TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PURWANCHAL CAMPUS

A
PROJECT REPORT
ON
SPEECH EMOTION RECOGNITION USING RAVDESS,CREMA-D AND TESS AUDIO DATASETS & CNN AS DEEP LEARNING MODEL

SUBMITTED BY:

Amrit Poudel (PUR075BCT010)
Anil Karki (PUR075BCT011)
Dilip Khadka (PUR075BCT028)
Jeevan Raj Panta (PUR075BCT041)

SUBMITTED TO:

DEPARTMENT OF COMPUTER AND ELECTRONICS ENGINEERING
PURWANCHAL CAMPUS,DHARAN

March,2023

Introduction

- Speech emotion recognition (SER) is the process of automatically identifying or detecting emotions conveyed through speech signals.
- It involves analyzing various acoustic features of speech, such as pitch, intensity, rhythm, and spectral characteristics, to determine the emotional state of the speaker.
- It is used in variety of applications, including speech therapy, human-robot interaction, and customer service, education, forensics and medical analysis.
- Its goal is to classify the emotional state of the speaker into one or more predefined emotional categories, such as happiness, sadness, anger, fear, or surprise.

Objectives

- To improve the accuracy, efficiency, and effectiveness of human-machine interactions.
- To enhance the understanding of human emotions and behavior.
- To classify various audio speech files into different emotions such as happy, sad, anger and neutral.
- To accurately detect and identify the emotional state of the speaker.
- To monitor psycho physiological state of a person.

Challenges

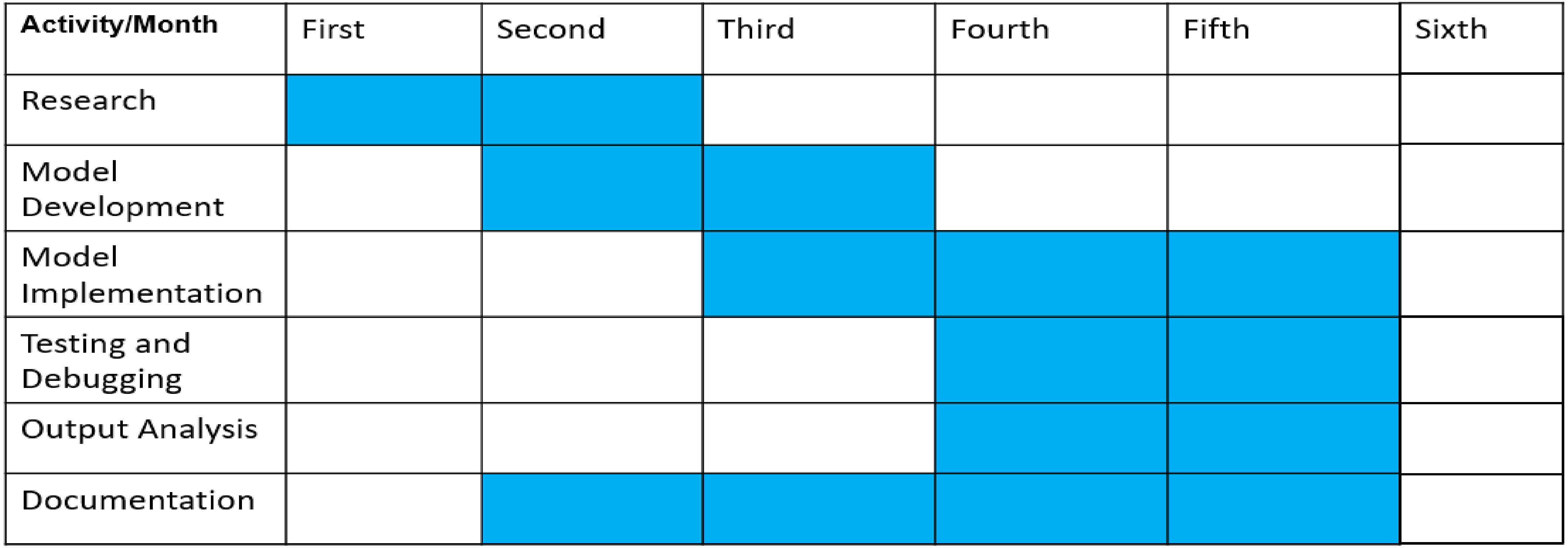
- Lengthy time for training model(more than hour)
- Inadequate data due to which model is not that much robust
- Model might not be able to predict efficiently in cross-language domain

Future Enhancement

- The training can be implemented by acquiring more datasets
- Can expand more emotion categories like bore, confuse and so on
- Can use more advanced machine learning models to increase accuracy

METHODOLOGY

Project Timeline



Activity/Month	First	Second	Third	Fourth	Fifth	Sixth
Research	■	■				
Model Development		■	■			
Model Implementation			■	■	■	
Testing and Debugging				■	■	
Output Analysis				■	■	
Documentation		■	■	■	■	

Tools Used

- Numpy : for linear algebraic operations
- Pandas :for data handling, data cleaning, data analysis and integration with other tools and libraries such as NumPy
- Librosa : to extract audio features
- Keras :for building and training deep learning models
- Matplotlib: to create 2D plots and graphs

