REVA UNIVERSITY

Emotion Detection With Speech Analytics

Project Submitted by: Krishna Goswami

Under Guidance of : Dr. J. B. Simha

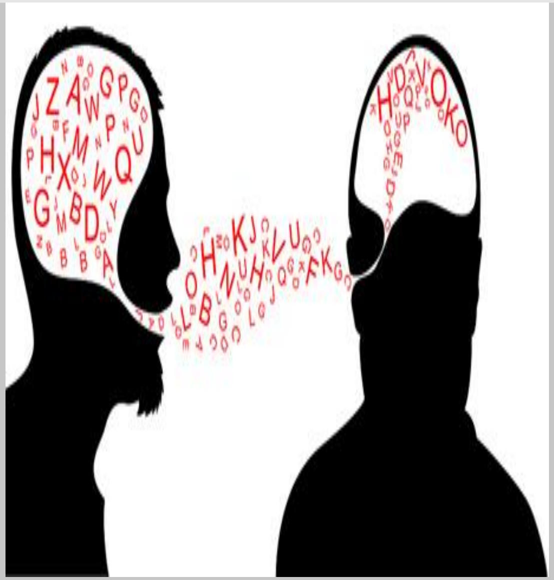
AGENDA

- Introduction
- Problem Statement
- Solution Overview
- Project Methodology
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Analysis and Results
- Future scope
- References
- GitHub Link

Introduction

- Speech is defined as the expression of thoughts and feelings by articulating sounds. Speech is the best, old and natural mode of communication among the humans.

- Speech analytics is the process of analysing the actual speaking voice of a person using different types of technology and gaining insights from the conversations.



- Emotion is feelings of human which he/she wants to express to other person through speech like happiness, fear, anger or frustration etc.
- Emotions are reflected from speech, hand and gestures of the body and through facial expressions and it is subjective in nature.
- Speech signal contains emotional state of the speaker, conveys the mood of the speaker by variations in pitch, loudness, intonation, pause and other such features.
- Emotion detection is the process of identifying human emotion.

Problem Statement

- 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.
- As emotions play a vital role in communication, the detection and analysis of emotion is of vital importance in today’s digital world of remote communication.
- According to the 7-38-55 Rule of Personal Communication, there are three elements involved about how humans express their feelings during communication:
  - Lexical features (the vocabulary used) account for 7%
  - Acoustic features (sound properties like pitch, tone 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.

Solution Overview

Limitations of Existing Speech-to-Text Solutions :

- Traditional speech systems use many heavily engineered processing stages, including specialized input features, acoustic models, and Hidden Markov Models (HMMs). To improve these pipelines, domain experts must invest a great deal of effort tuning their features and models.

- The introduction of deep learning algorithms has improved speech system performance, usually by improving acoustic models.

- While this improvement has been significant, deep learning still plays only a limited role in traditional speech pipelines and to improve performance in the scenario like recognizing speech in a noisy environment, one must laboriously engineer the rest of the system for robustness.

About Deep Speech:

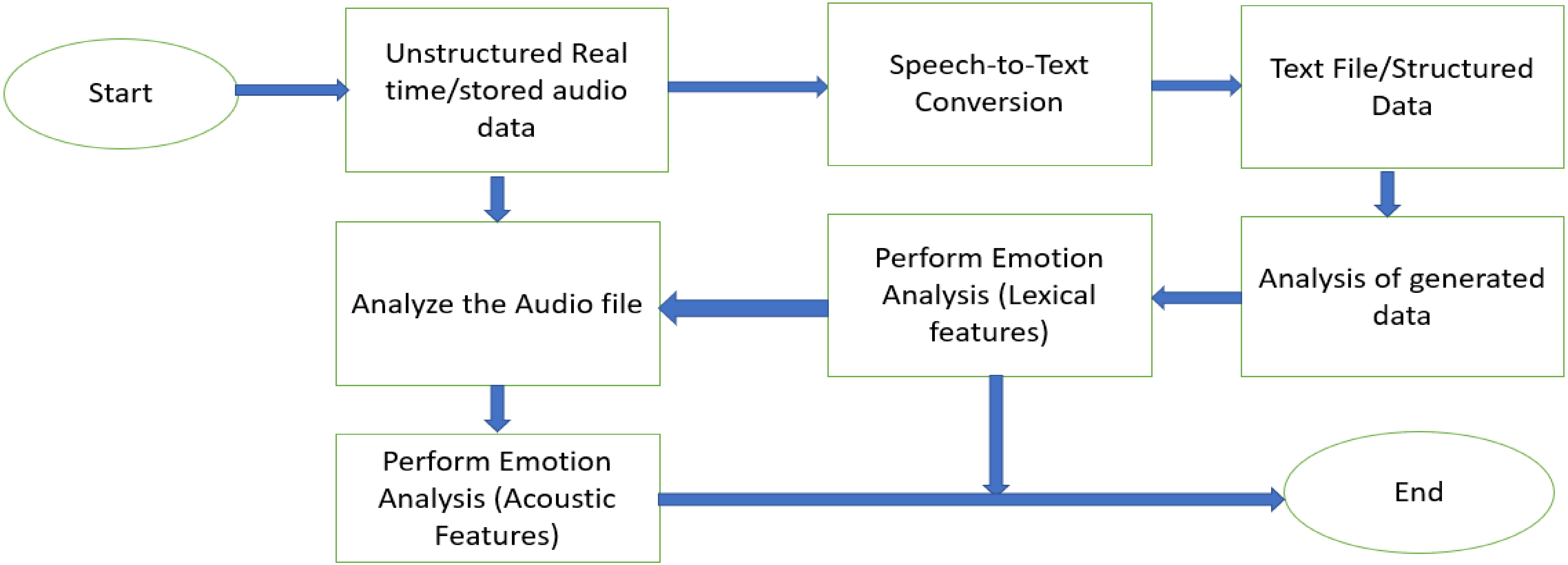
- Deep Speech is an open source Speech-To-Text engine, using a model trained by machine learning techniques based on Baidu's Deep Speech research paper. Project DeepSpeech uses Google's TensorFlow to make the implementation easier.
- In Deep Speech Recognition system , deep learning supersedes the processing stages.
- Combined with a language model, this approach achieves higher performance than traditional methods on hard speech recognition tasks(e.g. noisy background). These results are made possible by training a large recurrent neural network (RNN) using multiple GPUs and thousands of hours of data.
- This English DeepSpeech model was trained on 3816 hours of transcribed audio coming from Common Voice English, LibriSpeech, Fisher, Switchboard. The model also includes around 1700 hours of transcribed WAMU (NPR) radio shows.

Solution Approach:

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

- Mozilla’s Deep Speech model is used for Speech-to-Text conversion and audio input file is converted into text file.
- 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.

Emotion Detection Solution Flow Overview


Start

Unstructured Real time/stored audio data

Speech-to-Text Conversion

Text File/Structured Data

Analysis of generated data

Perform Emotion Analysis (Lexical features)

Analyze the Audio file

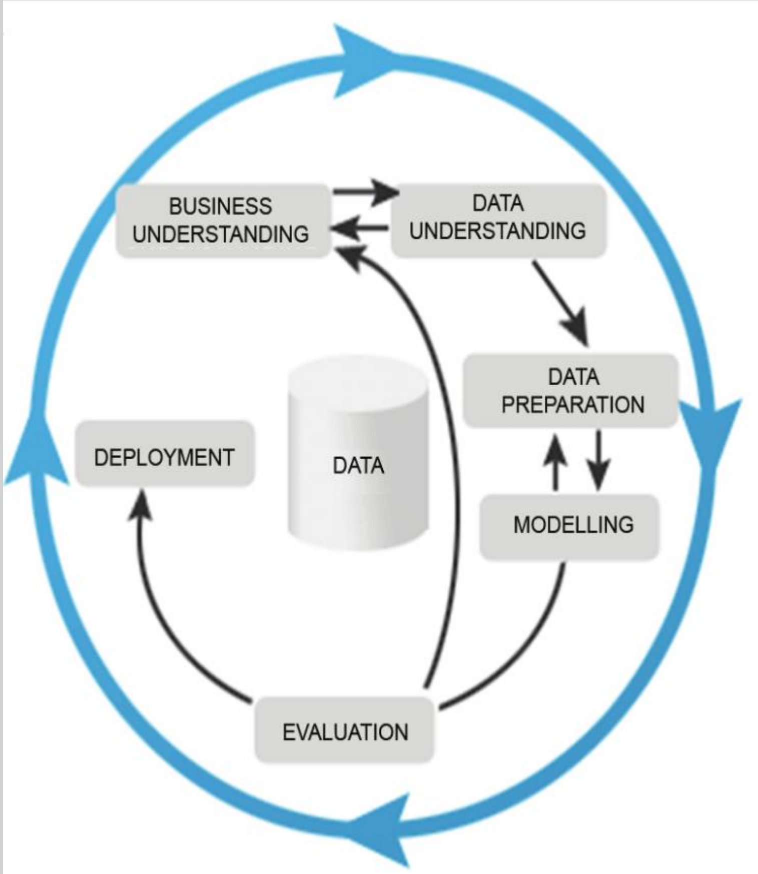
Perform Emotion Analysis (Acoustic Features)

