# Speech Emotion Recognition

Shubham Wasnik, IISc Bangalore

*Abstract*—
**In today's world speech is the ideal way to interact with people. Speech emotion recognition (SER) has an increasingly significant role in the interactions among human beings and computers. For improving human machine interaction, it is very ideal to recognize emotions automatically because attention is aimed at study of the emotions. In this paper I compare different approaches for emotions recognition task. CNN based classifier is used to classify seven emotions. High quality dataset like Ravdess and TESS databasets are explored. CNN mode1 was overfitting. I experimented with Dropout and Batch Normalization and got better results.**

***Index Terms*—SER, CNN, RAVDESS TESS**

## I. Introduction

Emotion recognition in spoken dialogues has been gaining increasing interest all through current years. Speech Emotion Recognition (SER) is a hot research topic in the field of Human Computer Interaction (HCI). It has a potentially wide applications, such as the interface with robots, banking, call centers, car board systems, computer games etc. For classroom orchestration or E-learning, information about the emotional state of students can provide focus on enhancement of teaching quality. For example teacher can use SER to decide what subjects can be taught and must be able to develop strategies for managing emotions within the learning environment. That is why learner's emotional state should be considered in the classroom.

## II. Literature Survey

In general, the SER is a computational task consisting of two major parts: feature extraction and emotion machine classification. The questions that arise here: What is the optimal feature set? What combination of acoustic features for a most robust automatic recognition of a speaker's emotion? Which method is most appropriate for classification? Thus came the idea to compare a RNN method with the basic method MLR and the most widely used method SVM. And also all previously published works generally0 use the berlin database. To our knowledge the Vector Machine (SVM)(A. et al., 2013),(G.S. et al., 2016),(Pan et al., 2012), (Peipei et al., 2011), Neural Networks (NN) (Sathit, 2015) and Recurrent Neural Networks (RNN) (Alex and Navdeep, 2014), (Lim et al., 2017), (Chen and Jin, 2015). Some other types of classifiers are also proposed by some researchers such as a modified brain emotional learning model (BEL) (Sara et al., 2017) in which the Adaptive Neuro-Fuzzy Inference System (ANFIS) and Multilayer Perceptron (MLP) are merged for speech emotion recognition. Another proposed strategy is a multiple kernel Gaussian process (GP) classification (Chen and Jin, 2015), in which two similar notions in the learning algorithm are presented by combining the linear kernel and radial basis function (RBF) kernel. The Voiced Segment Selection (VSS) algorithm also proposed in (Yu et al., 2016) deals with the voiced signal segment as the texture image processing feature which is different from the traditional method. It uses the Log-Gabor filters to extract the voiced and unvoiced features from spectrogram to make the classification.. In order to demonstrate the high effectiveness of the MFCC for emotion classification in speech, we provide results on two open emotional databases (RAVDESS and TESS).

## III. Emotional Speech data

The performance and robustness of the recognition systems will be easily affected if it is not well-trained with suitable database. Therefore, it is essential to have sufficient and suitable phrases in the database to train the emotion recognition system and subsequently evaluate its performance. In this section, we detail the two emotional speech databases used in our experiments: RAVDESS and TESS.

1) RAVDESS

This dataset contains 1440 audio files in wav format from 24 different actors (12 male, 12 female) where these actors records short audios in 8 different emotions i.e 1= neutral, 2= calm, 3= happy, 4= sad, 5= angry, 6= fearful, 7= disgust, 8= surprised. with 60 recordings from each actor. Each file name consists of a 7- part numerical identifier. Filename Identifiers :

- Modality (01=full-av, 02=video-only, o3=audio-only, 03 = audio-only
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).Link to dataset: https://zenodo.org/record/1188976/files/Audio_Speech_Actors_01-24.zip?download=1

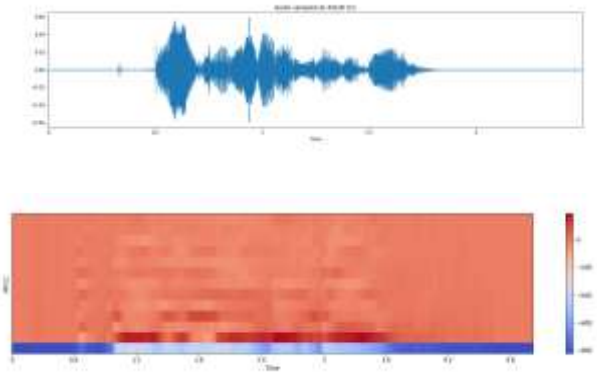
2) TESS Dataset
There are a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total. The dataset is organized such that each of the two female actor and their emotions are contain within its own folder. And within that, all 200 target words audio file can be found. The format of the audio file is a WAV format.
Link to dataset:
https ://tspace.library.utoronto.ca/handle/1807/24487

## IV. Features Extraction

The Speech signal contains large number of parameters that reflect the emotional characteristics. One important factor in SER is what features are used. In recent years, many common features are extracted such as energy, pitch, formants and some spectrum features such as Linear Prediction Coefficients (LPC), Mel-Frequency Cepstrum Coefficients (MFCC) and Modulation features. In this work I have selected MFCC to extract emotional features. MFCC features are the most commonly used representation of spectral property of voices signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. Usually, the process of calculating MFCC is shown in Figure 1.


