SPEECH EMOTION RECOGNITION USING AUDIO FEATURE EXTRACTIONS

A report submitted in partial fulfillment of the requirements for the Award of Degree of

Bachelor of Technology

In

Electrical Engineering

Submitted By:

SUBHANKIT PRUSTI (1801106553)
SANSKRUTI MOHAPATRA (1801106442)
KISAN BEHERA (1921106054)

Under the Guidance of:
DR. ULLASH KUMAR ROUT
Associate Professor, EE

DEPARTMENT OF ELECTRICAL ENGINEERING

ODISHA UNIVERSITY OF TECHNOLOGY AND RESEARCH,

BHUBANESWAR-751029,2022

DEPARTMENT OF ELECTRICAL ENGINEERING

ODISHA UNIVERSITY OF TECHNOLOGY AND RESEARCH, BHUBANESWAR

CERTIFICATE

This is to certify that the project report entitled **“Speech Emotion Recognition Using Audio Feature Extraction”** submitted by *Mr. Subhankit Prusti (Registration No.1801106553), Ms. Sanskruti Mohapatra (Registration No. 1801106442), Mr. Kisan Behera (Registration No. 1921106054)* of the Department of Electrical Engineering, fulfills the requirement of the regulation relating to the nature and standard of the work for the award of the degree of Bachelor of Technology, in Electrical Engineering for academic year 2021-22.

Prof. (Dr) Ajit Kumar Barisal	Dr. Ullash Kumar Rout
	Associate Professor, EE
HOD, Electrical Engineering	**Supervisor**

DEPARTMENT OF ELECTRICAL ENGINEERING

ODISHA UNIVERSITY OF TECHNOLOGY AND RESEARCH, BHUBANESWAR

DECLARATION

We do hereby declare that, the major project entitled, "**Speech Emotion Recognition Using Audio Feature Extraction**" is a bona-fide work of study carried out by us under the guidance of Dr. Ullash Kumar Rout, Associate Professor, Department of Electrical Engineering, College of Engineering and Technology, Bhubaneswar. It has been prepared for the fulfilment of the requirements of the degree of 'Bachelor of Technology in Electrical Engineering'. The work has not been submitted for any other purpose.

PLACE: Bhubaneswar

DATE:

Subhankit Prusti (1801106553)

Sanskruti Mohapatra (1801106442)

Kisan Behera (1921106054)

ACKNOWLEDGEMENT

We are really indebted and want to specific my feeling to my supervisor Dr. Ullash Kumar Rout for his inspiration, glorious steering, continued encouragement, and unwavering confidence and support throughout each stage of this endeavor while not that, it might not be doable on behalf of me to finish this endeavor with success. We additionally give thanks for his perceptive comments and suggestions that regularly helped to boost my understanding and knowledge.

We are additionally significantly duty-bound to the Head of the Department of Electrical Engineering Dr. Ajit Kumar Barisal, OUTR Bhubaneswar for providing all doable facilities towards my research work. we'd wish to due to all different college members within the department.

We wholeheartedly extend our gratitude to my parents and friends for their constant encouragement, love, desires, and support. Above all, we give thanks to the almighty who bestowed blessings upon us.

Subhankit Prusti (1801106553)

Sanskruti Mohapatra (1801106442)

Kisan Behera (1921106054)

ABSTRACT

In today's environment, detecting emotions is one of the most significant marketing methods. Emotion contains a great deal of information about a person's mental state. This has spawned a brand-new field of study known as automatic emotion recognition. There are various uses for detecting people's emotions, such as in robot interfaces, audio surveillance, contact centres, and so on. As a result, we decided to work on a project where we could detect a person's emotions simply by listening to their voice, which would allow us to manage a variety of AI-related applications. The goal of a voice emotion recognition system is to classify a speaker's utterances into emotional states automatically. This project is a classification issue in which various emotions are classified, features are extracted, models are developed, and their accuracy is compared. The human interface's purpose is to accurately recognise the user's emotional state. The most essential difficulty in the speech emotion recognition study is the effective parallel usage of suitable speech feature extraction and an adequate classification engine.

LIST OF FIGURES

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

As human beings speaking is one of the most natural means for expressing themselves. We are so dependent on them that we recognize their importance by resorting to other forms of communication such as emails or text messages where we often use emoticons to express the emotions associated with the messages.

Since emotions play a vital role in communication, their detection and analysis are vitally important in today's digital world. Finding emotions is a difficult task because emotions are subjective. There is no consensus on how these can be measured or categorized.
There are various techniques of communication, but the speech signal is one of the quickest and most natural. As a result, voice can also be a quick and efficient way of human-machine contact . Humans have the innate ability to employ all of their senses to get maximum awareness of the message they are receiving. People are able to perceive their communication partner's emotional condition using all of their senses. Humans are naturally good at detecting emotions, but machines have a hard time doing so. As a result, the goal of an emotion detection system is to improve human-machine communication by utilising emotion-related knowledge.

We define a SER (Speech Emotion Recognition)system as a set of methodologies that treat and classify voice signals to detect the emotions they contain. Such a system can be used in a wide variety of application areas such as interactive vocal assistant or caller-agent conversation analysis. In this study we attempt to detect underlying emotions in recorded speech by analysing the acoustic features of the audio data of recordings.

A speech emotion recognition system's design must consider three key factors. The first is the selection of appropriate speech representation features. The design of an acceptable categorization method is the second issue, and the right construction of an emotional speech database for evaluating system performance is the third issue. Communication will be improved as a result of these issues.

There are three categories of features in a speech: lexical, visual, and acoustic. The problem of recognizing the emotions of speech can be solved by analysing one or more of these characteristics. Choosing to follow the lexical features would require a transcript of the speech, which would require an additional step in extracting the text from speech if one wants to predict the emotions of the audio in real time.

