**SPEECH EMOTION RECOGNITION**

Arpitha Pradeep, Swathi H L, Ujwal Patil, Yashvanth D

Under the guidance of Dr. H.S Jayanna ( Professor ) and

Mr. Praveen Kumar (Research and scholar)

Information Science and Engineering

Siddaganga Institute of Technology, Tumakuru, India

***Abstract*** - Emotion plays a crucial role in day-to-day interpersonal human interactions. Recent findings have suggested that emotion is integral to rational and intelligent decisions. It helps us to relate with each other by expressing our feelings and providing feedback. This important aspect of human interaction needs to be considered in designing human-machine interfaces (HMIs). To build interfaces that are more in tune with the user’s needs and preferences, it is essential to study how emotion modulates and enhances the verbal and nonverbal channels in human communication. In this paper, it presents an approach to Recognizing the emotion underlying the speech signal and extract features like MFCC, ZCR, and RMS for classifying the emotions into their respective classes using an appropriate classifier like RNN.

***Keywords***- emotion recognition, MFCC, ZCR, RMS, RNN, Energy, HCI.

# Introduction

Emotion is such a unique power of human trial that plays a vital role in distinguishing human civilization from others. Emotion helps us to understand each other in a better way. To express emotions, people use body postures, facial expressions, and vocalizations.

Speech emotion recognition (SER) is one of the challenging scenarios in the real-life world. For example, if we consider the customer call service which has been done through robots, a customer is facing an issue and calls the services for help, but the robot is unable to understand one’s situation. This may frustrate the customer. Here, SER can be implemented for human-like behaviour in robots which may help to improve the services. and, in the field of medical science SER plays an important role. During the pandemic, we have seen the very huge impact of technology which made it possible to virtualize our routine. Likewise, now we have applications to get doctor’s appointments, virtual meetings with doctors, and many others. SER can be helpful in these applications to understand the patient's emotions and for keeping their good mental health while conducting mental therapies to examine the patient's actual mindset and emotions.

One of the most common and efficient methods for feature extraction is the Mel frequency Cepstral Coefficient (MFCC). The objective of Speech emotion Recognition is to pre-process the input signal, extract the required features, and characterize and recognize the speaker's emotions. In this paper, it attempted to detect underlying emotions in a recorded speech by analysing the acoustic features of the audio data of recordings.

# System design

Speech Emotion Recognition build with python language using Django framework. Django is a trending framework for construct web-based projects. SER front end static pages are handled by the HTML, CSS, and JS. The back-end control is taken by the python libraries like librosa, pyaudio, PyDub and Django web handler. SER includes several data pre-processing steps and the pre-processed data will use to extract the Graphical user interface, diagram, application, Teams

Description automatically generatedfeatures. MFCC, ZCR, and RMS are the main features to extract the energy level of the audio sample.

In the Fig.1 describes the block diagram of the project. Initially, the input speech will be taken in the form of user voice data or a single input mp3 file. This speech is fragmented into the number of ‘n’ fragments based on the window size and sample rate. These fragments are optimized and pre-processed to extract the features. Each fragment is treated as a small input audio file. These extracted features are passed to a trained and tested RNN model which will output the frequency of each emotion. Based on the highest frequency, the emotion of the speech is obtained.

## Data Collection

Training a model is also one of the important steps in the process. To train a model requires suitable data. In this project TESS (Toronto Emotional Speech Set), CREMA (Crowd-Sourced Emotional Multimodal Actors), SAVEE (Surrey Audio-Visual Expressed Emotion), and RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) are considered. Each dataset contains a minimum of 6 different emotions. Each dataset contains audio files which express each emotion by considering various scenes. Table shows how the database is integrated. There are almost 8 different emotions, ~500 Different Scenes, 34 different Actors, and a total of 14962 files are used to create the customized dataset.

### RAVDEES (Ryerson Audio-Visual Database of Emotional Speech and Song)

RAVDEES is a validated multimodal database of emotional language and songs. The database consists of 24 professional actors vocalizing lexically matched utterances in a neutral North American accent with gender balance. Voices include expressions of calm, happiness, sadness, anger, fear, surprise, and disgust, and songs contain emotions of calm, happy, sad, anger, and fear.

Each expression produces two levels of emotional intensity and neutral expression. All conditions are available in face and voice, face-only, and voice-only formats. The set of 7356 recordings, each rated 10 times on emotional validity, intensity, and verity. Ratings were provided by 247 untrained research participants from North America. An additional set of 72 participants provided test-retest data. High levels of emotional validity and test-retest concordance reliability were reported.