End

Project Methodology

CRISP-DM

- **Understanding Business Requirements** — What is the specific ask from business ? Convert the same into a measurable and specific goal and formalize as a problem statement.
- **Data Understanding** — This phase revolves around data gathering, exploration and comprehension.
- **Data Preparation** - This phase includes final data set selection, Cleaning, Transforming data etc.
- **Data Modeling** - Modelling is the heart of data analytics. One can think of a model as a black box which takes relevant data as input and gives a ML model as an output.
- **Model Evaluation** - The essence of model evaluation is dependent on the work that has happened in previous 4 steps. If the results obtained from model evaluation are not satisfactory, we reiterate the whole process, otherwise we can move on to implementation of the model. Evaluation is necessary to ensure that your model is robust and effective.
- **Deployment** - After model evaluation the final model should go through a thorough testing and then it should be deployed.



Business Understanding

Importance of Speech and Emotion Analytics in an organization

- With the complexity of customer interactions, the different available customer channels and increased customer demand, there is now more pressure for businesses to deliver a better customer experience.

- 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

- Emotion Analytics can be used in contact center, to train the employees based on the feedback mechanism by analyzing the conversations between the agents and the customers.

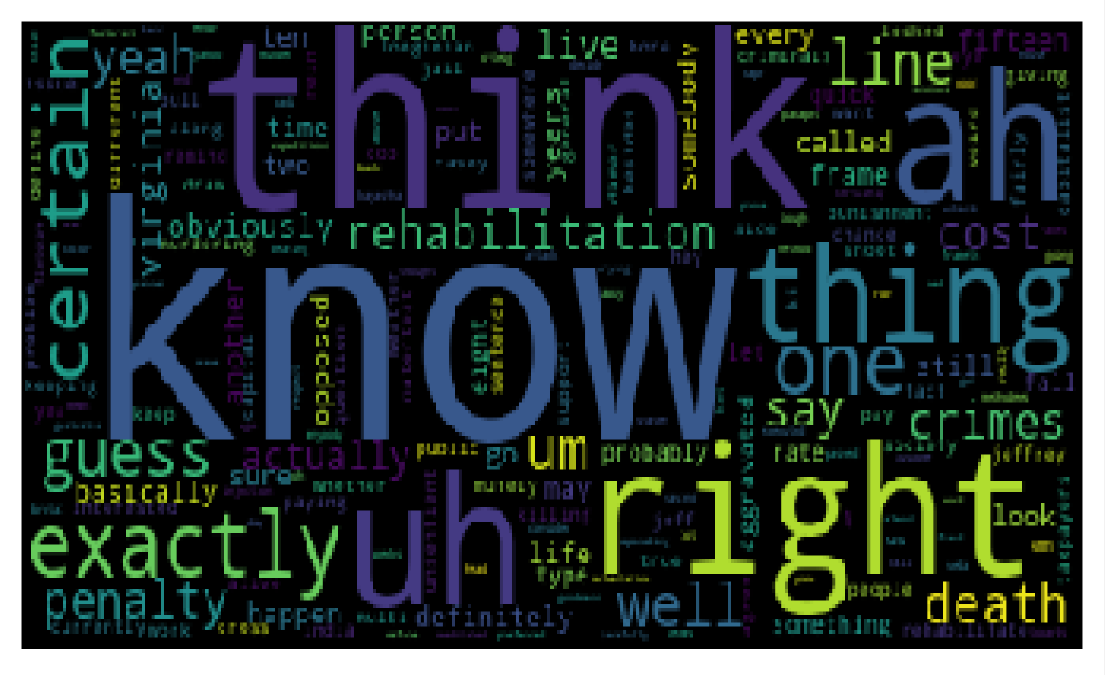
Data Understanding

Audio File Details

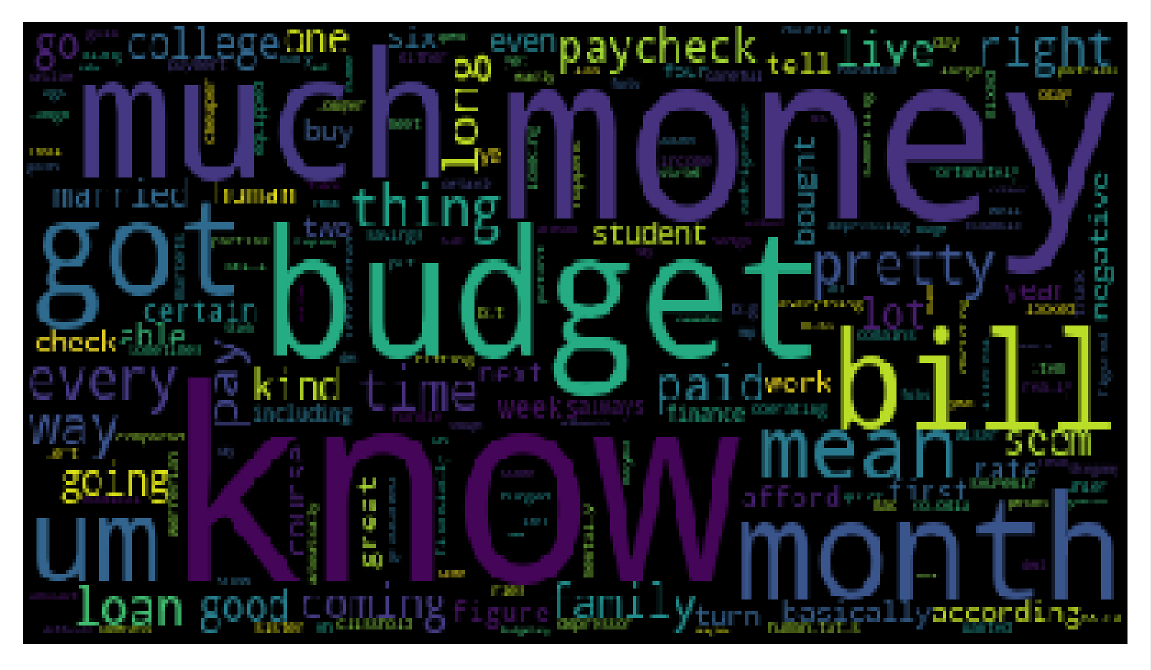
- 2 audio files are downloaded and used from the below CallHome English Corpus Link: https://ca.talkbank.org/access/CallHome/eng.html
- The CallHome English corpus of telephone speech was collected and transcribed by the Linguistic Data Consortium primarily in support of the project on Large Vocabulary Conversational Speech Recognition (LVCSR), sponsored by the U.S. Department of Defense. CallHome English corpus consists of 120 unscripted telephone conversations between native speakers of English.
- All speakers were aware that they were being recorded. They were given no guidelines concerning what they should talk about. Once a caller was recruited to participate, he/she was given a free choice of whom to call. Most participants called family members or close friends overseas.
- Each Audio file is of length approximate 5 mins.
  - Audio File1 contains 777 words.
  - Audio File2 contains 853 words