Similarly, analyzing visual characteristics would require excessive video of conversations which may not be feasible in all cases while the analysis of acoustic features can be done in real time during the conversation because we only need audio for our task. That is why we chose to analyse the acoustic features of this work

1.1 : NEED FOR SPEECH RECOGNITION SYSTEM

High-performance personal computers have recently gained popularity as a result of technological advancements in the information society. As a result, computer-human interactions have been actively shifting toward a bidirectional interface. As a result, a greater understanding of human emotions is required. It has the potential to improve human-machine interaction systems. Emotion recognition has been a popular study area in signal processing for the reasons stated above. The purpose of the human interface is to accurately assess the user's emotional state and provide personalized media based on that condition.

Speech signals are among the most natural forms of human communication, and they have the advantage of being easily measured in real time. They include the speaker's emotions as well as implicit paralinguistic information and linguistic content. Identifying and extracting different emotion-related speech components is a difficult issue when creating a speech emotion identification system. It's crucial to integrate appropriate audio characteristics in speech emotion identification since feature selection impacts categorization performance.

1.2 : PROBLEM STATEMENT

Mouse, keyboard, and other traditional methods of interface with robots or computers are used. With the evolution of technology, some techniques such as hand or facial gesture detection have been added to the Human Computer Interface. However, emotion recognition in computers is still a new field of study, and they have yet to demonstrate adequate emotional reactions.

As we know that Speech is the most significant mode of communication among human beings and hence a potential method for human-computer interaction (HCI) by using a microphone sensor. But Quantifiable emotion recognition directly using these sensors from speech signals is still a topic of research. Extraction of certain features for particular emotions is an emerging area of research in HCI in order to classify emotions and improve the HCI in a natural way.

1.3 : OBJECTIVE

The objective here is detection and classification of emotions from Speech taking into account the features of speech in frequency domain using MFCC extraction method and building SER system using ML algorithm models.

The main focus is on the feature extraction of speech signals in frequency domain using MFCC extraction technique which gives better frequency resolution in the frequency region where the difference between the frequency of the signals is barely noticeable.

Earlier, the work was more focused on heuristics-based approaches for solving SER, but now the approaches are geared towards more complex amalgamation of the features using spectrograms, MFCCs, and other acoustic features.

1.4 : LITERATURE STUDY

As per various literature surveys it is found that for implementing this project
three basic steps are required to be performed.

i. Preprocessing
ii.Feature extraction
iii.Emotion classification

PREPROCESSING:

Preprocessing of speech signals is considered a crucial step in the development of a robust and efficient speech or speaker recognition systemThe first step in working with audio files is to **turn the audio wave into numbers** so that you can feed this into our machine learning algorithm.

FEATURE EXTRACTION:

We retrieved two main elements from the audio data that were employed in this study: MFCC (Mel Frequency Cepstral Coefficients) and Mel Spectrogram. In order to extract them, the Python implementation of the Librosa package was employed.

Choice of features:

1. In research papers and open source projects, MFCC was by far the most explored and used feature.
2. On a "Mel" scale, a Mel spectrogram depicts amplitude on a frequency versus time graph. Because the project is about emotion identification, which is a completely subjective topic, we decided to plot the amplitude on the Mel scale, which converts the recorded frequency to "perceived frequency."

EMOTION CLASSIFICATION:

- These are broadly the 7 classes of emotions that are taken into account for our classification i.e- happy,fear,neutral,disgust,anger,sad and calm.
- The preprocessed data is used for our ML classification purpose.
- k cross validation test to be performed on diff ML classification algo and the model giving the best classification results among others. to be chosen
- To calculate the accuracy of our model,**accuracy_score()** function to be imported from sklearn. We will also use the **confusion matrix and classification report** for a better understanding of our model predictions

1.5 : PLANNING

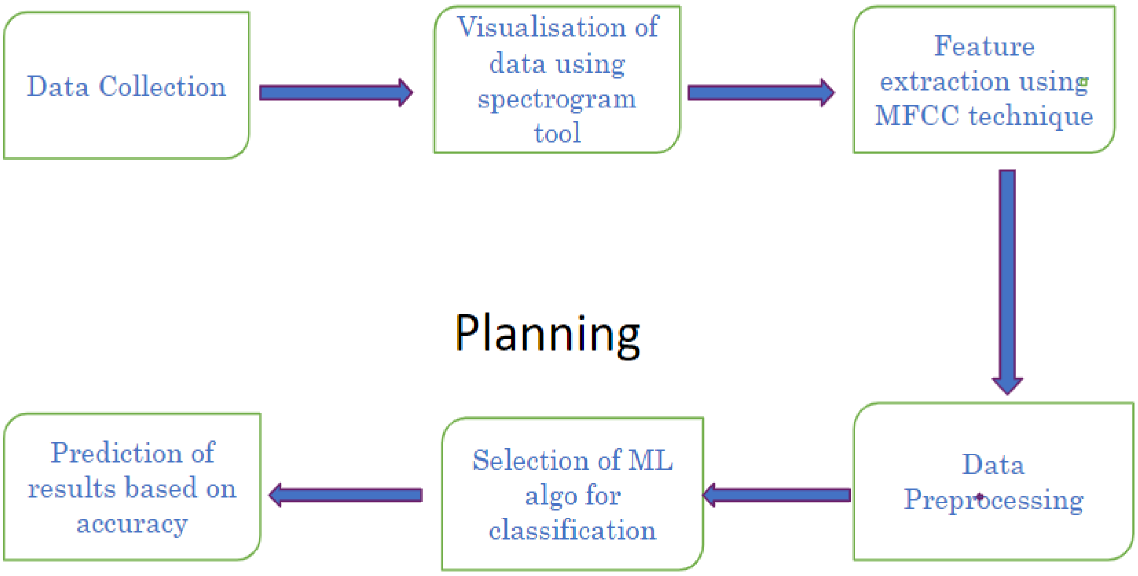

Data Collection
Visualisation of data using spectrogram tool
Feature extraction using MFCC technique
Planning
Data Preprocessing
Selection of ML algo for classification
Prediction of results based on accuracy

