Identification of Human Emotions from Speech Using CNN

Kunal Bhapkar[1], Krati Patni[2], Praddyumn Wadekar[3], Shweta Pal[4], Dr. Rubeena A. Khan[5], Mahesh Shinde[6]

[123456]*Department of Computer Engineering, Modern Education Society's College of Engineering, Pune-01*

---***---

Abstract - Emotion is something which is derived from the environmental factors. In the writing of discourse of Speech Emotion Recognition (SER), numerous methods have been used to extricate emotions from speech. In the proposed framework for the acknowledgment of emotions from speech CNN algorithm is used. To identify emotions, feature extraction and preprocessing are done so that the undesirable clamor gets filtered out. These separated features were carried forward and sent to the CNN classifier model. The dataset will go through the extraction system which would settle on a choice with respect to the basic emotion perceives in the sound.

***Key Words*:** Speech Emotion Recognition (SER), Convolution Neural Network (CNN), Feature Extraction, Preprocessing

1. INTRODUCTION

Many features of the human vocal system such as speech, tone, pitch and many others, convey information and context [1]. One of the most convenient and effective form of the exchange of the information is speech. Today, a lot of assets and endeavors are being placed into the advancement of artificial intelligence, and smart machines, for the key role of working on human existence. On the off chance that the machine can perceive the basic feeling in human discourse, it will bring both useful reaction and communication. Regardless, there are numerous issues in previous systems that should be appropriately tended to, especially as these frameworks move from lab testing to true application. Hence, efforts are needed to take care of such issues and accomplish better feeling acknowledgment by machines. In a paper, by Peng Shi, when compared to Artificial Neural Networks (ANNs) and support vector machines (SVMs), the Deep Belief Networks (DBNs) have about 5% higher accuracy rate than the traditional methods [2]. The yield shows that the features which are separated by Deep Belief Networks is obviously superior to the first element. J. Umamaheshwari and A. Akila outlined the process for preprocessing and feature extraction was explained [3]. M.S. Likitha et al. in their paper, discussed that the most widely used feature extraction method is Mel-cepstral coefficients (MFCC) [4]. Tian Kexin, Huang Yongming, Zhang Guobao and Zhang Lin, in their paper, outlined SVM (Support Vector Machine) for extracting the features [5]. Ye Sim Ulgen Sonmez and Asaf Varol, they discussed the growing scope of SERs like in the field of signal processing, pattern recognition, psychiatry [6]. There exists generous exploration work on improving the precision of the characterization results by carrying out various classifier models. SERs are trending in the Machine Learning field. This paper centers around the CNN algorithm for recognizing the emotion from speech. The proposed technique can be coordinated in existing frameworks. The rest of the paper is coordinated as follows. Section 2 peeks to the related works that are carried out regarding the same study topic. Area 3 gives an outline of the framework. In segment 4, the strategy executed in include extraction and classifier preparing is clarified. Segment 5 comprises the analysis of experimental results. In Section 6, we make results for the real time input. The conclusion based on the research is discussed in Section 7.

2. RELATED WORK

In the past few years, most of the papers use to extract different features such as MFCC (Mel Frequency Cepstral Coefficient), GLCM (Gray Level Co-occurrence Matrix), and LPCC (Linear Predictive Cepstral Model)[3][5]. Features such as instantaneous fundamental frequency (F0) using Zero Frequency Filtering (ZFF), signal energy, formant frequencies, and dominant frequencies are also used [7]. After extraction of these features, various methodologies are used such as Deep Belief Network (DBN), Support Vector Machine, Pattern Recognition Neural Network (PRNN), K-Nearest Neighbors (KNN), Hidden Markov Model (HMM), Artificial Neural Network (ANN), Bayesian Network (BN), Deep Neural Network (DNN) and Gaussian mixture models (GMMs) [1][2][6][8].

M.S. Likitha et al. in [4] recognized feeling, based on the training of its characteristics, like Sound, format, phoneme. Speech Emotion Recognition is also done for the unlabelled voice using t-distributed neighbor embeddings (t-SNE) to analyze visualizations of different representations [9].

