



A Project Report on  
**Emotion Detection with Speech Analytics**

Submitted in partial fulfilment for award of degree of

**MBA**  
**In Business Analytics**

Submitted by

**Krishna Gopal Goswami**  
R17DM006

Under the Guidance of

**Dr. J B Simha**  
Chief Technology Officer, ABIBA Systems

REVA Academy for Corporate Excellence

**REVA University**  
Rukmini Knowledge Park, Kattigenahalli,  
Yelahanka, Bangalore – 560064

**September 2020**



### **Candidate's Declaration**

I, Krishna Gopal Goswami hereby declare that I have completed the project work towards the MBA in Business Analytics, at REVA University on the topic entitled Speech Analytics under the supervision of Dr J.B. Simha. This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year 2020.

Place: Bengaluru

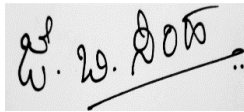
Name of the Student: Krishna Gopal Goswami

Date: 06/10/2020

Signature of Student:

## Certificate

This is to Certify that the PROJECT work entitled Emotion Detection and Analysis using Speech Analytics, carried out by Krishna Gopal Goswami with SRN R17DM006, is a bonafide student of REVA University, is submitting the project report in fulfilment for the award of MBA in Business Analytics during the academic year 2020. The Project report has been tested for plagiarism and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of PROJECT work prescribed for the said Degree.



Dr. J. B. Simha  
Guide

<Signature of the Director>

<Name of the Director>

Director

External Viva

Names of the Examiners

1. <Ravi Shukla> <Sr. Advisor and Data Scientist> <Signature>
2. <Krishna Kumar Tiwari> <Senior Data Scientist> <Signature>

Place: Bengaluru

Date: 06/10/2020





## **Acknowledgement**

It would not have been possible to complete this project without the kind support and help of many individuals and I am thankful to all of them.

First, I would like to thank the honourable Vice Chancellor, Dr Shyam Raju (REVA University) for giving me the opportunity to work in this project.

Also, I would like to thank Dr. S.Y Kulkarni, Ex-Vice Chancellor, Dr. K. Mallikarjuna Babu, Vice-Chancellor, Dr. M. Dhanamjaya, Dr. Shinu Abhi, Director Corporate Training and Management Team, REVA University for their help and support.

I thank my mentor Dr. J.B. Simha for his support and guidance during the tenure of my project.

Krishna Gopal Goswami  
(R17DM006)

Place: Bengaluru  
Date: 06/10/2020



## Similarity Index Report

Title of the Thesis: Emotion Detection with Speech Analytics

Total No. of Pages: 40

Name of the Student: Krishna Gopal Goswami

Name of the Guide(s): Dr. J.B. Simha

This is to certify that the above thesis was scanned for similarity detection. Process and outcome is given below.

Software Used: Turnitin

Date of Report Generation: 18/09/2020

Similarity Index in %: 14%

Total word count: 4344

Place: Bengaluru

Krishna Gopal Goswami

Name of the Student:

Date: 06/10/2020

Signature of Student

Verified By:

Signature

Dr. Shinu Abhi, Director, Corporate Training

## List of Abbreviations

Sl. No	Abbreviation	Long Form
1	ML	Machine Learning
2	ASR	Automatic Speech Recognition
3	HMM	Hidden Markov Models
4	GMM	Gaussian mixture models
5	CRISP-DM	Cross-Industry Standard Process for Data Mining
6	NRC	National Research Council
7	RNN	Recurrent Neural Networks
8	MFCC	Mel-Scale Frequency Cepstral Coefficients
9	PCA	Principal Component Analysis
10	SER	Speech Emotion Recognition

## List of Figures

Fig No.	Name	Page No.
Figure No. 1	Basic Process flow of Speech Analytics	12
Figure No. 2	CRISP-DM High Level Steps	18
Figure No. 3	Word Cloud Analysis on Audio Clip1	21
Figure No. 4	Word Cloud Analysis on Audio Clip2	22
Figure No. 5	Data-Preparation Steps for Video/Audio Files	23
Figure No. 6	Data-Preparation Steps for Text Files	24
Figure No. 7	Deep Speech Architecture	25
Figure No. 8	High-Level Flow Diagram of audio-to-text conversion	26
Figure No. 9	High-Level Flow Diagram of Emotion Analysis using Lexical Features	27
Figure No. 10	High-Level Flow Diagram of Emotion Analysis using Acoustic Features	28

Figure No. 11	Emotion Detection and Analysis using NRC Lexicon (Audio Clip 1)	31
Figure No. 12	Emotion Detection and Analysis using NRC Lexicon (Audio Clip 2)	32
Figure No. 13	Basics of Sound Vibrations	33
Figure No. 14	Frequency Visualization of Audio Clip1	34
Figure No. 15	Frequency Visualization of Audio Clip2	34
Figure No. 16	Spectrogram Analysis of Audio Clip1	35
Figure No. 17	Spectrogram Analysis of Audio Clip2	36

### List of Tables

No.	Name	Page No.
Table No. 1	Deep Speech model performance comparison	29



## **Abstract**

Speech is the best, old and natural mode of communication among the human and it is the most natural way to express ourselves. Emotions are biological states associated with our nervous system and through speech, it helps us express our feelings, thoughts and degree of pleasure or displeasure (wiki, n.d.). Emotions can have different forms like happiness, fear, anger, joy etc. We use emojis to express ourselves in other forms of communications (e.g. emails, text messages) also and this reflects the importance of emotion in today's world.(Wadhwa, 2020)

Emotion recognition from speech has become one of the active research themes in speech analytics. However, Emotion detection is very challenging because emotions are subjective in nature and there is no definite guide about how to measure emotions.

As part of this study, we have explored how emotion is perceived by listener and how can we recognize the emotional state in a speech. We have defined a Speech Emotion Detection system as a collection of methodologies that can process any speech signals and detect emotions associated with it . (Laurence Devilliers, 2005)This study covers how we can perform emotion detection and analysis on any video or audio input, using Open-Source Machine Learning and Deep Learning Algorithms.

This study explains the methodologies for the implementation of Audio to text conversion using Mozilla's Deep Speech Algorithm and Emotion Analysis using NRC Lexicon Library and Python's Librosa library.

***Keywords:* Speech Analytics, Text Analytics, Emotion Analytics, NRC Lexicon, DeepSpeech, Speech Recognition, Emotion Recognition, Spectrogram Analysis, Tone Analysis, Speech Emotion Recognition.**

## Contents

Candidate's Declaration.....	2
Certificate.....	3
Acknowledgement .....	5
Similarity Index Report.....	6
List of Abbreviations .....	7
List of Figures .....	7
List of Tables .....	8
Abstract.....	9
Chapter 1: Introduction .....	11
Chapter 2: Literature Review .....	14
Chapter 3: Problem Statement .....	16
Chapter 4: Objectives of the Study .....	17
Chapter 5: Project Methodology .....	18
Chapter 6: Business Understanding .....	20
Chapter 7: Data Understanding.....	21
Chapter 8: Data Preparation.....	23
Chapter 9: Data Modeling.....	25
Chapter 9: Data Evaluation .....	29
Chapter 10: Deployment .....	30
Chapter 11: Analysis and Results .....	31
Chapter 12: Conclusions and Recommendations for future work .....	37
Bibliography .....	38
Appendix.....	41
Plagiarism Report.....	41
Publications in a Journal/Conference Presented/White Paper .....	45

## **Chapter 1: Introduction**

Speech is defined as the expression of thoughts and feelings by articulating sounds.

Speech analytics can be defined as the process of analysing the actual speaking voice of a person using different types of technology and gaining insights from the conversations. Speech Analytics has become an increasingly popular topic in recent years, and it has been used in many areas like healthcare, marketing, call centre, digital assistants etc.

