# Human Speech Emotion Recognition Using Deep Neural Network

Mrinmoy Sikdar
M.Sc. in Big Data Analytics
Roll Number: B1930066
Department of Computer Science
RKMVERI (Belur Math, Howrah)
Summer Project Report

**Abstract**

As human beings speech is amongst the most natural way to express ourselves. As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today's digital world of remote communication. Emotion detection is a challenging task, because emotions are subjective. We define a Speech Emotion Recognition (SER) system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. Such a system can find use in a wide variety of application areas like interactive voice based-assistant or caller-agent conversation analysis. In this work we attempt to detect underlying emotions in recorded speech by analysing the acoustic features of the audio data of recordings. This work proposes an emotion recognition system based on speech signals in two-stage approach, namely feature extraction and classification engine.

## 1 Introduction

Emotion plays a significant role in human life. It helps us to match and understand feelings of others by conveying our feelings and giving feedback to others. Human speech is an important medium of expressing emotion. For example, when a person speaks about a particular subject, the emotion related to the subject gets revealed by that person's speech. These emotional displays convey considerable information about a human's mental state. Human speech [1] is nothing but audio signals. Compared to other biological signals (e.g., electrocardiogram) of human's, speech audio signals can be acquired more easily and economically. That's why it's comparatively easy to detect human's emotion from speech audio signals which leads to Speech Emotion Recognition (SER). The aim of the SER is to recognize the underlying emotional state of a speaker from her voice. There are many applications of detecting the emotion of the persons like in the interface with robots, audio surveillance, web-based E-learning, commercial applications, clinical studies, entertainment, banking etc. For example, in case of E-learning, information about the emotional state of students can provide focus on the enhancement of teaching quality.

Three important factors need to be taken care of to build a successful Speech Emotion Recognition System, which are, (1) choice of good and balanced emotion speech database, (2) extraction of effective features from the speech audio signals and (3) construction of a reliable classifier model using machine learning or deep learning algorithms. Emotional feature extraction is the main issue in a Speech Emotion Recognition system. There are variety of temporal and spectral features [2] that can be extracted from human speech. Some relevant features for SER are – Pitch, Mel-Frequency Cepstral Coefficients (MFCC), Chroma features etc. The statistics related to these features are the inputs of the classification algorithms.

The last step of speech emotion recognition is classification. It involves classifying the raw data in the form of utterance or frame of the utterance into a particular class of emotion on the basis of features extracted from the data. In recent years in speech emotion recognition, researchers proposed many classification algorithms, such as Gaussian mixture model[4], Hidden Markov model[5], Support vector machine (SVM)[6], Neural networks, and Recurrent neural networks (RNN)[7]. In this work we will use Convolutional Neural Networks (CNN) and LSTM to classify emotions. The emotion recognition accuracy of these experiments allows us to explain which features carry the most emotional information.

## 2 Datasets

To build a SER model, working with a balanced dataset is essential. We referred two datasets **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)** dataset and **Surrey Audio-Visual Expressed Emotion (SAVEE)** dataset.

| **Name of the dataset** | **Number of Audio Samples in the dataset** | **Different emotion labels present in the dataset** |
|---|---|---|
| RAVDESS | 2000 | Calm, Happy, Sad, Angry, Fearful, Surprise and Disgust |
| SAVEE | 480 | Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise |

Table 1: Description of the datasets

We made customized dataset by using these two datasets. We have less samples of ‘calm’ emotion. Hence, we eliminate ‘clam’ samples to balance the dataset. Our dataset contains 7 folders, each represents the different emotion. Contain Separate emotion's voice/speech in each separate folder.

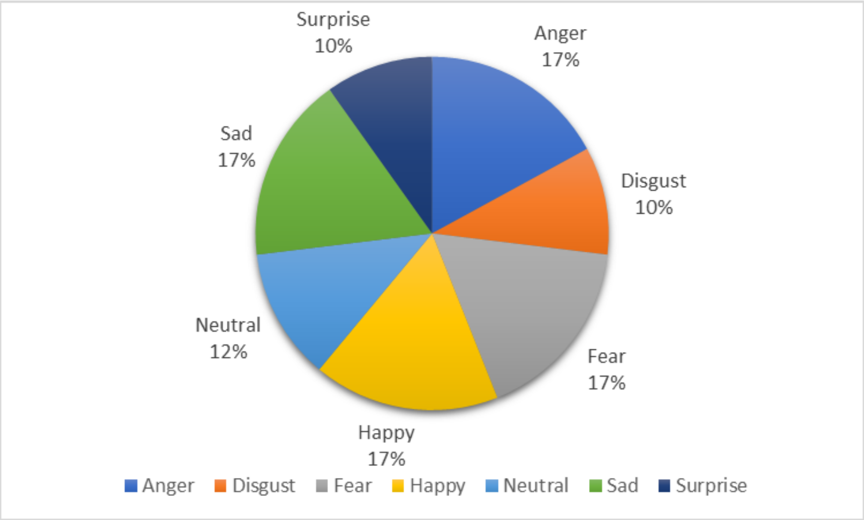

Figure 1: Different emotion labels in the final customised dataset

From Figure 1 it is evident that the above mentioned customised data is a balance dataset. Using this customised dataset further feature extraction has been done.

