**Final Project Report**

**ITCS 6050/8050 PSYCH 6099 Topics in Intelligent Systems: Computational Human Behavior Modeling**

Emotion Detection in Speech

Prerana Singh

Vaijyant Tomar

# Introduction

The characteristics of human voice such as the pitch, timbre, loudness, and tone make human voice a versatile to communicate. It can be observed that humans can also express their emotions by varying the stated characteristics. This allows for identifying human emotion by analyzing speech.

With different emotions and moods, not only does the tonal quality vary, but the associated speech patterns change too. For instance, people may tend to talk in loud voices when angry and use shrill or high-pitched voices when in a scared or panicked emotional state. Some people tend to ramble when they get excited or nervous. On the contrary, when in a pensive emotional state, people tend to speak slowly and make longer pauses, thereby indicating an increase in time spacing between consecutive words of their speech [9].

Having an artificial agent understand raw human emotion will contribute to enhancing current state of virtual agents. Apart from making an artificial agent understand human emotion, speech sentiment analysis can also be employed in making humans more aware of the emotion of the person talking to them. Sound characteristics of speech can be used in scenarios where face to face communication is not feasible or where there are a language constraint and proper model for lexicon-based speech analysis not readily available. Following is such scenarios where speech characteristics can serve as a tool for identifying human emotion:

* + 1. Playing music and changing the ambient room’s lighting as per the tone of the conversation.
    2. Implementation in social science research
    3. Customers service centers can gather insights on their customer satisfaction by simply analyzing the speech of their customers. Also, the scores received as a part of this analysis can be used to assess the overall opinion of a company/product/services.

# Background

There has been extensive research in extracting sentiment from transcribed speech however, there has been little work on speech sentiment analysis merely from acoustic attributes of the sound. It is observed that the characteristic of the sound changes with a change in emotional state of the speaker. The speaker may say a very positive sentence but may feel otherwise of what he is speaking.

Researchers have explored classification methods including the Neural Network (NN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Maximum Likelihood Bayes classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN), Support Vector Machine (SVM) [11].

An algorithm was developed to account for three emotions namely, normal, angry, panicked. There has been an attempt to formulate an algorithm to discern the emotion associated with human speech. The analysis was carried out in MATLAB and Wavepad. The proposed algorithm investigates four vocal parameters viz., pitch, sound pressure level (SPL), timbre (ascend and descend time), and time gaps between consecutive words of speech. The authors were able to discern the emotion by retrieving the quantitative value associated with each vocal parameter [9].

There has also been the formation of various library module in python which can process audio streams. One such library in *pyAudioAnalysis*. *pyAudioAnalysis* provides packages for feature extraction and classification (including implementation of Support Vector Machines and kNN classifier). It also has provision for regression, segmentation, and visualization. What makes *pyAudioAnalysis* different from other audio libraries is that it has provision for extracting general feature which is linked to machine learning components. It also has provision for baseline techniques which are implemented for audio analysis task [8].

Sentiment analysis involving speech can possess many challenges [1].

These are as follows:

* Presence of background noise,
* foreign accents,
* generation of speech in real time,
* presence of a diverse range of topics.

There are two ways a speech can be analyzed. First, analyzing the text transcript of the speech. This would strip off the acoustic characteristics and in turn the underlying sentiment associated with speech. Second, identifying and analyzing the acoustic characters hidden in the speech. This would preserve the acoustic characteristics. Also, in this study, we consider three different datasets. Most of the studies limit themselves to only one source of data. We also implement various models and compare them.

# Project

## Approach

The proposed approach for speech sentiment analysis is as follows:

1. The Ryerson Audio-Visual Database of Emotional Speech and Song [2] (RAVDESS) data set.
2. Berlin Database of Emotional Speech dataset
3. Surrey Audio-Visual Expressed Emotion (SAVEE) dataset
4. Extract the acoustic features over the full sample. To extract the features, pyAudioAnalysis is used. [8].
5. After feature extraction, this data will be feed into training models Support Vector Machines (SVM), Random Forest, K-Nearest Neighbor and Recurrent Neural Networks.
6. To further analyze the dataset, the dataset will be splitted into male speech only samples and female speech only samples.
7. A subset of the dataset is also taken which consists of only four emotions viz., angry, happy, sad and neutral.