As per Wiki, Emotions are biological states associated with the nervous system brought on by neurophysiological changes variously associated with thoughts, feelings, behavioural responses, and a degree of pleasure or displeasure(wiki, n.d.). Emotions are reflected from speech or through hand and gestures of the body or through facial expressions. Speech signal contains information about the text that is spoken, the language in which it is spoken and the emotional state of the speaker. It also conveys the mood of the speaker by variations in pitch, loudness, intonation, pause and other such features.

As Emotions play an important role in human life, emotion detection is very important in today's world. Emotions are subjective in nature and there is no universal descriptor for different types of emotions. This has made the recognition of emotions in a speech signal, a very challenging and complex task. (Bertero & Fung, 2017)(Wadhwa, 2020) .

In this study, we have followed a novel approach for Emotion Detection and Analysis on any input speech, using open source technologies and algorithms.

1. Mozilla's Deep Speech model is used for Speech-to-Text conversion and audio input file is converted into text file.
2. For Emotion Detection, two different techniques are used:

- National Research Council (NRC) Lexicon for analysing Lexical features (words, vocabulary) on the text file.
- Python Librosa Library for analysing Acoustic features (loudness, pitch etc.) on the Audio input.

The below diagram shows the process flow of our approach for Emotion Detection using Speech Analytics:

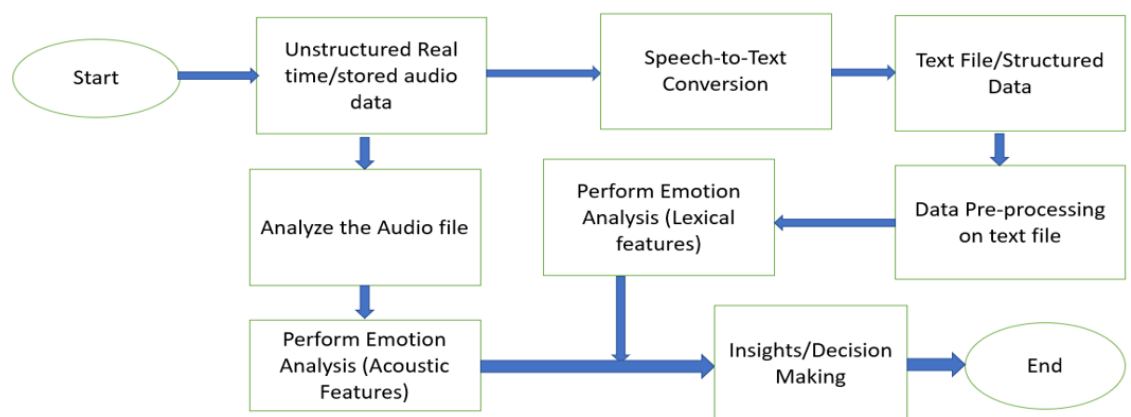


Fig1: Process flow of Emotion Detection

Deep Speech is open source Speech-To-Text engine and model is trained and built using deep learning techniques. It is based on Baidu's Deep Speech research paper and it uses Google's TensorFlow to make the implementation easier.(Mozilla, 2020)

The NRC Emotion Lexicon is also called EmoLex, it is a list of English words and their associations with eight basic emotions. It has 14,182 unigrams (words) and approximately 25000 senses associated with the 8 emotions. The emotion types are anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. The annotations were done manually by crowdsourcing on Mechanical Turk.(Mohammad, 2020)

Librosa is a very popular python package used for audio analysis and provides the building blocks required to create audio information retrieval systems. (Librosa, 2013)

## Chapter 2: Literature Review

Speech Analytics and Emotion Detection have been active research area for more than a decade and various technologies and methods are invented in the same area. Emotions play an important role in human communication.

For last few decades, Speech Analytics is one of the main driving forces behind many machine learning (ML) techniques. The machine learning techniques includes Hidden Markov model, discriminative learning, structured sequence learning, Bayesian learning etc. (Deng & Li, 2013)

For Speech processing, both deterministic and stochastic models are used, and Hidden Markov model (HMM) is one of the stochastic models.(Rabiner, 1989) . HMM is mainly used in Acoustic Modeling. Hidden Markov models are based on simple set of concepts and assumptions. Bayesian network is another way of expressing and computing with probability distributions. It is used in Speech Recognition to explicitly model factors such as acoustic context, speaking rate etc.(Nefian et al., 2002).

Mel-frequency cepstral coefficients (MFCC) is another audio extraction method. In Speech recognition system using MFCC, only 16 coefficients of MFCC corresponding to the Mel scale frequencies are used. They are extracted from spoken word database. After extraction, they are analysed statically by using PCA.(Ittichaichareon, 2012).(Tiwari, 2010). MFCC along with Support Vector Machine (SVM) can be used in Speech Emotion Recognition system. Features are extracted using pitch, formats, MFCC and speaker dependent SER can be improved by comparing the results with different kernels of Support Vector Machine classifier.(Dahake et al., 2017)

Traditional Speech Recognition Technologies are based on heavily engineered processing stages e.g. acoustic models, HMM etc. (Awni Hannun, 2014)Tuning of their features and models require lots of efforts by domain experts. (Devillers et al., 2005)

Use of deep learning algorithms has improved the performance of Speech Recognition systems, by improving acoustic models. However, to improve performance during speech recognition in a noisy environment, there is still need of lot of efforts required to build the overall system for robustness. (Awni Hannun, 2014)

Mozilla's Open Source Deep Speech approach is simpler to use and in it has achieved higher performances than any traditional systems in noisy and speaker variation environment also.(Hannun et al., 2014).(Awni Hannun, 2014) It applies deep learning end-to-end using RNN and takes advantages of the deep learning systems trained on large datasets.

Spectrograms can be considered as heat wave representation of speech signals and are commonly used to display frequencies of sound waves produced by humans, machinery, animals, whales, jets, etc., as recorded by microphones and it has been used extensively in audio file analysis in many applications. (Yenigalla et al., 2018)

However, the existing tools and technologies of Speech Emotion Detection do not combine the analysis of emotion detection based on both Acoustic features (tone, pitch etc.) and Lexical features(vocabulary) of the speech signal. As part of our approach, we have used text analytics for emotion detection based on Lexical Feature and used Spectrogram and Audio Wave Analysis for emotion detection based on Acoustic Features.

### Chapter 3: Problem Statement

According to 7-38-55 Rule of Personal Communication, there are three elements involved in about how humans express their feelings during communication : (Mehrabian, n.d.)

- Lexical features (the vocabulary used) account for 7%
- Acoustic features (pitch, tone, jitter, etc.) account for 38%
- Visual features (the expressions the speaker makes) accounts for 55%

During face-to-face communication, we are covering only one channel of communication (Visual features). We also need to analyze other two channels for emotion detection and as part of this project, we are trying to capture both **Lexical features and Acoustic features channels** of personal communication.



## **Chapter 4: Objectives of the Study**

The project objective is to develop an end-to-end methodology and Speech Emotion Recognition system using open-source tools and technologies.

There are many commercially available Speech-to-Text software available. However, in this project, we are using Mozilla's Deep Speech Model, which is one of the best Speech-to-Text technologies in the current market.

For Emotion Detection based on Lexical Feature, we are using NRC library.

For Emotion Detection based on Acoustic features, we are using Python's Librosa library.

## Chapter 5: Project Methodology

For this project, we have used CRISP-DM project methodology.

