation is done on the

files, Dimension for CNN Model is changed by applying activation function as soft-max units.

After loading the model from disk, we perform predicting emotion on test data. In order to further assess the SER performances, we give the confusion matrix at the end of CNN based on LSTM NETWORK. We observe that on RAVDESS datasets, Sadness receives the lowest recognition rate, whereas anger receives the highest.

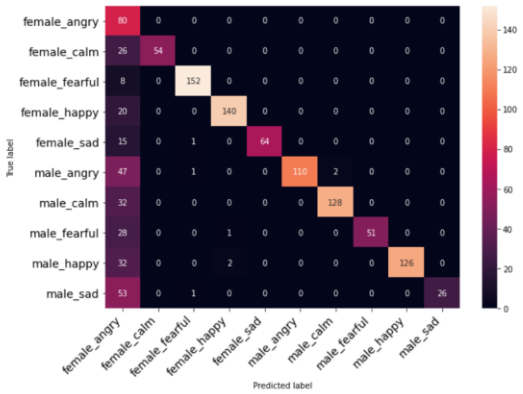

Figure 5: Confusion matrix plot for ten different classes

A max-pooling layer is placed following each convolutional layer in the suggested design. Both convolutional and fully connected layers use Rectified Linear Units (ReLU) for activation functions to establish nonlinearity in the model.

Batch normalization is a technique for improving the neural network stability, It makes the output of the preceding activation layer more consistent by lowering the amount of hidden unit values that wander about, allowing each layer in the network to train independently.

A dense layer is utilized when all the neurons in the layer above are connected to the neurons in the layer below, resulting in a fully connected layer.

To calculate the classes' probability distribution, the SoftMax unit is employed. The number of SoftMax to be utilized is determined by the number of classes to be used to categorize the emotions. It took between 8 and 12 hours to train the model. CPUs take a long time to train a model; however, GPUs can be employed to speed up the training process. The GPU's many cores increase speed and save time.

700 epochs were used to train the model.80 percent of the dataset is used for training, while 20 percent is used for testing. The model has been taught to differentiate between the 10 various classes of male and female voice. For this model the accuracy is approx. 89%. The below figures show the Model loss graph of it.

13. MODEL LOSS PLOT

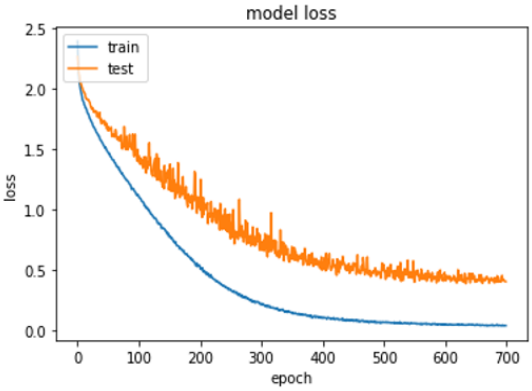

Figure 6: Loss vs. epoch plot

A Loss curve during training a popular chart for debugging a neural network. It offers a glimpse of the network's learning trajectory and training process.

The quantitative loss measure at the specified epoch is delivered by the loss function, which computes over all data items over the course of an epoch.

In other scenario, the model has 700 training epochs. 20% of the dataset is utilized for testing, while 80% of the dataset is used for training. The model has been taught to distinguish between the two types of male voice, namely male positive and male negative. For this model the accuracy is **93.75%** which is a very good result, and **93.37%** for 3 class positive, negative, neutral. The below figures show the emotion distribution and confusion matrix of it.

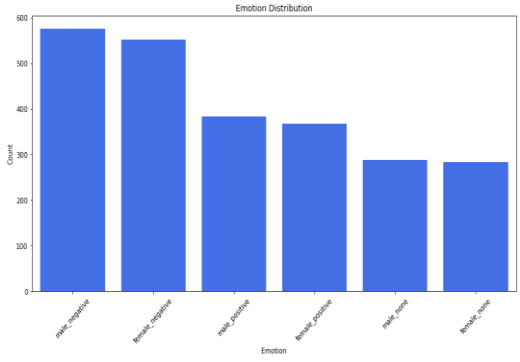

Figure 7: Emotion Distribution plot for two class (Positive, negative)

1148	male_calm	male_calm
1149	male_calm	male_calm
1150	male_calm	male_calm
1151	male_calm	male_calm
1152	male_calm	male_calm
1153	male_calm	male_calm
1154	male_happy	male_happy
1155	male_happy	male_happy
1156	male_happy	male_happy
1157	male_happy	male_happy
1158	male_happy	male_happy
1159	male_happy	male_happy
1160	male_happy	male_happy
1161	male_happy	male_happy
1162	male_sad	male_sad
1163	male_sad	male_sad
1164	male_sad	male_sad
1165	male_sad	male_sad
1166	male_sad	male_sad
1167	male_sad	male_sad
1168	male_sad	male_sad
1169	male_sad	male_sad