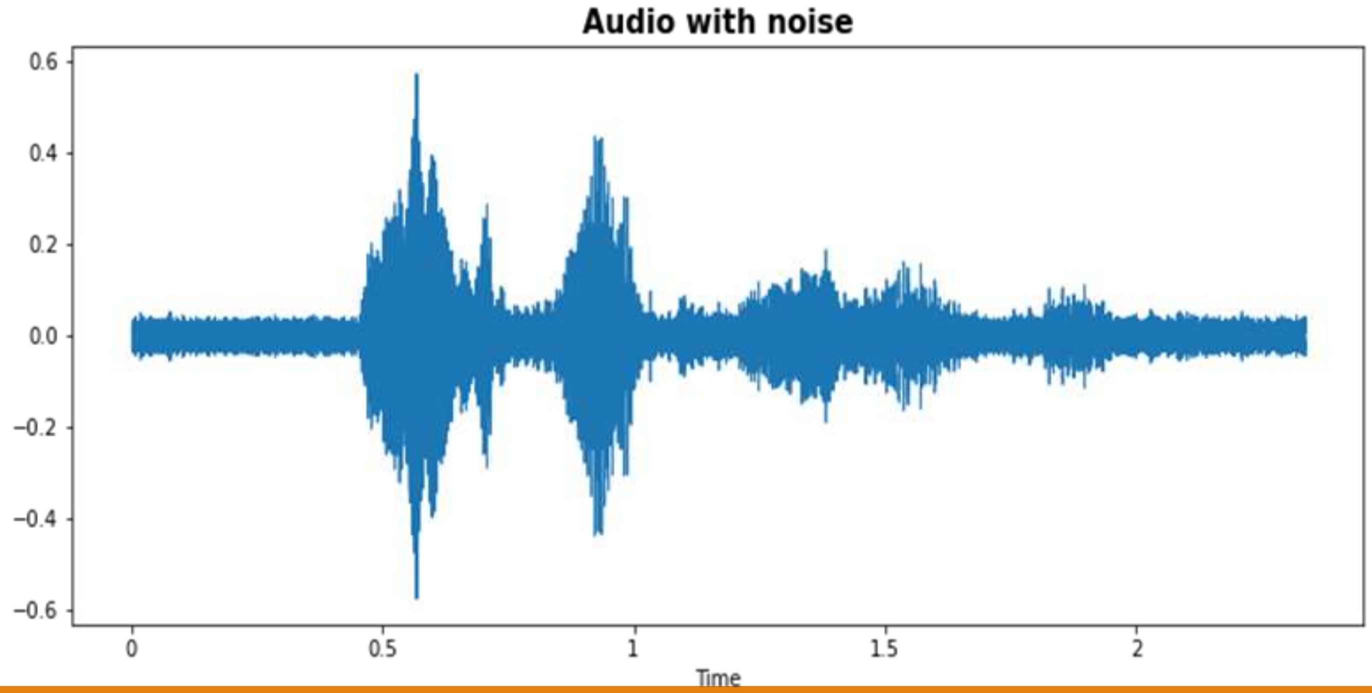
Data Acquisition Techniques Used

- ❖RAVDESS Dataset : contains 24 professional actors (12 female, 12 male) and 1440 files, 60 trials per actor
- ❖CREMA Dataset : contains 7,442 original clips from 91 actors (48 Male,43 Female) of different races and ethnicities
- ❖TESS Dataset: contains 2800 audio files in total.200 target words were spoken from 26 and 64 years aged two actresses, and recording were made portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral)

