JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Human Emotion Detection and Identification using Voice Signals

[1]Rutuparna Kudtarkar, [2]Keyur Modh, [3]Bishal Nakoda, [4]Nilambari Narkar

[1]Student, [2]Student, [3]Student, [4]Supervisor
[1]Computer Engineering,
[1]Affiliation (Computer Engineering, Xavier Institute of engineering, Mumbai University)

Abstract : In recent years, the personal touch among people has been decreased and has been reducing furthermore due to Covid-19 pandemic. It has become much more difficult to understand what the other person might be emotionally going through. Emotions assume a critical role in human life. It is an instrument of explanation of one's point of view or one's mental state to other people. Emotions can be perceived from one’s body language, facial expressions and their voice. So, the primary objective for developing this system is to recognise the emotions a person is experiencing just from his/her speech. Thus, the aim is to create Speech Emotion Recognition by using CNN. Essentially, Speech Emotion Recognition can help to further develop a man-machine interface. Speech Emotion Recognition can be defined as extraction of the emotional state of the speaker from his or her speech signal. With the fruitful use of profound learning in the fields of picture and speech recognition, researchers have begun to attempt to utilize it in Speech Emotion Recognition and have proposed many profound learning-based Speech Emotion Recognition calculations. There are few universal emotions- such as Normal, Angry, Sad, Happy in which any intelligent framework with limited computational assets can be trained to distinguish as required. Speech Emotion Recognition can help to solve this particular problem statement by recognizing the emotions of a person by utilizing their vocal expression. Speech recognition, as one of the media of human-computer interaction, has gradually become the key with the progress and development of technology in that area.

***IndexTerms* - CNN, Data augmentation, Speech Emotion Recognition.**

1. INTRODUCTION

Emotions play an important role in a person’s life as it is one's viewpoint or one's psychological state to other people. There are many universal emotions such as Neutral, Angry, Sad, Happy, etc. Emotions are generally spontaneous and are weakly communicated, eclectic and are difficult to recognise from one another. In the literature, emotional statements are termed as positive and negative based on how it is expressed by a character. Other experiments imply that the listener-based acted emotions are much firmer and precise than natural emotions, which may indicate that actors overemphasize emotional dialogues. With the help of Speech Emotion Recognition, systems can be developed to improve Human Computer Interaction. Thus, it has a wide variety of applications like in games, augmented reality, virtual reality, etc. Usually, Speech Emotion Recognition can be divided into two phases: feature extraction and feature classification. But, the model that is proposed has 3 phases: data augmentation, feature extraction and feature classification.

2. RESEARCH METHODOLOGY

2.1. SPEECH EMOTION RECOGNITION USING LSTM

To manage pointless data in non-speech voice fragments, two methodologies were proposed. The first was to utilize a voice activity detector (VAD) to eliminate the quiet piece of the speech. The second way to deal with i.e. to manage a quiet voice piece is to overlook it as opposed to eliminating it. The attention model in deep learning is one that attempts to overlook quietness, outlines, and different pieces of the content which don’t convey emotions. To start with, the model eliminates quiet inside speech sound, then includes extraction from those voice fragments. The acquired components are trained with neural networks to further classify emotions. Attention model is included in this network to track down the significant yield just on the past layer. At last, the assessment is given by contrasting execution from entire speech and speech only fragments, with and without attention model. The common primary blocks of the pattern recognition work process including the Speech Emotion Recognition task comprise two stages: feature extraction and classification. The commitment of the proposed technique is the expansion of silence elimination before including extraction of features and the utilization of the attention model with bidirectional long short-term memory network (LSTM) for classification.

2.1.1 ADVANTAGES AND DISADVANTAGES OF LSTM

2.1.1.1 ADVANTAGES

- LSTMs provide an enormous scope of boundaries like learning rates, and input and output biases.
- Bidirectional Long Short-Term Memory Recurrent Neural Network (BLSTM-RNN) has been demonstrated to be exceptionally viable for displaying and foreseeing sequential information.
- Non-decaying error backpropagation.

2.1.1.2 DISADVANTAGES

- BLSTM has double LSTM cells, which makes it costly.
- Training time is very slow, also not very interpretable.

2.2. SPEECH EMOTION RECOGNITION USING DEEP LEARNING

Deep Learning methods have been as of lately proposed as an option to conventional methods in Speech Emotion Recognition. The paper discussed here presents an outline of Deep Learning methods. Deep Learning has been viewed as an arising research field in AI and has acquired consideration lately. Deep Learning strategies for Speech Emotion Recognition have a few upper hands over conventional techniques, including their capacity to identify the perplexing construction and provisions without the requirement for manual feature extraction and tuning propensity toward extraction of low-level features from the given raw information, and the capacity to manage unlabeled information. Deep learning techniques contain different nonlinear parts that perform calculation on a parallel basis. However, these strategies should be organized with more deeper layers of design to defeat the constraints of different methods. Ongoing advancement brings about image recognition and speech recognition have created a gigantic interest in this field in light of the fact that applications in numerous different areas giving enormous information appear to be conceivable. On a disadvantage, the numerical and computational strategy underlying deep learning models is exceptionally difficult, particularly for interdisciplinary researchers.

2.2.1 ADVANTAGES AND DISADVANTAGES OF DEEP LEARNING

2.2.1.1 ADVANTAGES

- Flexible with changing voice characteristics and the system is robust.
- Fast to train the model.

2.2.1.2 DISADVANTAGES

- The quality of generated speech can be degraded.
- Acoustic features can get over-smoothed, making speech sounds muffled.

2.3. EXISTING SYSTEM

Currently, the majority of Speech Emotion Recognition models use Gaussian mixture model, K- Nearest Neighbours or Hidden Markov Model.

2.3.1 GAUSSIAN MIXTURE MODEL

A Gaussian mixture model is a probabilistic model that expects every one of the data points that are produced from a combination of a limited number of Gaussian distributions with obscure specification. In the Gaussian mixture model, all the training and testing equations are based on the supposition that all vectors are independent therefore the Gaussian mixture model cannot form a temporal structure of the training data. A maximum accuracy of 78.77% could be achieved using the Gaussian mixture model.