It involves 6 steps which are captured below:

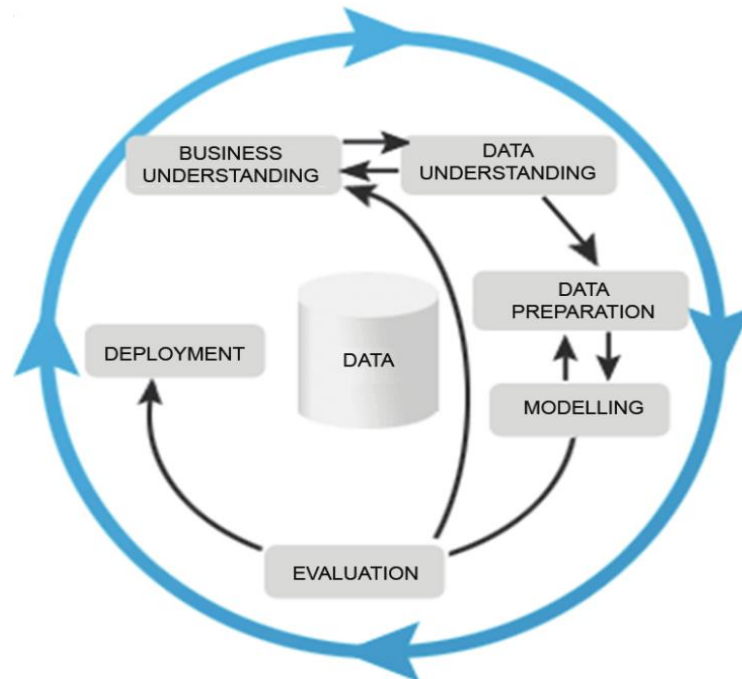


Fig 2: CRISP-DM High Level Steps

**Business Understanding** — The goal of this stage is to understand the business goal and then convert it into a measurable and specific project goals and then formalize it as problem statement.

**Data Understanding** — The goal of this stage is to gather data and then explore and comprehend the data.

**Data Preparation** – The goal of this stage is to select the final data which will be relevant to the data mining objectives, clean and transform the data.

**Data Modeling** - The goal of this stage is, to apply the modelling techniques and record it.

**Model Evaluation** – The goal of this stage is, to assess the degree to which model meets the business requirements and also to test the model in the real applications.(sv-europe.com, n.d.)

**Deployment** - The goal of this stage is to determine the model deployment strategy based on evaluation result and plan for monitoring and maintenance of models in the business environment, (sv-europe.com, n.d.)

## Chapter 6: Business Understanding

Automatic Speech Recognition has become an increasingly popular concept in recent years, and it is currently used in multiple domains like healthcare, marketing, digital assistants, finance, customer care etc. It has been active research theme for many years.

Why Speech and Emotion Analytics are important for organizations:

- There is more pressure for businesses to deliver a better customer experience due to increased customer demand and availability of different customer channels.
- If Speech Emotion Analytics is applied correctly and if integrated well with overall strategy and used effectively, it can help businesses drive product & process innovation leading to significant growth and market differentiation.
- Some of the major benefits of an organization, by implementing Speech Analytics are:
  - Think Contextually to get to the root cause
  - Integrate Speech Analytics with Quality monitoring process
  - Better Segmentation of customers
- With new technologies e.g. speech and emotion analytics, enterprises can create better customer experiences. This is applicable for any enterprise, e.g. retail industry, media industry, travel & tourism etc.

## Chapter 7: Data Understanding

We have used 2 audio files for our study as part of this project.

The audio files are downloaded from the below CallHome English Corpus Link:

<https://ca.talkbank.org/access/CallHome/eng.html>

The corpus of telephone speech was collected and transcribed by the Linguistic Data Consortium and sponsored by the U.S. Department of Défense. The corpus consists of 120 unscripted telephone conversations between native speakers of English.(LDC, n.d.)

**Note:** We can also take multiple audio/video files and perform the analysis, there is no limitation on this.

Each Audio file is of length approximate 5 mins.

Audio File1 contains 777 words.

Audio File2 contains 853 words.

### Word Cloud Analysis:



Fig 3: Word Cloud Analysis on Audio Clip1



## Chapter 8: Data Preparation

### Data Preparation for Video/Audio Files:

The input for Deep Speech model should be in 16000 Hz .wav format. So, any video/audio file needs to be converted into 16000 Hz .wav format file.

Following diagram shows the high-level architecture of data preparation for Video/Audio file:

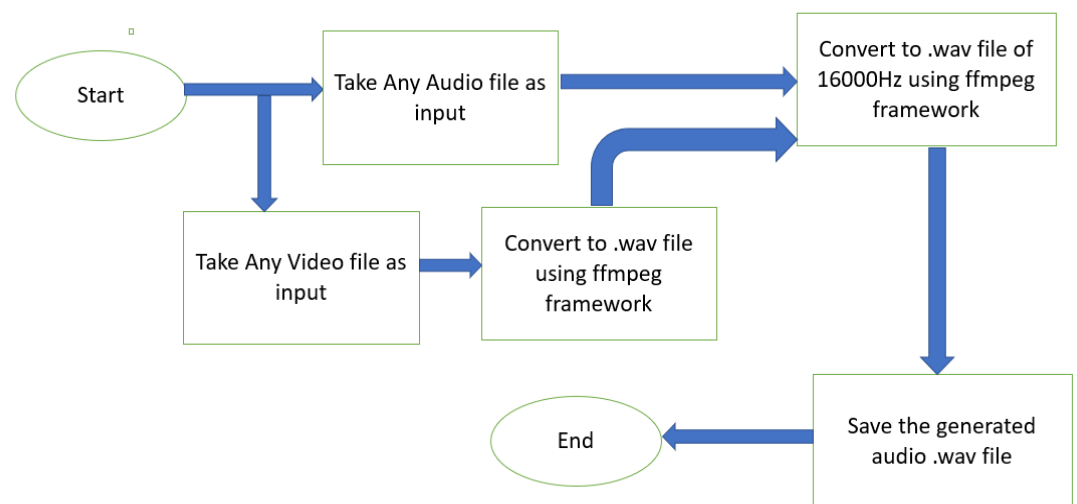


Fig 5: Data-Preparation Steps for Video/Audio Files

- Step 1: If the input file is mp4 video file, first we need to convert the file into .wav file using ffmpeg. ffmpeg is open-source multimedia framework and this can be used to decode, encode, transcode almost everything that humans and machines have created.
- Step 2: Change the sampling rate of the audio file to 16000Hz using ffmpeg.

### Data Preparation for Text Files:

- The text file generated in the previous step as part of Speech-to-Text is used as input for data preparation step of text file.
- The Text file is tokenized using python nltk library
- The text file is cleaned using python nltk library (stop words removal, punctuation removal, special character removal etc.)
- Cleaned texts is saved as text file, so that it can be used for further analysis.