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

Word Cloud Analysis



Audio Clip1



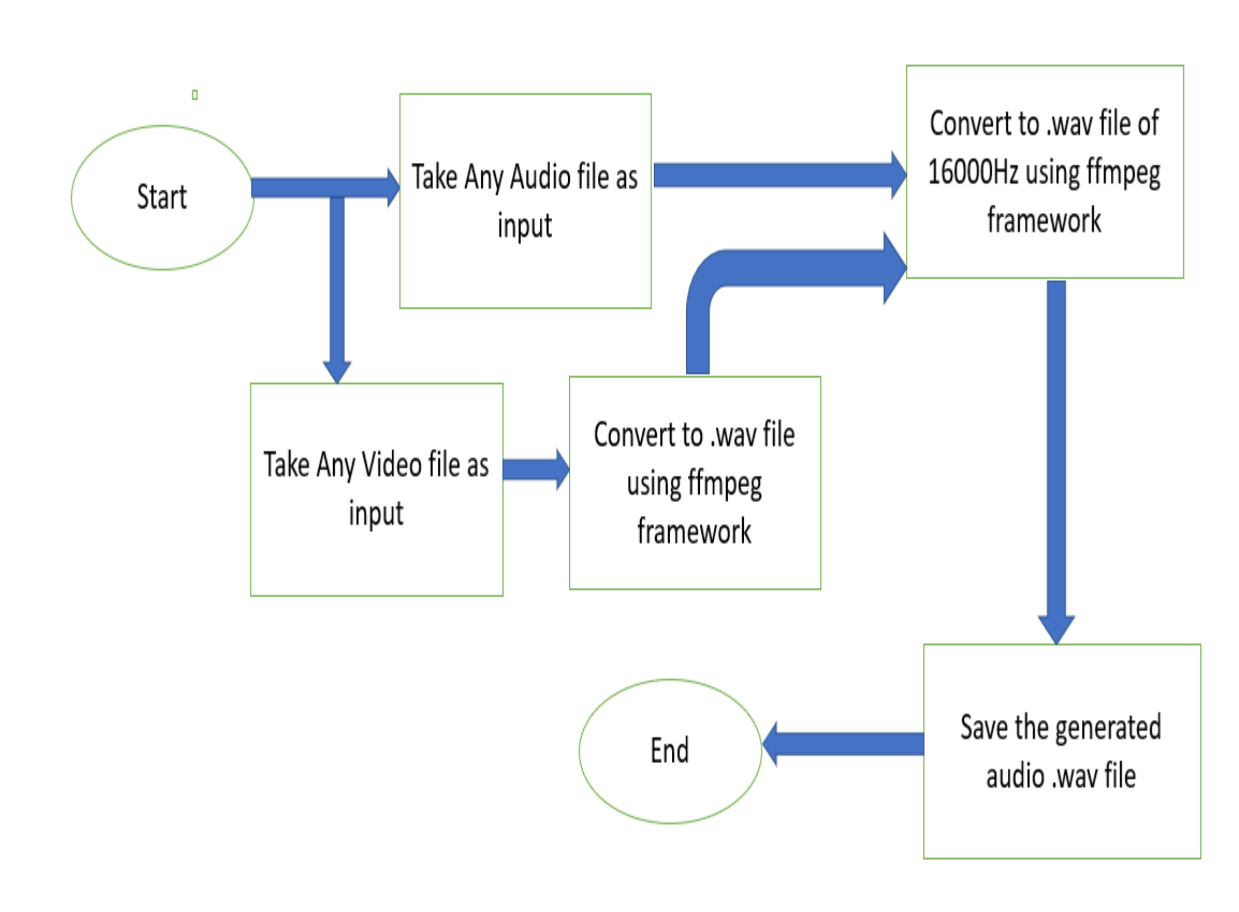
Audio Clip2

Word cloud analysis says that in audio clip1, the speakers discussed about crime, death, penalty etc. and in audio clip2, the speakers discussed about money, budget, loan etc.

Data Preparation

Data Preparation for Video/Audio Files

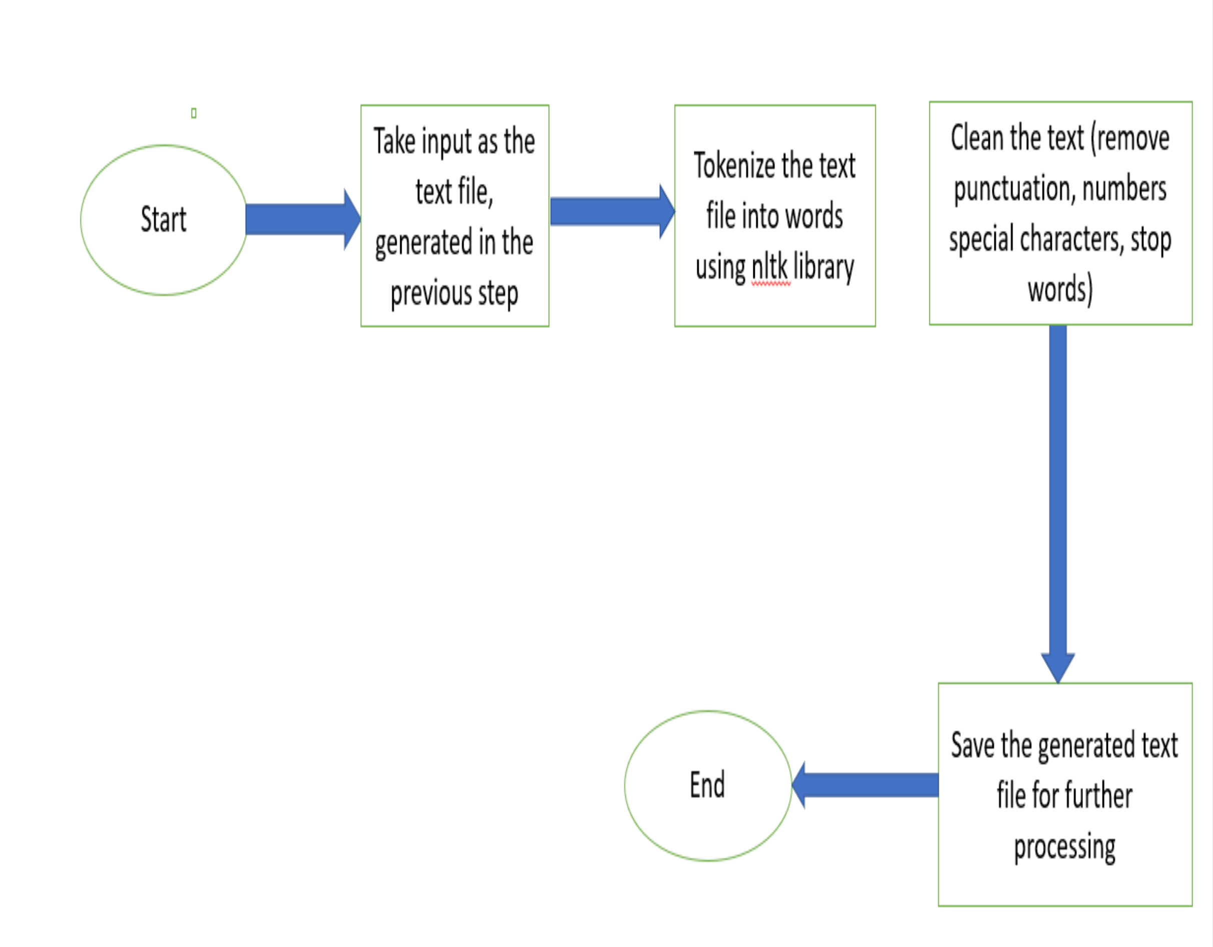
- The input for Deep Speech model should be in 16000 Hz .wav format. So, any video/audio file needs to be converted into 16000 Hz .wav format file.
- Following diagram shows the high-level architecture of data preparation:
  - ➢ Step 1: If the input file is mp4 video file, first we need to convert the file into .wav file using ffmpeg. ffmpeg is open-source multimedia framework, which can be used to decode, encode, transcode and play pretty much anything that humans and machines have created.
  - ➢ Step 2: Change the sampling rate of the audio file to 16000Hz using ffmpeg.