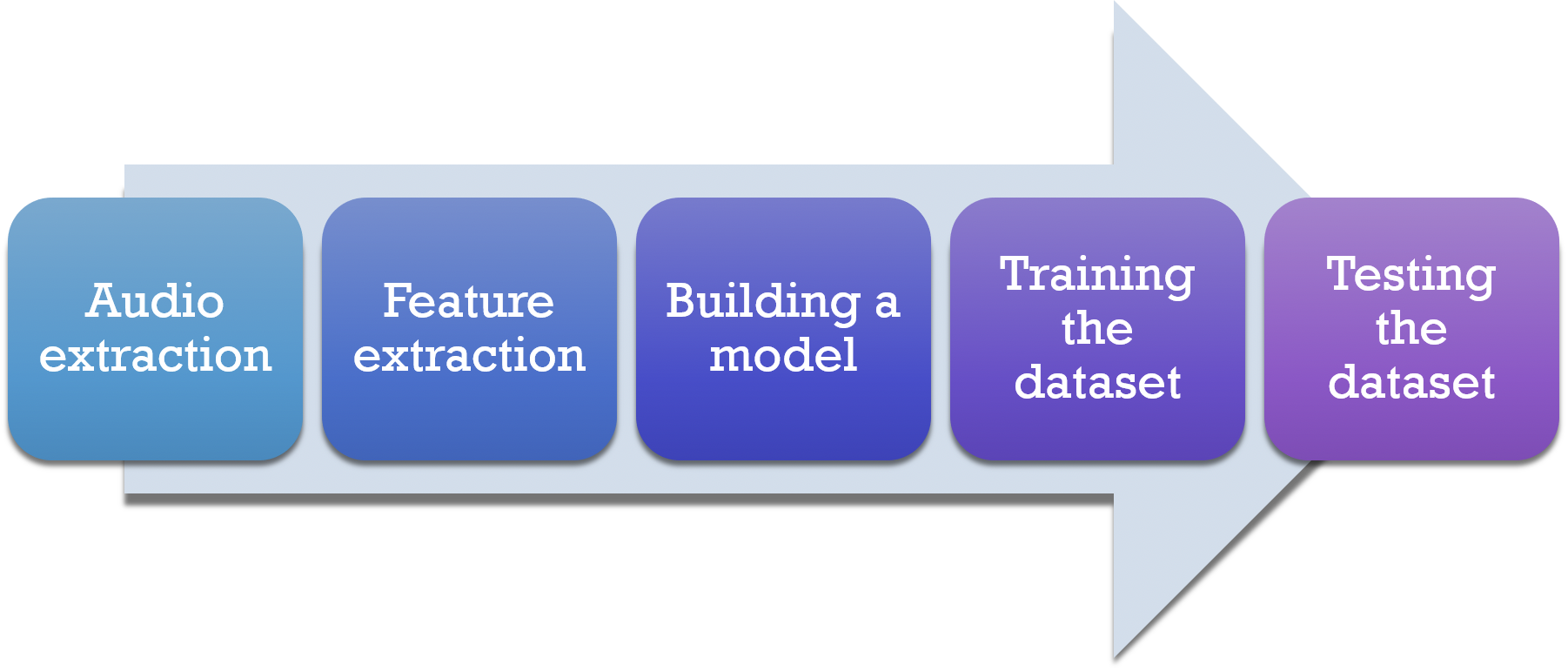


Figure III.1. Approach

## Data Collection

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is used as the database. It is a database comprising of voice samples of 24 actors (12 male, 12 female). The speech of these are in various emotions and all the actors speak in North American English accent. The dataset also contains 7,356 high-quality video recordings which correspond to the audio dataset. The dataset voice samples for eight emotional expressions viz., neutral, calm, happy, sad, angry, fearful, surprise, and disgust. The data set for song samples for six emotional expressions viz., neutral, calm, happy, sad, angry, and fearful. All emotion expressions, except neutral, are expressed at two levels of emotional intensity normal and strong. The database has been validated in by 297 participants [2].

