International Journal of Electrical and Electronics Engineers
Volume 15, Issue No. 01, Jan-June 2023
ISSN (O) 2321-2055
ISSN (P) 2321 -2045

Speech based Emotion Recognition using MachineLearning

[1] Palvadi SaiLakshmi, [2] Samudrala Lokesh, [3] Kasaragadda HymaNagaLakshmi, [4] Kondrathi Manikanta, [5] kanchupati Sasikanth , [6]Venkata Hanuman

[1,2,3,4,5] UG Students , [6]Associate Professor

Department of Electronics and Communication Engineering

Tirumala Engineering College, Jonnalagadda , Andhra Pradesh

ABSTRACT

Communication through voice is one of the main components of affective computing in human- computer interaction. In this type of interaction, properlycomprehending the meanings of the words or thelinguistic category and recognizing the emotion included in the speech is essential for enhancing the performance. In order to model the emotional state, the speech wavesare utilized, which bear signals standing for emotions such as boredom, fear, joy and sadness. This project is aiming to design and develop speech based emotional reaction (SER) prediction system, where different emotions are recognized by means of Convolutional Neural Network (CNN) classifiers. Spectral features extracted is Mel-Frequency Cepstral (MFCC). LIBROSApackage in python language is used to develop proposedalgorithm and its performance is tested on taking Ryerson Audio- Visual Database of Emotional Speech and Song (RAVDESS) samples to differentiate emotions such ashappiness, surprise, anger, neutral state, sadness, fear etc.Feature selection (FS) was applied in order to seek for the most relevant feature subset. Results show that the maximum gain in performance is achieved by using CNN.

Keywords—CNN, Audio Feature Extraction, LIBROSA,RAVDES, SER, MFCC.

I. INTRODUCTION

The human brain is an intricate organ that has been a lasting inspiration for research in Artificial Intelligence (AI). The neural networks in brain had the capability of learning all concepts from experiencing low level information and is remembers them which are processed by sensory periphery Theapproach for speech emotion recognition (SER) primarily comprises two phases known as feature extraction and featuresclassification phase. The first phase Feature extraction is the key part in the Speech Emotion Recognition. The quality of thefeatures directly influences the accuracy of classification results. Typically, the Feature Extraction method designs handcraft features based on acoustic features of speech. The second phase includes feature classification using linear and non-linear classifiers. The most commonly used linear classifiers for emotion recognition include the

Maximum Likelihood Principle (MLP) and Support Vector Machine(SVM) and Convolution Neural Network (CNN). Usually, the speech signal is considered to be non-stationary. Hence, it is considered that non-linear classifiers.

T. Pao, C. Wang and Y. Li, in their paper, dated 2012 discussed 78 features extractable from a speech signal and classified a 13 feature set as being most suitable for a particularclassifier[1]. M. S. Likitha, S. R. R. Gupta, K. Hasitha and A.

U. Raju, outlined the process to determine the MFCC coefficients and checked deviation to determine amongst 3 emotions[2]. In the paper, by Chen and Luo, a text-dependent speaker verification system was implemented with the purposeof recognizing an imposter voice against an authentic user[3]. This was done through training the SVM with the help of speaker model and imposter model after extracting MFCC coefficients from the password spoken by the user.