Fig. 1.1: Layout of the project plan

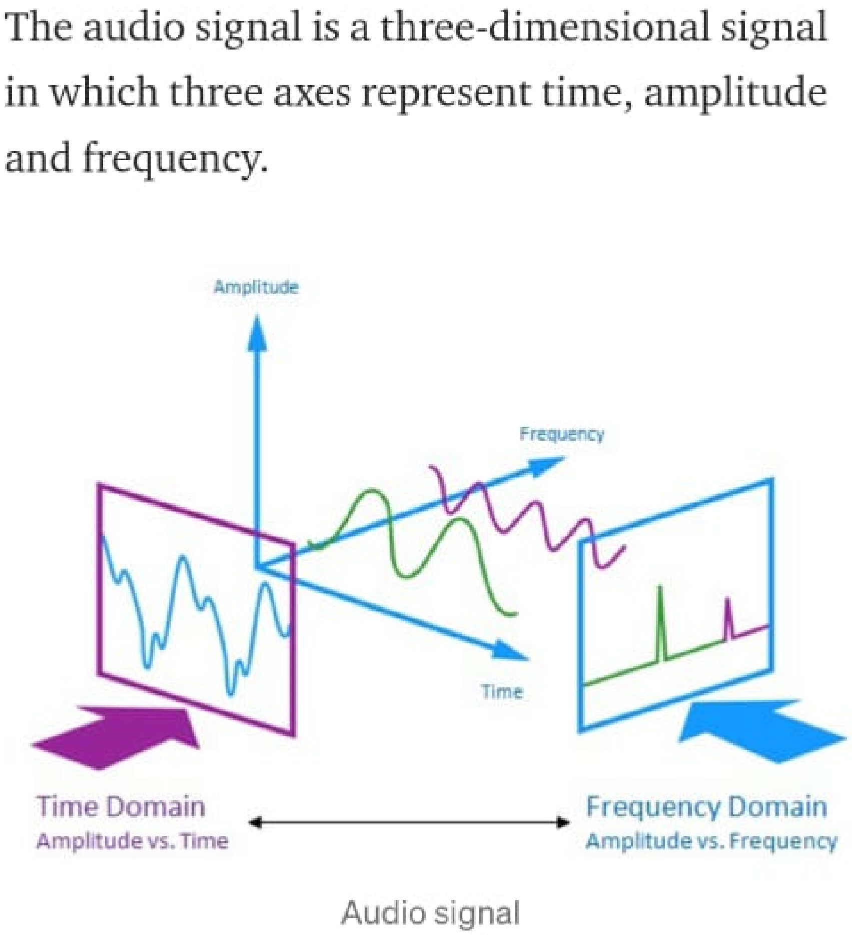
Here is the project planning – Ryerson audio-visual database of emotional speech and song(RAVDESS)data containing 7356 audio files was collected from the site kaggle.com, visualization of data(EDA) was performed using a waveplot which represented the time-domain representation of the speech signal .

Much more info was extracted when we performed visualization with a spectrogram tool which represented frequency of signals with time.

Feature extraction was further performed using MFCC technique which presented frequency domain features in lower level dimensionality.After this emotions were classified and data was preprocessed to obtain the final input data,ML algo selected and overall accuracy was calculated

CHAPTER 2: THEORY

2.1. REPRESENTATION OF AUDIO SIGNAL



The audio signal is a three-dimensional signal in which three axes represent time, amplitude and frequency.

Fig. 2.1: Pictorial representation of a speech signal

The pressure variations in the air falling on the eardrum results in the formation of audible sound. The human auditory system is responsive to sounds in the frequency range of 20 Hz to 20 kHz as long as the intensity lies above the frequency dependent "threshold of hearing". The human audible intensity range is approximately 120 dB. Audio signal analysis usually assumes that the signal properties change relatively slowly with time.

Signal parameters, or features, are estimated from the analysis of short windowed segments of the signal, and the analysis is repeated at uniformly spaced intervals of time. Sound is vibrations that propagate as an acoustic wave. The first step in working with audio files is to turn the audio wave into numbers so that you can feed this into our machine learning algorithm.

2.2. LOADING DATA

Librosa supports a wide range of audio codecs when it comes to loading audio into Python. When it comes to audio data processing, the.wav(lossless) format is commonly employed. The libROSA library needs to be installed and imported into a jupyter notebook and then simply feed the file path to the librosa.load() function to read an audio file.

The function librosa.load() —> returns two things: 1. an array of amplitudes. 2. The rate at which samples are taken. The Sampling frequency' utilised while recording the audio file is referred to as the sample rate.

2.3.VISUALISING AUDIO

WHAT IS BASICALLY DATA VISUALISATION?

The graphical depiction of information and data is known as data visualisation. Data visualisation tools make it easy to examine and comprehend trends, outliers, and patterns in data by employing visual elements like charts, graphs, and maps.

Data visualisation tools and technologies are critical in the Big Data environment for analysing enormous volumes of data and making data-driven decisions.

Data visualisation is a type of visual art that piques our curiosity and keeps our gaze fixed on the message We can rapidly spot trends and outliers while looking at a chart. If we can see something, we immediately assimilate it. It's purposeful storytelling.

We have got amplitudes and sampling-rates from librosa. We can easily plot these amplitudes with time. LibROSA provides a utility function waveplot().