Start
Take Any Audio file as input
Convert to .wav file of 16000Hz using ffmpeg framework
Take Any Video file as input
Convert to .wav file using ffmpeg framework
Save the generated audio .wav file
End

Data Preparation for Text Files

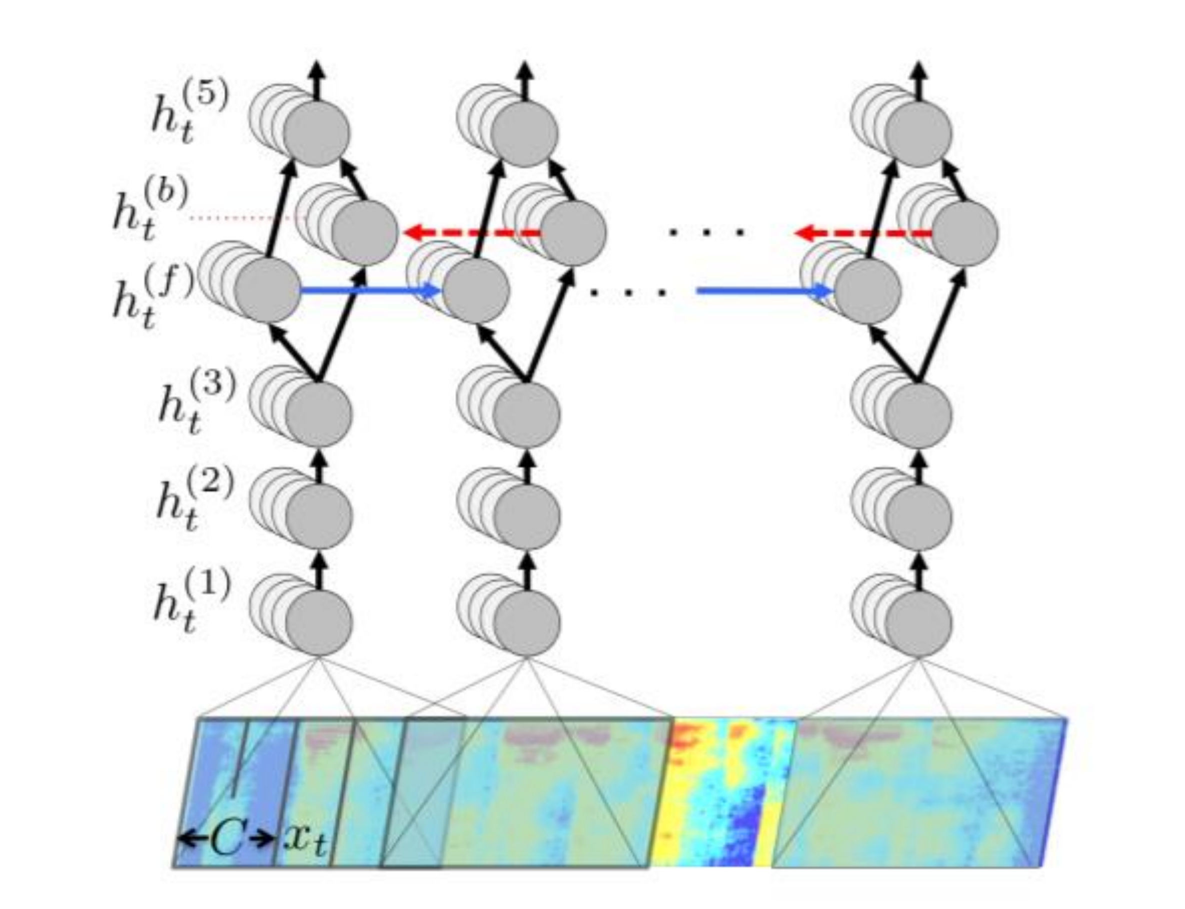
- The text file generated in the previous step as part of Speech-to-Text is used as input for data preparation step of text file.
- The Text file is tokenized using python nltk library
- The text file is cleaned using python nltk library (stop words removal, punctuation removal, special character removal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


Start
Take input as the text file, generated in the previous step
Tokenize the text file into words using nltk library
Clean the text (remove punctuation, numbers special characters, stop words)
Save the generated text file for further processing
End

Modeling

Deep Speech Architecture

- RNN model is composed of 5 layers of hidden units.
- The first three layers are not recurrent.
- The fourth layer is a bi-directional recurrent layer, This layer includes two sets of hidden units: a set with forward recurrence h (f) , and a set with backward recurrence h (b)
- The fifth (non-recurrent) layer takes both the forward and backward units as inputs.

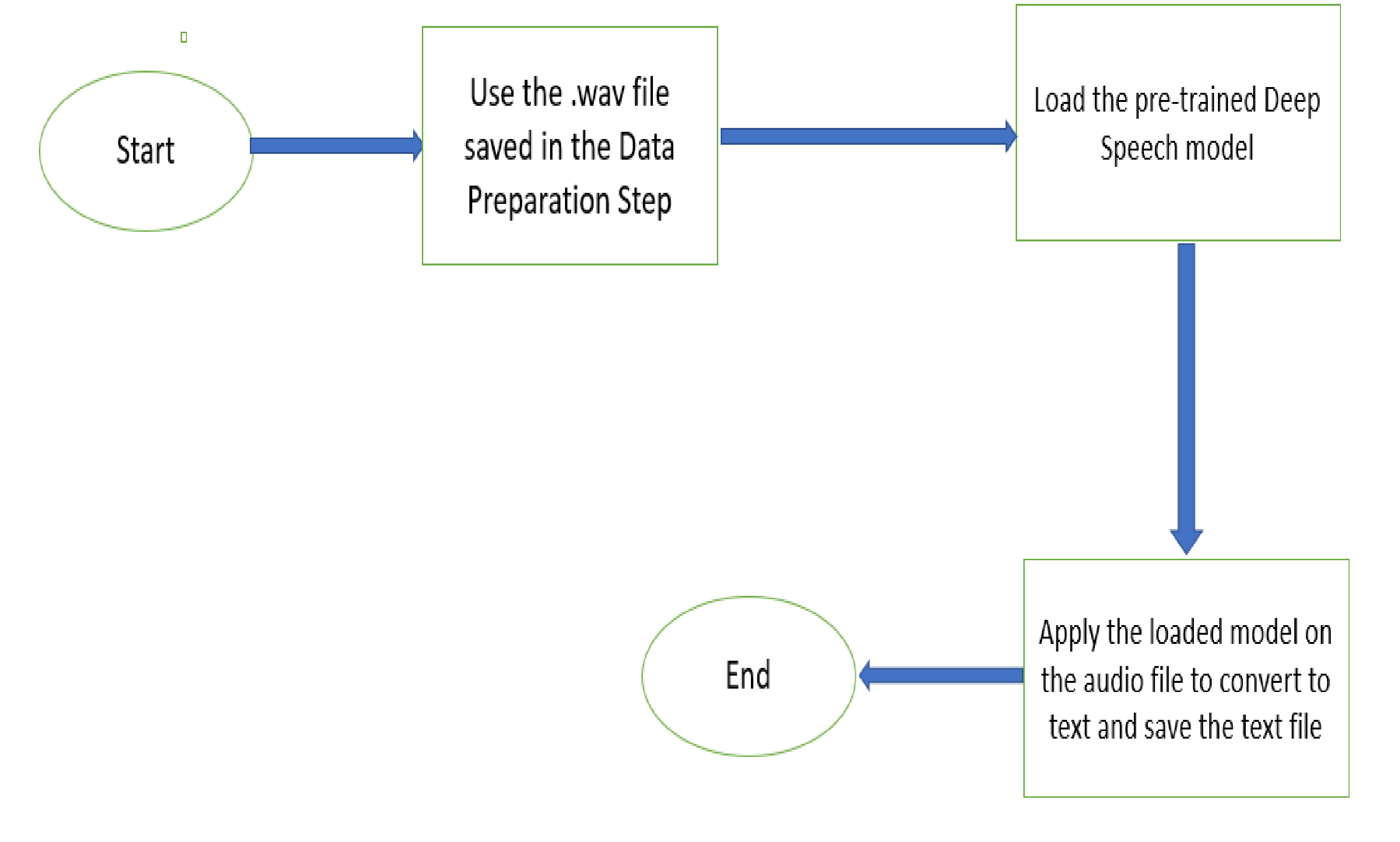


Audio to Text conversion (Using Deep Speech Algorithm)

Step 1: Use the .wav file saved in the Data Preparation Step

Step 2: Load the DeepSpeech pre-trained model.

Step 3: 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.