Following diagram shows the high-level architecture of data preparation for text file:

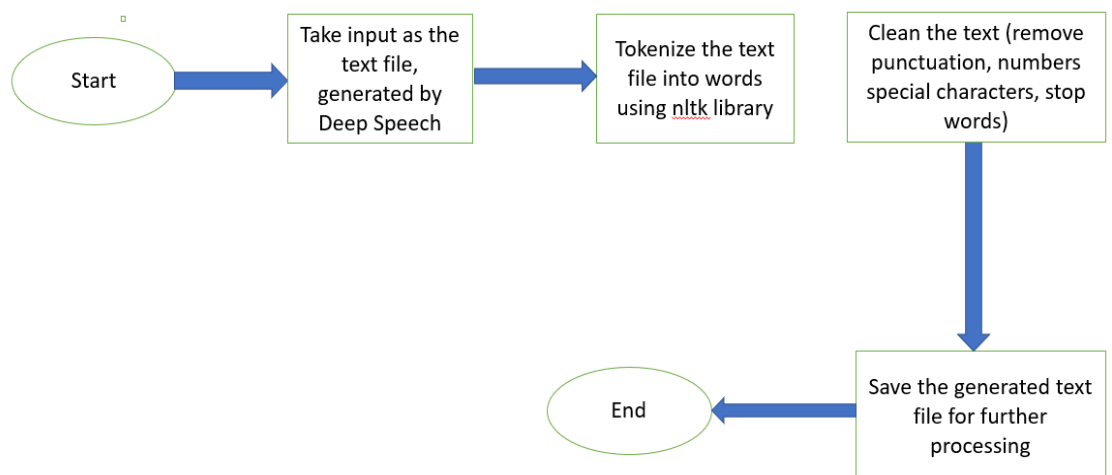


Fig 6: Data-Preparation Steps for Text Files



## Chapter 9: Data Modeling

As part of data modelling, we have used 3 approaches:

- Using Deep Speech Algorithm to convert from Speech to Text.
- Perform Emotion Analytics using Lexical features on the generated texts using NRC lexicon.
- Perform Emotion Analytics using Acoustic features (Tone Analysis) on the Audio/Video file.

### Deep Speech Architecture

The English DeepSpeech model was trained on 3816 hours of transcribed audio.

It was collected from LibriSpeech, Fisher, Common Voice English, Switchboard.

The model also includes 1700 hours of transcribed WAMU (NPR) radio shows.(Morais, 2019)

Below diagram shows the Deep Speech Architecture:

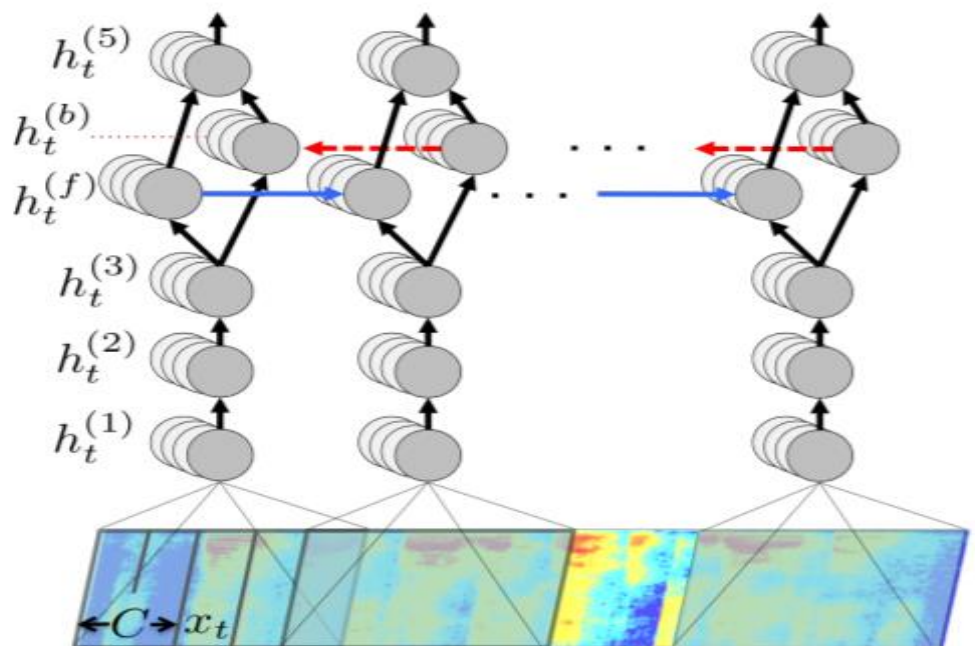
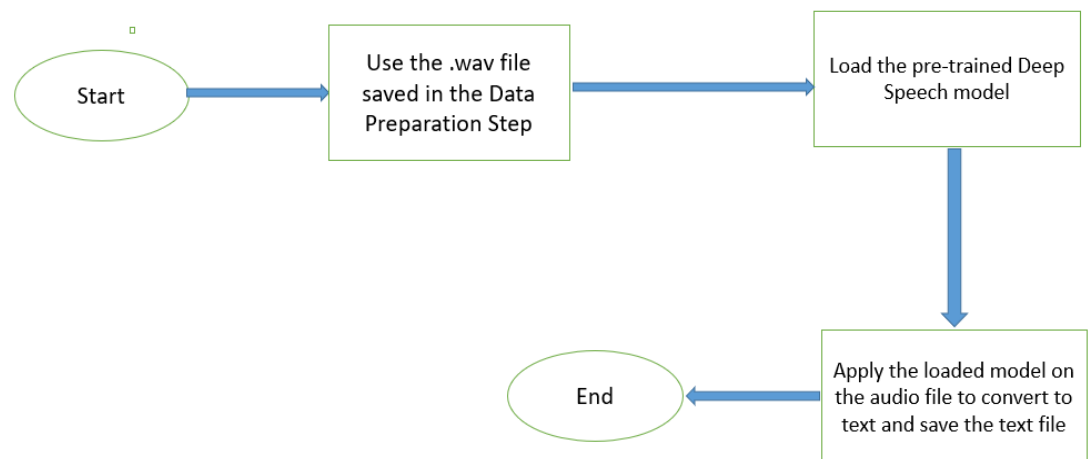


Fig 7: Deep Speech Architecture

- RNN model consists of 5 layers of hidden units.
- The first 3 layers are not recurrent.
- The 4th layer is a bi-directional recurrent layer. This layer includes two sets of hidden units- forward recurrence  $h(f)$ , and backward recurrence  $h(b)$ (Awni Hannun, 2014)
- The 5<sup>th</sup> layer is a non-recurrent layer and its inputs are both the forward and backward units.(Awni Hannun, 2014)

### **Audio to Text Conversion using Deep Speech Algorithm**

The high-level flow diagram for the conversion of Audio file to Text file is given below:



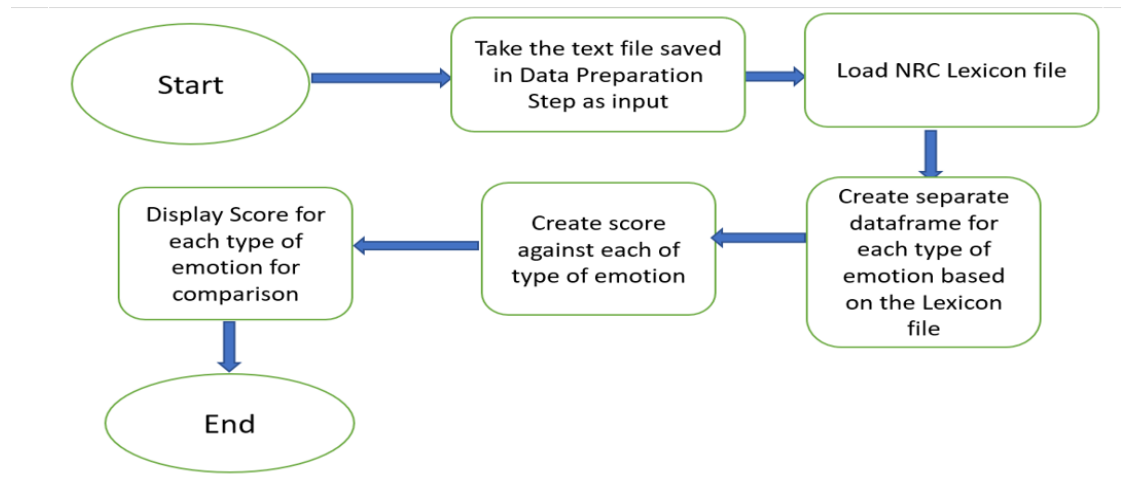
**Fig 8: High-Level Flow Diagram of audio-to-text conversion**

Step 1: Load the DeepSpeech pre-trained model.