Figure 8: Actual values vs. predicted values for 10 class

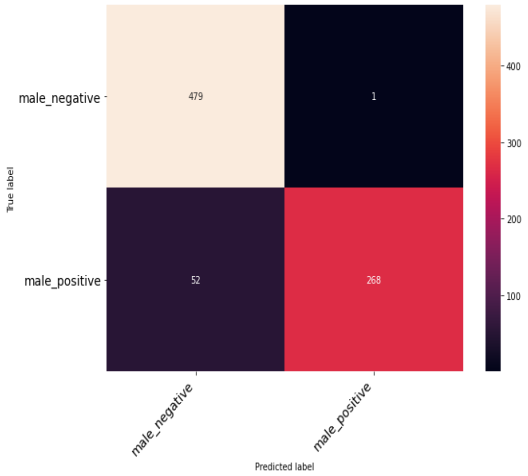

Figure 9: Confusion matrix plot for two class (Positive, negative)

14. NETWORK ARCHITECTURE COMPARISON

CNN is used in this section to perform SER on the MFCC as the number of convolutional layers increases. When compared to IEMOCAP and EMODB Datasets, CNN architecture based on LSTM which delivers the RAVDESS performance is at its peak when there are nine convolutional layers. The findings demonstrate that the amount and kind of the training data have a substantial impact on the best SER design, this finding is extremely important for developing SER systems on new datasets. Furthermore, we investigated CNN with more conv layers, but we did not see any improvement in UAR.

15. CONCLUSION

Automatic emotion recognition is a problem that has caused widespread concern and has an impact on the understanding of human behavior and interaction. High-dimensional characteristics on a curated data set are typically used in the creation of an emotion identification system. The disadvantage of this method is that the data set is limited, and it is difficult to analyze in the high-dimensional feature space. Our proposed system is based on a 9-layer CNN design which act as a CNN-LSTM Network, and the proposed work accuracy is 89.00%. Here, we use the North American English data set, the accuracy of the traditional system is 77%, so our proposed work is improved by approx. 12%.

ACKNOWLEDGEMENT

The key contribution of this study is to use a collection of data to train an end-to-end deep neural network model on the temporal and frequency domain information of raw voice files. Previous end-to-end learning models [16, 17] added some preprocessing to the input data and a chain of post-processing procedures to the network prediction. The suggested model, which employs information from the temporal and frequency domains, performs admirably in speech emotion recognition tasks. In fact, the addition of frequency domain data serves as a kind of data augmentation to address the issue of over-fitting caused by the training phase's limited data amount. Another noteworthy aspect of our neural network model is that it was trained on CNN layers in a manner similar to the state-of-the-art end-to-end emotion recognition model of Darshan K.A. [17], while with only the RAVDESS Dataset in this study.

REFERENCES

[1] Apoorv Singh, Kshitij Kumar Srivastava, Harini Murugan "Speech Emotion Recognition Using Convolutional Neural Network (CNN)" International Journal of Psychosocial Rehabilitation 2020.

[2] Ali Bakhshi and Aaron S. W. Wong and Stephan Chalup, "End-To-End Speech Emotion Recognition Based on Time and Frequency Information Using Deep Neural Networks" ECAI 2020.

[3] Tripathi, Samarth, and Homayoon Beigi. "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning." arXiv preprint arXiv:1804.05788 (2018).

[4] Chernykh, Vladimir, Grigoriy Sterling, and Pavel Prihodko. "Emotion recognition from speech with recurrent neural networks." arXiv preprint arXiv:1701.08071 (2017).

[5] Jerry Joy, Aparna Kannan, Shreya Ram, S. Rama "Speech Emotion Recognition using Neural Network and MLP Classifier'' IJESC, April 2020.

[6] Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang. "Automatic speech emotion recognition using recurrent neural networks with local attention." In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227-2231. IEEE, 2017.

[7] Darshan K.A, Dr. B.N. Veerappa "Speech Emotion Recognition" IRJET 2020.

[8] Christopher Olah, Understanding LSTM Networks — colahs blog, 2015.

[9] Sepp Hochreiter and J J Urgen Schmidhuber, LONG SHORT-TERM MEMORY, MEMORY Neural Computation, vol. 9, no. 8, pp. 1735 — 1780, 1997

[10] Min Seop Lee, Yun Kyu Lee, Myo-Taeg Lim, and Tae-Koo Kang, "Emotion Recognition Using Convolutional Neural Network with Selected Statistical Photoplethysmogram Features" Adpl 2020.

[11] Yuki Yamazaki, Masaya Tamaki, C. Premachandra, C. J. Perera, S. Sumathipala, B. H. Sudantha "Victim Detection

Using UAV with On-board Voice Recognition System" IEEE International Conference on Robotic Computing (IRC)2019.

[12] Panikos Heracleous ID, Akio Yoneyama "A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme" PLOS ONE 2019.

[13] k Ashok Kumar, J L Mazher Iqbal "Machine Learning Based Emotion Recognition using Speech Signal" (IJEAT) 2019

[14] Wisal Hashim Abdulsalam, Rafah Shihab Alhamdani, and Mohammed Najm Abdullah "Facial Emotion Recognition from Videos Using Deep Convolutional Neural Networks" International Journal of Machine Learning and Computing 2019.

[15] Yuanchao Li, Tianyu Zhao, Tatsuya Kawahara "Improved End–to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning" Interspeech 2019.

[16] Bagus Tris Atmaja, Masato Akagi." Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model" 2019

[17] Darshan K.A, Dr. B.N. Veerappa." Speech Emotion Recognition", International Research Journal of Engineering and Technology (IRJET), 2020.