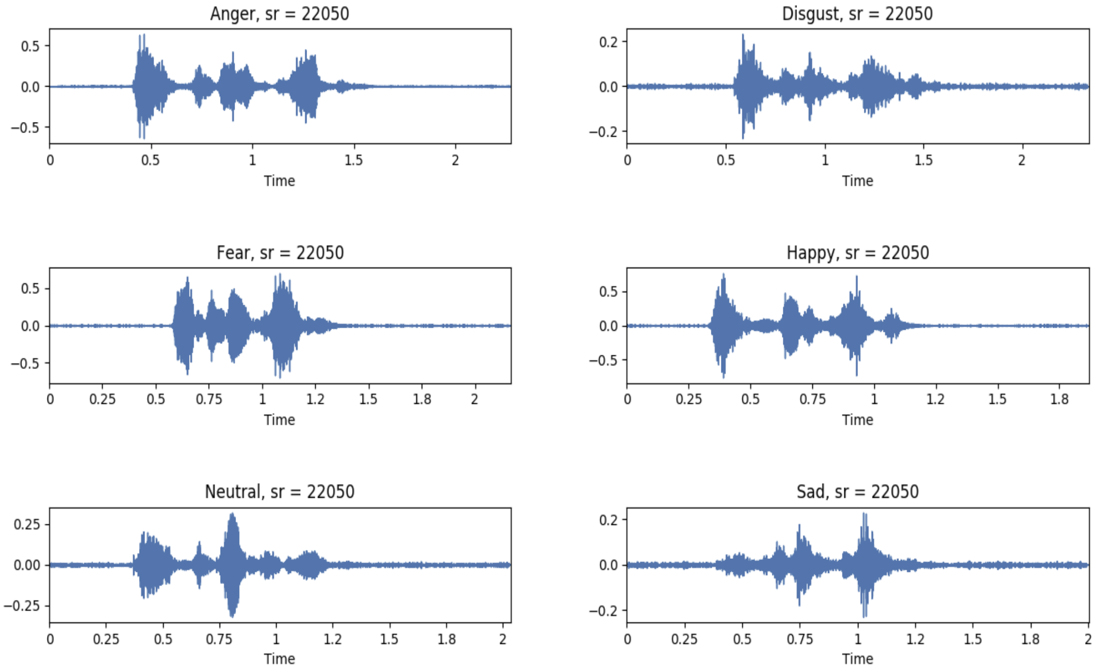


Fig. 2.2: Visualisation for raw audio time series for various emotions (Waveplot)

It is evident that Anger, Fear, and Happy emotions are touching a higher amplitude of 0.5 because people are generally loud with such emotions. But this also shows that these emotions will have a significant overlap in the time domain. Emotions like Sad, Neutral, Disgust have a lower amplitude (< 0.25) Also there is more breathing noise in these emotions as compared to Fear and Anger.Therefore spectral features are more suited for Speech Emotion Recognition as those are invariant to such time-domain peculiarities.

These amplitudes aren't very useful because they simply refer to the volume of an audio recording. It is required to convert the audio signal into the frequency domain in order to better understand it. A signal's frequency-domain representation reveals which frequencies are present in the signal. The

Fourier Transform is a mathematical technique that allows you to convert a continuous signal from time to frequency.

2.4. FOURIER TRANSFORM (FT)

An audio signal is a complex signal made up of multiple single-frequency sound waves' that move together in the medium as a disturbance (pressure change). We only record the resultant amplitudes of those numerous waves when we record sound. The Fourier Transform is a mathematical technique that allows you to break down a signal into its individual frequencies. The Fourier transform not only determines the frequencies present in a signal, but it also determines the magnitude of each frequency.

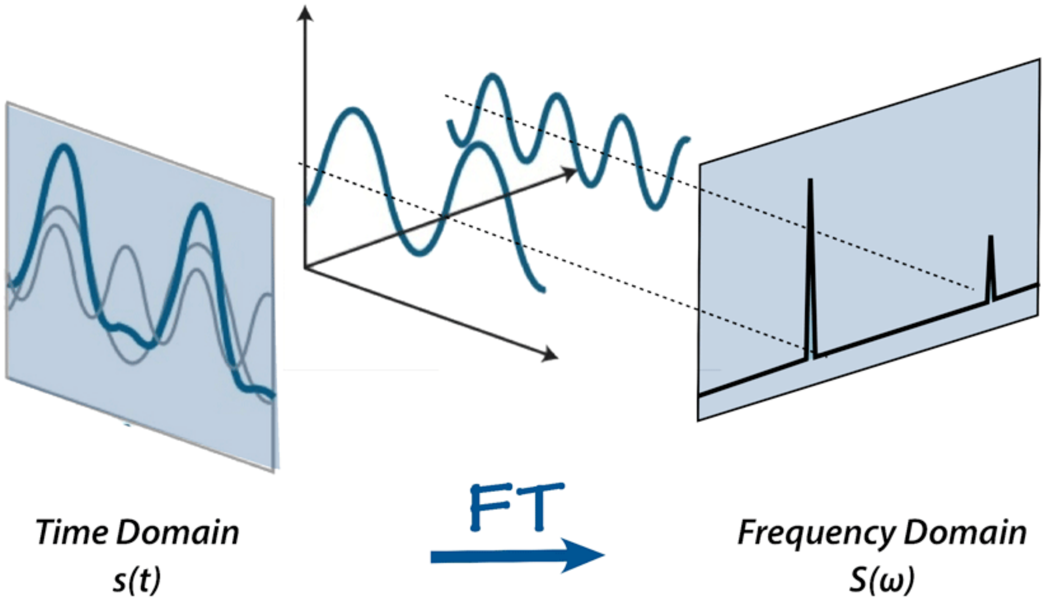


Fig. 2.3: Fourier Transformation

The Inverse Fourier Transform is just the opposite of the Fourier Transform. It takes the frequency-domain representation of a given signal as input and does mathematically synthesize the original signal.

The Fast Fourier Transform (FFT) is a mathematical approach for computing the Discrete Fourier Transform (DFT) of a sequence. The main difference between FT (Fourier Transform) and FFT is that FT considers a continuous signal as input, whereas FFT examines a discrete signal. DFT, like FT for continuous signals, converts a sequence (discrete signal) into its frequency elements. We have a sequence of amplitudes sampled from a continuous audio stream in our scenario. This time-domain discrete signal can be converted to a frequency-domain signal using the DFT or FFT algorithms.

However, when we used FFT to analyse our data, we just got frequency numbers and lost track of time. If we employ these frequencies as features, our system will no longer be able to determine which was spoken first. We must devise a new method of calculating characteristics for our system, one that includes frequency values as well as the time at which they were detected. Spectrograms are used in this case.