2.3.2 K-NEAREST NEIGHBORS

K-Nearest Neighbors calculates a way to deal with data classification that gauges how possible a data point is to be an individual from one group or the other relying upon what group the data points closest to it are in. In K- Nearest Neighbours, according to the emotional state of the k utterances, the classifier allocates an utterance to an emotional condition. The classifier can classify all the utterances in the design set properly, if k equals to 1, however its performance on the test set will be reduced. Utilizing the information of pitch and energy contours, the K- Nearest Neighbours classifier attains an accurate classification rate of 64% for four emotional states.

2.3.3 HIDDEN MARKOV MODEL

Hidden Markov Model is a statistical Markov model in which the system being modeled is assumed to be a Markov process with non-observable/hidden states. The Hidden Markov Model is generally used in Speech Emotion Recognition because of its physical relation with the speech signals production mechanism. For speech emotion recognition, typically a single Hidden Markov Model is trained for each emotion and an unknown sample is classified according to the model which illustrates the derived feature sequence better. Hidden Markov Model has the important advantage that the temporal dynamics of speech features can be caught second accessibility of the well established procedure for optimizing the recognition framework. The main problem in building the Hidden Markov Model based recognition model is the features selection process. Because it is not enough that features carries information about the emotional states, but it must fit the Hidden Markov Model structure as well.he accuracy rate of the speech emotion recognition by using Hidden Markov Model classifier is observed as 76.12%

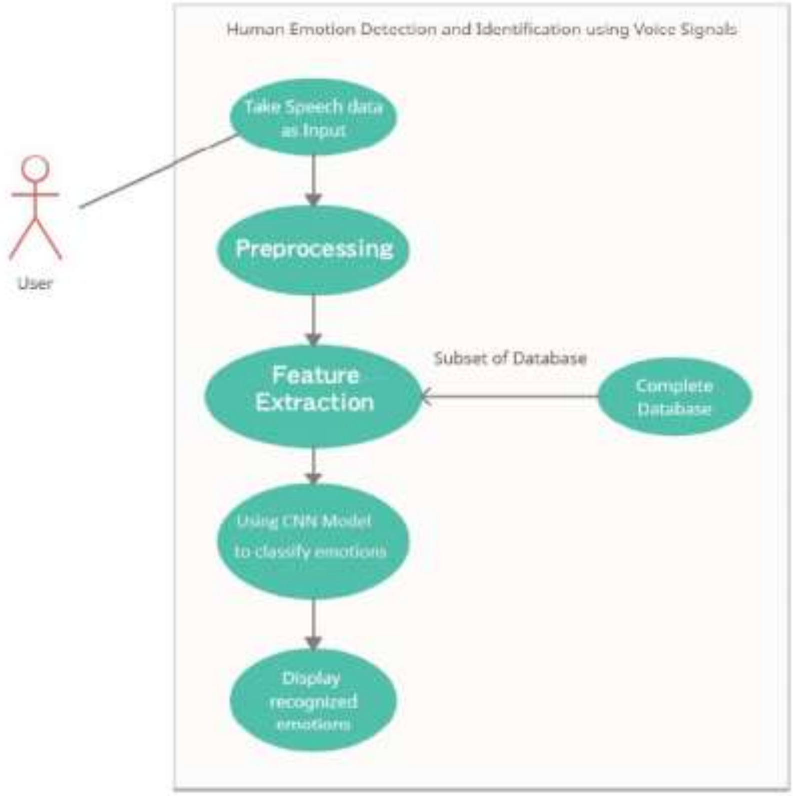
3. Theory and Calculation

3.1 Analysis

In the previous sections, problems have been clearly defined and surveyed the existing literature and approaches.

3.1.1 Use Case Diagram

Figure 1. Use Case Diagram of Proposed System


Human Emotion Detection and Identification using Voice Signals
Take Speech data as Input
Preprocessing
Feature Extraction
Subset of Database
Complete Database
Using CNN Model to classify emotions
Display recognized emotions
User

3.1.1 Feasibility Study

- Cost of data acquisition

– How difficult is it to acquire data?
– How much data will be required?
– How expensive is data labelling?

- Cost of wrong Prediction

– How frequently does the system need to be right to be useful?

- Accessibility of good published writeups about comparable issues

– Has the issue been decreased to rehearse?
– Is there sufficient Literature on the problem?

- Computational Resources available for both training and inference

– Will the model be sent in a resource-constrained platform?

3.1.2.1. Cost of Data Acquisition

- Ready made datasets were available on the internet. Thus, the available datasets known as CREMA-D (an emotional multimodal actor data set of 7,442 original clips from 91 actors), Surrey Audio-Visual Expressed Emotion (SAVEE) database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total, Toronto emotional speech set (TESS) consists of recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral) there are 2800 stimuli in total, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent.

- The golden rule of Machine Learning is that the more the data, the more accurate the results will be. So, to increase the size of the dataset, data augmentation has been implemented which would change the pitch, tone, etc.

3.1.2.2. Cost of wrong Prediction

- For the Speech Emotion Recognition project, the metrics which are used to judge the results is how accurate the prediction is based on the output given in above mentioned datasets.
- Still from the perspective of probabilistic thinking, the algorithm should be correct 7/10 times to satisfy the metric expressed previously
- Also, the performance of the model has been judged based on loss and accuracy of the model.

3.1.2.3. Availability of good published work about similar problems

- Currently, it has been implemented to some extent, but the issue is the insufficiency of the research in some key areas related to this.
- As the research is still at this point in its beginning, the associated literature in such a niche area is also limited.

3.1.2.4. Computational Resources available for both training and inference

- The previous works and research in this area, all had an issue that they required a lot of data as well as computational resources. The attempt has been made to improve the algorithm to work in an resource constrained computational environment just as giving acceptable outcomes even with little datasets.

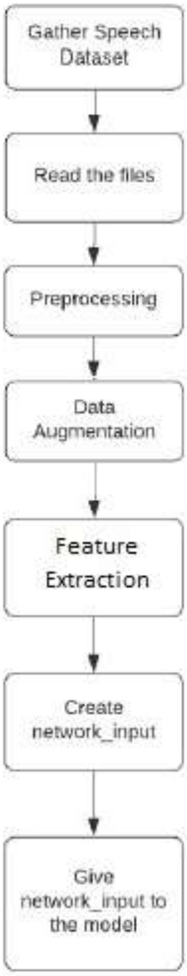
3.2 Design

3.2.1 Flowchart

In this section, the high-level steps involved during the various phases of the model using flowcharts have been explained.