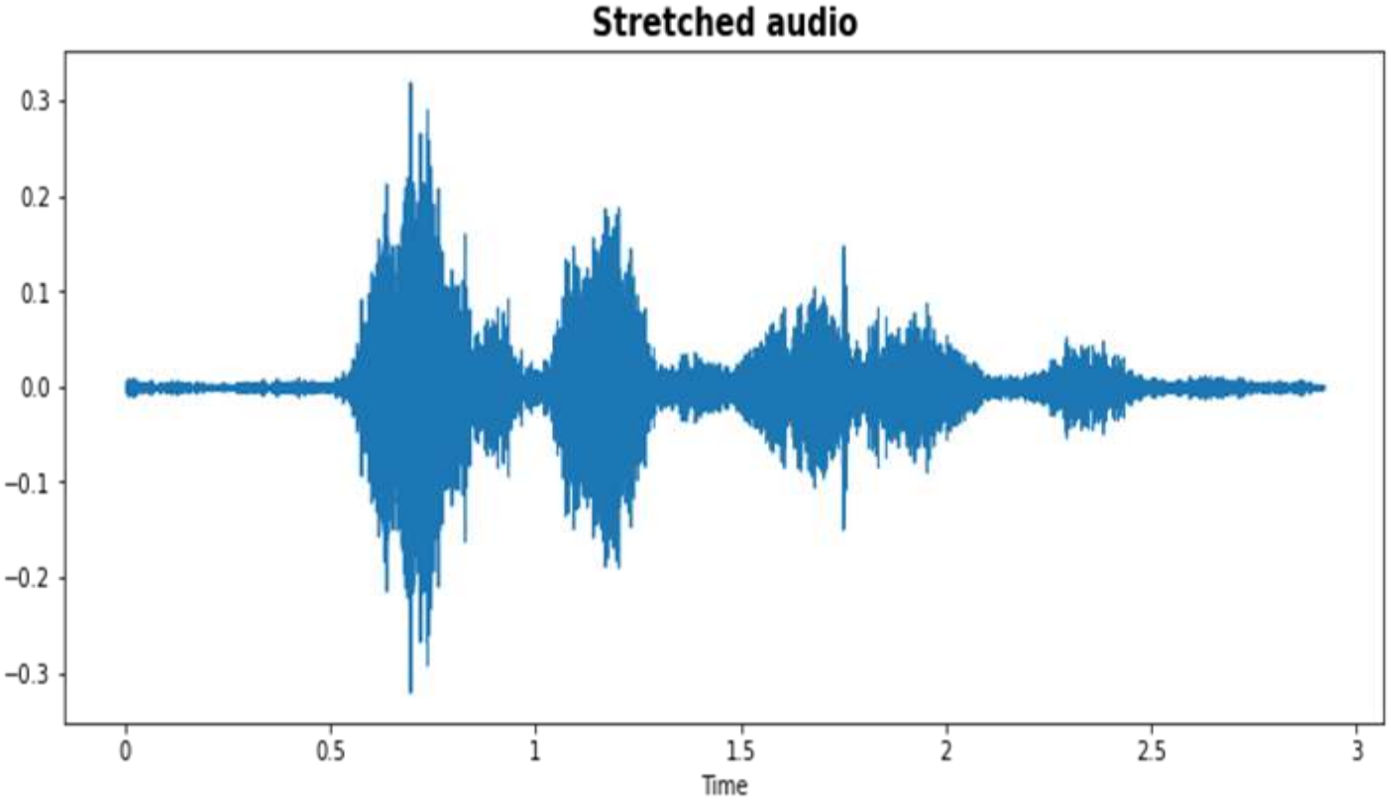
Data Augmentation

- Data augmentation is a powerful tool for improving the performance of machine learning models, especially when the size of the dataset is small.