# 3 Methods

This work proposes an emotion recognition system based on speech signals in two-stage approach, namely feature extraction and classification engine. Figure 2 gives a summarised view of the implemented methodology of this work.

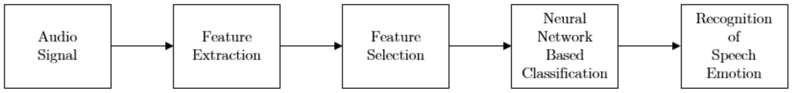
Audio Signal
Feature Extraction
Feature Selection
Neural Network Based Classification
Recognition of Speech Emotion

Figure 2: Methodology of building a Speech Emotion Recognition system

## 3.1 Feature Extraction

Feature extraction plays the most important in SER system. We tested out one of the audio file of emotion 'Angry' to know its features by plotting its waveform and spectrogram. We used Librosa library in Python to process and extract features from the audio files. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.

### 3.1.1 Time Domain Representation of Audio Data

The time-domain representation of sound is very complex, and in its original form, it does not provide very good insight into key characteristics of the signal. We map this time domain representation into more telling features. The most straightforward technique involves determining the average energy of the signal. This metric, along with total energy in the signal, indicates the "volume" of the speaker. Duration also offers insights into emotion, as do statistics like the maximum, minimum, range, mean, and standard deviation of both the signal and spectrum. These may indicate fluctuations in the volume or pitch that can be useful in determining emotion. For both the signal and spectrum, we also derive skewness, the measure of departure of horizontal symmetry in the signal, and kurtosis, the measure of height and sharpness of central peak, relative to a standard bell curve.

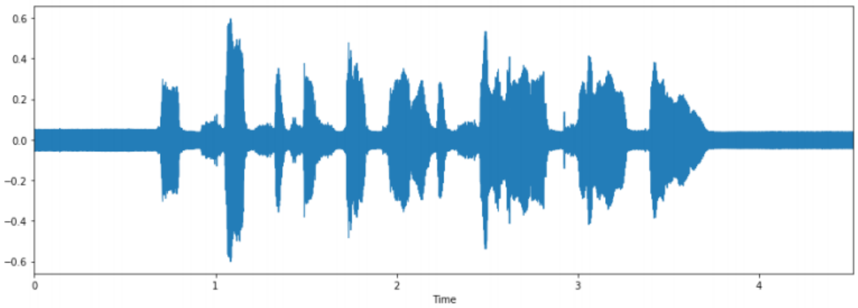

Figure 3: Time domain plot of the sample speech signal

### 3.1.2 Frequency Domain Representation of Audio Data

We also process the signal in the frequency domain through the *Fourier Transform.* We use windowed samples to get accurate representations of the frequency content of the signal at different points in time. Taking the square value of the signal at each window sample, we can derive the power spectrum. We use the values of the power spectrum as features, but we also find the frequencies that have the greatest power. We obtain the three largest frequency peaks for each

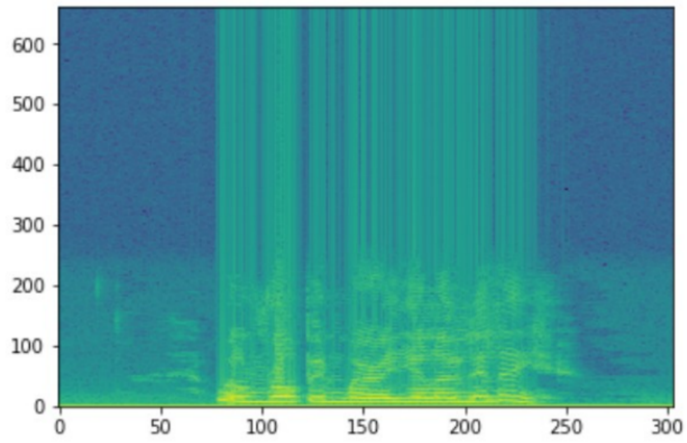

Figure 4: Frequency domain plot of the sample speech signal

window and add those to the feature vector. In addition, we find the maximum and minimum frequencies with substantial power for each time frame, and use these values to determine the frequency range for each frame. The auditory spectrum can be derived by mapping the power spectrum to an 4auditory frequency axis by combining the Fast Fourier Transform bins into equally spaced intervals.