The widest dataset used in papers is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Berlin Database of Emotional Speech. The German Copus (Berlin Database of Emotional Speech) has more than 250 audio files distributed into 20 milliseconds each. Radim Burget et al. in [10] used this dataset and removed 3098 silent segments to cut up the information into testing, validation, and training sets.

Unfortunately, we did not find any paper that used Convolutional Neural Network (CNN) to recognize speech emotions on four different datasets.

3. SYSTEM DESCRIPTION

The system has been trained on four different datasets, Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Survey Audio-Visual Expressed Emotion (SAVEE), Toronto Emotional Speech Set (TESS). Dataset collection is a task of priority and requires genuine scenarios [7]. To use these datasets together, a dataset integration was performed. Fig.1 depicts the overall concept of the proposed system.

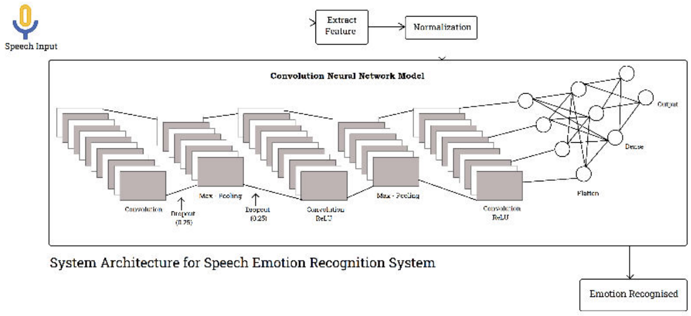


Fig.1. System Architecture

Feature is extracted from the integrated dataset and is split into training and testing parts. Deep neural network allows multi-layer models that learns representation of data with multiple levels and for this purpose CNNs are used [8]. Later, a model is saved considering the characteristics of voice and adding suitable numbers of ReLU activation layers to it.

The User uploads a voice clip in wav file format. To pass it further to the trained model, we extracted the features. The system categorizes the voice in suitable emotions such as sad, angry, happy, disgust, neutral, surprise and fear for male and female genders after loading the model.

The voice clip uploaded by the user is input data and is a critical step in emotion detection and recognition systems as it may be corrupted by unwanted noise [9]. Then, a wave plot portraying the pitch and frequency of voice is generated. The wave plot displays the sampling rate with respect to time. Also, it makes it easier to visualize and check whether the voice clip is successfully loaded.

4. METHODOLOGY

4.1. Dataset Description

In the proposed system, four different datasets are used. These are stated below:

- Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)
- Survey Audio-Visual Expressed Emotion (SAVEE)
- Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)
- Toronto Emotional Speech Set (TESS)

Name of Dataset	Total No. of files	Actors
RAVDESS	1440	24
SAVEE	480	4
CREMA-D	7442	91
TESS	2800	2

Table -1: Dataset Description

Table I shows the distribution of audio files among the actors for different datasets. These datasets consist of various actors speaking English statements in different emotions. It holds different emotions such as disgust, angry, happy, fear, sad, surprise and neutral. We segmented the files which can be given to the trained model resulting in 12162 segments.

4.2. Feature Extraction

A wave plot decides the type of sound. If system recognizes the sound, it maps it according to the emotion it had. MFCCs helps to represent sound correctly. It uses human belief to sound signals and sensitivity of sound to convert it into mel scale. MFCC is one of the features based on spectrum which is used to predict the correct emotion. In this phase, mean is used as a feature, one can use min and max, etc. as well. Since we have used four different datasets for the system, hence the feature extraction phase takes about an approximate of 10 minutes.

4.3 Preprocessing

After the feature extraction of each audio file, preprocessing is a mandatory step to be done. In the proposed system, we used data normalization for preprocessing purposes. Missing values are caused during data collection, by normalizing it, we fill NaN values to zero (0). There are much more options to it, e.g., Either it could be discarded or it can be replaced by mean via interpolation methods. Normalization is also used to improve the accuracy of model in our proposed system. It is also performed on training and testing data. To fit in the model, we expand the array shape of training and testing parts.