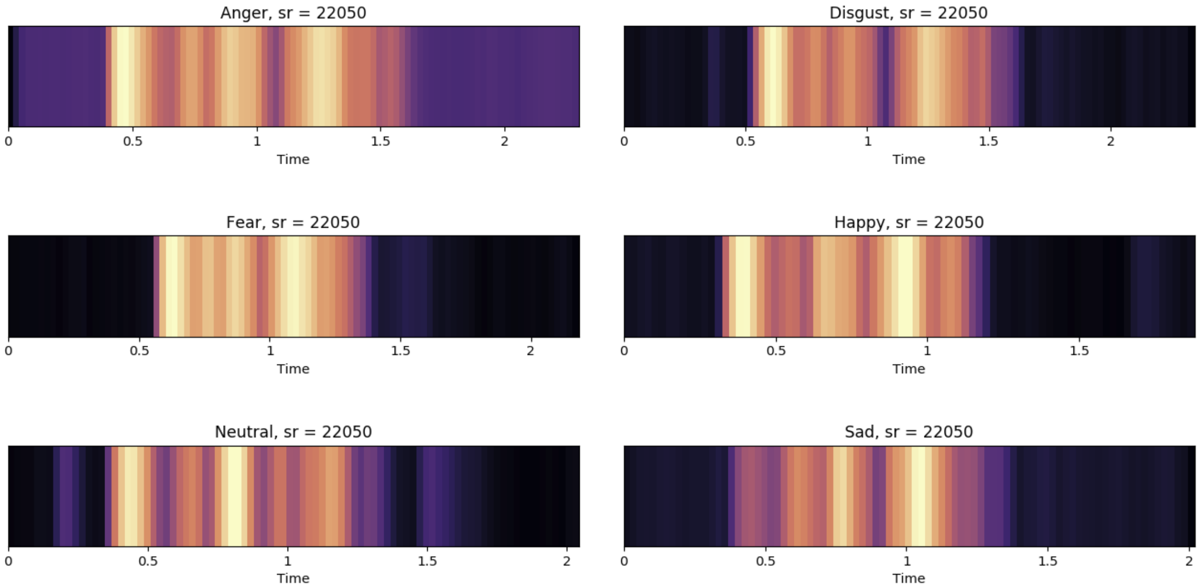


Fig. 2.4: Visualisation for MFCC feature plot (n_mfcc = 1) for various emotions

For emotions like Fear, Happy the sound level (dB) goes real low and we can observe much more black regions. For emotions like Anger, Neutral the sound level (dB) doesn't go very low and some purple areas are clearly visible. This shows that emotions like Fear or Happiness are impulsive and occur for a shorter duration whereas Anger, Neutral and Sad emotions are much more spread out across the timeline. Fear and Anger emotions have much more yellow regions close to the 0 dB reference point which shows that these emotions are much louder as compared to other emotions. MFCC features represent better distinctions between various emotions in terms of loudness, the spread of emotions, and frequency ranges.

2.5. SPECTROGRAM

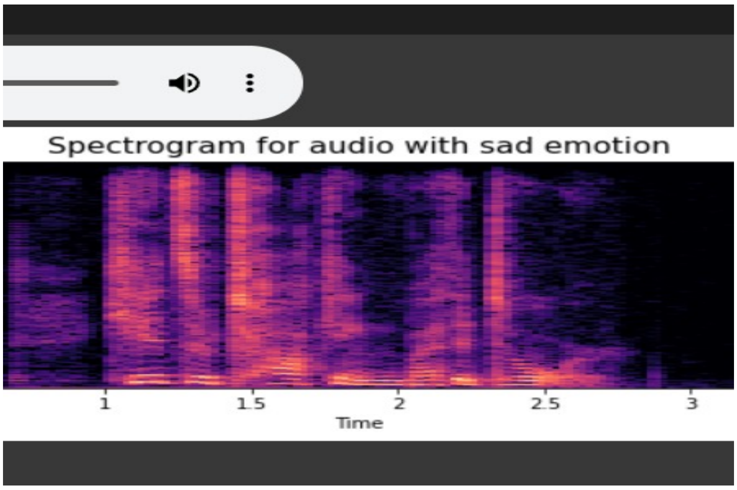


Fig. 2.5: Visualisation of data using spectrogram tool

- A spectrogram is a visual way of representing the signal strength, or “loudness”, of a signal over time at various frequencies present in a particular waveform.
- Spectrograms are basically two-dimensional graphs, with a third dimension represented by colors. Time runs from left (oldest) to right (youngest) along the horizontal axis.
- Here is a representation of both freq and time domain features(2-D graphs), each color represents how much freq level is contributed at a particular time frame.