Step 2: Once the model is loaded, we applied it on the audio file to convert from audio to text file. The generated text file is stored for further text processing.

### **Emotion Analysis using Lexical Features on Extracted Text File**

The high-level flow diagram for the conversion of Emotion Analysis and Comparison using NRC Lexicon is given below.



**Fig 9: High-Level Flow Diagram of Emotion Analysis using Lexical Features**

NRC Lexicon model is used for the purpose of performing emotion analysis in the generated text files. Scoring can be generated for the 8 types of emotions and they are: anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

Once the score is generated, we can use the compare the different emotion types associated with the audio file for single speaker or multiple speakers.

### **Emotion Analysis using Acoustic Features on the Audio inputs**

Acoustic features is analyzed using python Librosa package.

The high-level steps involved are:

1. Take the original audio file as input and load it into as floating-point time series, using Librosa load package.

2. Apply Librosa display.waveplot package to display the time-frequency graph. This graph will help in the analysis of tone of the audio input file.

3. Apply Librosa display.specshow to show the Spectrogram. This will help in the analysis of loudness(volume) of the audio input file.

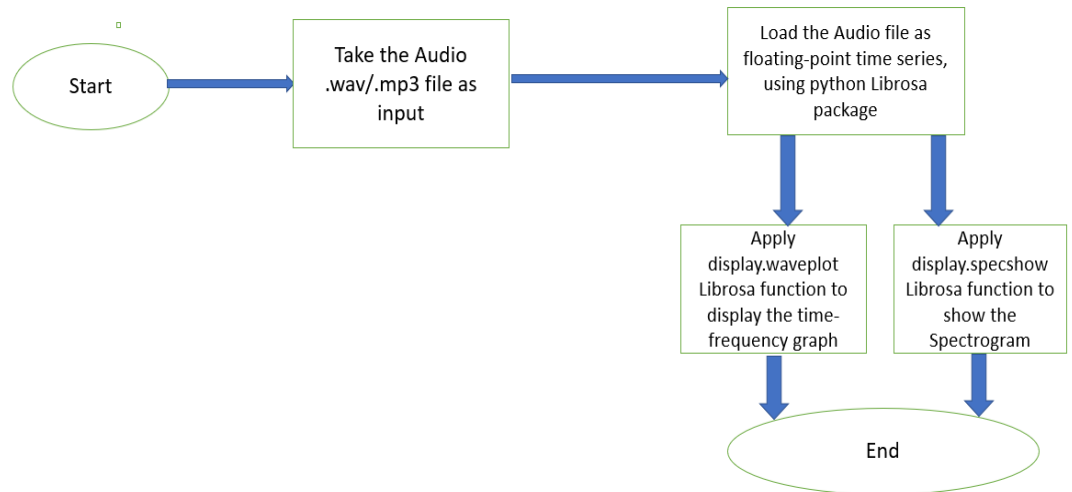


Fig 10: High-Level Flow Diagram of Emotion Analysis using Acoustic Features

## Chapter 9: Data Evaluation

Below Data Evaluation results are captured from Deep Speech Paper.(Hannun et al., 2014)

The dataset used for training the model consists of 5000 hours of read speech from 9600 speakers. Two RNNs were trained, one on 5000 hours of data without noise and the other one on 5000 hours of noisy data.

On the 100 audio clips without noise, both models perform about 9.2 % WER (Word Error Rate). On the 100 noisy audio clips, the noisy RNN model achieves 22.6% Word Error Rate over the clean model's 28.7% Word Error Rate, a 6.1% improvement.(Awni Hannun, 2014)

The below table compares the Word Error Rate of different Speech Recognition Systems:

System	Clean (94)	Noisy (82)	Combined (176)
Apple Dictation	14.24	43.76	26.73
Bing Speech	11.73	36.12	22.05
Google API	6.64	30.47	16.72
wit.ai	7.94	35.06	19.41
<b>Deep Speech</b>	<b>6.56</b>	<b>19.06</b>	<b>11.85</b>

Table 1: Deep Speech model performance comparison

## **Chapter 10: Deployment**

After performing Data Processing, any audio/video inputs can be feed into the Deep Speech model and it can be converted into text file.

Once the text file is generated, we can perform Emotion Analysis on Lexical features using python and NRC lexicon library.

Emotion Analysis on Acoustic features can be performed on the audio input files, using python Librosa package.

The details about the result are captured in the below section Analysis and Results.

## Chapter 11: Analysis and Results

### Emotion Analysis using NRC Lexicon:

Below we have captured the emotion detection and analysis of the 2 audio clips using NRC Lexicon.

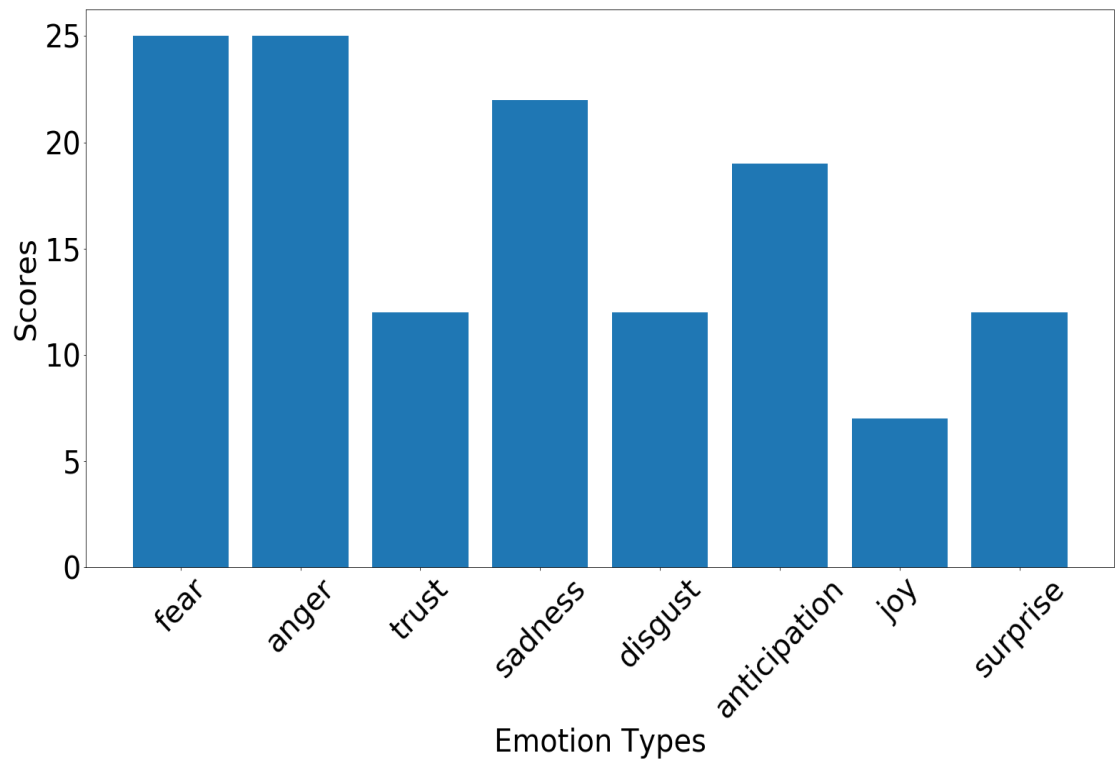


Fig 11: Emotion Detection and Analysis using NRC Lexicon  
(Audio Clip 1)