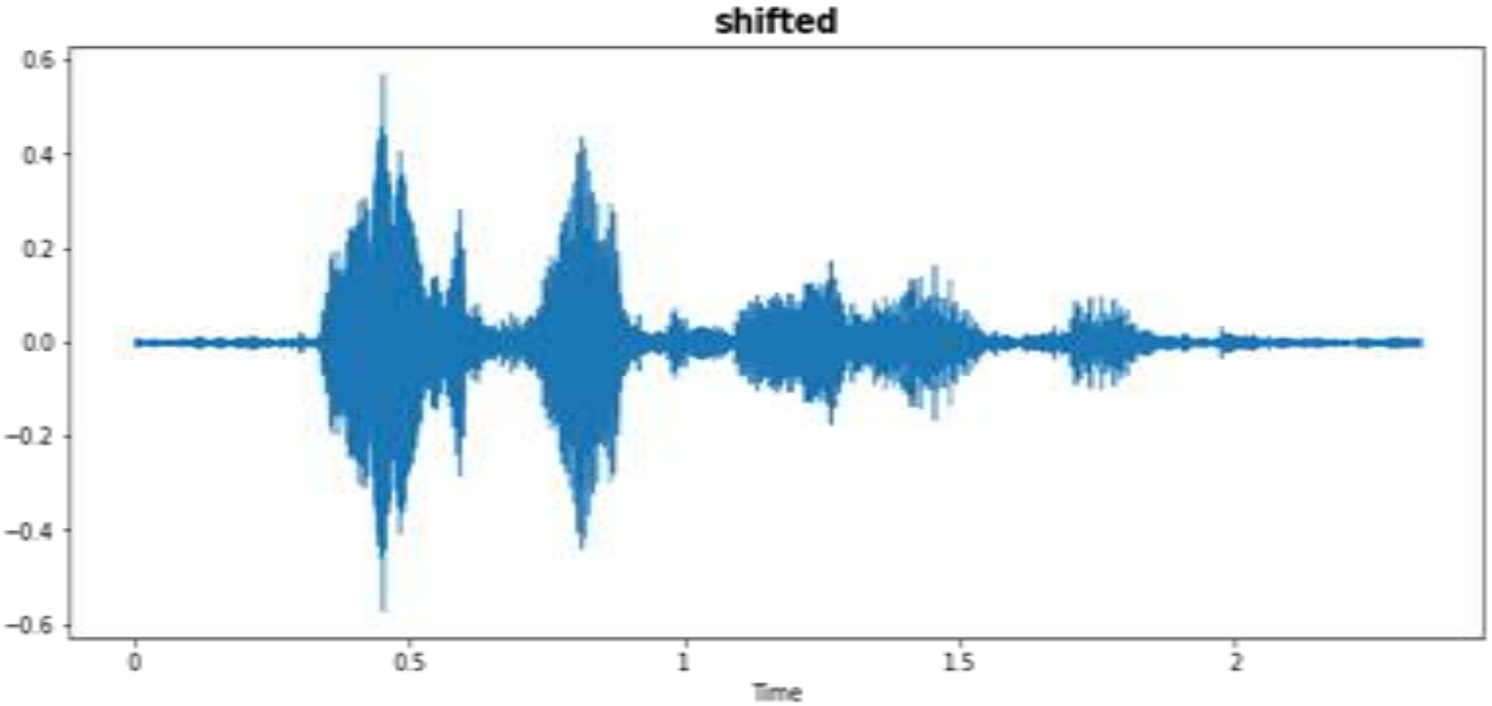
1. **Noise Injection**



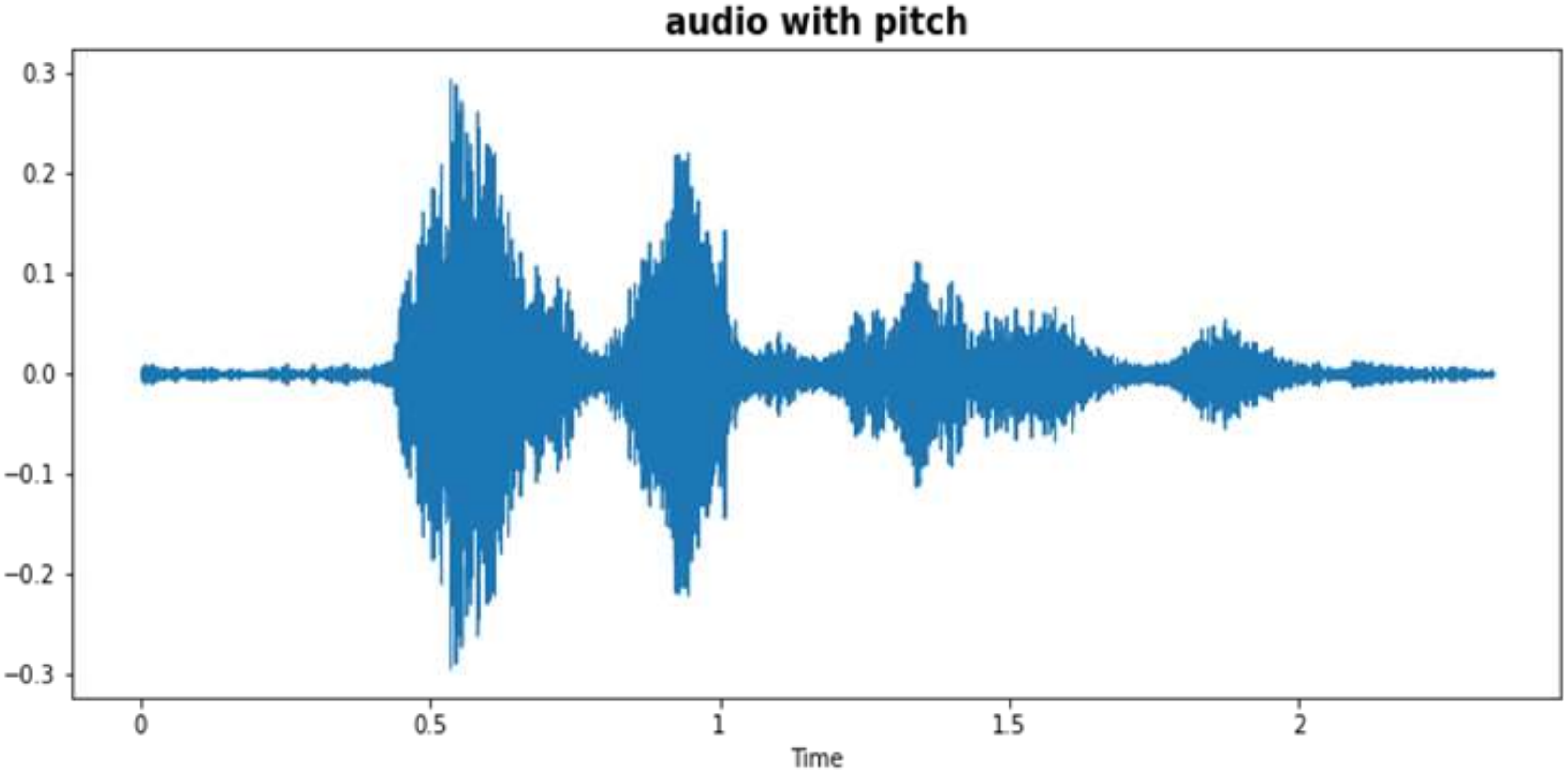
2. Stretching



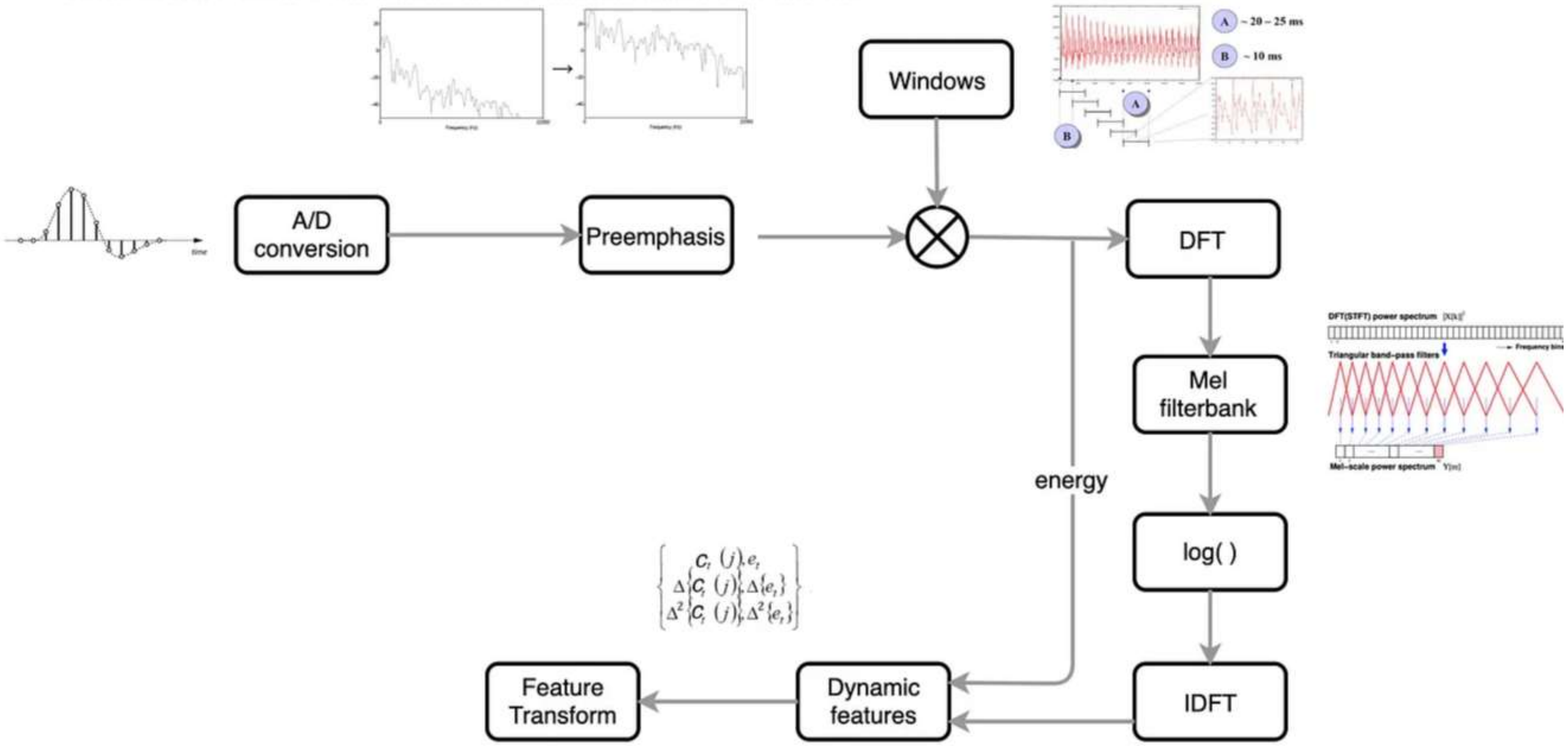
3.Shifting



4.Pitching



Mel Frequency Cepstral Coefficients (MFCC)



A/D conversion

Preemphasis

Windows

DFT

Mel filterbank

log( )

IDFT

energy

Dynamic features

Feature Transform

$$\left\{\begin{matrix} C_t(j), e_t \\ \Delta\{C_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{C_t(j)\}, \Delta^2\{e_t\} \end{matrix}\right\}$$

time
A ~ 20 – 25 ms
B ~ 10 ms
A
B
DFT(STFT) power spectrum $|X(k)|^2$
Frequency bins
Triangular band-pass filters
Mel-scale power spectrum $Y(m)$