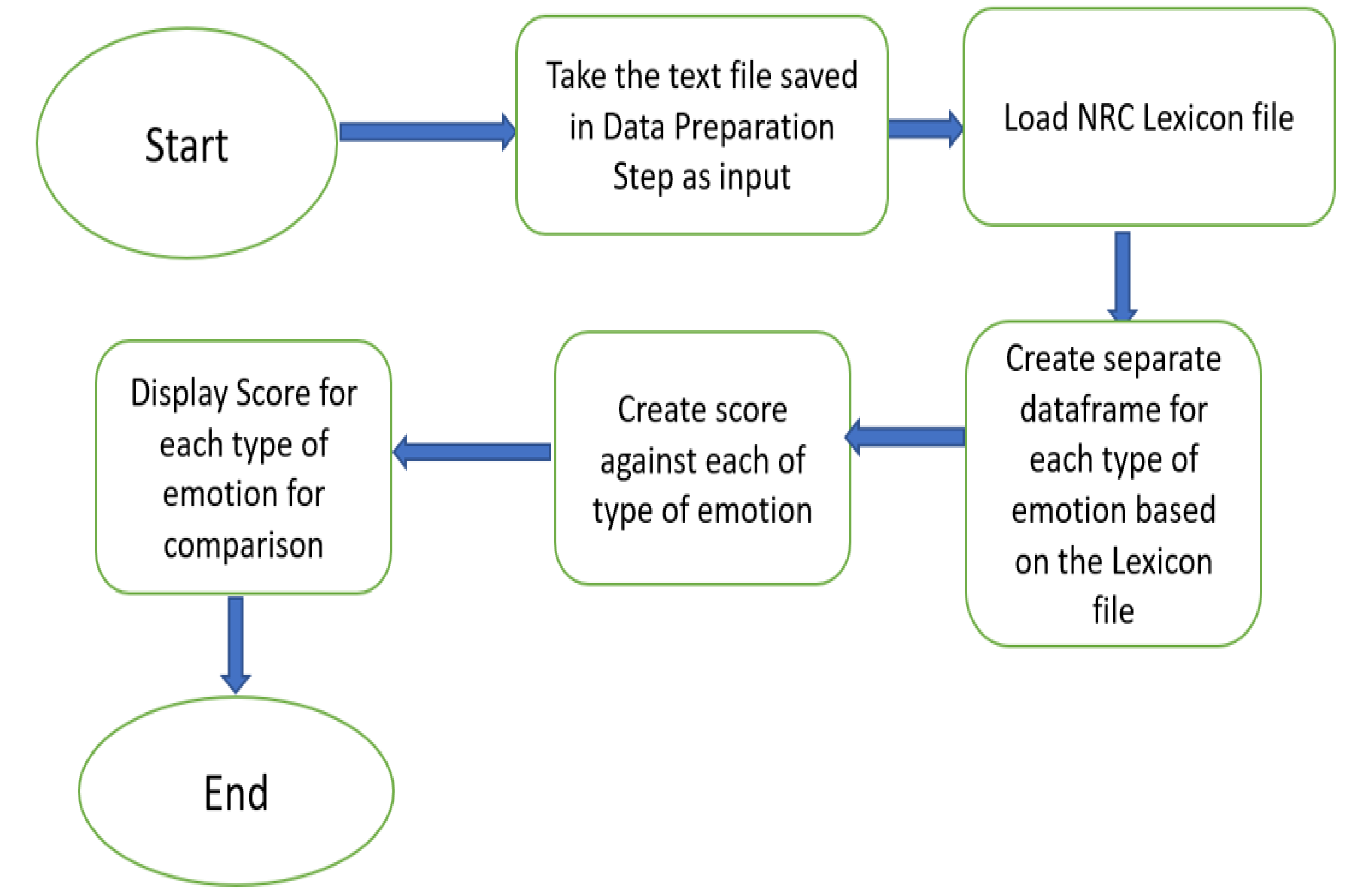

Start
Use the .wav file saved in the Data Preparation Step
Load the pre-trained Deep Speech model
Apply the loaded model on the audio file to convert to text and save the text file
End

REVA UNIVERSITY

Emotion Analysis(for Lexical Features) on Extracted Text File(Using NRC Lexicon)

- With the help of NRC Lexicon, scoring can be generated for 8 different types of emotions: 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.

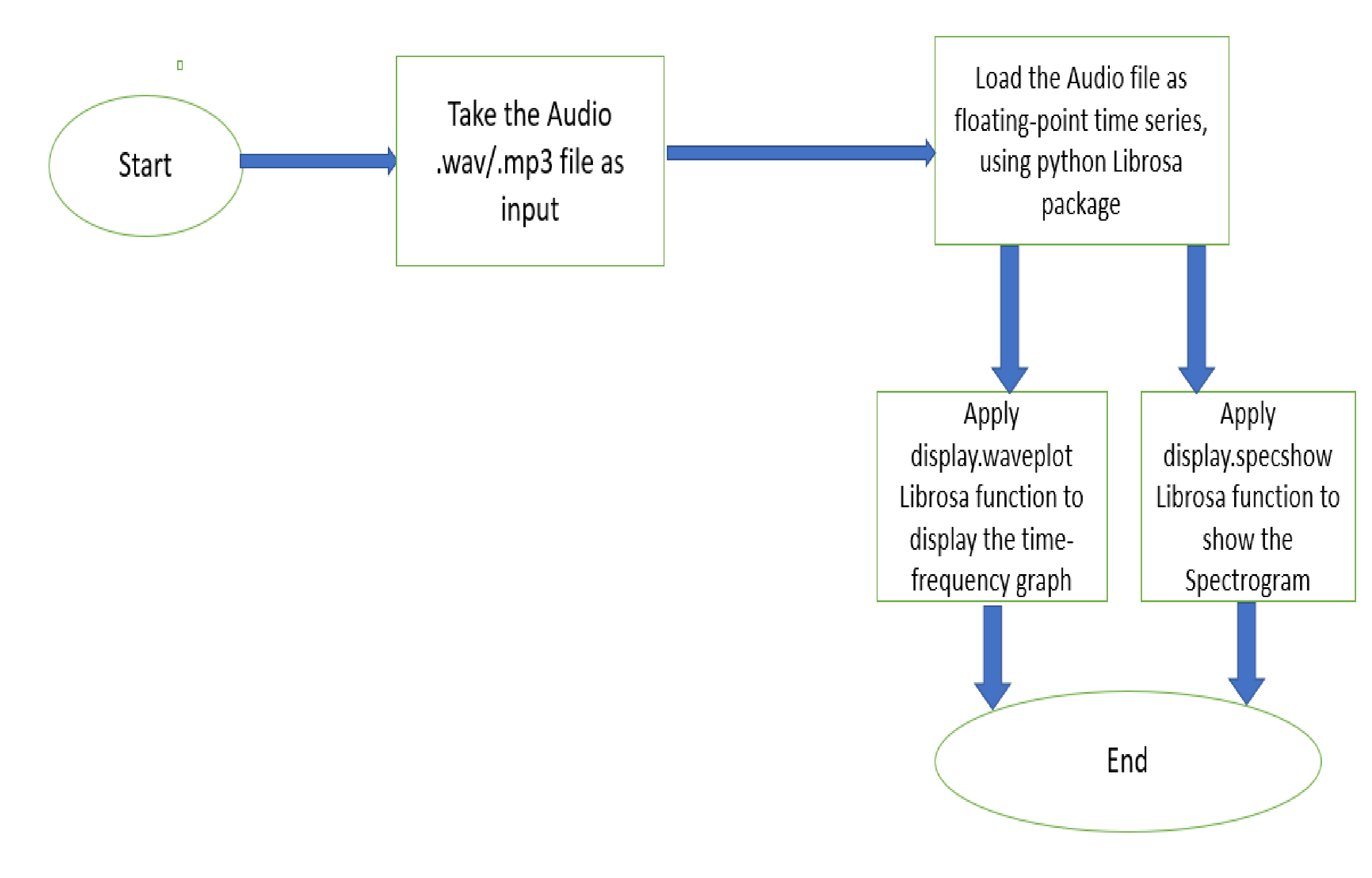

Start
Take the text file saved in Data Preparation Step as input
Load NRC Lexicon file
Create separate dataframe for each type of emotion based on the Lexicon file
Create score against each of type of emotion
Display Score for each type of emotion for comparison
End

Emotion Analysis(for Acoustic Features) using Librosa library

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

- 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.