3.2.1.1 FLOWCHART DETAILING THE TRAINING PROCESS

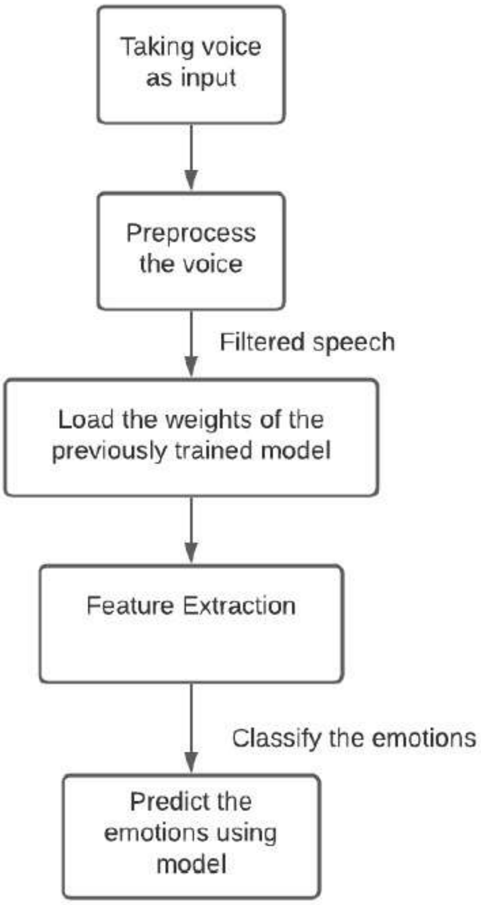
Figure 2. Flowchart detailing the training process



In simple terms, speech as an input is taken then pre-processed i.e removing noise and silence, etc. Then the improved and processed speech is used to extract features from speech. After extracting relevant features, these extracted features have been given to the model for training purposes.

3.2.1.2 Flowchart detailing the Human Emotion Detection and Identification using Voice Signals

Figure 3. Flowchart detailing the Human Emotion Detection and Identification using Voice Signals



In Speech Emotion Recognition, the first step is to input the voice signal. Afterwards, pre-process the voice signal then use the pre-trained model to predict the emotions.

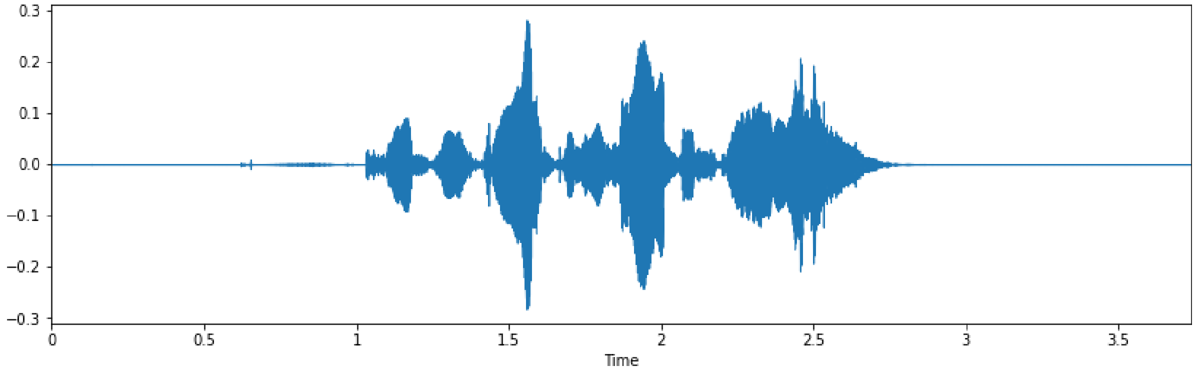
3.3 Explanation of the Algorithm

Three phases have been considered below for the algorithm namely Data Augmentation, Feature Extraction and Classification. Five emotions have been considered i.e. angry, calm, happy, sad and surprised.

3.3.1 Data Augmentation

Data augmentation is the cycle by which new data sets are introduced by including little annoyances on the initial training set. To generate synthetic data for audio, try changing pitch and speed, injecting noise, and time shifting. The goal is to make the model invariant to those irritations and upgrade its capacity to generalize. In order for this to work adding the perturbations must conserve the same label as the original training sample. In this project, data augmentation methods implemented are as follows:

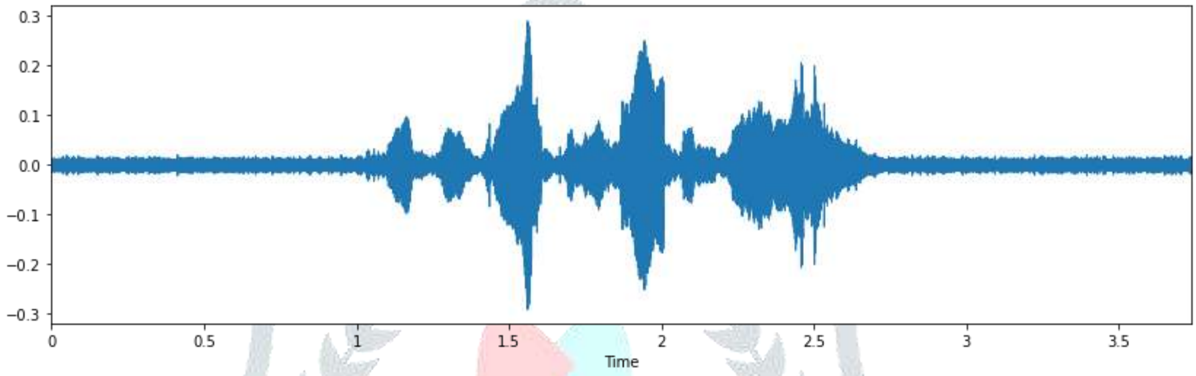
Figure 4. Simple Audio



a. Noise Injection

The technique for injecting noise alludes to adding noise misleadingly to the ANN input information during the training phase.