Agile Model

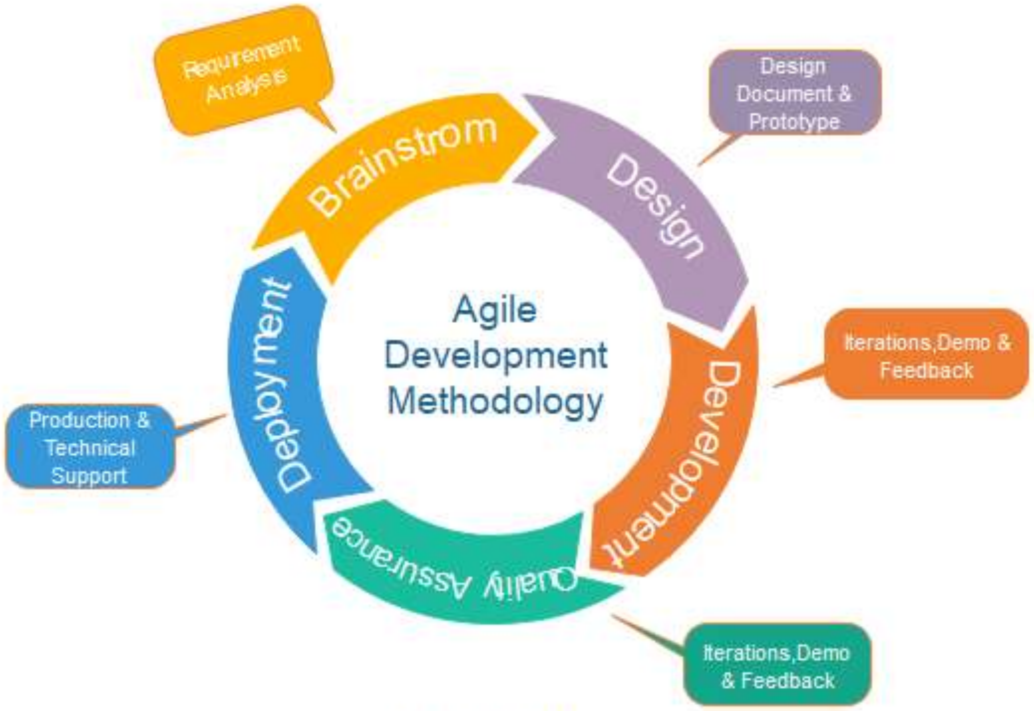
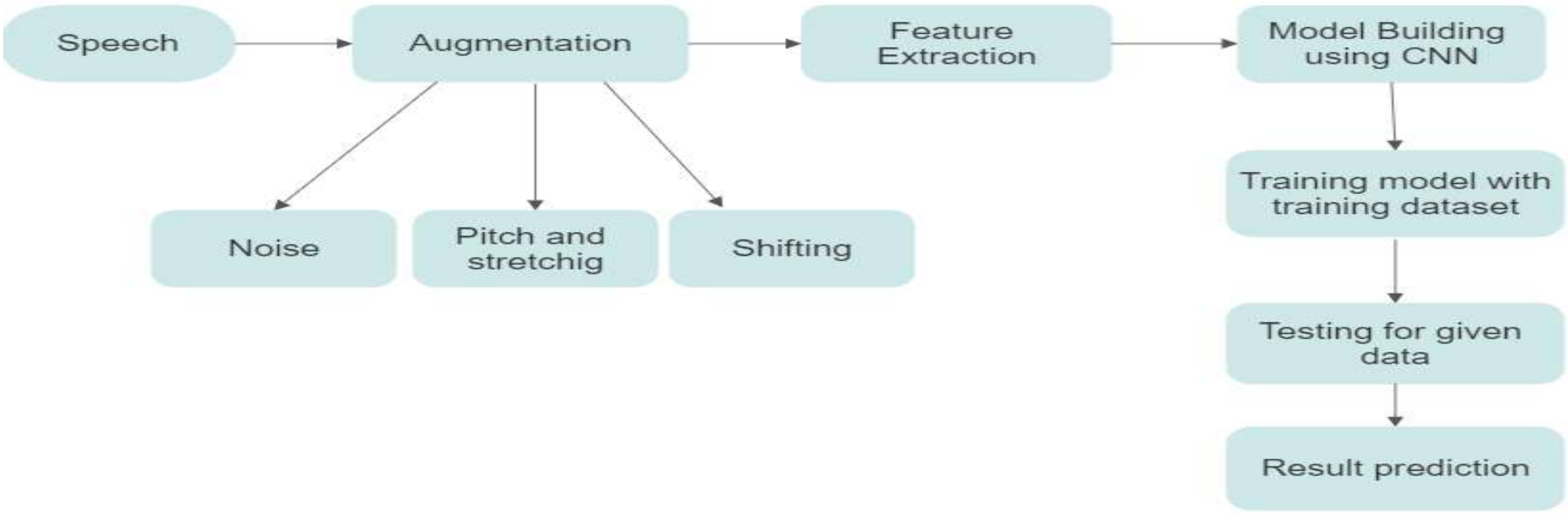
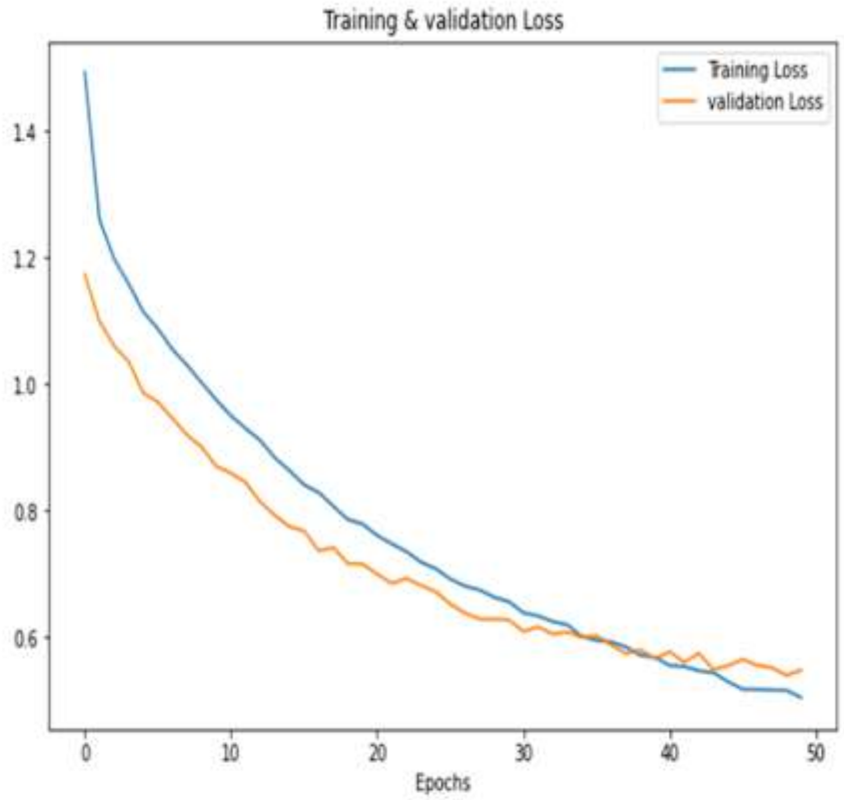

Requirement Analysis
Brainstrom
Design
Design Document & Prototype
Agile Development Methodology
Development
Iterations,Demo & Feedback
Quality Assurance
Iterations,Demo & Feedback
Deployment
Production & Technical Support