Here, Xinzhou Xu [3] et al generalized the Spectral Regressionmodel exploiting the joins of Extreme Leaning Machines (ELMs) and Subspace Learning (SL) was expected for overlooking the disadvantages of spectral regression based Graph Embedding (GE) and ELM. Using the GSR model, in theexecution of Speech Emotion Recognition (SER) we had to precisely represent theses relations among data. The system output can be improved by exploring embedded graphs at more precise levels. Only Least-Square Regression along with l2-norm minimization wasconsidered in the regression stage. Zhaocheng Huang[4] et al uses a heterogeneous token-used system to detect the speech depression. Abrupt changes and acoustic areas are solely and collectively figured out in joins among different embedding methods. Contributions towards the detection of depression were used and probably various health problems that would affects vocal generation. Landmarks are used to pull out the information particular to individual type of articulation at a time. This is a hybrid system. LWs and AWs hold various information. AW holds section of acoustic area into single token per frame, and on the contemporary the abrupt changes inspeech articulation are shown by LWs. The hybrid join of the LWs and AWs permits exploitation of various details, more specifically, articulatory dysfunction into conventional acoustic characteristics are also incorporated. Peng Song [5] offers Transfer Linear Subspace Learning (TLSL) framework for cross corpus recognition of speech. TLSL approaches, TULSLand TSLSL were taken in count. TLSL aims to extract robust characteristics representations over corpora into the trainedestimated subspace. TLSL enhances the currently used transferlearning.

II. SYSTEM DESCRIPTION

Fig. 1 illustrates the overall system. The system is divided broadly into dataset formation, pre-processing, feature extraction and classification. The entire system is programmedusing MATLAB R2014a. The Ryerson Audio-VisualDatabaseof Emotional Speech and Song (RAVDESS)[4],

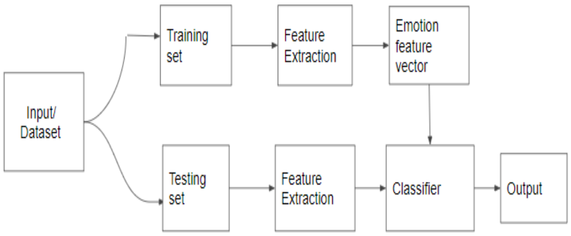
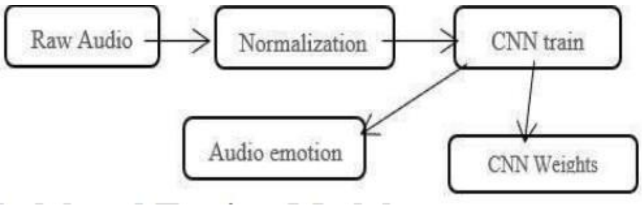


Fig. 1. System block Diagram

III.METHODOLOGY

3.1 Convolution Neural Network: The speech emotion recognition application is executed using CNN. Followingis the architecture of the system:



3.2 Training Model and Testing Model: A training data isfetched to the system which consists the expression label andWeight training is also provided for that network. An audiois taken as an input. Thereafter, intensity normalization is applied over the audio. A normalized audio is used to train the Convolutional Network, this is done to ensure that the impact of presentation sequence of the examples doesn't affect the training performance. The collections of weights come out as an outcome to this training process and it acquires the best results with this learning data. While testing, the dataset fetches the system with pitch and energy, and based on final 9network weights trained it gives the determined emotion. The output is represented in a numerical

value and type of Emotion.

Algorithm (Training):

1. Loading the data, dividing it in train and test
2. Using the LIBROSA, a python library we extract theMFCC (Mel-Frequency Cepstral Coefficient
3. There after constructing a CNN model to train the dataset
4. Save Model

Testing and prediction

- The sample audio file(.wav) is provided as input.
- Using the LIBROSA, a python library we extract theMFCC(Mel Frequency Cepstral Coefficient)
- Predicting the human voice emotion from that traineddata(predicted value)

identifier (e.g., 03-01-05-01-01-01-01.wav) 60.

Audio Feature Extraction:

The Shape of the Speech signal determines what sound comes out. If the shape is determined accurately, then the correct representation of the sound being generated is obtained. The job of Mel Frequency Cepstral Coefficients' (MFCC's)is to correctly represent it. MFCCs is used as input feature. Loading and converting audio data into MFCCs format is done by python package librosa.

LIBROSA:

Librosa is a Python package for music and audio analysis. Librosa is basically used when we work with audio data like in music generation (using LSTM's), Automatic Speech Recognition. It provides the building blocks necessary to create the music information retrieval systems. Librosa helps to visualize the audio signals and also do the feature extractions in it using different signal processing techniques. It is the starting point towards working with audio data at scale for a wide range of applications such as detecting voicefrom a person to finding personal characteristics from an audio.