Fig 1: Schema of MFCC extraction (Srinivasan et al., 2014)

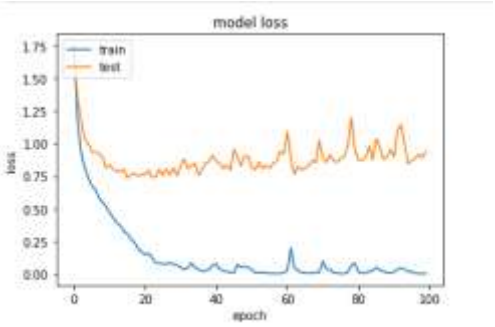
## V. Experimentation and Results



MFCC features

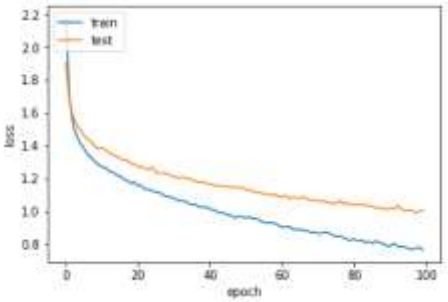
Experiment 1
In this experiment I tried Adam optimizer and Cross Entropy Loss (CSE). I trained the model for following Parameters.

| Batch Size | Lr. Rate | Epochs | Optimizer | Loss |
|---|---|---|---|---|
| 20 | 0.0001 | 100 | Adam | CSE |



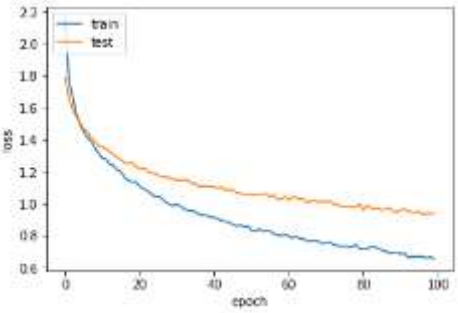

Experiment 2
In this experiment I tried new model with Adam optimizer and Cross Entropy Loss (CSE). I trained the model for following Parameters. Here I used dropout

| Batch Size | Lr. Rate | Epochs | Optimizer | Loss |
|---|---|---|---|---|
| 16 | 0.0001 | 100 | Adam | CSE |




Experiment 3
In this experiment I tried previous with Adam optimizer and Cross Entropy Loss (CSE). I trained the model for following Parameters. Here I used dropout(p=0.25) and also added batch Normalization.

| Batch Size | Lr. Rate | Epochs | Optimizer | Loss |
|---|---|---|---|---|
| 16 | 0.0001 | 100 | Adam | CSE |



## VI. Conclusion

A lot of uncertainties are still present for the best algorithm to classify emotions. Different combinations of emotional features give different emotion detection rate. The researchers are still debating for what features influences the recognition of emotion in speech. I calculated the MFCC features for each audio file in RADESS dataset and TESS dataset, concatenated the two datasets after calculating features for classification. For model1.The model was overfitting and after experimenting with drop out and batch normalization and Epoch and optimizers, I got better accuracy on both training and test dataset. Some more Feature extraction techniques like MS feature and combination of these can also be used for better classification. Data augmentation techniques like adding noise, stretching can be used.

Link to code
https://github.com/shubhamGwasnik/Speech_Emotion-_Recognition

## VII. References

1) Kerkeni, Leila & Serrestou, Youssef & Mbarki, Mohamed & Raoof, Kosai & Mahjoub,
Mohamed. (2018). Speech Emotion Recognition: Methods and Cases Study. 175-182.
10.5220/0006611601750182.

2) Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and
Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in
Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational
Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

3) E. Ramdinmawii, A. Mohanta and V. K. Mittal, "Emotion recognition from speech
signal," *TENCON 2017 - 2017 IEEE Region 10 Conference*, 2017, pp. 1562-1567, doi:
10.1109/TENCON.2017.8228105.

4) S.R. Kadiri, P. Gangamohan, V. Mittal and B. Yegnanarayana, "Naturalistic audio-visual emotion database", *11th International Conference on Natural Language Processing*, pp. 206, 2014.