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


Start
Take the Audio .wav/.mp3 file as input
Load the Audio file as floating-point time series, using python Librosa package
Apply display.waveplot Librosa function to display the time-frequency graph
Apply display.specshow Librosa function to show the Spectrogram
End

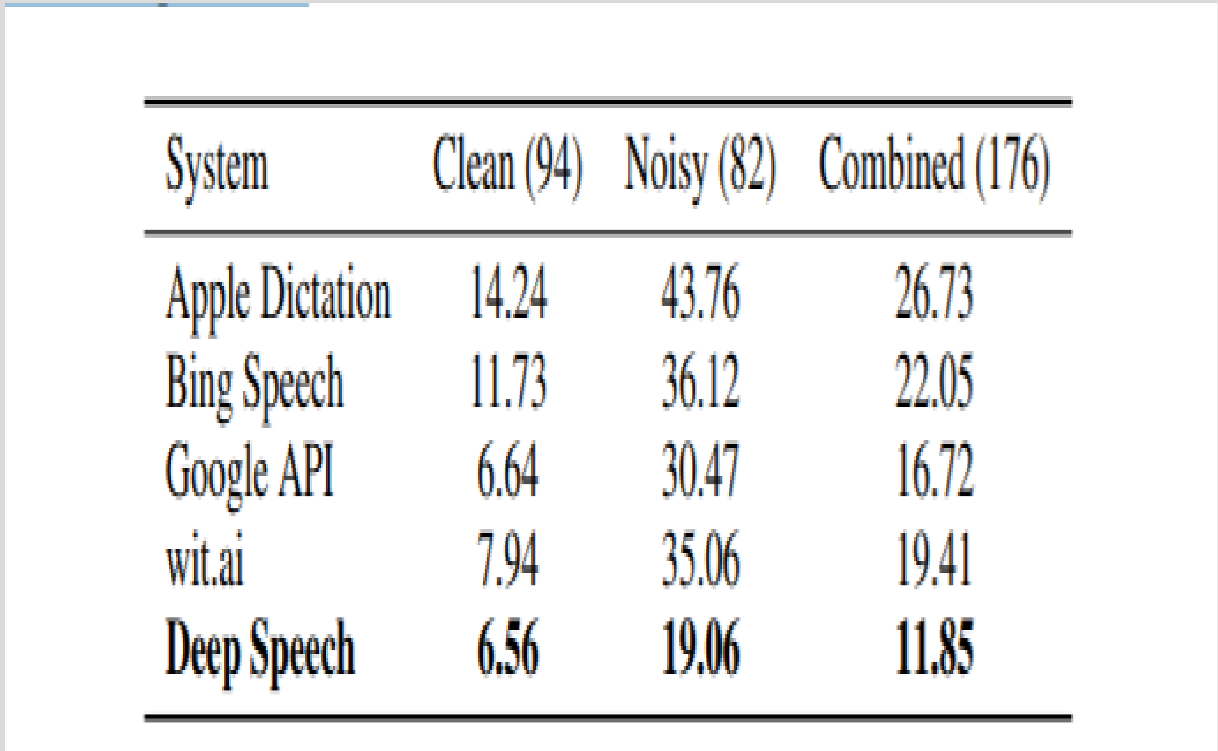
Model Evaluation

Deep Speech Model Performance

- Deep Speech system was to several commercial speech systems: (1) wit.ai, (2) Google Speech API, (3) Bing Speech and (4) Apple Dictation.

- To evaluate the efficacy of the noise synthesis techniques, two RNNs were trained, one on 5000 hours of raw data and the other trained on the same 5000 hours plus noise. On the 100 clean utterances both models perform about the same, 9.2% WER and 9.0% WER for the clean trained model and the noise trained model respectively.

- However, on the 100 noisy utterances the noisy model achieves 22.6% WER over the clean model's 28.7% WER, a 6.1% absolute and 21.3% relative improvement.

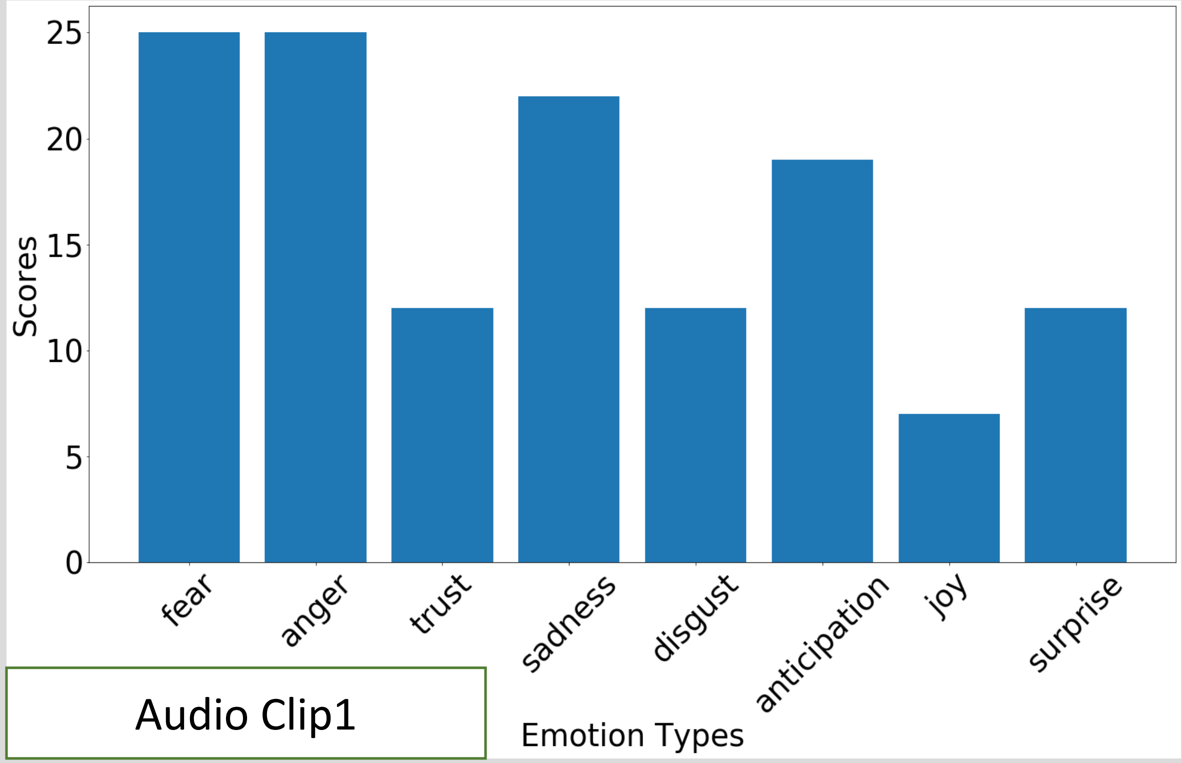


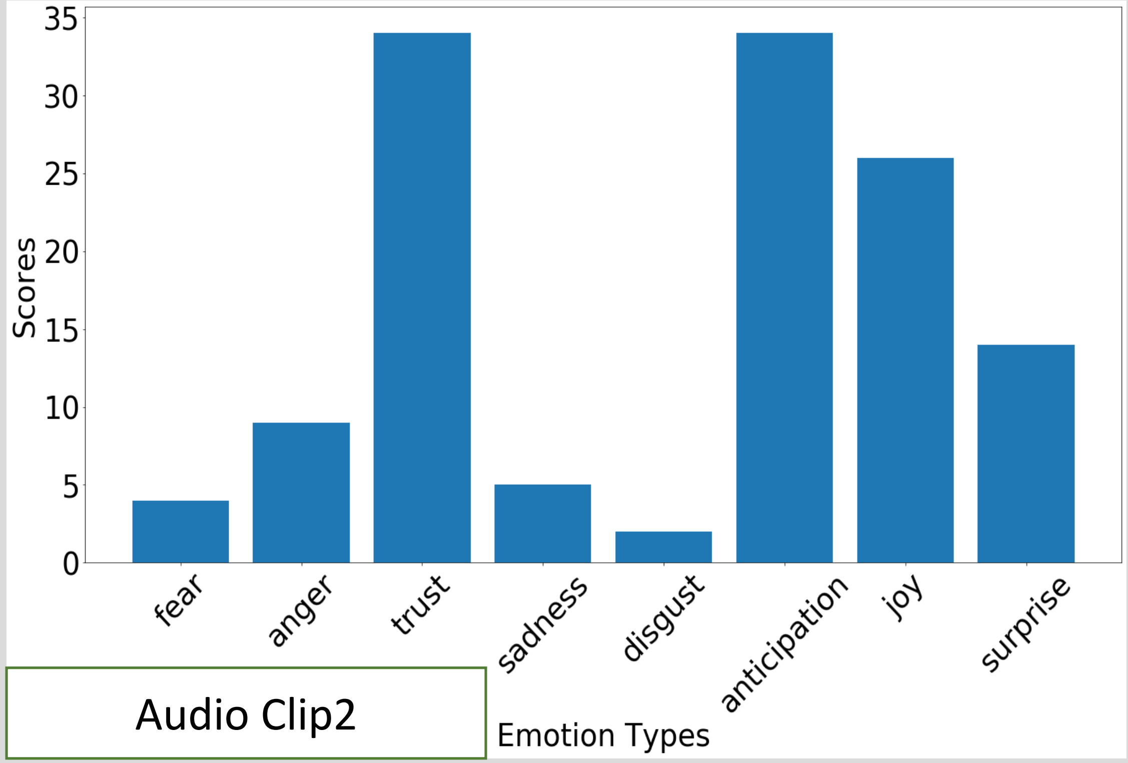
System	Clean (94)	Noisy (82)	Combined (176)
Apple Dictation	14.24	43.76	26.73
Bing Speech	11.73	36.12	22.05
Google API	6.64	30.47	16.72
wit.ai	7.94	35.06	19.41
Deep Speech	**6.56**	**19.06**	**11.85**

Note: This result is captured from Deep Speech Research Paper.