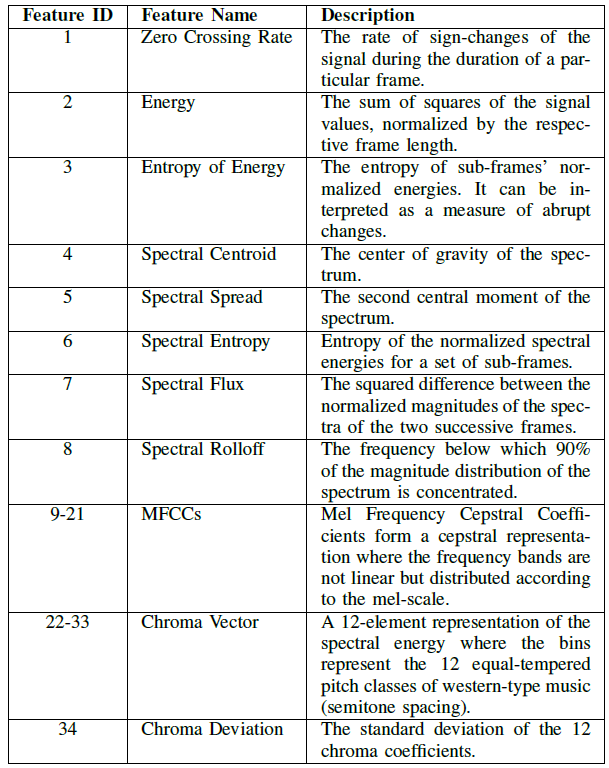
Apart from this, two more database – Berlin Database (in the German Language) and SAVEE Dataset (in British English) were also taken to compute the accuracy of the model. Data was cleaned by removing the metadata of each speech sample file by using MP3Tag.

## Feature Extraction

*pyAudioAnalysis* is the tool which is used to extract features from the speech. *pyAudioAnalysis* extracts these features in two stages. These are as follows:

* + Short-term feature extraction: pyAudioAnalysis implements this in stFetureExtraction() function of the audioFeatureExtraction.py file. It splits the audio file into short term windows (frames) and then computes features for every frame. This result in the formation of short-term feature vector for the entire audio signal.
  + Mid-term feature extraction: pyAudioAnalysis implements this in mtFeatureExtraction() function of the audioFeatureExtraction.py file. It computes the mean and standard deviation over each short-term feature vector of the audio signal [8].

Table III.1. List of features implemented [8]



## Model Implementation

The model initially employed in this study is Support Vector Machines (SVM). The model was trained with all the speech samples (RAVDESS dataset). This accounted for 1,361 speech samples. Later on , the data set was trained on KNN , Random Forest and Recurrent Neural Network.

To further investigate the data, the model was trained with only speech samples of males and females individually. There were 725 male speech samples and 636 female speech samples in the dataset that were taken for each model. These samples were again trained pm SVM, KNN, Random Forest and Recurrent Neural Network to compare their accuracy.

A subset of RAVDESS which consisted of lesser emotions (i.e. happy, angry, say and neutral) was also considered for the above mentioned models so as to compare the accuracy with the original dataset.

Data was also taken from a dataset which was having a combination of speech samples from different languages (RAVDESS + Berlin + SAVEE) so as to compare the performance of the model over it. This experiment was conducted to draw out two observations:

* Taking more data would be able to train the model more efficiently
* The model used for prediction is independent of the languages used.

## Results

All the speech samples from RAVDESS were trained on SVM initially. The accuracy obtained with this approach was 49.6% when regularization parameter *C* was set to 0.500. KNN approach yielded an accuracy of 53.7%. Later a random forest approach was also employed which yielded an accuracy of 54.5%. For RNN we found accuracy to be 39.93% after tuning the hyperparameters (activation = 'tanh', hidden\_layer\_sizes = (64, 64, 64), learning\_rate = 'constant', learning\_rate\_init = 0.001, max\_iter = 100000).

To further investigate the data, the model was trained with only speech samples of males. The accuracy obtained with this approach with SVM was 49.9% when regularization parameter C was set to 0.500, with KNN an accuracy of 52.3% was obtained and for Random Forest the accuracy obtained was 56.2%. For RNN we found accuracy to be 47.24% after tuning the hyperparameters (activation = 'tanh', hidden\_layer\_sizes = (64, 64, 64), learning\_rate = 'constant', learning\_rate\_init = 0.001, max\_iter = 100000).