MEL-FREQUENCY CEPSTRUM :

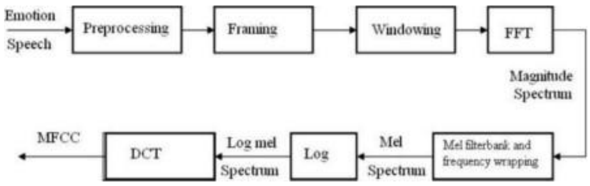
Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal spectrum. This frequency warping can allow for , better representation of sound, for example, in audio compression. MFCCs are commonly used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken into a telephone. MFCCs are also increasingly finding uses in music information retrieval applications such as genre classification, audio similarity measures, etc

To train the model for accuracy calculation:

Within this module we train the model for accuracy estimations. First, import necessary modules and then import the dataset. We will receive the sampling rate value with librosa packages and MFCC function. Thereafter this value holds other variables. Now audio files and MFCC value holda variable consequently it will add a list. Then zip the list and hold two variables x & y. Then we have represented (x, y) shape values with the use of numpy package.

Dataset

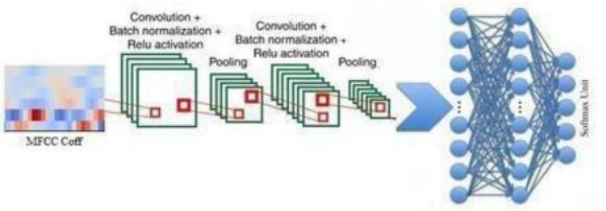
We are making use of RAVDESS dataset. It is downloaded from kaggle.com. It holds "12,800 files: 1800 audio- files/emotion multiplied with 7 different emotions = 12800 trials". The RAVDESS consists of 24 professional voices (12 feminine, 12 masculine), speaking2 lexically- matched sentences in the even North-American accent. Happy, sad, angry, fearful, calm, disgust and surprise are the variousspeech emotion expressions used. Every expression is generated in 2 levels of emotional intensity (light, bold), witha neutral expression. Every file out of 12800 files has an unique filename. The filename holds a 7-part



Implementation process of CNN model:

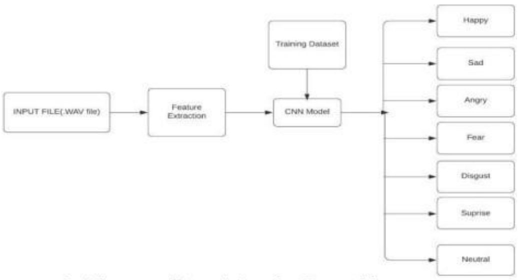
The deep neural network architecture actualized is convolutional neural network. In the proposed architecture

after each convolutional layer max-pooling layer is placed. To establish non linearity in the model, for activation function Rectified Linear Units (ReLU) is used in both convolutional and fully connected layers. Batchnormalization is used to improve the firmness of neural network, which normalizes the result of the preceding activation layer by reducing the number by what the hidden unit values move around and allows each of the layer in a network to learn by itself. Dense layer is used; in which all the neurons in a layer are connected to neurons in the next layer and it is a fully connected layer. SoftMax unit is used to compute probability distribution of the classes. The number of SoftMax to be used depends on number of classesto classify the emotions.


Convolution + Batch normalization + Relu activation
Pooling
Convolution + Batch normalization + Relu activation
Pooling
MFCC Coff
Softmax Unit

In our CNN model we have four important layers: • Convolutional layer: Identifies salient regions at intervals, length utterances that are variable and depicts the feature map sequence. • Activation layer: A non-linear Activation layer function is used as customary to the convolutional layer outputs. In this we have used corrected linear unit (ReLU) during our work. • Max Pooling layer: This layer enables options with maximum value to the Dense layers. It helps tokeep the variable length inputs to a fixed sized feature array.