### 3.1.3 Mel-frequency Cepstral Coefficients (MFCCs)

The Mel-frequency cepstral coefficients (MFCCs) of a signal are a small set of features (usually about 10–20) which concisely describe the overall shape of a spectral envelope. It models the characteristics of the human voice. Pitch is one of the characteristics of a speech signal and is measured as the frequency of the signal. *Mel scale* is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear (humans are better at identifying small changes in speech at lower frequencies). A frequency measured in Hertz (f) can be converted to the Mel scale using the following formula :

$$Mel(f) = 2595 \times \log(1 + f/100) \tag{1}$$

Here is an intuitive example of what the mel scale captures. The range of human hearing is 20Hz to 20kHz. Imagine a tune at 300 Hz. This would sound something like the standard dialer tone of a land-line phone. Now imagine a tune at 400 Hz (a little higher pitched dialer tone). Now compare the distance between these two howsoever this may be perceived by your brain. Now imagine a 900 Hz signal (similar to a microphone feedback sound) and a 1kHz sound. The perceived distance between these two sounds may seem greater than the first two although the actual difference is the same (100Hz). The mel scale tries to capture such differences.

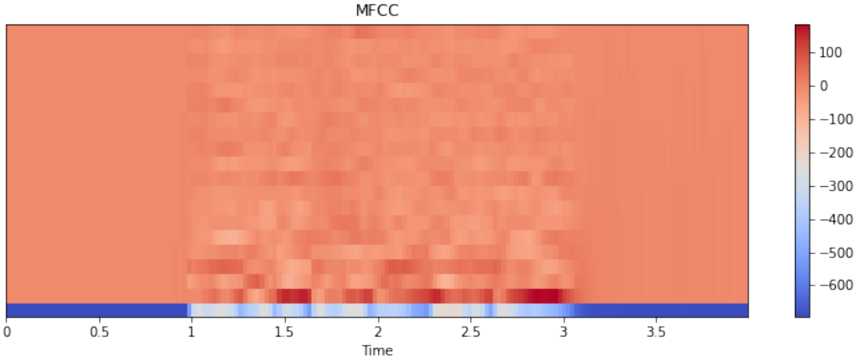

Figure 5: MFCC plot of the sample speech signal

## 3.2 Feature Selection

After we processed the original sound signal to extract features, the high variance of our algorithm revealed that we needed to filter the many features to determine which contribute most to the classifier. Our input speech signals were windowed, with approximately 72 windows per audio sample, and each of these windowed samples provided a total of 577 features. In total, we extracted **41,558** features. This large number of features (much larger than the number of examples) resulted in a very high variance. Clearly, we needed to extract the most important features. Because of the large number of features, we used heuristics to score each feature.

## 3.3 Construction of Classifier Model

To construct our classification model for speech emotion recognition, we have used Convolutional neural network (CNN) [8] and Long short-term memory (LSTM) reccurrent neural network [9]. CNN is used to perform the feature learning and classification. In order for features to better reflect temporal continuity, LSTM is used to increase the information between adjacent frames. LSTM is naturally suitable for speech recognition due to its ability to take advantage of dynamically changing time information. The structure of the classifier model is described in the following Figure 6.

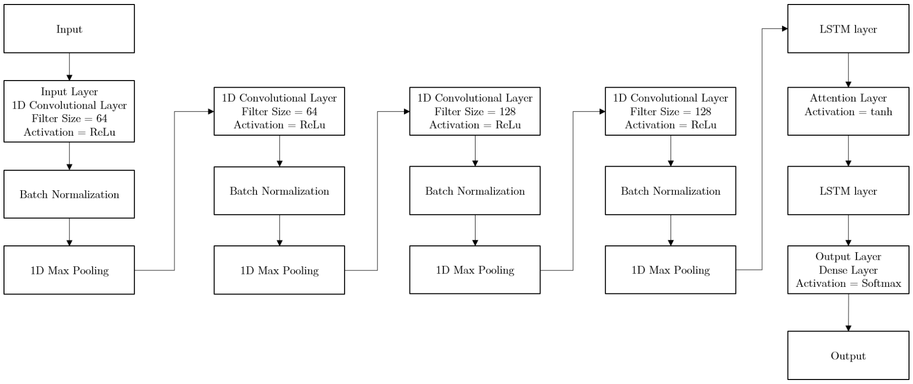

Figure 6: Block diagram of the classifier model (CNN & LSTM )

After constructing the classification model, we have complied the model and then fit the model. Details of model compilation and model fitting are given in the table 2.

| | |
|---|---|
| Training data | 80% of the dataset |
| Test data | 20% of the dataset |
| Opitimizer | Stochastic gradient descent |
| Learning rate | 0.001 |
| Loss function | Categorical Crossentropy |
| Metric | Accuracy |
| Number of epochs | 400 |

Table 2: Model compilation and fitting

# 4 Results

We got the accuracy of the training model about 96% but the accuracy of cross validation about 60%. Our dataset is over fitting the model. Dataset could be modified to get better accuracy. The below Figure 7 shows the training and validation accuracy on our dataset.