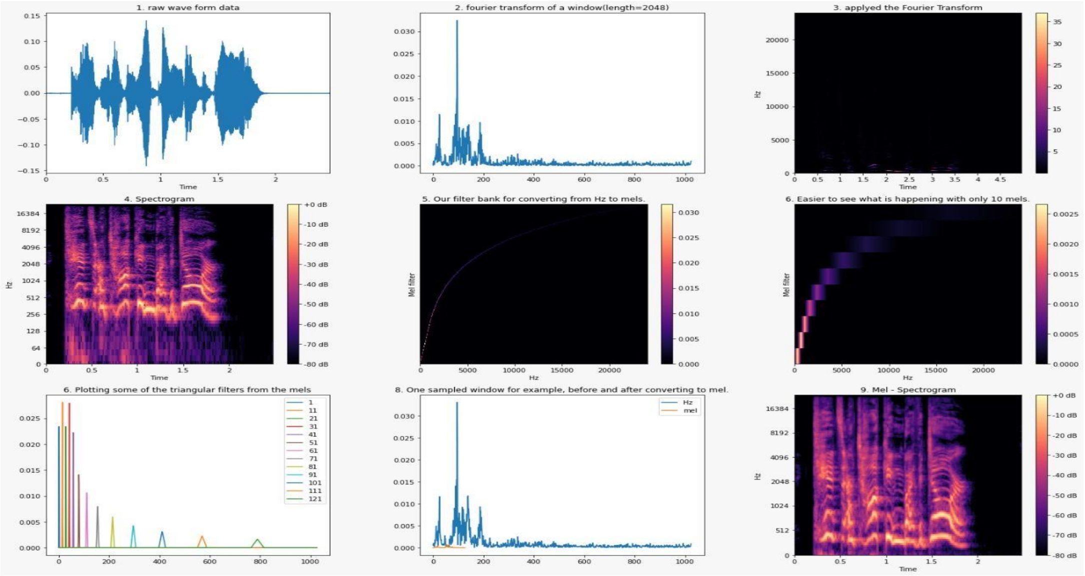


Fig. 2.6: Representation of raw audio data frame to Mel-spectrogram

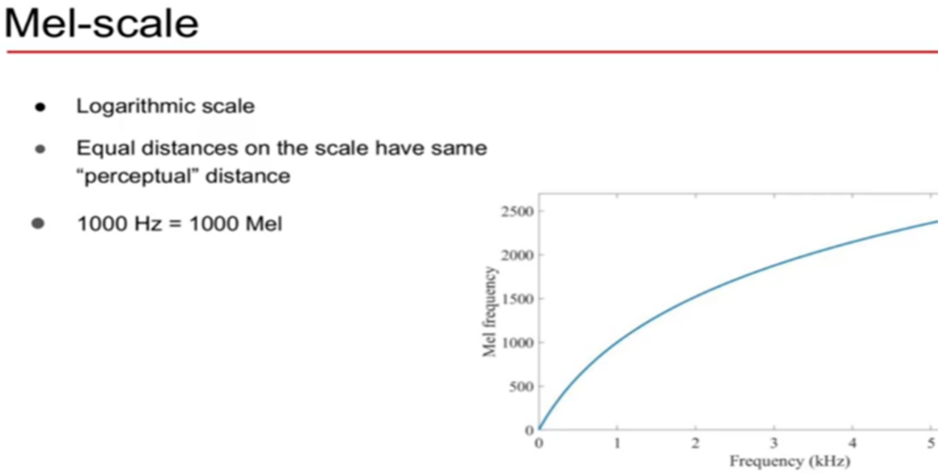
Here we can see that first we have represented the raw audio waveform in time domain. Then to convert it to freq domain we have performed a FFT on the signal. Then it has been represented in both time and frequency domain ,i.e. Spectrogram. Then the spectrogram is passed to a mel filter bank to convert it into a mel spectrogram plot.

Mel-Spectrogram applies a frequency-domain filter bank to audio signals that are windowed in time.Basically spectrogram is a visual representation of sound where we plot the intensity of frequency over time. Frequency refers to the number of sound waves per unit time. At any point in time, a sound is a combination of multiple frequencies such as 10 Hz, 50 Hz, 300 Hz, or others, and the intensity is different at each frequency. A spectrogram represents intensity variation in frequencies over a period.

2.6.FEATURE EXTRACTION

- Speech recognition is a task that must be learned under supervision. The audio signal will be the input for the speech recognition issue, and we must predict the text from the audio signal. Because there will be a lot of noise in the audio stream, we can't use it as an input to our model.
- The purpose of feature extraction is to eliminate redundancy, minimise computing time, and make feature distribution modelling easier.

- It is clear that extracting features from the audio signal and applying them as input to the basic model produces significantly better results than utilising the raw audio signal as input.
- Mel Frequency Cepstral Coefficients, or MFCC, is a widely used approach for extracting characteristics from audio signals (physical feature extraction)



Mel-scale

- Logarithmic scale
- Equal distances on the scale have same "perceptual" distance
- 1000 Hz = 1000 Mel

Fig. 2.7: Mel-Scale graph

The Mel scale is a scale that compares the perceived frequency of a tone to the frequency that can be measured. It adjusts the frequency to more precisely match what the human ear can hear (humans are better at identifying small changes in speech at lower frequencies).

In Fourier analysis, the cepstrum is the result of computing the inverse Fourier transform of the logarithm of the estimated signal spectrum. Cepstral coefficients are commonly used as compact representations of audio signals.

Generally, they are calculated after an audio signal is passed through a filter bank and the energy in the individual filters is summed.

2.7. MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

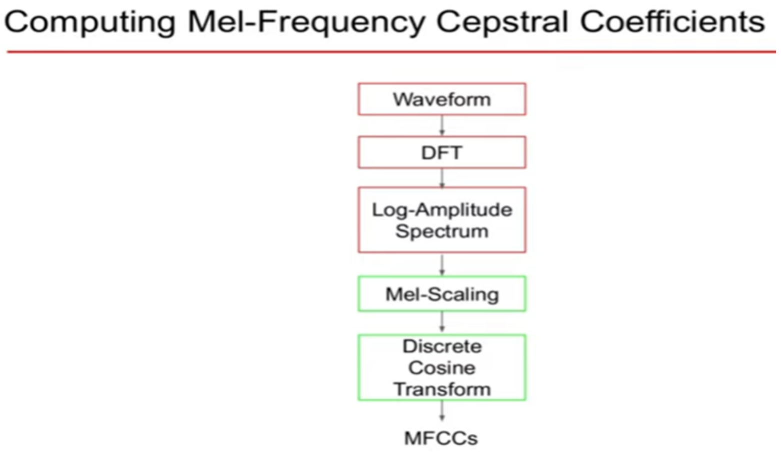

Computing Mel-Frequency Cepstral Coefficients
Waveform
DFT
Log-Amplitude Spectrum
Mel-Scaling
Discrete Cosine Transform
MFCCs

Fig. 2.8: Computation method for MFCC

The MFCC technique aims to develop the features from the audio signal which can be used for detecting the phones in the speech.The motivating idea of mel frequency cepstral coefficients is to compress information about the vocal tract (smoothed spectrum) into a small number of coefficients based on an understanding of the cochlea.

Windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by using the inverse DCT are all part of the MFCC feature extraction technique.

The MFCC summarises the frequency distribution across the window size, so it is possible to analyse the frequency characteristics of the sound. These audio representations will allow us to identify features for classification.