Analysis and Results

Emotion Detection and Analysis using NRC Lexicon

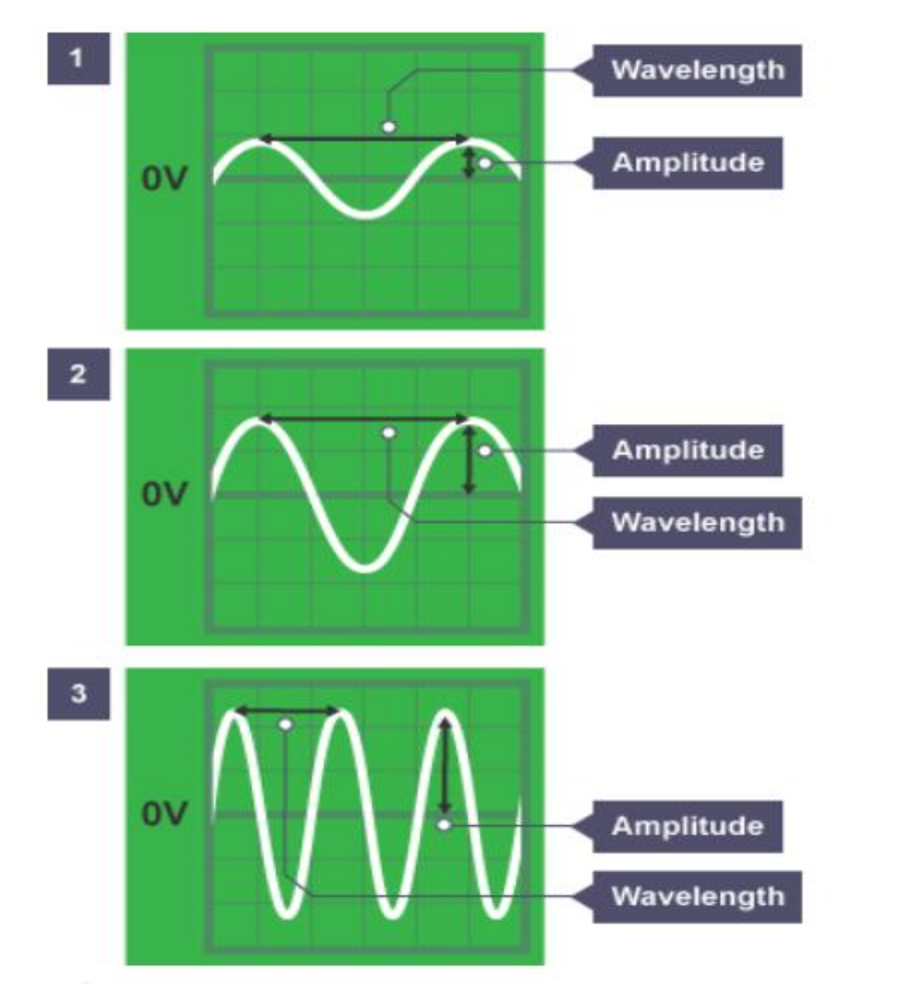




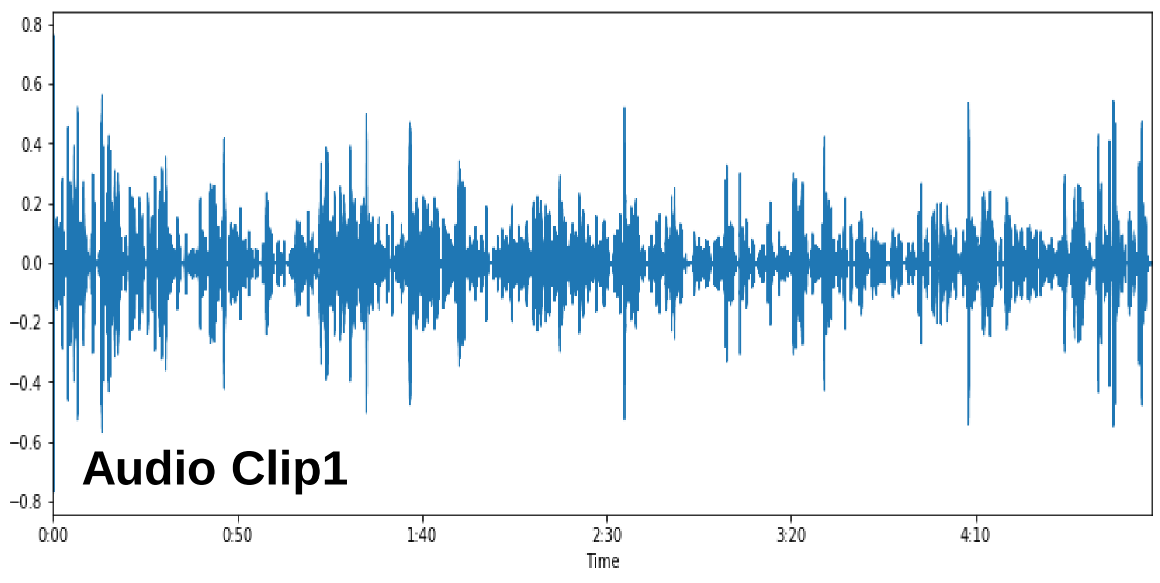
We can see that, for Audio Clip1, there is more fear, anger, sadness and very little joy and
for Audio Clip2, there is trust, anticipation and joy but very little fear and anger.

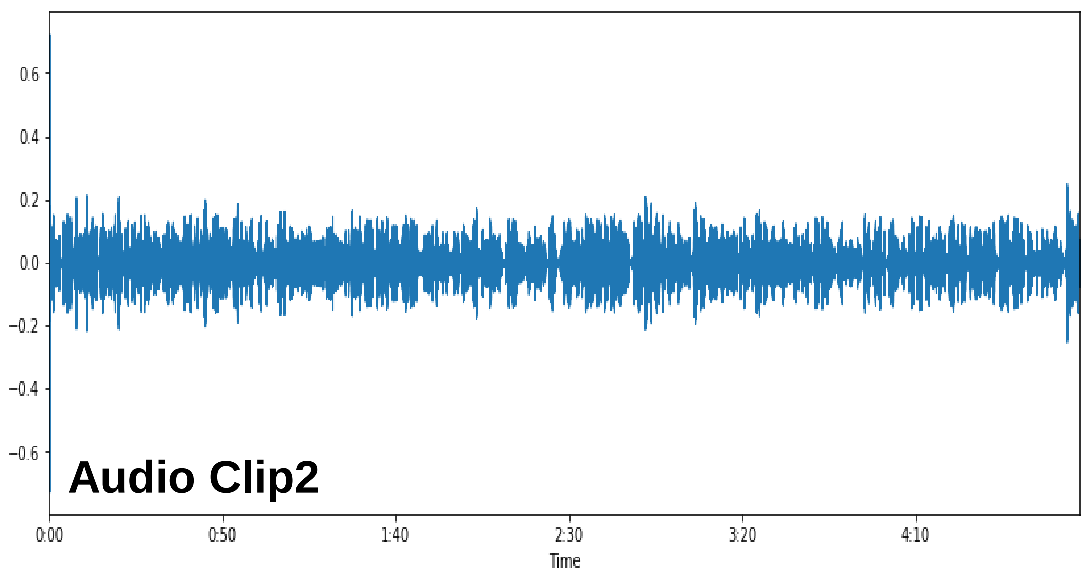
Amplitude and Pitch

- 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. Pitch can also refer to the degree of highness or lowness with which one speaks. A high pitch has high frequency, and a low pitch has low frequency.
- 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).