### 2. TESS (Toronto Emotional Speech Set)

This dataset would be in the service of a very good training dataset for the emotion classifier in terms of generalization. The dataset consists of 200 target words spoken by two actresses and recordings were made of the set portraying each of seven emotions (disgust, happiness, surprise, sadness, anger, fear, and neutral). There is a total of 2800 audio files.

The dataset is organized such that each of the two female actors and their emotions is contained within its own folder. And within that, all 200 target word audio files can be found. The audio file is in the form of a WAV format.

### 3. CREMA (Crowd-Sourced Emotional Multimodal Actors)

CREMA dataset is a very good dataset to use to ensure the model does not overfit. This dataset of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of origins and ethnicities. Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions and four different emotion levels.

### 4. SAVEE (Surrey Audio-Visual Expressed Emotion)

This dataset was recorded from four native English male speakers, postgraduate students, and researchers at the University of Surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: anger, fear, disgust, surprise, happiness, and sadness. We attached neutral to provide recordings of 7 emotion categories. The text material consisted of 15 sentences per emotion: 3 common, 2 emotion-specific, and 10 generic sentences that were different for each emotion and phonetically balanced. The 3 common and 2\*6 = 12 emotion-specific sentences were recorded as neutral to give 30 neutral sentences.

## Data preparation

A continuous speech signal may contain surrounding disturbance, silence or sometimes the same speech segments overlapped with others. This may increase the complexity of speech processing. So, the speech signal is split into short-time frames because the frequency of the signal changes over time and it will be difficult for transforming from the time domain to the frequency domain, to get individual frequency signals. Thus, the taken frame length is **2048** samples. Hop length is the length of the non-intersecting portion in the audio fragment. Hop length considered here is **512** samples. These two parameters are necessary for obtaining detailed processing of audio signals.

Total audios hops = Max length/hop length = 314818/512 = 615

## Data Pre-processing

Before the extraction of the features of the user voice input, this voice data is manipulated by using pre-processing methods.

1. Raw data : It is the voice signal taken from the user directly. This voice data includes many unwanted and lack of usable precise data.
2. Normalized signal : It is a strategy for modifying the volume of sound to a standard level. Normalization is used the signal sequence divided by highest value of the signal to ensure that each sentence has a comparable volume level. Audio files where the speech is loud in some portions and quiet in others. Having this variance in volume can hinder transcription. Luckily, PyDub's effects module has a function called normalize() which finds the maximum volume of an AudioSegment, then adjusts the rest of the AudioSegment to be in proportion. This means the quiet parts will get a volume boost. headroom is how close to the maximum volume to boost the signal up o (specified in dB).
3. trim : Trim refers to changing the level of a signal (up or down) to align it more appropriately for a particular device. A trim control may apply some “gain” to raise a signal level, or it may apply some negative gain, or attenuation to lower a signal.
4. Padding : Padding reduce signal levels before the active amplification process to avoid overloading the microphone circuitry.
5. Noise : Noise reduction is the process of removing noise from a signal. Noise reduction techniques exist for audio and images. Noise reduction algorithms may distort the signal to some degree. All signal processing devices, both analog and digital, have traits that make them susceptible to noise.

## Feature Extraction

1. Mel-frequency cepstral coefficients (MFCC): MFCC is a speech emotion feature parameter that is an inverse spectral coefficient extracted in the frequency domain of the Mel scale, a feature widely used in automatic speech and speaker recognition. The Mel scale is very accurate in describing the nonlinear characteristics of human ear frequency. The computational relationship between it and the frequency can be expressed in equation (1).

(Eq.4.1)

Where :   
 M🡪 Meier frequency function  
 f 🡪 linear frequency

1. Zero-Crossing Rate (ZCR) : As the name suggests the zero-crossing rate is the measure of the rate at which the signal is going through the zeroth line more formally signal is changing from positive to negative or vice versa. Mathematically it can be measured as.

(Eq. 4.2)

Where   
 s 🡪 signal  
 T🡪 length of signal

1. Root mean square (RMS) : The square root of the mean of the square. RMS is a meaningful way of calculating the average of values over a period. With audio, the signal value (amplitude) is squared and averaged over a period, then the square root of the result is calculated. The result is a value, that when squared, is related (proportional) to the effective power of the signal. RMS of a signal is a just a value used to calculate average, or continuous, power.

Unfortunately, calculating the RMS value of anything but a simple sine wave (.707 of peak) is very difficult. The further a signal gets in harmonic content from a sine wave, the less accurate RMS values will be. It is nearly impossible to get even close to a true RMS value for a dynamic signal like most music.