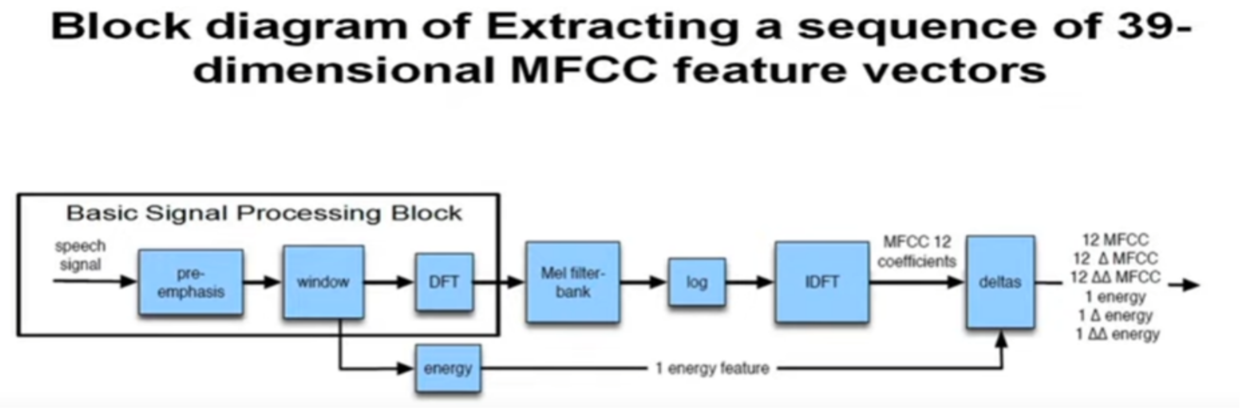


Fig. 2.9: Block diagram of MFCC technique

The MFCC model takes the first 12 coefficients of the signal after applying the idft operations. Along with the 12 coefficients, it will take the energy of the signal sample as the feature. It will help in identifying the phones.

Along with these 13 features, the MFCC technique will consider the first order derivative and second order derivatives of the features which constitute another 26 features.

Derivatives are calculated by taking the difference of these coefficients between the samples of the audio signal and it will help in understanding how the transition is occurring

So overall MFCC technique will generate 39 features from each audio signal sample which are used as input for the speech recognition model.

2.8.CLASSIFICATION

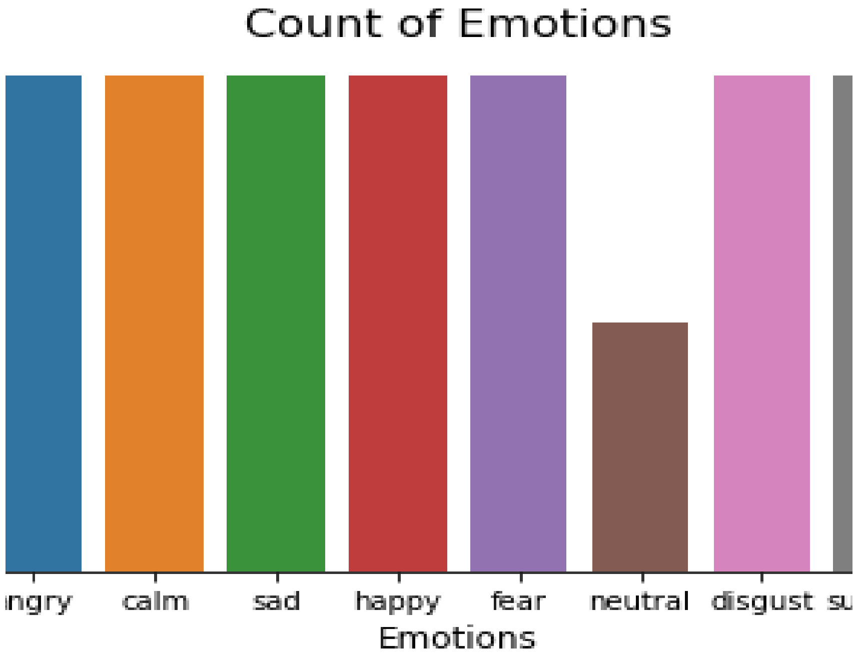
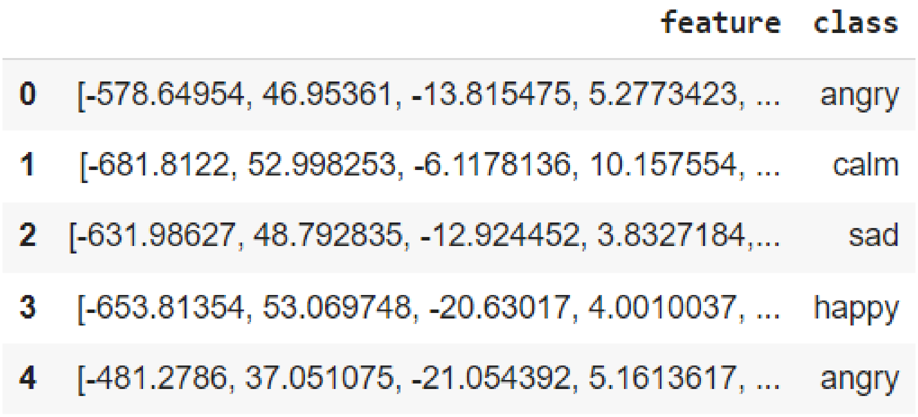


Fig. 2.10: Classes of emotions

These are broadly the 7 classes of emotions that are taken into account for our classification i.e- happy,fear,neutral,disgust,anger,sad and calm.

CHAPTER 3 : RESULTS

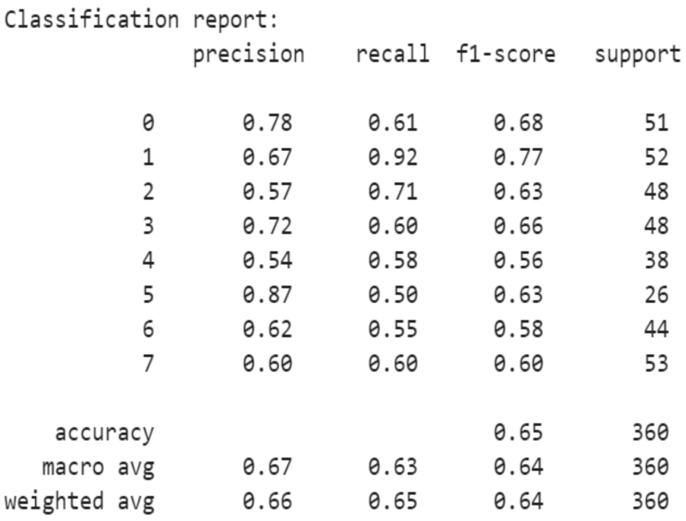


	feature	class
0	[-578.64954, 46.95361, -13.815475, 5.2773423, ...	angry
1	[-681.8122, 52.998253, -6.1178136, 10.157554, ...	calm
2	[-631.98627, 48.792835, -12.924452, 3.8327184,...	sad
3	[-653.81354, 53.069748, -20.63017, 4.0010037, ...	happy
4	[-481.2786, 37.051075, -21.054392, 5.1613617, ...	angry

Fig. 3.1: Preprocessed input data

So this is the particular pre-processed data we are using for our ML classification purpose.