The most important preprocessing step is encoding the labels where the human-readable labels are converted in a machine-readable format. The model will consider these as a key-value, where new labels are assigned (say paired) with the old labels and hence model classifies accordingly.

4.4. Convolution Neural Network

In Deep Learning, Convolution Neural Network (CNN) is a type of technique which is used for classification. The most commonly classification technique used for data classification and pattern recognition CNN is used. It consists of three layers and two parameters which are Convolutional layer, Pooling layer, Fully-connected (FC) layer, dropout layer and activation function.

The purpose of the Convolution layer is to extract the features from the given input. The feature map is created from this layer and given to the next layer that is Pooling layer which is used to decrease the size of feature map. In the fully connected layer actual classification takes place after the both layers get executed. Dropout is used for regularization on whole feature vector before the classification takes place [10]. It drops some neurons during training from neural network which results in reduced size of model. Activation function is used to supply non linearity to the neural network.

In the proposed system, the addition of Convolution layer is done followed by Max Pooling layer. To supply non-linearity in the model, a ReLU Activation Function is used and to avoid overfitting Dropout layer is used which drops 25% of neurons from the model. Then this data is fed to Fully connected layer that is Dense Layer which predicts the labels.

5. EXPERIMENTAL ANALYSIS

For the proposed system, the performance is measured by finding out the accuracy, f-measure, recall and precision. The accuracy is calculated by finding out the ratio between the correctly predicted emotion to the total emotion given as an input. The model finds the emotion even for the real time input. In this model, the dataset is split into training and testing parts as 75% and 25% respectively. The emotion classes are classified as disgust, angry, happy, fear, sad, surprise and neutral for the respective gender.

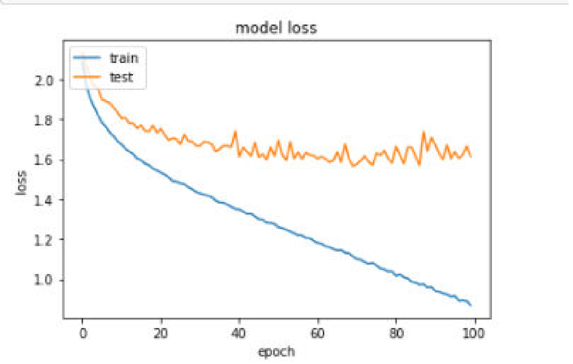


Fig.2. Model Loss

The model loss for training and testing part is shown in the fig. 2.

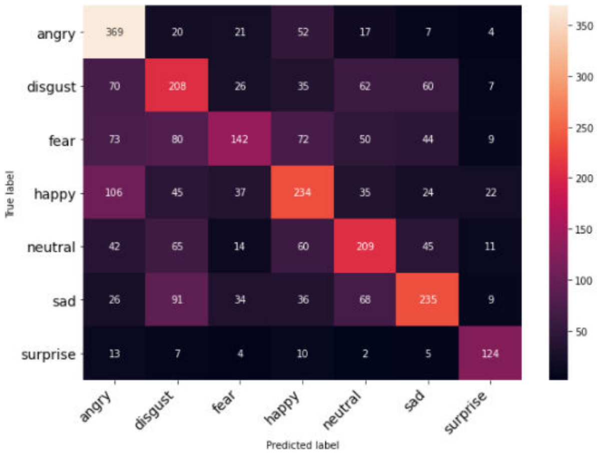


Fig.3. Confusion Matrix for individual emotion

Fig. 3 represents confusion matrix of individual emotion for the proposed model.