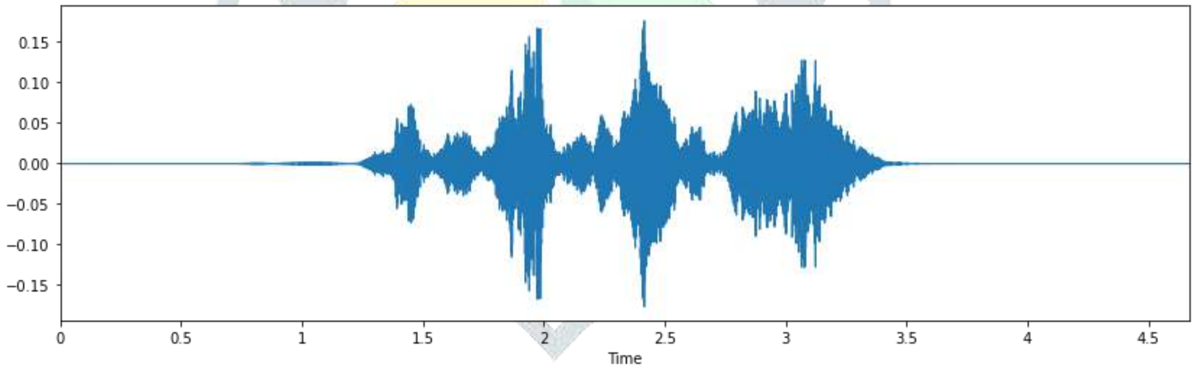
Figure 5. Applying Noise Injection



b. Stretching

Time-stretch a sound series by a decent rate.

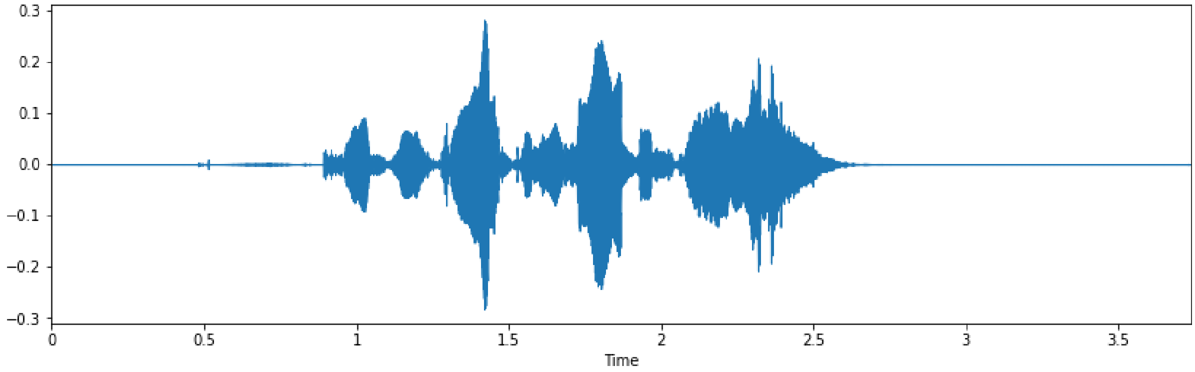
Figure 6. Applying Stretching



c. Time Shifting

It just shifts audio to left/right with a random second.note that at the same time, its qualities are generally not modified. This implies that the time-moving activity brings about the difference in the situation of the signal without influencing its abundance or range.

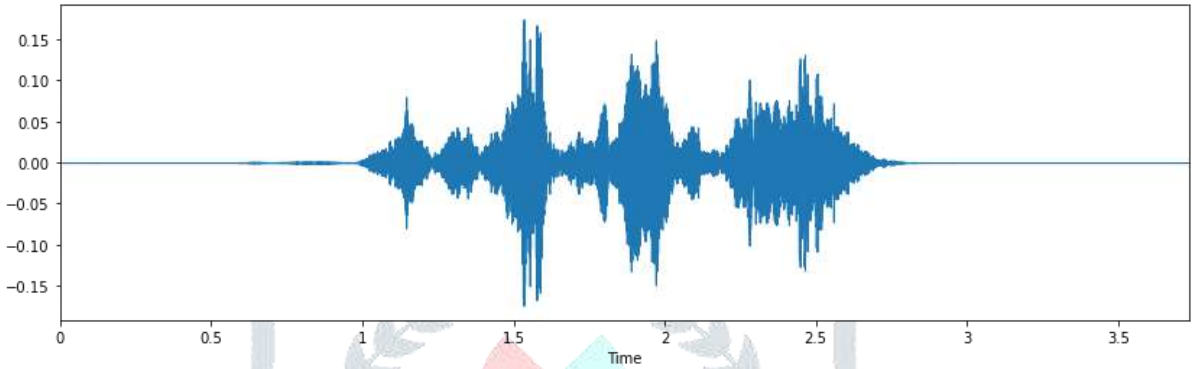
Figure 7. Applying Time Shifting



d. **Changing Pitch**

This augmentation is a wrapper of the librosa function. It changes pitch randomly.

Figure 8. Changing the pitch



3.3.2 Feature Extraction

Features Extraction is a critical part in examining and finding relations between different things. The input data given i.e. speech signals cannot be comprehended by the models straightforwardly so it is needed to change them into a justifiable format which uses feature extraction.
In this project, 5 features have been extracted:

a. **Zero Crossing Rate**
The zero-crossing rate (ZCR) is the rate at which a signal changes from positive to zero to negative or from negative to zero to positive.

b. **Chroma Short Term Fourier Transformation**
Compute a chromagram from a waveform or power spectrogram

c. **MFCC**
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").

d. **RMS(root mean square) value**
The RMS value is the square root of the mean (average) value of the squared function of the instantaneous values.

e. **Mel Spectrogram to train our model.**
A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale.

For creating the dataset, extract the above given five features for normal audio and stack every feature horizontally in a numpy array. After extracting features, repeat the process for augmented data and stack them vertically in a numpy array.

3.3.3 Classification

After Feature Extraction, several steps are implemented to normalize and split our data for training and testing. As this is a multiclass classification problem OneHotEncoder (Encode categorical features as a one-hot numeric array) has been used. Transformer takes array-like input of strings and integers, that depicts categorical feature values. For features, One-Hot Encoding also known as dummy encoding or one-of-K encoding scheme is used. Dense array or a sparse matrix is returned and a binary column is created for each category.