• Dense layer: In Dense layer, for the perceptron we assign each node with input, weight and bias and we compute the function f(x)=activation (Weights * Input + bias). Before allthis we perform ravel for data flattening. After computing thefunction for each node its output is given as input for the next layer or hidden layer and so on.. till the final output is received and then any of the activation functions are used and this output is compared to the real o input and we find the error of the output. Now we apply backpropagation to set the weights so that the error decreases resulting in good accuracy of the model. This isnothing but epoch.



Architecture of Speech Emotion Recognition

Inference

The various Automatic speech recognition (ASR) in noise surrounding the person requires a multichannel improvement of speech with a mic array. Using the beam formation, the multichannel speech improvement can be approached. We can lay focus on speech that comes from one direction and noise is cancelled from the other direction which are basically the spatial information. This approach improved the results ofimprovised ASR in Chime Challenge with the help of this approach. There are many varieties of beam forming for e.g. Minimum Variance Distortion-less Response (MVDR), Multi-Channel Wiener Filtering (MWF), Generalised Side Lobe Cancelling (GSC) and Generalized Eigen-Value (GEV), performed at the time-frequency (TF) domain. Although DNN-based beam forming performs good in handful and regulated in demonstration environments, it possesses two big issues in world. Firstly, due to over fitting to the training data having many pairs of noisy and disturbed speech spectrograms and Ideal Binary Mask (IBM) have resulted in low performance of ASR under unknown environments. Secondly the physical meanings and generative processes of characteristics such as Inter-Channel Level and Phase Variants (ILDs and IPDs) are not taken intoconsideration and they are kept simply as an in input to DNNs.

Modules

In our CNN model we have four important layers:

1.Convolutional layer: Identifies salient regions at intervals, length utterances that are variable and depicts the feature mapsequence.

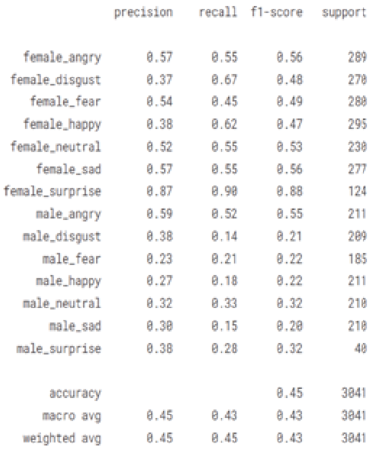
2. Activation layer: A non-linear Activation layer function is used as customary to the convolutional layer outputs. In this we have used corrected linear unit (ReLU) during our work
3. Max Pooling layer: This layer enables options with the maximum value to the Dense layers. It helps to keep the variable length inputs to a fixed sized feature array.
4. Dense layer Audio Feature Extraction and Visualizations. (module01) Characteristics extraction is required for classification and depiction.

The audio signal is a 3D signal in which 3 axes indicate time, amplitude and frequency. We will use librosa to analyze and extract characteristics of any audio signal. (.load) function pulls an audio file and decrypts it into a 1D array which is of time series x, and SR is actually sampling rate of x. By default SR is 22 kHz. Here I will show one audio file display with the use of (IPython.display) function. Librosa.display is important to represent the audio files in various forms i.e. wave plot, spectrogram and colormap. Wave plots use loudness of the audio at a particular time. Spectrogram displays various frequencies for a particular time with its amplitude.