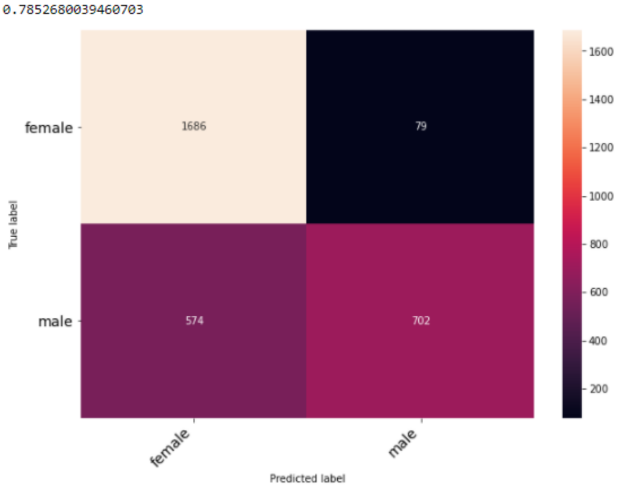


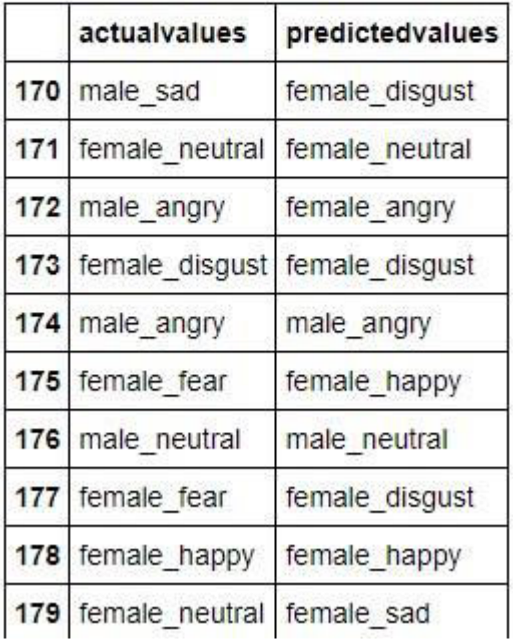
Fig.4. Confusion matrix on basis of gender

Fig. 4 stands for the confusion matrix for the overall model depending on the gender.

6. RESULTS

6.1. Accuracy Result

In order to recognize emotion from speech, 12162 samples are given as an input and a CNN model is built which is used to name the emotion. Table II shows 10 samples from the training dataset along with the actual and predicted emotions.



	actualvalues	predictedvalues
170	male_sad	female_disgust
171	female_neutral	female_neutral
172	male_angry	female_angry
173	female_disgust	female_disgust
174	male_angry	male_angry
175	female_fear	female_happy
176	male_neutral	male_neutral
177	female_fear	female_disgust
178	female_happy	female_happy
179	female_neutral	female_sad

Table -2: Predicted and Actual Emotions.

For the proposed model, overall accuracy is calculated using only one feature that is MFCC from the speech. Initially the accuracy is calculated for each emotion along with gender then the accuracy is found to be 44% only, but when the accuracy for each emotion is calculated then the accuracy becomes 50%. The final accuracy of model is calculated based on the gender distribution which is found to be 78.52%.

6.2. Real Time Input

For determining the emotion for real time voice input, user uploads an audio file and the emotion is predicted. For this unit, we have tested the system for three different audio files of individuals and the system predicted it with ease. Three various emotions, i.e., male_disgust, male_neutral, female_sad. As per the given input, system identified the correct emotions along with gender classification.
Below figures represents the wave plot uploaded by an individual.

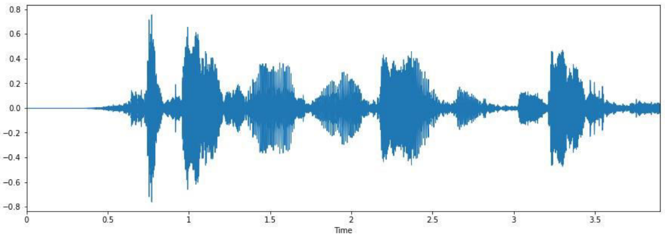


Fig.5. Waveplot of Audio file of actor 1.