By default, the encoder derives the categories based on the unique values in each feature. Categories can be alternatively specified manually. To feed categorical data in more scikit-learn estimators, this encoding is required (such as standard kernels SVMs and linear models).

Then finally scaling data using StandardScaler (Standardize features by removing the mean and scaling to unit variance). Many machine learning estimators require a standardization of a dataset they might behave badly if the individual features do not more or less look like standard normally distributed dataset

3.3.4 Model Architecture of Convolutional Neural Network

The model is a sequential model with three one-dimensional convolution layers with a kernel size of five each followed by a one-dimensional max pooling layer with a max pool window size of five. The activation function for convolution layers used is a rectified linear unit (ReLU). Number of filters for the convolution layers used are 256, 128 and 64. Padding used is 'same' for convolution and max pooling layer which evenly pads in horizontal and vertical direction such that it has similar size as of input given. Strides for convolution and max pooling layer used are 1 and 2 respectively. A Flatten is also added followed by a Dense layer with a unit of 32 and an activation function as rectified linear unit (ReLU). Two dropout layers are added with a rate of 0.5 and 0.25 where the first dropout layer is added after the fourth layer and the second dropout layer after the ninth layer. The last layer is a Dense layer with 5 units and the activation function used is Softmax function.

Table no. 1. Model Summary

Layer (type)	Output Shape	Param #
conv1d_8 (Conv1D)	(None, 162, 256)	1536
max_pooling1d_8 (MaxPooling1D)	(None, 81, 256)	0
conv1d_1 (Conv1D)	(None, 81, 128)	163968
max_pooling1d_1 (MaxPooling1D)	(None, 41, 128)	0
dropout (Dropout)	(None, 41, 128)	0
conv1d_2 (Conv1D)	(None, 41, 64)	41024
max_pooling1d_2 (MaxPooling1D)	(None, 21, 64)	0
flatten (Flatten)	(None, 1344)	0
dense (Dense)	(None, 32)	43040
dropout_1 (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 5)	165

Total params: 557,189
Trainable params: 557,189
Non-trainable params: 0

3.3.5 Training the Convolutional Neural Network

As already mentioned above, a combination of datasets is used i.e RAVDESS, TESS, SAVE and CREMA-D and data augmentation, which results in a large dataset which will be used for training. The training and testing dataset is split as 75:25. For training, Adam optimizer is used with a minimum learning rate of 0.0000001. The loss function used is Categorical Crossentrophy. The model was trained for 100 epochs with a batch size of 64.

4. Results and Discussion

4.1. Waveplots and Spectrogram

Waveplots depicts the loudness of the audio at a given time, whereas a spectrogram is a visual representation of the spectrum of frequencies of sound or other signals as they vary with time. It's a graphical view of frequencies alternating with respect to time for signals of sound/music.