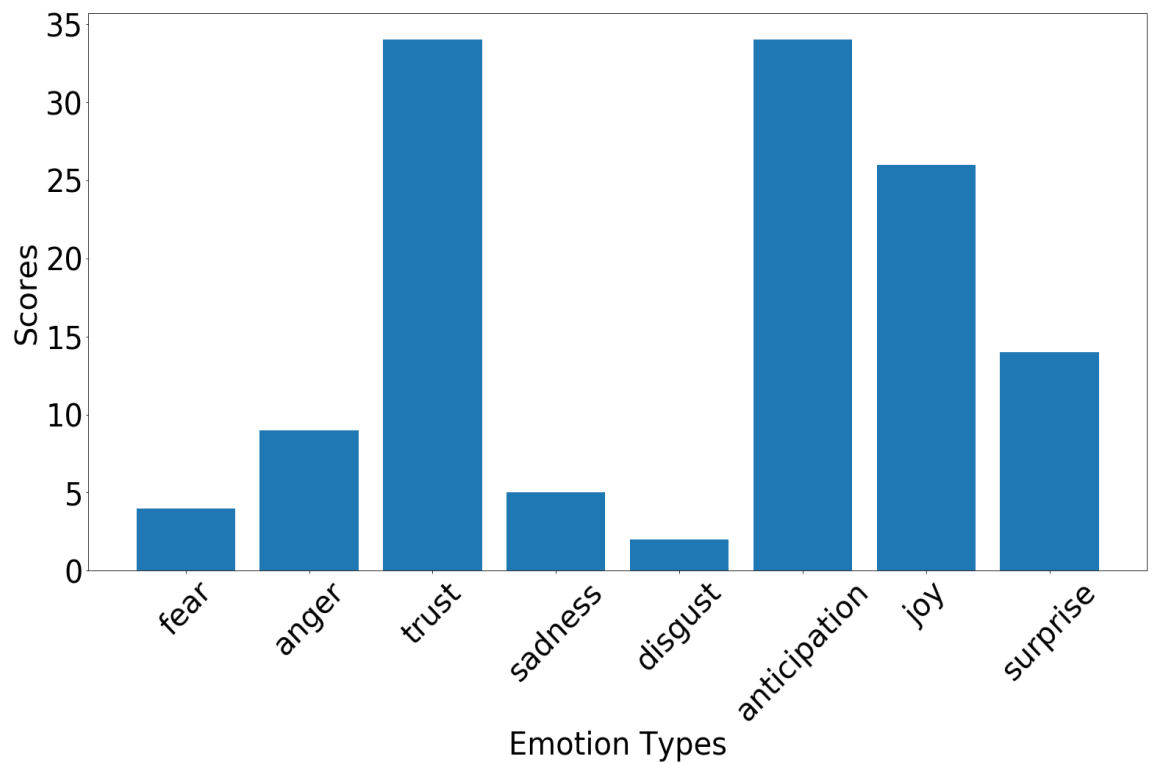


Fig 12: Emotion Detection and Analysis using NRC Lexicon (Audio Clip 2)

From the above graph, it is noticed that:

For Audio Clip1, there is more fear, anger, sadness and very little joy.

For Audio Clip2, there is trust, anticipation and joy but very little fear and anger.

### **Visualizing the Audio Files**

#### **Basics of Speech Waveform:**

Before visualizing and analysing the audio files, below we have summarized few basic terms associated with any audio file:



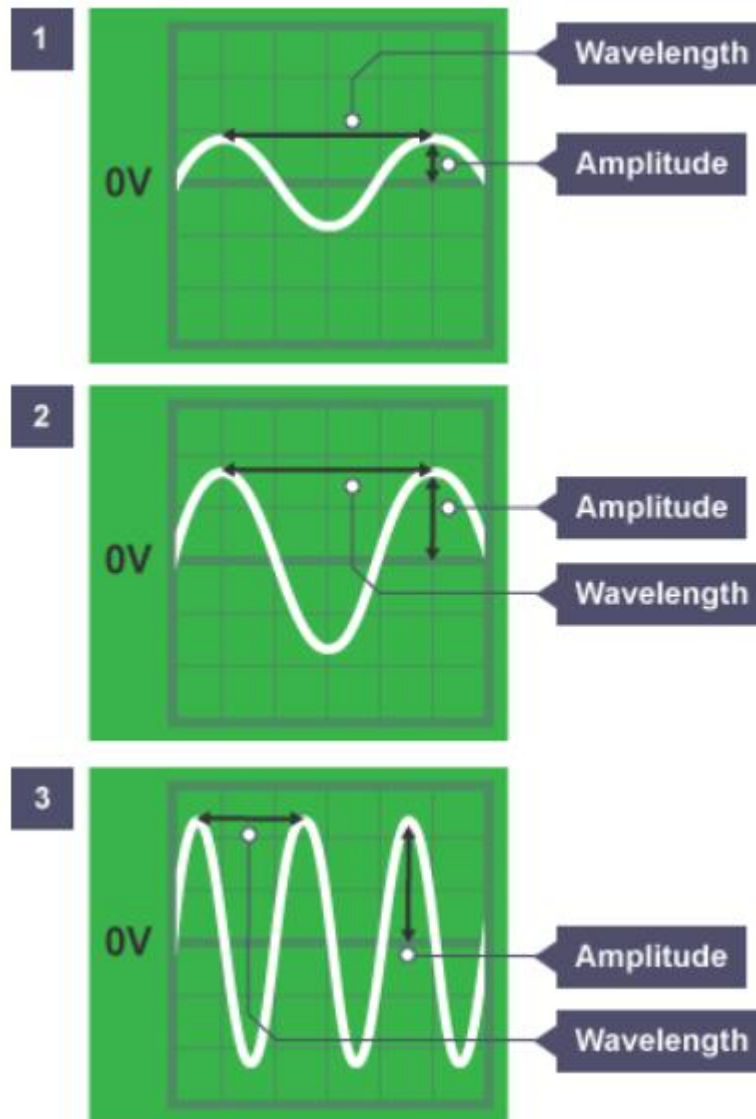


Fig 13: Basics of Sound Vibrations

Amplitude (Volume) is shown by the height of the waves. Higher waves means higher loudness(volume).

Pitch (Frequency) is shown by the spacing of the waves and refers to the degree of high or low of the speech signal. A high pitch means it has high frequency and a low pitch means it has low frequency.

In the above diagram, Sound 2 has higher amplitudes than Sound1, it means Sound2 has more loudness/volume.

Sound2 and Sound3 has same amplitudes(volume) but higher pitch(frequency).

Below diagram plots the amplitude envelope of the waveform using `display.waveplot` function of python Librosa library.