.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Accuracy was calculated for the classification of emotions by mode and mean method by using American english speech corpus. Here, 80% of the database was given to training set and 20% to testing. Accuracy for real time input was also calculated. Regional language dataset in Hindi and Marathi languages was created by recording the audio input of speakers in the age range 18-25. The emotion classes were anger, happiness, and sadness.



```
                  precision    recall  f1-score   support

    female_angry       0.57      0.55      0.56       289
  female_disgust       0.37      0.67      0.48       270
     female_fear       0.54      0.45      0.49       280
    female_happy       0.38      0.62      0.47       295
  female_neutral       0.52      0.55      0.53       230
      female_sad       0.57      0.55      0.56       277
 female_surprise       0.87      0.90      0.88       124
      male_angry       0.59      0.52      0.55       211
    male_disgust       0.38      0.14      0.21       209
       male_fear       0.23      0.21      0.22       185
      male_happy       0.27      0.18      0.22       211
    male_neutral       0.32      0.33      0.32       210
        male_sad       0.30      0.15      0.20       210
   male_surprise       0.38      0.28      0.32        40

        accuracy                           0.45      3041
       macro avg       0.45      0.43      0.43      3041
    weighted avg       0.45      0.45      0.43      3041
```

Fig. 3 represents signal on applying pre-emphasis high pass filter.

Fig. 4 represents de-silenced signal. When we compare pre-emphasized signal with de-silenced one, we realise that the silence part of signal at the beginning as well as end is removed and hence the whole signal is shifted towards the Y- axis.

Fig. 5 represents plot of a single frame. It is observed that there are 1200 samples in the frames which matches with the calculated value.

Fig. 6 represents the frame after applying hamming window. It can be observed that the shape is similar to that of the hamming window with maximum attenuation towards the endsand minimum at the centre.

Fig. 7 represents the Mel filter bank of 26 filters overlappedon each other.

Fig. 8 represents the plot of the extracted energy signal (short-term energy) of an input audio signal.

Fig 1.3:Confusion matrix:

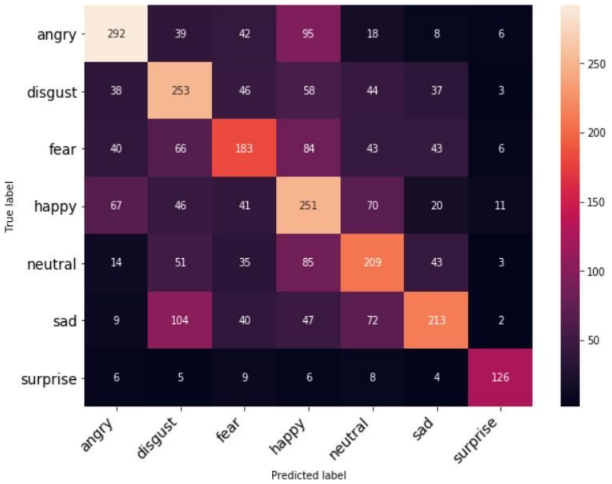


Figure 1.1: Audio Frequency

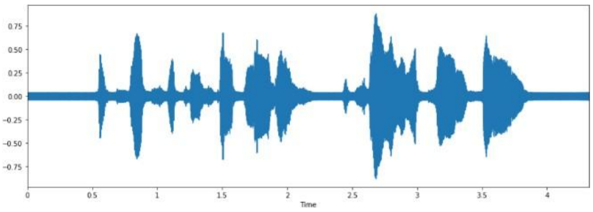
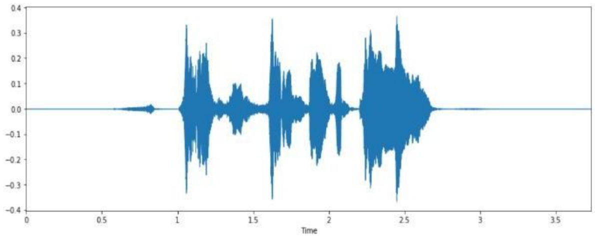


Figure1.2: Pre-Emphasis Engery signal:



Aftertraining various models it came out with the most optimum accuracy of 82% with SoftMax activation layer, “rmsprop” activation layer,18 layers, and with 1000 epoc.

V. CONCLUSION

In this paper, three emotions- anger, happiness, and sadness, were classified using three feature vectors. Pitch, Mel-frequencycepstral coefficients, Short Term Energy were the three featurevectors extracted from audio signals. Open source North American English acted speech corpus and recorded natural speech corpus were used as input. The data set used for trainingand testing consisted of audio samples in male and female voice and divided in ratio 4:1.