Figure 9. Waveplot for audio with fear emotion

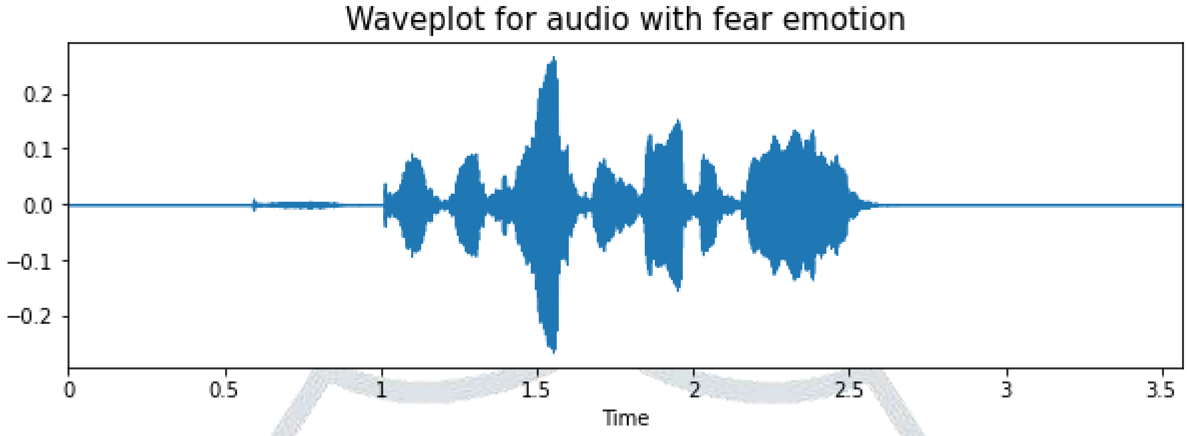


Figure 10. Spectrogram for audio with fear emotion

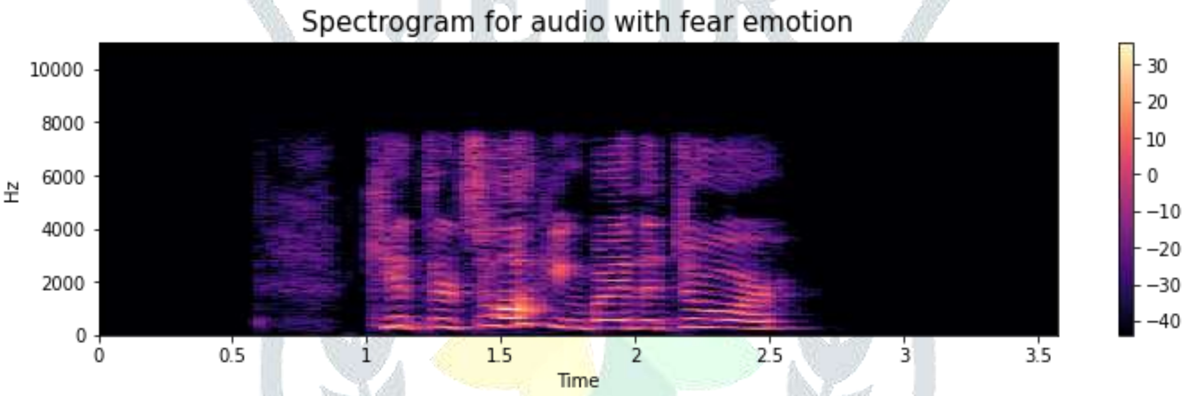


Figure 11. Waveplot for audio with angry emotion

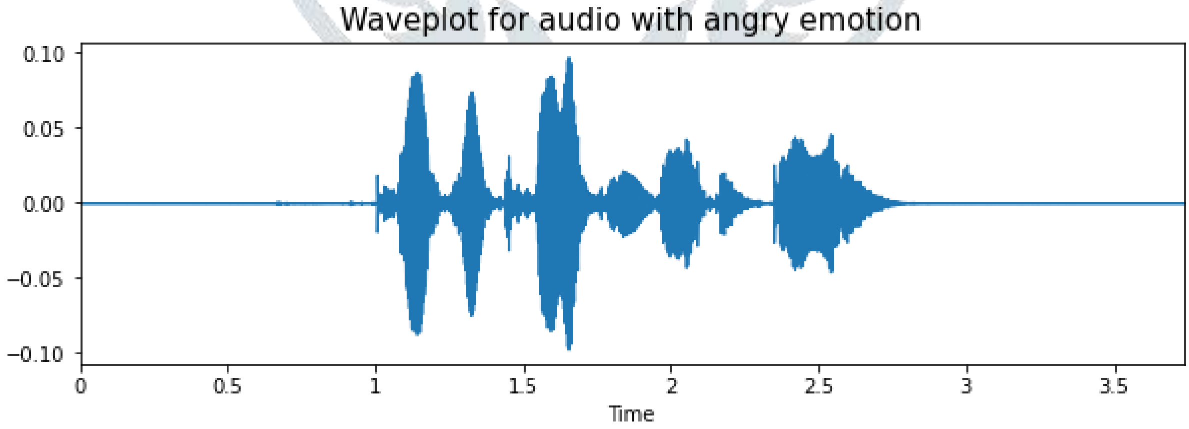
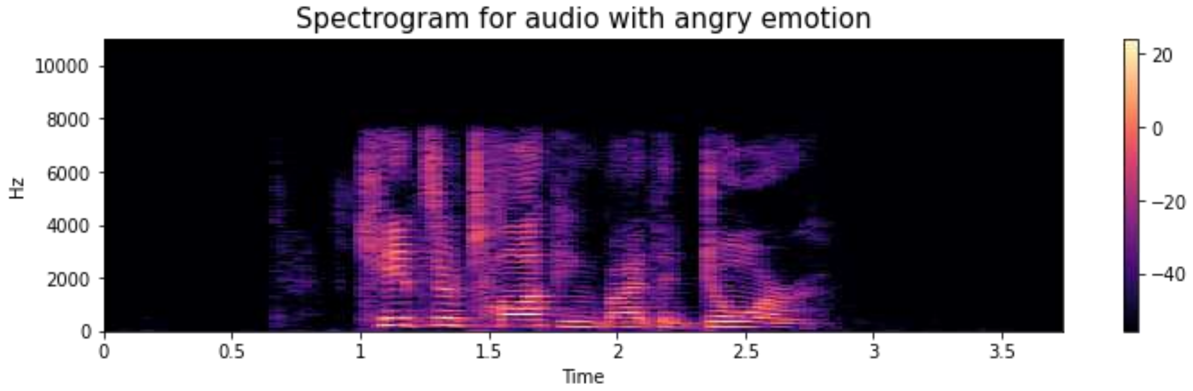
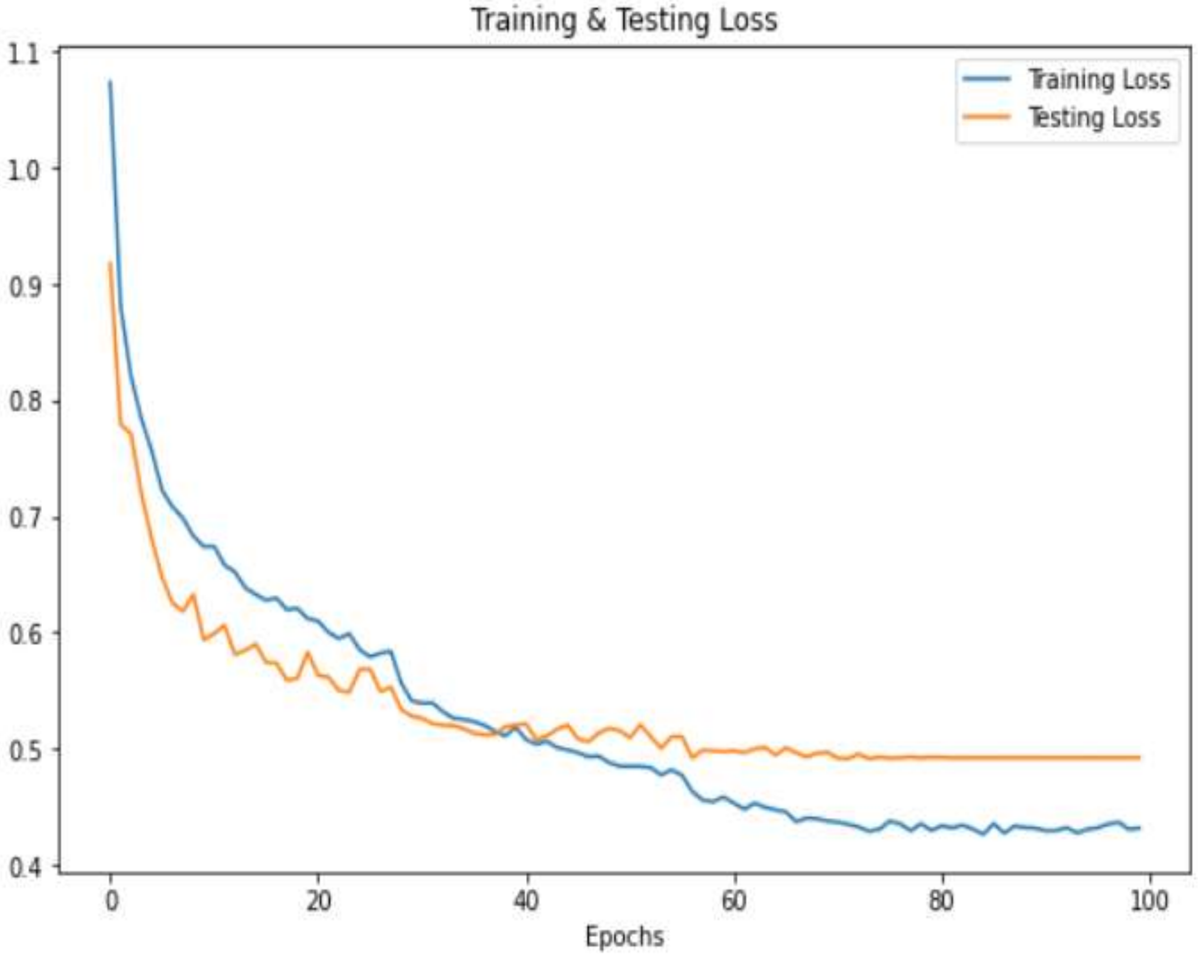