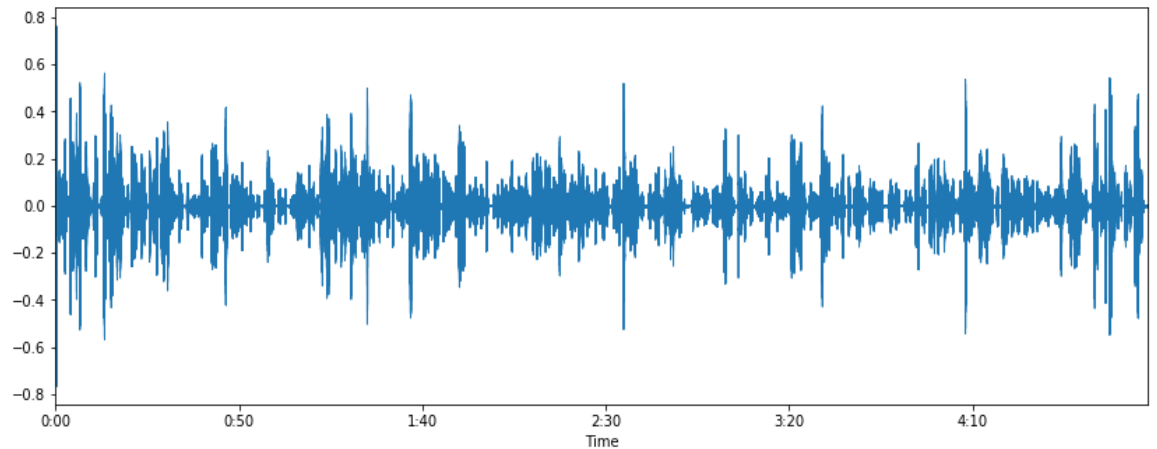


Fig 14: Frequency Visualization of Audio Clip1

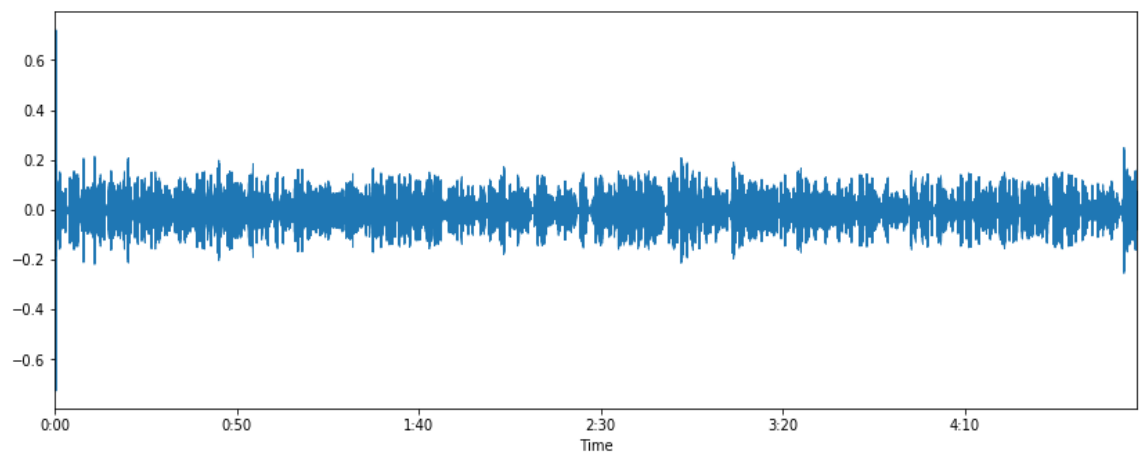


Fig 15: Frequency Visualization of Audio Clip2

Above visualization shows that there is high frequency means high pitch in the conversation for Audio Clip1. Emotions has an effect on the pitch e.g., A person speak may speak in a higher pitch than usual during sudden emotions like anger, fear etc. Similarly, the pitch of a tired person may be lower.

As per NRC Lexicon Analysis, for Audio Clip1, there is high degree of Anger and Fearness and this is also reflected in the frequency visualization of Audio Clip1.

### **Spectrogram Analysis:**

Amplitude means the distance between the resting position and the maximum displacement of the wave.

A spectrogram can be considered as heat map of sound signal. It represents time, frequency and amplitude all in one graph. In the spectrogram view, X-axis represents time, Y-axis represents frequency, and brightness represents amplitude.

In Spectrogram analysis, High Amplitudes means brighter color and low amplitudes means colors are less bright. For example, very high amplitudes will be displayed with colors close to white, and very low amplitudes (silent parts of the sound) will be displayed with colors close to black.

In the below diagrams, we have displayed the spectrogram analysis of the 2 audio clips:

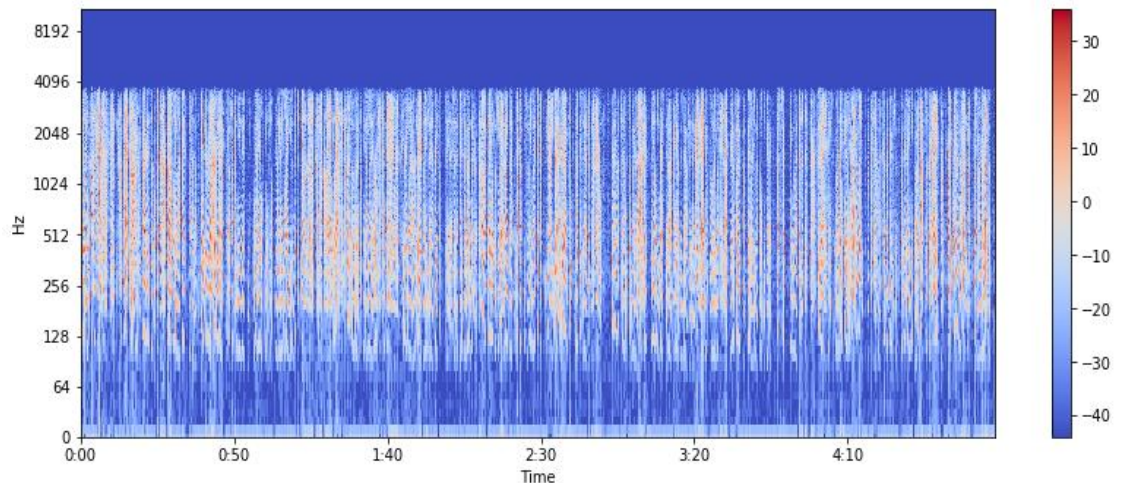


Fig 16: Spectrogram Analysis of Audio Clip1

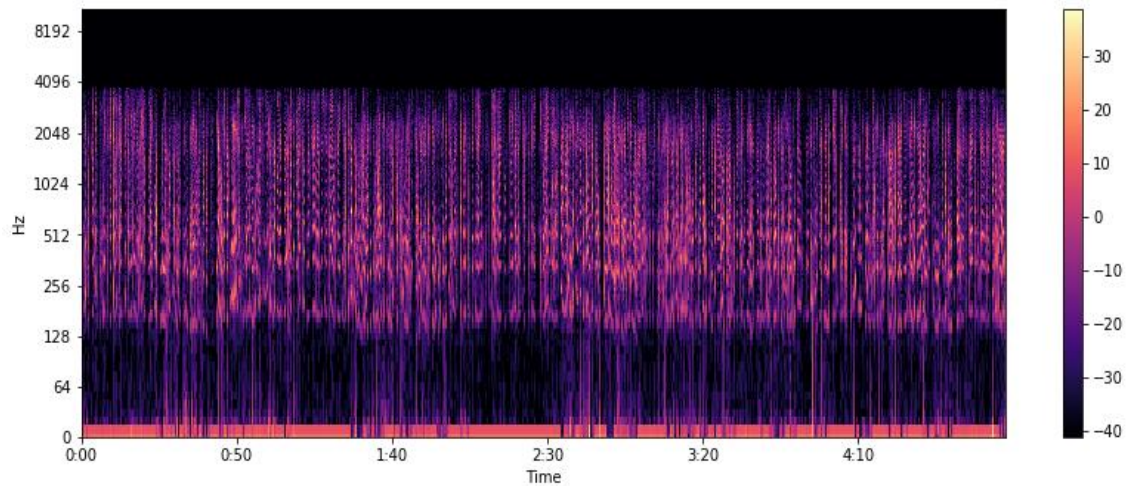


Fig 17: Spectrogram Analysis of Audio Clip2

High amplitudes means high loudness and low amplitudes means low loudness (also there is more silence).

In the above Spectrogram analysis, it can be observed that for Audio Clip1, as the color is white, so it means the conversation had very high amplitudes (means more loudness/volume). For Audio Clip2, as the color is black, so it means the conversation had low Amplitudes (means less loudness/volume).

So, it can be concluded that in Audio Clip1, which has high degree of anger and fearness, also has high pitch and high amplitudes (loudness/volume).