The mean method provided greater accuracy over the modemethod. Mean of a set of data values takes into consideration every value present while there can be two or more values at the highest frequency.

The classification accuracy for all three emotions was foundto have increased by 20% by using three features as against using two features. Anger and happiness emotions classificationaccuracy increased by 15%-20% with the help of STE feature vector. The emotion sadness did not improve its accuracy of classification despite using STE feature vector. Sadness is more susceptible in being misclassified as happiness emotion. Happiness is misclassified as angry due to close values for STE,and sadness due to its lower pitch.

After constructing various models, we got the better CNN model for the emotion distinction task. We reached 71%accuracy from the previously available model. Our model would’ve performed better with more data. Also our model performed very well when distinguishing among a masculine and feminine voice. Our project can be extended to integrate with the robot to help it to have a better understanding of the mood the corresponding human is in, which will help it to havea better conversation as well as it can be integrated with 73 various music applications to recommend songs to its users according to his/her emotions, it can also be used in various online shopping applications such as Amazon to improve the product recommendation for its users. Moreover, in the upcoming years we can construct a sequence to sequence modelto create voice having different emotions.

Classification of emotions for regional Indian languages namely Hindi and Marathi was implemented. The accuracy of classification of real time input audio for regional language Hindi was obtained 100%. Further, accuracy could be affected according to the natural tone of speaker. Analysis on the basis of age group of speakers is another area of future work.

VI. REFERENCES

[1] T. Pao, C. Wang and Y. Li, "A Study on the Search of the Most Discriminative Speech Features in the Speaker Dependent Speech Emotion Recognition." *2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming*, Taipei, 2012, pp. 157-162.

[2] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC." *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, 2017,pp. 2257-2260.

[3] Chen, S.-H. and Y.-R. Luo (2009), “Speaker Verification Using MFCC and Support Vector Machine,” in *Proceedings ofthe International Multi-Conference of Engineers and ComputerScientists*, March 18 - 20, 2009, Vol I IMECS 2009.

[4] Livingstone SR, Russo FA (2018) 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 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[5] Thomas Zawistowski & Paras Shah, “An Introduction to Sampling Theory.”Internet: http://www2.egr.uh.edu/~glover/applets/Sampling/Sampling.h tml, [Feb.24, 2019].

[6] Akhilesh Chandra Bhatnagar, R. L. Sharma, Rajesh Kumar, “Analysis of Hamming Window Using Advance Peak Windowing Method.” *International Journal of ScientificResearch Engineering &Technology,* vol.1 issue 4, pp 015- 020, July 2012.

[7] Practical Cryptography, "Mel Frequency CepstralCoefficients (MFCC) tutorial. "Internet:http://practicalcryptography.com/miscellaneous/machi ne-learning/guide-mel-frequency-cepstral-coefficients-mfccs/, [Feb.27, 2019].

[8] Naotoshi Seo, "Project: Pitch Detection." Internet:http://note.sonots.com/SciSoftware/Pitch.html, [Feb.25, 2019].

[9] Vocal Technologies, "Pitch Detection using Cepstral Method." Internet: https://www.vocal.com/perceptual- filtering/pitch detection/, [Feb.25, 2019].

[10] Sunil Ray, Analytics Vidhya, “Understanding SupportVector Machine algorithms from examples.” Internet:https://www.analyticsvidhya.com/blog/2017/09/understaing- support-vector-machine-example-code/, Sept.13, 2017 [Mar.10, 2019].

[11] Institute for Digital Research and Education, UCLA, "What is the difference between categorical, ordinal and interval variables?" Internet: https://stats.idre.ucla.edu/other/multpkg/whatstat/what -is-the-difference-between-categorical-ordinal-and-interval- variables/, [Feb.27, 2019].

[12] Laerd Statistics, "Measures of Central Tendency." Internet:https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php, [Feb.27, 2019].