Figure 12. Spectrogram for audio with angry emotion



4.2 Training and Testing Loss

Figure 13. Training and Testing Loss



4.3 Training and Testing Accuracy

Figure 14. Training and Testing Accuracy

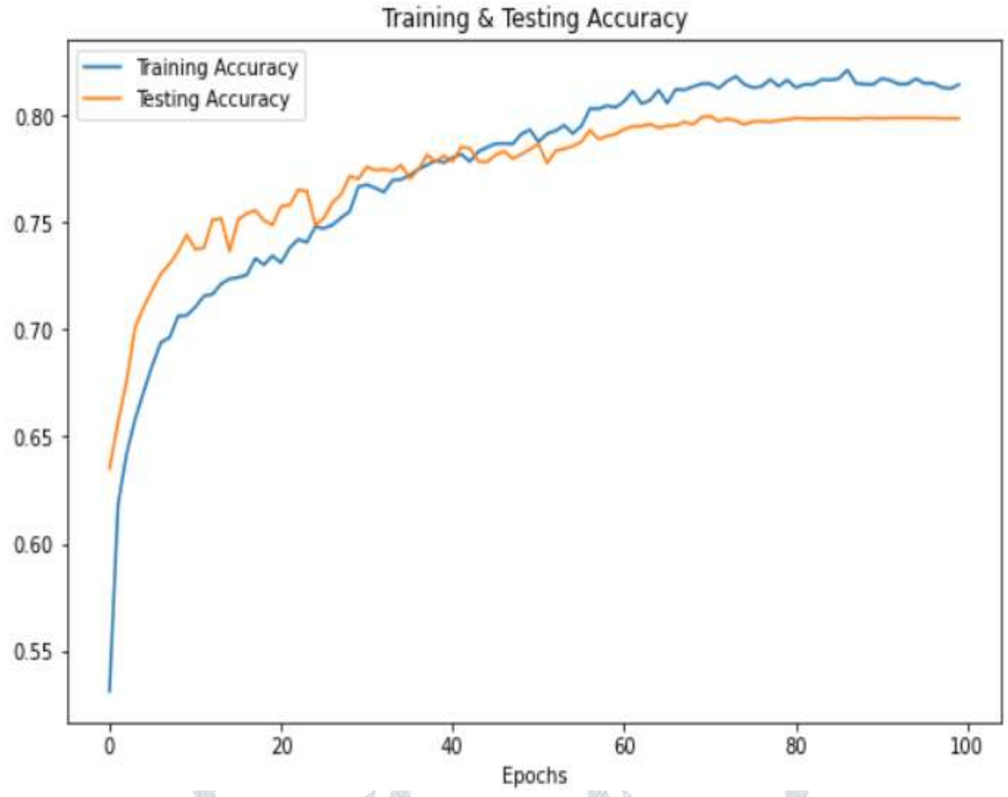
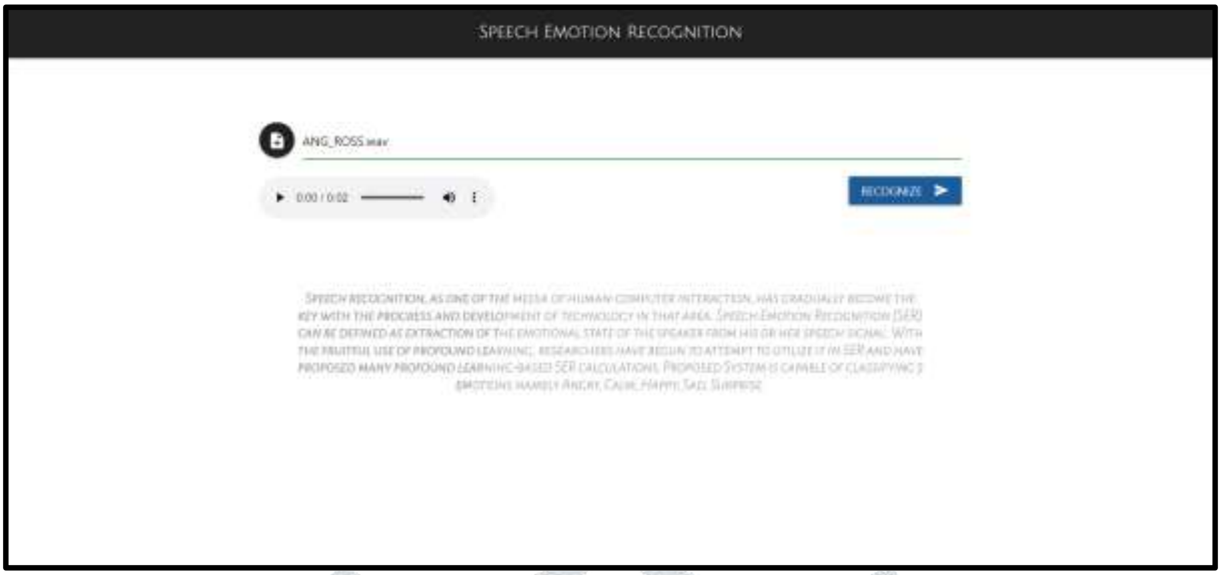