Frequency Visualization of Audio Files


Audio Clip1


Audio Clip2

Observations:

- There is high frequency (means high pitch) in the conversation for Audio Clip1. Emotions can affect the pitch of the voice, e.g, sudden emotions like anger, surprise, joy can make a person speak in a higher pitch than usual.
- 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.

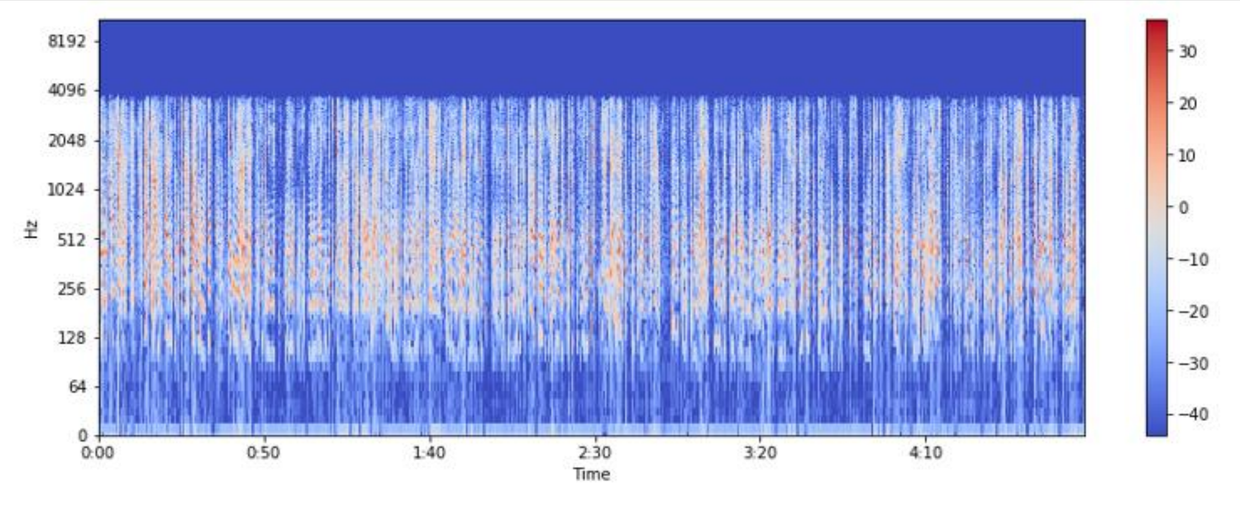
What is Spectrogram Analysis

- A spectrogram can be considered as heat map of sound signal. It is a visual way of representing the signal strength, or “loudness”, of a signal over time at various frequencies present in a waveform.

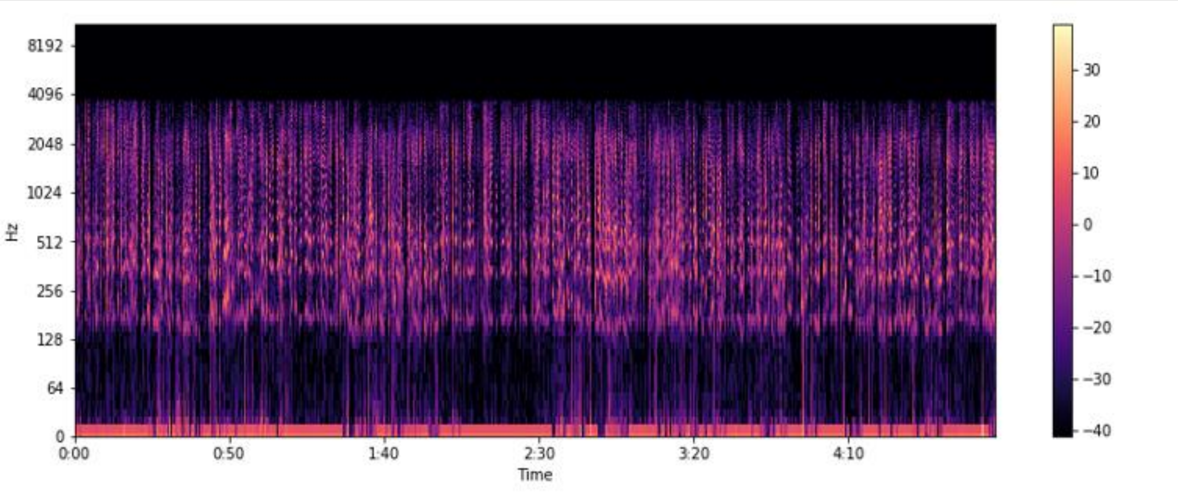
- It represents time, frequency and amplitude all in one graph. In the spectrogram view, vertical axis represents frequency in Hertz, horizontal axis represents time and brightness represents amplitude.

- 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.

Spectrogram Analysis of Audio Files



Audio Clip1



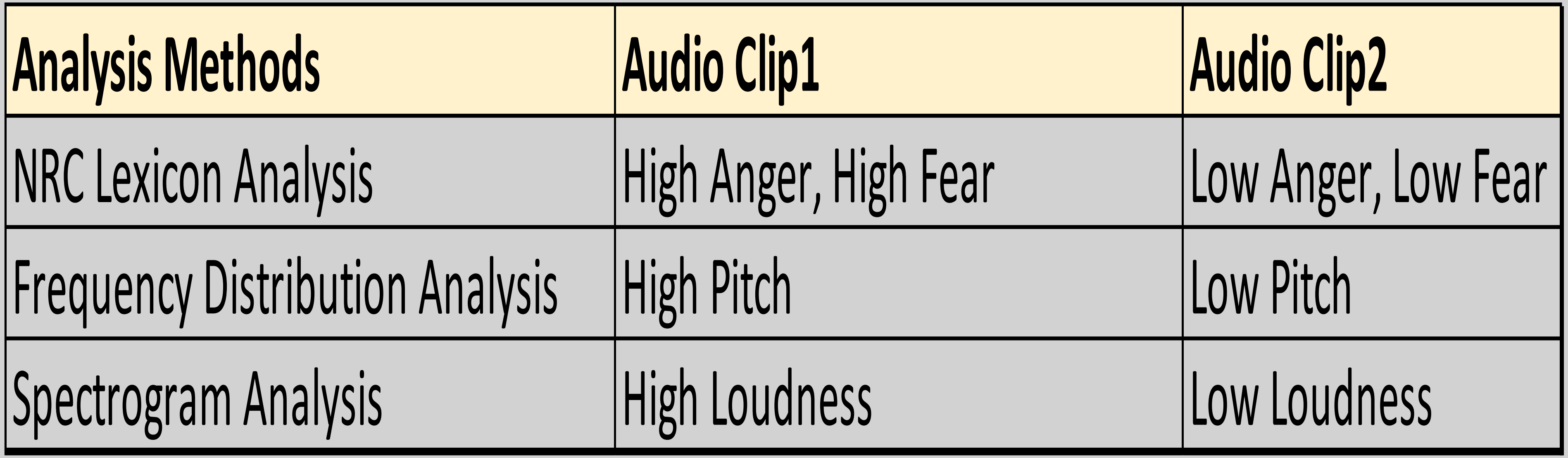
Audio Clip2

Observations:

- 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).

Emotion Analysis Summary

Below is the summary of Emotion Analysis of the two audio clips:



Analysis Methods	Audio Clip1	Audio Clip2
NRC Lexicon Analysis	High Anger, High Fear	Low Anger, Low Fear
Frequency Distribution Analysis	High Pitch	Low Pitch
Spectrogram Analysis	High Loudness	Low Loudness

Future Scope

- 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.

References

References

- Boruah, S., & Basishtha, S. (2013). A study on HMM based speech recognition system. *2013 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2013*. https://doi.org/10.1109/ICCIC.2013.6724147
- 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
- Dalya Gartzman. (2019). *https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0*.
- 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

References

- 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
- Gupta, S. C. (2020). *https://www.satishchandragupta.com/tech/python-speech-to-text-asr-transcriber-with-mozilla-deepspeech.html*.
- 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

GitHub Link

- The following GitHub link contains the code base used for the project:

https://github.com/kgopal1982/AnalyticsProjects/tree/master/VideoAnalysis

THANK YOU

Q & A