Next, the model was trained with only speech samples of females. The accuracy obtained with this approach was 61.9% when regularization parameter C was set to 0.500, with KNN an accuracy of 58.0% was obtained and with Random Forest the accuracy obtained was 60.1%. For RNN we found accuracy to be 51.56% after tuning the hyperparameters (activation = 'tanh', hidden\_layer\_sizes = (64, 64, 64), learning\_rate = 'constant', learning\_rate\_init = 0.001, max\_iter = 100000).

When a subset of the RAVDESS database, was considered an accuracy of 65.6% was obtained with SVM classifier, 68.2% for KNN classifier, 66.4% was obtained with Random Forest classifier and of 59.66% for RNN classifier.

For male only subset, SVM classifier obtained an accuracy of 67.1%, KNN classifier obtained an accuracy of 69.9%, Random Forest obtained an accuracy of 70.2%, and RNN obtained an accuracy of 63.77%.

For female only subset, SVM classifier obtained an accuracy of 73%, KNN classifier obtained an accuracy of 73.9%, Random Forest obtained an accuracy of 72.8%, and RNN obtained an accuracy of 68.63%.

As evident from the above analysis, it is seen that Random Forest model had the best accuracy for the complete dataset. It also had a better classification for Random forest model trained only on male speech dataset. However, SVM model was only slightly better than random forest for female only dataset.

The subset dataset performed better than the dataset containing all the emotions. It is only natural because the complexity of emotions required to identified was reduced. KNN model performed better than other models when all the samples were taken account. However, Random Forest model outperformed when samples of only male speech were taken separately.

It was noted that the consolidated dataset (RAVDESS + Berlin + SAVEE) gave an accuracy of 51.3% on SVM Model, 62.8% on KNN, 54.05% on RNN. Thus, our first hypothesis was proved as increasing the size of data added valuable insights. Whereas, our second hypothesis was held true as the performance our models increase over the previous dataset taken.

We also considered emotion subset of the consolidated data. This gave us an accuracy of 75.1% on KNN and 75.78% on RNN.

Table III.2. Summary of Results (in percentages)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DataSet | Category | SVM | RF | KNN | RNN |
| RAVDESS | Male + Female | 49.6 | 54.5 | 53.7 | 39.93 |
| Male | 49.9 | 56.2 | 52.3 | 47.24 |
| Female | 61.9 | 60.1 | 58.0 | 51.56 |
| Male + Female (Subset) | 65.6 | 66.4 | 68.2 | 59.66 |
| Male (Subset) | 67.1 | 70.2 | 69.9 | 63.77 |
| Female (Subset) | 73.0 | 72.8 | 73.9 | 68.63 |
| RAVDESS + Berlin + SAVEE | Male + Female | 51.3 | - | 62.8 | 54.05 |
| Male + Female (Subset) | 65.6 | - | 75.1 | 75.78 |

We do not report result for Random forest because the python module of Random Forest in *pyAudioAnalysis* couldn’t internally handle the dataset.

## Challenges

Initially, openSMILE was proposed as a tool to extract features from speech samples but, it was not getting configured in the system. Thus, the tool was replaced by pyAudioAnalysis, which initially was not getting loaded.This issue was resolved by installing python 2.7 instead of the higher versions. Apart from this, Libmagic library wasn’t loading, this was resolved by installing compatible lib magic module.

On processing the feature extraction though pyAudioAnalysis, chunks in \*.wav files weren’t being parsed. Hence, the metadata information of each file was being removed by Audacity. Audacity doesn’t have the ability of batch processing, thus making the data cleaning process very time-consuming and required a lot of manual effort. To reduce the manual effort and time in data cleaning, another tool, MP3Tag was taken which was focused more on batch processing.

The data taken was prompted and it is not feasible to take emotional data from human beings when naturally speaking.

## Roles

|  |  |  |
| --- | --- | --- |
|  | Team Member Name | Responsible For |
| 1 | Prerana Singh (800973733) | Literature Review, Data Collection, Implementing Models, Review, Documentation |
| 2 | Vaijyant Tomar (800990636) | Literature Review, Implementing Models, Feature Extraction, Review, Documentation |

# Summary

In this project, we aimed at extracting emotion from human speech samples. This area has not been explored. Hence, we explore and then provide an in-depth analysis of various approaches we can employ in extracting emotion from speech. This study can find its use in emotion recognition from acoustic signals picked up by Intelligent Agents. Current systems only predict the emotion based only on the sequence of words in a speech. There has been little importance given to underlying emotion associated with words in a speech. We explore the acoustic parameters of sound samples and predict what the emotional state a person is.