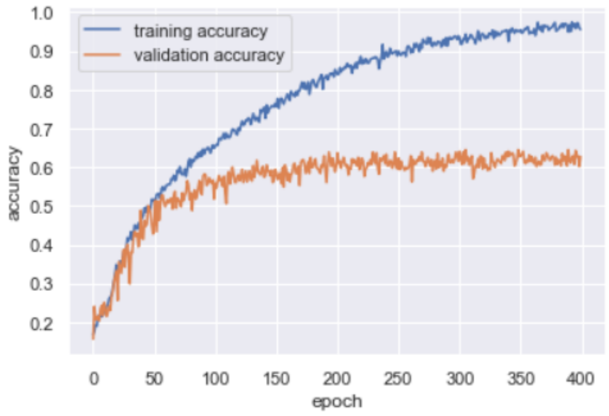

Figure 7: Model accuracy vs. Epochs plot

Figure 8 shows the training and test/validation loss on our dataset (since our test data has been used for validation).

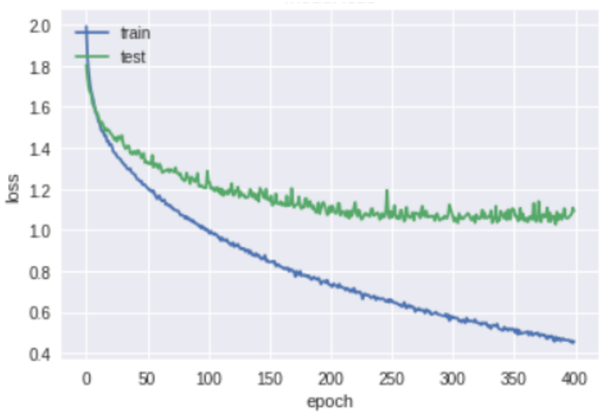

Figure 8: Model loss vs. Epochs plot

## 5 Discussion

Hence our project presents a new way to give the ability to machine to determine the emotion with the help of the human voice. It will give the machine the ability to have a better approach towards having a better conversation and seamless conversation like human does. Variation in voice tones as well as internal physiological changes while uttering a sentence (or even a single word) combine to generate the speaker emotional state. Perfect recognition of emotions is not easy even by human when listening to each other; sometimes the human cannot recognize his own innermost emotion. In fact some aspects of internal feeling remains hidden and not detectable from the speech, especially when the speaker need to suppress emotions. Therefor computer-based system cannot do beyond what is observed from the speech sample input [10]. Consequently, categorizing emotional speech samples is a serious challenge due to the long debate about the real meaning of "emotion" and the emotional classes that should be dealt with.

# 6 References

[1] Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing. 2018;273:271-280 - *Liu ZT, Wu M, Cao WH, Mao JW, Xu JP, Tan GZ.*
[2] Feature selection for classification: A review. Data Classification: Algorithms and Applications. 2014:37 - *Tang J, Alelyani S, Liu H.*
[3] Signal modeling techniques in speech recognition. Proc. IEEE 81, 1215–1247 (1993) - *J.W. Picone,*
[4] Recognition of emotions in German speech using Gaussian mixture models. LNAI. 2009;5398:256-263 - *Martin V, Robert V.*
[5] Speech emotion recognition using hidden Markov model and support vector machine. International Journal of Advanced Engineering Research and Studies. 2012:316-318 - *Ingale AB, Chaudhari D.*
[6] SVM scheme for speech emotion recognition using MFCC feature. International Journal of Computer Applications. 2013;69 - *Milton A, Sharmy Roy S, Tamil Selvi S.*
[7] Multi-Modal Dimensional Emotion Recognition using Recurrent Neural Networks. Australia: Brisbane; 2015 - *Chen S, Jin Q.*
[8] A CNN Approach for Audio Classification in Construction Sites - *Alessandro Maccagno, Andrea Mastropietro, Umberto Mazziotta, Michele Scarpiniti, Yong-Cheol Lee and Aurelio Uncini*
[9] Speech emotion recognition based on voiced speech using LSTM with attention model - *Bagus Tris Atmaja, Masato Akagi*
[10] Emotion recognition from speech – Tools and Challenges - *Abdulbasit Al-Talabani, Harin Sellahewa and Sabah A. Jassim*