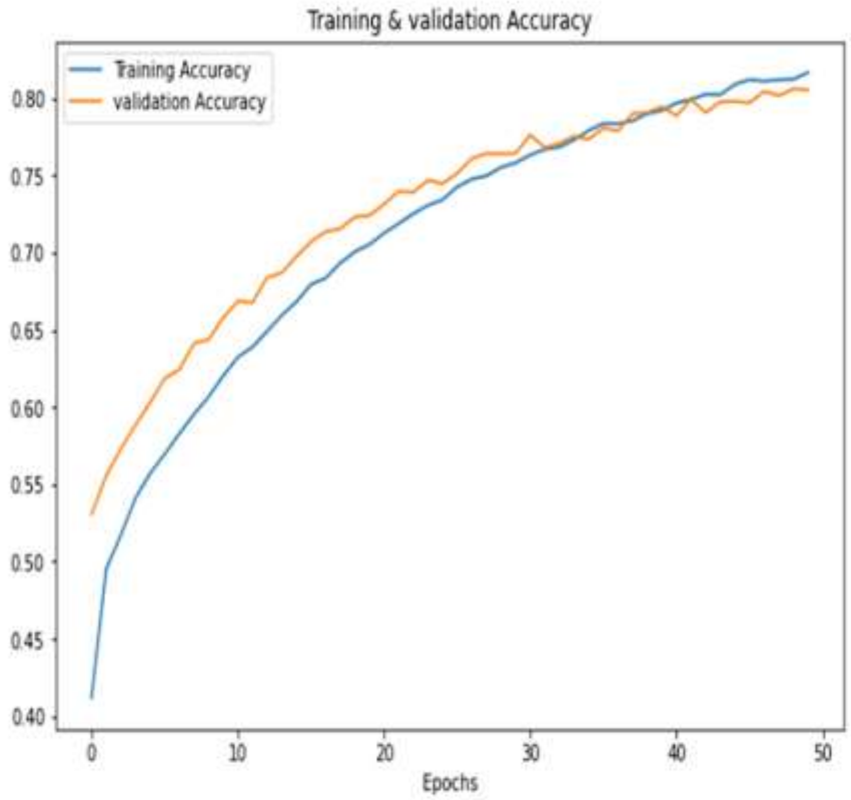
Fig. Agile Model

Overall System Design

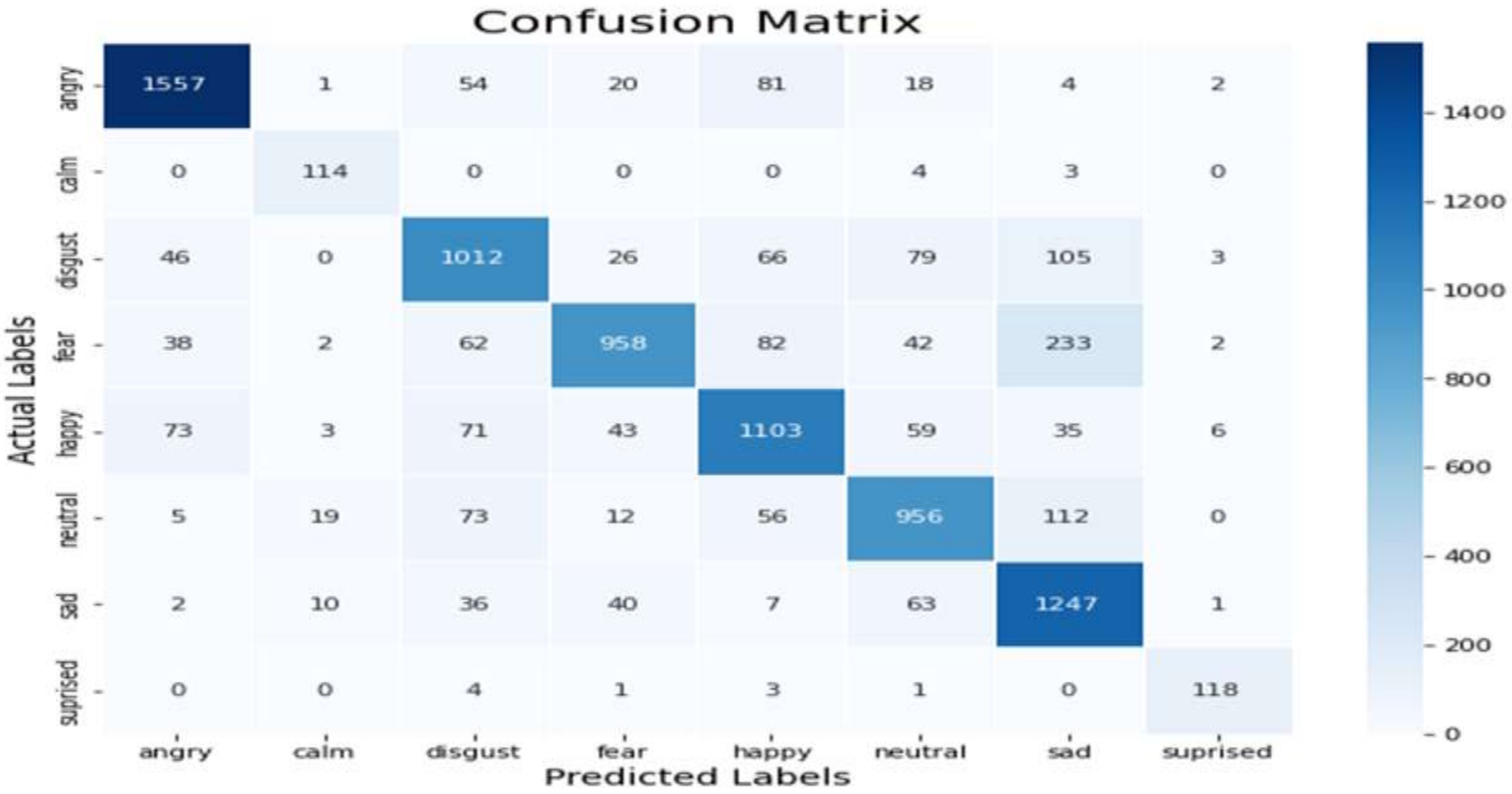


Training & Validation Accuracy





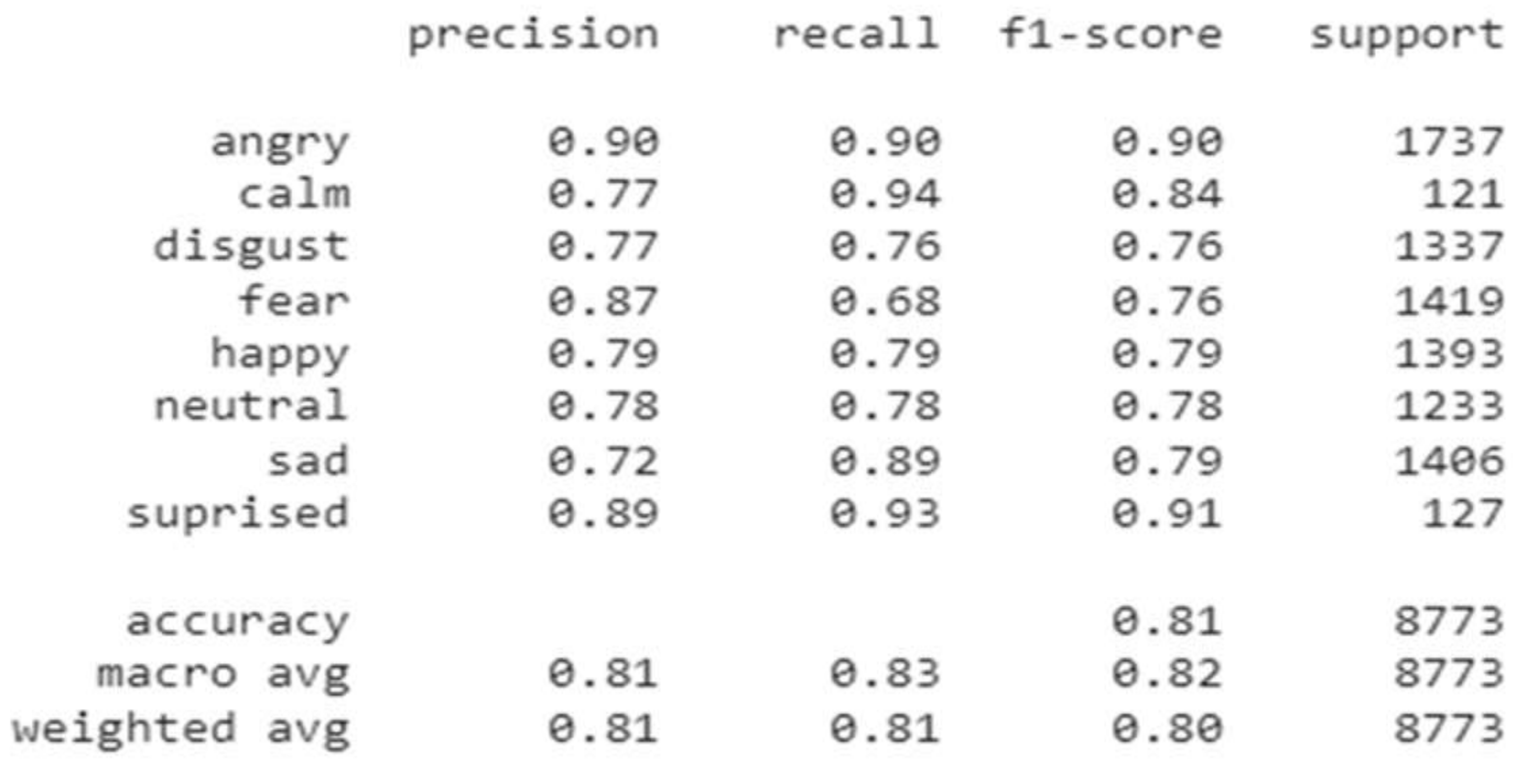
Confusion Matrix



Confusion Matrix

Actual Labels \ Predicted Labels	angry	calm	disgust	fear	happy	neutral	sad	suprised
angry	1557	1	54	20	81	18	4	2
calm	0	114	0	0	0	4	3	0
disgust	46	0	1012	26	66	79	105	3
fear	38	2	62	958	82	42	233	2
happy	73	3	71	43	1103	59	35	6
neutral	5	19	73	12	56	956	112	0
sad	2	10	36	40	7	63	1247	1
suprised	0	0	4	1	3	1	0	118

Classification Report



```
              precision    recall  f1-score   support

       angry       0.90      0.90      0.90      1737
        calm       0.77      0.94      0.84       121
     disgust       0.77      0.76      0.76      1337
        fear       0.87      0.68      0.76      1419
       happy       0.79      0.79      0.79      1393
     neutral       0.78      0.78      0.78      1233
         sad       0.72      0.89      0.79      1406
    suprised       0.89      0.93      0.91       127

    accuracy                           0.81      8773
   macro avg       0.81      0.83      0.82      8773
weighted avg       0.81      0.81      0.80      8773
```

References:

[1]. H. Cao, R. Verma, and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.
[2]. L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.
[3].T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.
[4]. S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2005.

Thank You!