In this study, we use three datasets namely, RAVDESS, Berlin, and SAVEE. We found that increasing the number of samples increases the accuracy of the results. We also found out that a model trained on audio samples from females yielded better accuracy. Also, it has been observed that a better accuracy for classification was obtained when fewer emotions were used. This is because the approach we have to consider is unable to discern a wider spectrum of emotions.

# Conclusion

The accuracy obtained in this study cannot be the compared to other studies as there is a difference in the datasets. Also, the dataset used here was originally intended to incorporate visual cues such as facial features which were recorded in video samples and contained songs sung by various actors. The accuracies obtained for data from female speakers is more than with the male speakers. This is in accordance with other studies which found that accuracies on a data consisting of only female speakers are higher than the data consisting of only male speakers.

Majority of the datasets available for Speech Sentiment Analysis have prompted emotions. That is it involves speech samples which are produced by identical utterances of a speech in a given emotion. As these speeches are a deliberate effort it may not always be like an unprompted speech which are more natural in nature. However, the major drawback in obtaining unprompted speech samples is that it would require more human effort and time. It would also involve recording speech samples all the time which may lead to privacy concerns.

Incorporating a mechanism which would in real time update the models could also be looked into. However, this approach would require the active interaction of a participant to annotate the data in real time.

Supplementing the audio dataset with visual cues like facial features may help in predicting the broader spectrum of emotion. This can be taken further as a future work in the research to improve the accuracy of the predictions. Also, tuning the parameters of the models further can also lead to better accuracy that can be taken into future work.

# References

[1] L. Kaushik, A. Sangwan, and J. H. Hansen, “Sentiment Extraction

from Natural Audio Streams”

https://www.utdallas.edu/˜john.hansen/Publications/CP-ICASSP13-

KaushikSangwanHansen-Sentiment-0008485.pdf

[2] S. R. Livingstone, K. Peck, and F. A. Russo (2012). RAVDESS: The

Ryerson Audio-Visual Database of Emotional Speech and Song. Paper

presented at the 22nd Annual Meeting of the Canadian Society for Brain,

Behaviour and Cognitive Science (CSBBCS), Kingston, ON.

[3] W. Medhat, A. Hassan, H. Korashy, “Sentiment analysis algorithms and

applications: A survey”, Ain Shams Engineering Journal, Sciencedirect,

27 May 2014

[4] Y. Chavhan, M. L. Dhore and P. Yesaware, “Speech Emotion Recognition

Using Support Vector Machine.” International Journal of Computer

Applications 1(20):69, February 2010. Published By Foundation of

Computer Science.

[5] Mp3tag. Florian Heidenreich. March 20, 2018. Retrieved March 20,

2018.

[6] Audacity® software is copyright ©1999-2018 Audacity Team. The name

Audacity® is a registered trademark of Dominic Mazzoni.

[7] C. Sobin and M. Alpert, “Emotion in Speech: The Acoustic Attributes

of Fear, Anger, Sadness, and Joy.” Journal of Psycholinguistic Research,

Vol. 28, No. 4, 1999

[8] T. Giannakopoulos (2015) pyAudioAnalysis: An Open-Source Python

Library for Audio Signal Analysis. PLoS ONE 10(12): e0144610.

https://doi.org/10.1371/journal.pone.0144610

[9] P. B. Dasgupta, “Detection and Analysis of Human Emotions through

Voice and Speech Pattern Processing.” International Journal of Computer

Trends and Technology (IJCTT) Volume 52 Number 1 October 2017

[10] R. Lyon and S. Shamma, “Auditory Representation of Timbre and Pitch”.

Harold L. Hawkins & Teresa A. McMullen. Auditory Computation.

Springer. pp. 22123. ISBN 978-0-387-97843-7, 1996

[11] S. S. Jarande, S. Waghmar, “Speech based Human Emotion Recognition

Using Hybrid Classifier Technique” International Journal of Advanced

Research in Computer Science and Software Engineering