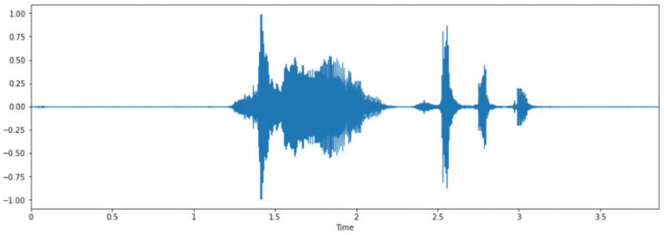


Fig.6 Waveplot of audio file actor 2.

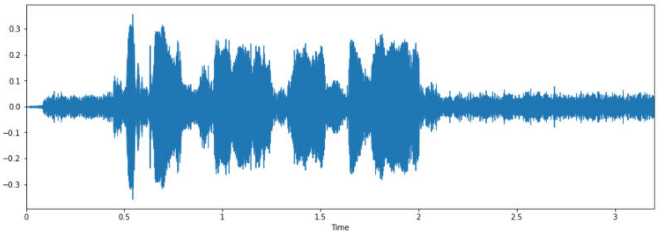


Fig.7 Waveplot of audio file actor 3.

7. CONCLUSION AND FUTURE SCOPE

The purpose of this paper is to find the emotion from a given audio file. Seven emotions for each gender are predicted as male-disgust, male-happy, male-angry, male-fear, male-sad, male-surprise, male-neutral, female-disgust, female-happy, female-angry, female-fear, female-sad, female-surprise and female-neutral using Mel frequency cepstral coefficients from audio file using four different datasets. The dataset is splitted into ratio of 3:1. To build the model various experiments are carried out by changing some features such as number of epochs, batch size and optimizer. The overall accuracy of the proposed system is found to be 78.52%. The performance of model is also calculated using precision, recall and f-score. The CNN model didn't own any knowledge about the context of what the actor says. The proposed system is open to modifications such as tweaking in epoch and batch-size, altering the optimizers from a range of options available such as SGD (Stochastic Gradient Descent), Adam and RMSprop (Root Mean Square Propagation). Since, the system has support of only one language i.e., English, it can be further trained to predict in another languages, based on standard dataset availability. Only wav files are supported by the system, but it could be made scalable when it also treats mp3, aac and other voice formats. For the purpose of study, one can build as many features to experiment on the accuracy of the results.

REFERENCES

[1] Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni, "Speech based Emotion Recognition using Machine Learning", Institute Of Electrical And Electronics Engineers, Mar. 2019.

[2] Peng Shi, "Speech Emotion Recognition Based on Deep Belief Network", Institute Of Electrical And Electronics Engineers, March 2018.

[3] J. Umamaheswari, A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN", Institute Of Electrical And Electronics Engineers, Feb 2019.

[4] Sri Raksha R. Gupta, M.S. Likitha, A. Upendra Raju and K. Hasitha "Speech Based Human Emotion Recognition Using MFCC", Institute Of Electrical And Electronics Engineers, March 2017.

[5] Tian Kexin, Huang Yongming, Zhang Guobao, Zhang Lin, "Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition", Institute Of Electrical And Electronics Engineers, Nov. 2019.

[6] Ye Sim Ülgen Sonmez, Asaf Varol, "New Trends in Speech Emotion Recognition", Institute Of Electrical And Electronics Engineers, June 2019.

[7] Esther Ramdinmawii, Abhijit Mohanta, Vinay Kumar Mittal, "Emotion recognition from speech signal", Institute Of Electrical And Electronics Engineers, Nov. 2017.

[8] PavolHarár, RadimBurget, Malay Kishore Dutta, "Speech emotion recognition with deep learning", Institute Of Electrical And Electronics Engineers, Feb. 2017.

[9] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, And Thamer Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review", Institute Of Electrical And Electronics Engineers, Aug. 2019.

[10] Michael Neumann, Ngoc Thang Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech", Institute of Electrical and Electronics Engineers, May 2019.