(Eq.4.3)

## Recurrent Neural Network (RNN)

RNN - Recurrent Neural Network is one of the most common Neural networks used for Sequential data. As we know the audio signal is continuous or sequential data, Output for a given input is based on the input at a given time and the previous input in hidden layers. RNN as the name refers to is recurrent, it has a recurrent connection to the hidden state. This looping constraint ensures that sequential information is captured in input data. LSTM (Long Short-Term Memory) keeps the information of current input and previous input for future calculations which helps in getting a pattern from sequential data.

The Activation Function used here is **softmax** any activation function in the recurrent neural network.

### 5.2 Training A Recurrent Neural Network

The backpropagation algorithm of an artificial neural network is modified to include the unfolding in time to train the weights of the network. This algorithm is based on computing the gradient vector and is called back propagation in time or BPTT algorithm for short. The pseudo-code for training is given below. The value of k can be selected by the user for training. In the pseudo-code below pt is the target value at time step t:

* + 1. Repeat till stopping criterion is met:
       - 1. Set all h to zero.
         2. Repeat for to
    2. Forward propagate the network over the unfolded network for k time steps to compute all h and y.
    3. Compute the error as:
    4. Backpropagate the error across the unfolded network and update the weights.

LSTM were also designed to address the vanishing gradient problem in RNNs. LSTM use three gates called input, output and forget gate. Like GRU, these gates determine which information to retain.

### 5.3 Testing and Reporting

After configuring the model, given a shot for the higher value of epoch, which gives an overall picture of the model towards overfitting and underfitting. The following graph and of training accuracy vs. validation accuracy over the 500 epochs.

Chart, line chart

Description automatically generatedChart

Description automatically generatedAs seen in the graph the model will get overfitted after 100 epochs. To avoid overfitting, the model must be trained with the epoch value less than or equal to 100. The following graph and of training accuracy vs. validation accuracy over the 100 epochs.

Figure 8 Loss function and Model accuracy graph for 500 epochs

These results show that our model can distinguish angry emotions better than other emotions. In the above confusion matrix, we can see 9 Surprised emotions get predicted as Happy, which in a way indicates that the model takes the loudness of speech to detect emotion into account. When it comes to Neutral and sad emotions, the model gets confused between Neutral, Sad, and calm because they are almost similar in terms of energy and pitch.

|  |  |  |
| --- | --- | --- |
| **Method** | **Dataset** | **Model accuracy / Validation accuracy** |
| Simple RNN | CREMA, RAVADEES, SAVEE, TESS | acc 60 – Val. acc 26 |
| CNN | CREMA, RAVADEES, SAVEE, TESS | acc 58 – Val. acc 38 |
| LSTM | CREMA | acc 65 – Val. acc 59 |
| LSTM | RAVDESS | acc 92 – Val. acc 20 |
| LSTM | TESS | acc 98 – Val. acc 36 |
| DNN (LSTM) | RAVDESS, TESS | acc 60 – Val. acc 49 |
| PCA | SAVEE | acc 30 – Val. acc 26 |
| LSTM | CREMA, RAVADEES, SAVEE, TESS | acc 68 – Val. acc 56 |

The above table { } shows the tabulation of values obtained from training and testing various model with different datasets. We used 8 different combinations of models and datasets. We mainly focussed on Simple RNN(Recurrent Neural Network), LSTM(Long-short term memory), CNN(convolution Neural Network) and PCA(Principal Component Analysis). The datasets considered are CREMA, RAVDESS, SAVEE and TESS. We obtained maximum accuracy of 68 and validation accuracy 56 from LSTM.

Icon

Description automatically generatedA picture containing chart

Description automatically generatedTable

Description automatically generated

Figure 10 Realtime emotion prediction summary

Figure 11 Softmatrix for tested data

# Conclusion

In recent years, SER technology as one of the key technologies in human-computer interaction (HCI) systems, has received a lot of attention from researchers at home and abroad for its ability to accurately recognize emotions and thus improve the quality of human-computer interaction. In this paper, we propose a Recurrent Neural Network (RNN) algorithm with fused features for SER. in terms of data processing, we quadruple processed all four standard datasets RAVDESS, CREMA, TESS, and SAVEE with 14,962 audio samples. For the network structure, we constructed LSTM Recurrent neural networks (LSTM RNN) to extract spatial features and a transform encoder network to extract temporal features to classify emotions from one of the eight categories. Taking advantage of the simple structure of LSTM RNN in spatial feature representation and sequence coding transformation, we obtained an accuracy of 68% on the holdout test set.