A k- cross validation test was performed on diff ML classification algo and it was found that the Extra Tree Classifier model gave the best classification result among others.

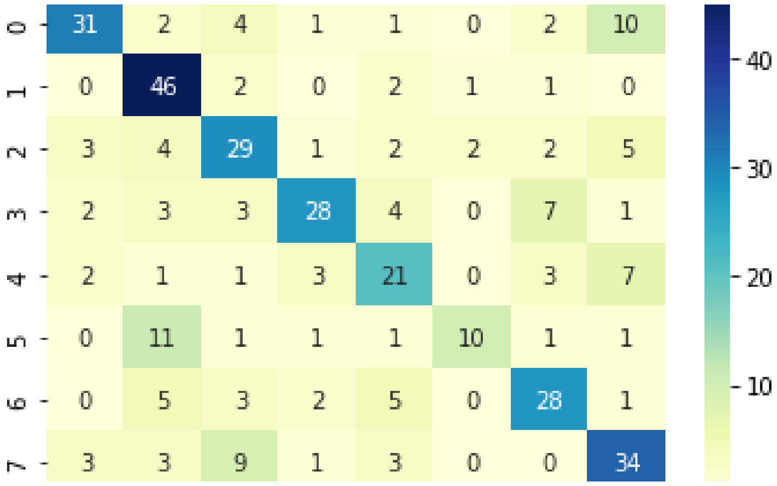


```
Classification report:
              precision    recall  f1-score   support

           0       0.78      0.61      0.68        51
           1       0.67      0.92      0.77        52
           2       0.57      0.71      0.63        48
           3       0.72      0.60      0.66        48
           4       0.54      0.58      0.56        38
           5       0.87      0.50      0.63        26
           6       0.62      0.55      0.58        44
           7       0.60      0.60      0.60        53

    accuracy                           0.65       360
   macro avg       0.67      0.63      0.64       360
weighted avg       0.66      0.65      0.64       360
```

Fig. 3.2: Evaluation metric depicting model accuracy(classification report)



0	31	2	4	1	1	0	2	10
1	0	46	2	0	2	1	1	0
2	3	4	29	1	2	2	2	5
3	2	3	3	28	4	0	7	1
4	2	1	1	3	21	0	3	7
5	0	11	1	1	1	10	1	1
6	0	5	3	2	5	0	28	1
7	3	3	9	1	3	0	0	34

Fig. 3.3: Evaluation metric(confusion matrix)

To calculate the accuracy of our model,accuracy_score() function to be imported from sklearn library Classification report and confusion matrix were imported as evaluation metrics to evaluate model accuracy

Model accuracy - 65%

CHAPTER 4 : CONCLUSION

4.1: CONCLUSION

So from the present progress of our project, we concluded that-

The frequency domain analysis of audio speech data was performed where the audio signal was represented using Spectrogram which was helpful in identifying the characteristics of non linear signals like audio.
Then concisely the shape of spectral envelope of signal using MFCCs feature extraction method was extracted
Then Machine Learning classification task was performed with an accuracy of 65% with Extra Tree Classifier model to classify the emotions

- The developed Extra Tree classifier model was able to predict the emotions from speech with an accuracy of 65% ,emotions like angry,sad,happy, neutral etc.

- MFCC feature extraction was preferred as it gives higher recognition accuracy and has less complexity over other feature extraction methods

- Similarly a spectrogram is helpful for vibrational analysis in a changing environment, it illustrates the patterns of energy change which may not be visible in FFT(Fast fourier transform)

- Thus speech recognition technology allows computers to take spoken audio and interpret it .Over time , the computer can learn to understand speech ,all possible due to advances in artificial intelligence and ML.

4.2 : FUTURE SCOPE

- As it can be seen that we got an accuracy of 65% from the classification of frequency domain features of audio signals. But furthermore extraction and use of other physical features of signal like Zero cross rate, spectral centroid, short time energy etc. can be done so as to get more data and perform better classification tasks.

- In terms of improving accuracy Neural Network models like CNN can be implemented to perform better classification of our models.

- Also, a front end interface can be developed where real time emotions can be classified and calculate the degree of affection such as love, truthfulness, and friendship of the person you are talking to.

- These models can be extremely useful for the individuals with Autism Spectrum Disorder to help guide conversation and create/maintain healthy relationships with others.

- SER(Speech Emotion Recognition) system can also be used in-car board system based on information of the mental state of the driver can be provided to the system to ensure his/her safety preventing accidents to happen

REFERENCES

[1] Dave, Namrata. "Feature extraction methods LPC, PLP and MFCC in speech recognition.." International journal for advance research in engineering and technology 1, no. 6 (2013): 1-4.

[2] Prasomphan, S., 2015, Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. International Conference on Systems, Signals and Image Processing (IWSSIP), September, IEEE-2015 (pp. 73-76), IEEE.

[3] Khalil, R.A., Jones, E., Babar, M.I., Jan, T., Zafar, M.H. and Alhussain, T., 2019. Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7, pp.117327- 117345

[4] Jin, Q., Li, C., Chen, S. and Wu, H.Speech emotion recognition with acoustic and lexical features,In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)

[5] Likitha, M.S., Gupta, S.R.R., Hasitha, K. and Raju, A.U., 2017, Speech based human emotion recognition using MFCC. In 2017 international conference on wireless communications, signal processing and networking (WiSPNET), March (pp. 2257-2260). IEEE.

[6] Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S. and Vepa, J., 2018, Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In Interspeech, September (pp. 3688-3692).

[7] Akçay, M.B. and Oğuz, K., 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, pp.56-76.