Table no. 2. Model Results

	precision	recall	f1-score
angry	0.83	0.75	0.79
calm	0.72	0.86	0.78
happy	0.70	0.74	0.72
sad	0.86	0.86	0.86
surprise	0.86	0.89	0.88
Total	0.80	0.82	0.81

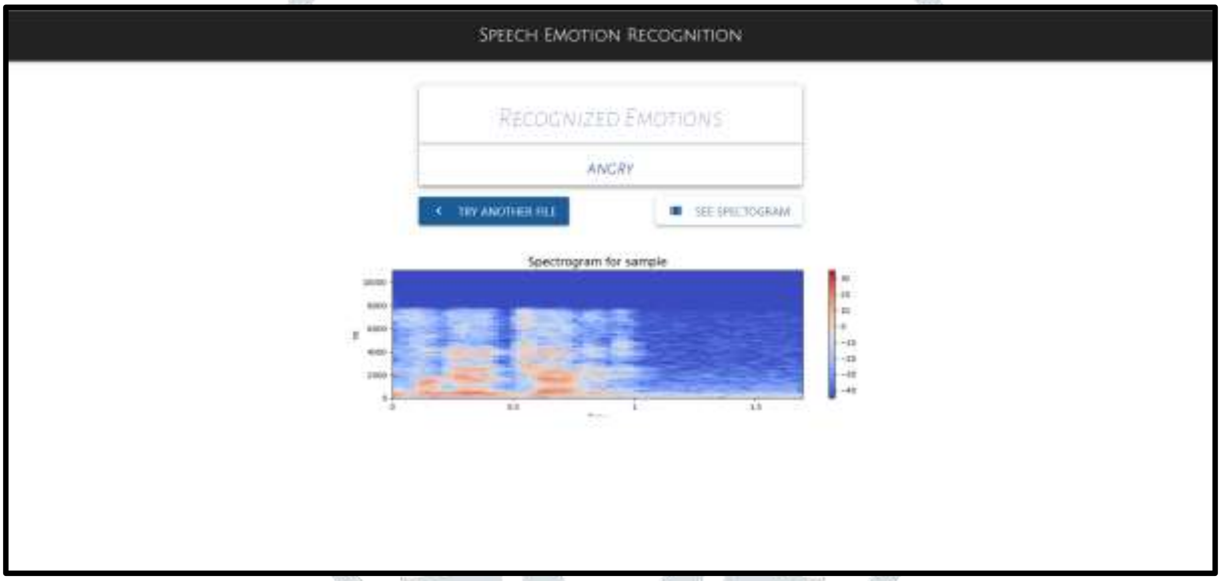
4.4 Screenshots

Figure 15. System`s landing page



Speech recognition, as one of the media of human computer interaction, has gradually become the key with the process and development of technology in that area. Speech Emotion Recognition (SER) can be defined as extraction of the emotional state of the speaker from his or her speech signal. With the fruitful use of profound learning, researchers have begun to attempt to utilize it in SER and have proposed many profound learning-based SER calculations. Proposed System is capable of classifying 5 emotions namely Angry, Calm, Happy, Sad, Surprise

Figure 16. Recognized emotions with Spectrogram



5. Conclusion

The proposed model predicts 5 emotions namely angry, calm, happy, sad and surprise with an accuracy of 81%. The model comprises Data Augmentation, Feature Extraction and Feature Classification. For Data Augmentation, 4 augmentation methods which are Noise Injection, Stretching, Time Shifting and Pitch Changing have been used whereas feature extraction uses 5 Feature Extraction methods namely Zero Crossing Rate, Chroma STFT, MFCC, RMS value and Mel Spectrogram. For the main model, one dimensional Convolutional Neural Network along with one dimensional Max Pooling have been used.

The man-machine connection has requested the smart patterns that machines need to respond to the human emotional levels. In this paper a strategy is proposed that utilizes the Convolutional Neural Network (CNNs), one of the profound neural networks, to extract the characteristics of various emotions from raw speech signals.

6. References

[1] Mingke Xu, Fan Zhang, Xiaodong Cui and Wei Zhang ,” Speech Emotion Recognition with multiscale area attention and data augmentation“ presented at the Nanjing Tech University, China ,3rd Feb 2021. DOI : 10.48550/arXiv.2102.01813

[2] Yunfeng Xu, Hua Xu, and Jiyun Zou, “Hgfm: A hierarchical grained and feature model for acoustic emotion recognition,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6499–6503. DOI : 10.1109/ICASSP40776.2020.9053039

[3] Darshana Priyasad, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes, “Attention driven fusion for multi-modal emotion recognition,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 3227–3231. DOI : 10.1109/ICASSP40776.2020.9054441

[4] Anish Nediyanchath, Periyasamy Paramasivam, and Promod Yenigalla, “Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7179–7183. DOI : 10.1109/ICASSP40776.2020.9054073

[5] Mingke Xu, Fan Zhang, and Samee U Khan, “Improve accuracy of speech emotion recognition with attention head fusion,” in 2020 10th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2020, pp. 1058–1064. DOI : 10.1109/ACCESS.2021.3067460

[6] Srinivas Parthasarathy and Carlos Busso, “Semi-supervised speech emotion recognition with ladder networks,” IEEE/ACM transactions on audio, speech, and language processing, 2020. DOI : 10.1109/TASLP.2020.3023632

[7] Bagus Tris Atmaja and Masato Akagi,” Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model” presented at the 2019 IEEE International Conference on Signals and Systems (ICSigSys), Nomi, Japan, 2019. DOI : 10.1109/ICSIGSYS.2019.8811080

[8] 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 “ presented at the 2019 IEEE International Conference on Signal Processing, 2019. DOI : 10.1109/ACCESS.2019.2936124

[9] Brian Mcfee, Colin Raffel, Dawen Liang, Daniel Ellis, and Oriol Nieto, “librosa: Audio and music signal analysis in python,” in Python in Science Conference, 2015. DOI : 10.25080/MAJORA-7B98E3ED-003

[10] Chenchen Huang, Wei Gong, Wenlong Fu, and Dongyu Feng, “A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM” presented at the Department of Computer, Communication University of China, Beijing 100024, China, 27 May 2014. DOI : 10.1155/2014/749604