## **Chapter 12: Conclusions and Recommendations for future work**

- This project can be further developed to include Video Analysis as well.
- Using all the 3 features of communications (Lexical, Acoustic and Visual), we can understand person's overall behavior during any presentation or conversations.
- This can further be developed as a tool to train call center employees, to provide feedback to the interviewee, to provide feedback to presenter etc.

## Bibliography

- Awni Hannun. (2014). <https://arxiv.org/abs/1412.5567>.
- Bertero, D., & Fung, P. (2017). A FIRST LOOK INTO A CONVOLUTIONAL NEURAL NETWORK FOR SPEECH EMOTION DETECTION Dario Bertero , Pascale Fung Human Language Technology Center Department of Electronic and Computer Engineering The Hong Kong University of Science and Technology , Clear Water. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2017*, 5115–5119. <https://doi.org/10.1109/ICASSP.2017.7953131>
- Dahake, P. P., Shaw, K., & Malathi, P. (2017). Speaker dependent speech emotion recognition using MFCC and Support Vector Machine. *International Conference on Automatic Control and Dynamic Optimization Techniques, ICACDOT 2016*, 1080–1084. <https://doi.org/10.1109/ICACDOT.2016.7877753>
- Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech and Language Processing*. <https://doi.org/10.1109/TASL.2013.2244083>
- Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*. <https://doi.org/10.1016/j.neunet.2005.03.007>
- Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). *Deep Speech: Scaling up end-to-end speech recognition*. 1–12. <http://arxiv.org/abs/1412.5567>
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *Ieee Signal Processing Magazine*. <https://doi.org/10.1109/MSP.2012.2205597>
- Ittichaichareon, C. (2012). Speech recognition using MFCC. ... *Conference on Computer ....* <https://doi.org/10.13140/RG.2.1.2598.3208>
- Laurence Devilliers. (2005). *Challenges in real-life emotion annotation and*

*machine learning based detection.*

LDC. (n.d.). <https://catalog.ldc.upenn.edu/LDC97S42>. LDC.

Librosa. (2013). <https://librosa.org/doc/latest/index.html>. Librosa.

Marwala, T. (2018). Gaussian Mixture Models. In *Handbook of Machine Learning*. [https://doi.org/10.1142/9789813271234\\_0013](https://doi.org/10.1142/9789813271234_0013)

Mehrabian, A. (n.d.).

[https://en.wikipedia.org/wiki/Albert\\_Mehrabian#:~:text=It%20becomes%20more%20likely%20that,38%25%2D55%25%20Rule%22](https://en.wikipedia.org/wiki/Albert_Mehrabian#:~:text=It%20becomes%20more%20likely%20that,38%25%2D55%25%20Rule%22).

Mohammad, S. (2020). <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm#:~:text=The%20NRC%20Emotion%20Lexicon%20is,were%20manually%20done%20by%20crowdsourcing>.

Morais, R. (2019). <https://hacks.mozilla.org/2019/12/deepspeech-0-6-mozillas-speech-to-text-engine/>.

Mozilla. (2020).

<https://github.com/mozilla/DeepSpeech#:~:text=About,to%20high%20power%20GPU%20servers>.

Nefian, A. V., Liang, L., Pi, X., Liu, X., & Murphy, K. (2002). Dynamic Bayesian networks for audio-visual speech recognition. *Eurasip Journal on Applied Signal Processing*.

<https://doi.org/10.1155/S1110865702206083>

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2), 257–286. <https://doi.org/10.1109/5.18626>

sv-europe.com. (n.d.). <https://www.sv-europe.com/crisp-dm-methodology/>.

Tiwari, V. (2010). MFCC and its applications in speaker recognition.

*International Journal on Emerging Technologies*.

Wadhwa, M. (2020). <https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/>.

wiki. (n.d.). <https://en.wikipedia.org/wiki/Emotion>.

Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., & Vepa, J. (2018). Speech emotion recognition using spectrogram & phoneme embedding. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.

<https://doi.org/10.21437/Interspeech.2018-1811>



## Appendix

### Plagiarism Report

Turnitin Plagiarism report screenshot and report is attached below for reference:

# Emotion Detection with Speech Analytics

*by* Krishna Goswami

---

**Submission date:** 18-Sep-2020 09:12PM (UTC+0530)

**Submission ID:** 1390490233

**File name:** inal\_Project\_Report\_Emotion\_Detection\_With\_Speech\_Analytics.docx (1.52M)

**Word count:** 4344

**Character count:** 24889

## Emotion Detection with Speech Analytics

### ORIGINALITY REPORT

14%	%	%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

### PRIMARY SOURCES

1	<a href="http://www.cswe.org">www.cswe.org</a> Internet Source	2%
2	Submitted to Sogang University Student Paper	1%
3	<a href="http://www.talkbank.org">www.talkbank.org</a> Internet Source	1%
4	Submitted to The Arthur Lok Jack School of Business Student Paper	1%
5	<a href="http://xn----8sbbdpbxapcd0afe2b6a1grc.xn--p1ai">xn----8sbbdpbxapcd0afe2b6a1grc.xn--p1ai</a> Internet Source	1%
6	<a href="http://www.termpaperwarehouse.com">www.termpaperwarehouse.com</a> Internet Source	1%
7	<a href="http://hacks.mozilla.org">hacks.mozilla.org</a> Internet Source	1%
8	Submitted to Indian Institute of Technology Student Paper	1%
9	<a href="http://www.analyticsinsight.net">www.analyticsinsight.net</a>	

	Internet Source	1 %
10	<a href="#">pnsn.org</a> Internet Source	<1 %
11	<a href="#">mafiadoc.com</a> Internet Source	<1 %
12	Submitted to King's College Student Paper	<1 %
13	Submitted to The American College of Greece Libraries Student Paper	<1 %
14	<a href="#">curriculum.new-albany.k12.oh.us</a> Internet Source	<1 %
15	Tie Qiu, Zhao Zhao, Tong Zhang, Chen Chen, C.L.Philip Chen. "Underwater Internet of Things in Smart Ocean: System Architecture and Open Issues", IEEE Transactions on Industrial Informatics, 2019 Publication	<1 %
16	Submitted to British University in Egypt Student Paper	<1 %
17	<a href="#">towardsdatascience.com</a> Internet Source	<1 %
18	Saptarshi Boruah, Subhash Basishttha. "A study on HMM based speech recognition system",	<1 %

2013 IEEE International Conference on  
Computational Intelligence and Computing  
Research, 2013

Publication

19	<a href="http://www.inmybangalore.com">www.inmybangalore.com</a> Internet Source	<1 %
20	Submitted to Lovely Professional University Student Paper	<1 %
21	Florian Stoffel, Wolfgang Jentner, Michael Behrisch, Johannes Fuchs, Daniel Keim. "Interactive Ambiguity Resolution of Named Entities in Fictional Literature", Computer Graphics Forum, 2017 Publication	<1 %
22	<a href="http://www.iowaactuariesclub.org">www.iowaactuariesclub.org</a> Internet Source	<1 %
23	"Communications, Signal Processing, and Systems", Springer Science and Business Media LLC, 2020 Publication	<1 %
24	<a href="http://mediateto.blogspot.com">mediateto.blogspot.com</a> Internet Source	<1 %
25	<a href="http://link.springer.com">link.springer.com</a> Internet Source	<1 %

Exclude quotes On

Exclude matches < 10 words

Exclude bibliography On



Plagiarism\_Report\_  
Emotion Detection wii

## **Publications in a Journal/Conference Presented/White Paper**

Submitted the paper in 'International Journal of Scientific & Engineering Research' ([www.ijser.org](http://www.ijser.org)) and the paper is also already accepted, publication is pending. The certificate screenshot is given below for reference:

