A

PROJECT REPORT ON

EMOTION DETECTION USING SPEECH RECOGNITION

SUBMITTED TO SAVITRIBAI PHULE PUNE UNIVERSITY
FOR PARTIAL FULFILLMENT
OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

BACHELOR OF ENGINEERING
In
Electronics and Telecommunication Engineering

By
AMEY BHILEGAONKAR B150053038
SAURABH ZINJAD B150053267

GUIDE
Dr. G.V.BANSOD

DEPARTMENT OF
ELECTRONICS AND TELECOMMUNICATION ENGINEERING
PUNE INSTITUTE OF COMPUTER TECHNOLOGY
PUNE – 43

Department of Electronics and Telecommunication Engineering
Pune Institute of Computer Technology, Pune – 43

CERTIFICATE

This is to certify that the Project Report entitled

EMOTION DETECTION USING SPEECH RECOGNITION

Has been successfully completed by

AMEY BHILEGAONKAR B150053038

SAURABH ZINJAD B150053267

Is a bona fide work carried out by them under the guidance of Prof. G.V.Bansod and it is approved for the partial fulfillment of the requirement of the Savitribai Phule Pune University, Pune for the award of the degree of the Bachelor of Engineering (Electronics and Telecommunication Engineering). This project work has not been earlier submitted to any other Institute or University for the award of any degree or diploma.

Dr. G.V.Bansod Guide	Prof. Dr. Y. Ravinder HOD, E&TC Dept	Prof. Dr. P.T. Kulkarni Principal, PICT

Place: Pune
Date :

ACKNOWLEDGEMENTS

The successful completion of this project depends on co-operation and the help of many people. We take this opportunity to express our sense of gratitude to those who have extended a helpful hand in our efforts to make it successful. We are grateful to Dr.D.Y.Ravinder Sir Head of Department and Prof.R.C.Jaiswal Sir and Dr.G.V.Bansod Sir our guide and Prof. Galande Sir for providing the necessary atmosphere and co-operation to make this project a successful one. We are highly in betted to our guide for consent and valuable technical guidance without whom this project may not have been possible.

At last we would like to thank all our friends and colleagues who directly or indirectly helped us in completion of our work.

Amey Bhilegaonkar

Saurabh Zinjad

CONTENTS

ABSTRACT

The human voice is very versatile and carries a multitude of emotions. Emotion in speech carries extra insight about human actions. Through further analysis, we can better understand the motives of people, whether they are unhappy customers or cheering fans. Humans are easily able to determine the emotion of a speaker, but the field of emotion recognition through machine learning is an open research area.

In this proposed project, we perform speech data analysis on speaker discriminated speech signals to detect the emotions of the individual speakers involved in the conversation. We are analyzing different techniques to perform speaker discrimination and speech analysis to find efficient algorithms to perform this task.

List of Figures

CHAPTER 1

Introduction

1.1 Background

- Although emotion detection from speech is a relatively new field of research, it has many potential applications. In human-computer or human-human interaction systems, emotion recognition systems could provide users with improved services by being adaptive to their emotions. In virtual worlds, emotion recognition could help simulate more realistic avatar interaction.
- The body of work on detecting emotion in speech is quite limited. Currently, researchers are still debating what features influence the recognition of emotion in speech. There is also considerable uncertainty as to the best algorithm for classifying emotion, and which emotions to class together.
- For a machine to understand the mindset/mood of the humans through a conversation, it needs to know who are interacting in the conversation and what is spoken, so we implement a speaker and speech recognition system first and perform speech analysis on the data extracted from prior processes.
- Understanding the mood of humans can be very useful in many instances. For example, computers that possesses the ability to perceive and respond to human non-lexical communication such as emotions. In such a case, after detecting humans' emotions, the machine could customize the settings according his/her needs and preferences.

.

1.2 Relevance

- In this project, we attempt to address these issues. We will use Convolutional Neural Networks and LSTM to classify opposing emotions. We separate the speech by speaker gender to investigate the relationship between gender and emotional content of speech.
- There are a variety of temporal and spectral features that can be extracted from human speech. We use statistics relating to the pitch, Mel Frequency Cepstral Coefficients (MFCCs) and Formants of speech as inputs to classification algorithms. The emotion recognition accuracy of these experiments allow us to explain which features carry the most emotional information and why.

- It also allows us to develop criteria to class emotions together. Using these techniques we are able to achieve high emotion recognition accuracy.

1.3 Literature Survey

Considering our project, we started off with finding the already done researches and successful projects and papers over the internet and on the IEEE official website. We did find out the surveys made by Maisy Wieman, Andy Sun. Their research paper specifies the correct and on point information about the vocal pattern analysis to detect the emotions. We also tried to find out the algorithms and methods used to determine the features from the vocal pattern. We were able to find out about the algorithm called as MFCC (, Mel Frequency Cepstral Coefficients).

1.4 Motivation

We got this idea when we were doing our hobby project 'Emotion Detection using Iamge Processing'. We got about 68% accuracy and we were trying to come with a good idea for increasing the accuracy. We tried increasing features and increasing quality of dataset, but we couldn't improve it to a great extent. Hence we come up with the idea that if we integrate the speech features and image features can we have a better accuracy for the model.

We searched over the internet and we found that this area is currently under research and nt so much work is done in this area. We also did watch Google's Duplex call video and were so impressed by that it inspired us to do this project.

1.5 Aim of the Project

To determine and classify the Emotions from the Speech/Vocal signals of human.

1.6 Scope and Objectives

To determine the emotions from the speech so as to give the machine a better approach to have a good conversation with humans.

1.7 Technical Approach

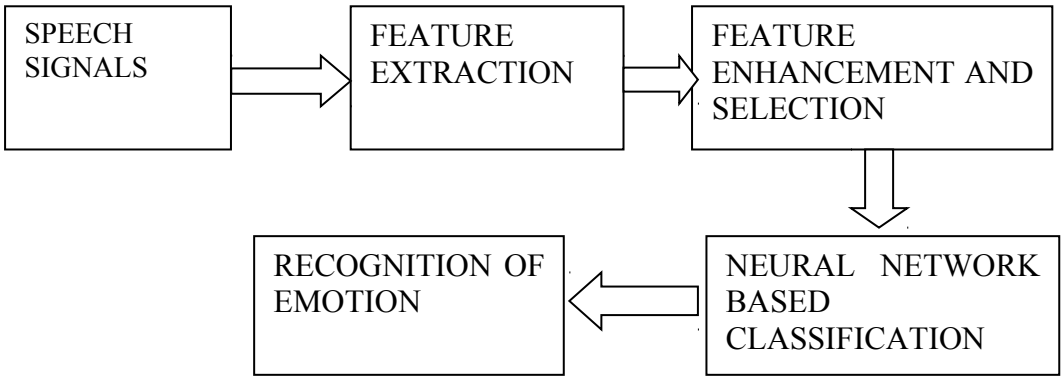


Fig 1: Steps for detecting Emotions from Speech data

CHAPTER 2

Theoretical Description of Project

2.1 Speech Signal Dataset:

We referred two dataset RAVDESS and SAVEE Dataset. Only took the audio data from this datasets.

The RAVDESS database is gender balanced consisting of 24 professional actors. Speech part of dataset includes calm, happy, sad, angry, fearful, surprise, and disgust expressions and song part of dataset contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. We got 2000 audio samples which were in the wav format.

The SAVEE dataset consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. 'DC', 'JE', 'JK' and 'KL' are the four male speakers recorded in the SAVEE database. There are 15 sentences for each of the 7 emotion categories, and one file for each sentence. It includes 'anger', 'disgust', 'fear', 'happiness', 'neutral', 'sadness' and 'surprise'.

We made customized dataset by using this two dataset. We have less samples of ‘calm’ emotion. Hence we eliminate ‘clam’ samples to balance the dataset. Our dataset contain 7 folders, each represents the different emotion. Contain Separate emotion's voice/speech in each separate folder.

anger : 436 samples
disgust : 252 samples
fear : 436 samples
happy : 436 samples
neutral : 308 samples
sad : 436 samples
surprise : 252 samples

Total : 2556 Samples

We tested out one of the audio file to know its features by plotting its waveform and spectrogram.

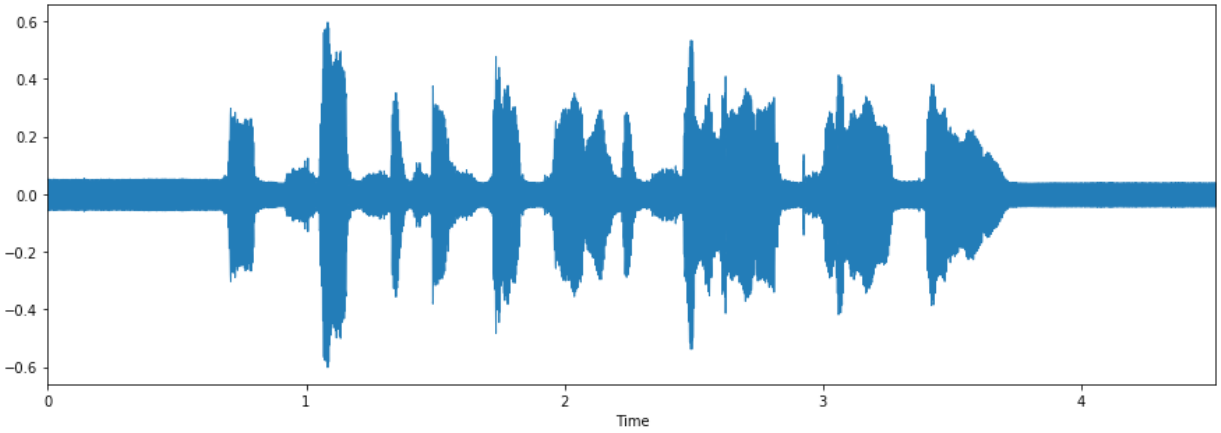


Fig 2: Time Domain Plot of the Speech signal

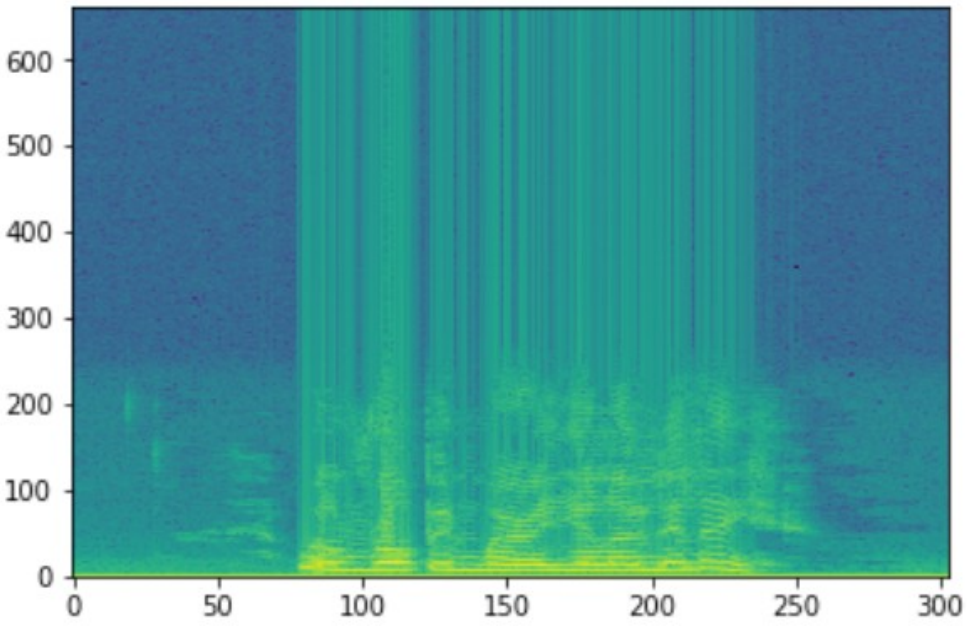


Fig 3: Frequency Domain Plot of the Speech signal

The next step involves organizing the audio files. Each audio file has a unique identifier which tells the emotion of the file which can be used to determine the label of the audio file. We have 7 different emotions in our dataset.

We used Librosa library in Python to process and extract features from the audio files. Librosa is a python package for music and audio analysis. It provides the building

blocks necessary to create music information retrieval systems. Using the librosa library we were able to extract features i.e MFCC(Mel Frequency Cepstral Coefficient). MFCCs are a feature widely used in automatic speech and speaker recognition. We also separated out the females and males voice by the using the identifiers provided in the website. This was because as experiment we found out that separating male and female voices increased by 15%. It could be because of the pitch of the voice was affecting the results.

Each audio file gave us many features which were basically array of many values. These features were then appended by the labels which we created in the previous step.

2.2 Feature Extraction:

The time-domain representation of sound is very complex, and in its original form, it does not provide very good insight into key characteristics of the signal. Because of this characteristic of sound signals, we map this time domain representation into more telling features. The most straightforward technique involves determining the average energy of the signal. This metric, along with total energy in the signal, indicates the “volume” of the speaker. Duration also offers insights into emotion, as do statistics like the maximum, minimum, range, mean, and standard deviation of both the signal and spectrum. These may indicate fluctuations in the volume or pitch that can be useful in determining emotion. For both the signal and spectrum, we also derive skewness, the measure of departure of horizontal symmetry in the signal, and kurtosis, the measure of height and sharpness of central peak, relative to a standard bell curve.

We also process the signal in the frequency domain through the (Fast) Fourier Transform. We use windowed samples to get accurate representations of the frequency content of the signal at different points in time. By taking the square value of the signal at each window sample, we can derive the power spectrum. We use the values of the power spectrum as features, but we also find the frequencies that have the greatest power. We obtain the three largest frequency peaks for each window and add those to the feature vector. In addition, we find the maximum and minimum frequencies with substantial power for each time frame, and use these values to determine the frequency range for each frame. The auditory spectrum can be derived by mapping the power spectrum to an

auditory frequency axis by combining the Fast Fourier Transform bins into equally spaced intervals.

The Mel-frequency Cepstrum captures characteristics of the frequency of the signal represented on the Mel scale, which closely aligns to the nonlinear nature of human hearing. By extension, the Mel-frequency Cepstrum Coefficients (MFCC) represent the "spectrum of the spectrum." MFCC's can be derived by mapping the powers of the frequency spectrum onto the mel scale, and then by taking the log of these

powers, followed by the discrete cosine transform. MFCC's are commonly used as features in many speech recognition applications.

Changes in pitch over time are measured on both a coarse time scale and a fine time scale. For coarse measurement, the signal is divided into 3 parts (beginning, middle, and end), and the part of the signal with the highest average pitch is used to determine whether the pitch rises or falls over time. For fine measurement, the dominant frequency of each windowed sample is compared to the dominant frequencies of the windowed samples immediately preceding and following. This difference is recorded in the feature vector.

2.3Feature Selection:

After we processed the original sound signal to extract features, the high variance of our algorithm revealed that we needed to filter the many features to determine which contribute most to the classifier. Our input speech signals were windowed, with approximately 72 windows per audio sample, and each of these windowed samples provided a total of 577 features. In total, we extracted 41,558 features. This large number of features (much larger than the number of examples) resulted in a very high variance. Clearly, we needed to extract the most important features. Because of the large number of features, we used heuristics to score each feature, rather than implement a brute force forward or backward search.

CHAPTER 3

CNN AND LSTM

We need three basic components to define a basic convolutional network.

1. The convolutional layer
2. The Pooling layer[optional]
3. The output layer

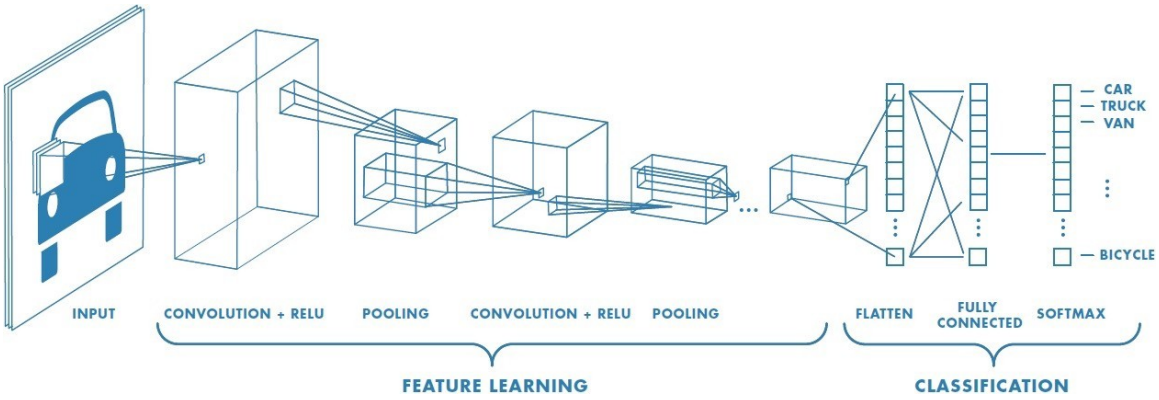


Fig 4: General Representation of CNN model

i. We pass an input image/audio to the first convolutional layer. The convoluted output is obtained as an activation map. The filters applied in the convolution layer extract relevant features from the input image to pass further.

ii. Each filter shall give a different feature to aid the correct class prediction. In case we need to retain the size of the image, we use same padding(zero padding), other wise valid padding is used since it helps to reduce the number of features.

iii. Pooling layers are then added to further reduce the number of parameters

iv. Several convolution and pooling layers are added before the prediction is made. Convolutional layer help in extracting features. As we go deeper in the network more specific features are extracted as compared to a shallow network where the features extracted are more generic.

v. The output layer in a CNN as mentioned previously is a fully connected layer, where the input from the other layers is flattened and sent so as the transform the output into the number of classes as desired by the network.

vi. The output is then generated through the output layer and is compared to the output layer for error generation. A loss function is defined in the fully connected output layer to compute the mean square loss. The gradient of error is then calculated.

vii. The error is then backpropagated to update the filter(weights) and bias values.

viii. One training cycle is completed in a single forward and backward pass.

LSTM NETWORKS:

Long Short Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter & Schmidhuber (1997), and were refined and popularized by many people in following work.[1] They work tremendously well on a large variety of problems, and are now widely used.

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

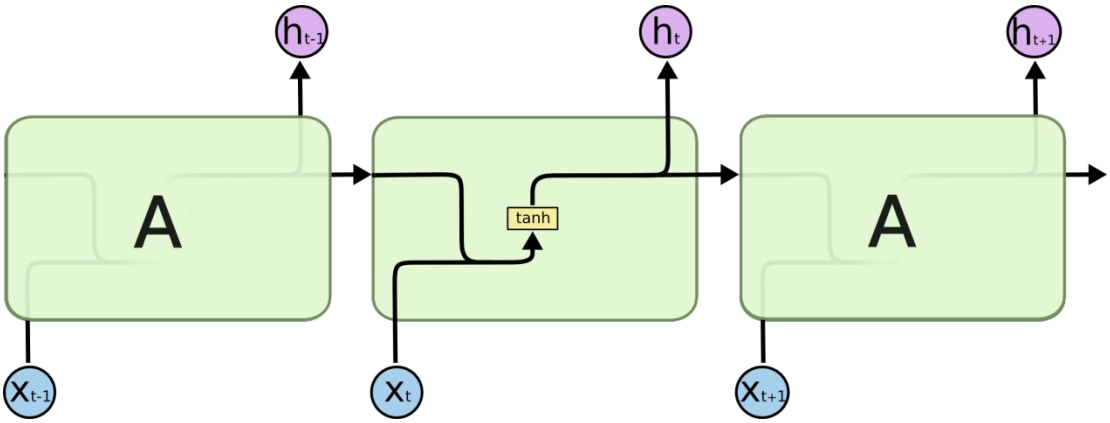


Fig 5 :The repeating module in a standard RNN contains a single layer.

LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.

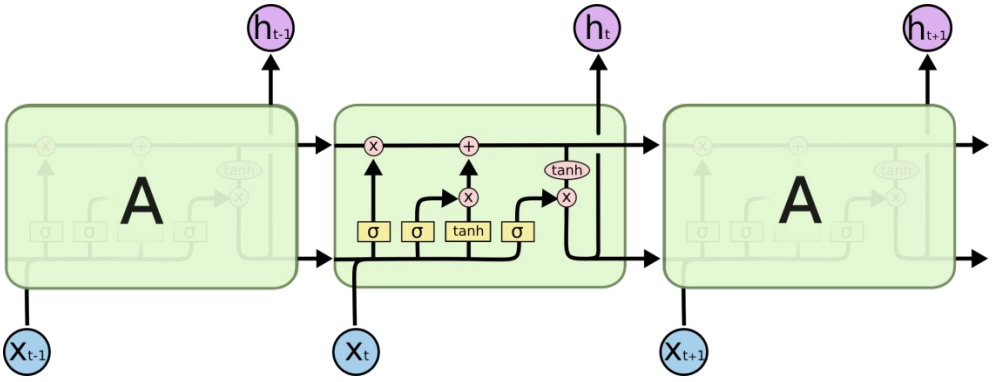
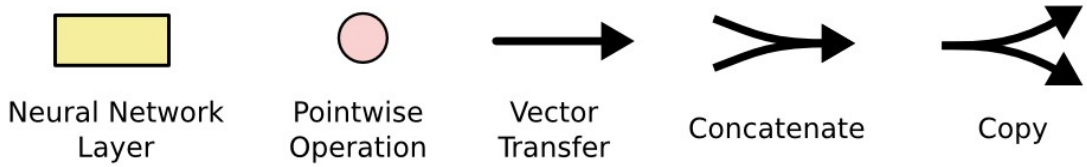


Fig 6: The repeating module in an LSTM contains four interacting layers.

Don’t worry about the details of what’s going on. We’ll walk through the LSTM diagram step by step later. For now, let’s just try to get comfortable with the notation we’ll be using.



In the above diagram, each line carries an entire vector, from the output of one node to the inputs of others. The pink circles represent pointwise operations, like vector addition, while the yellow boxes are learned neural network layers. Lines merging denote concatenation, while a line forking denote its content being copied and the copies going to different locations.

CHAPTER 4

Results and Discussion

. The below figure shows the training and testing loss on our dataset. As we can see from the graph that both training and testing errors reduces as number of epochs to the training model increases.

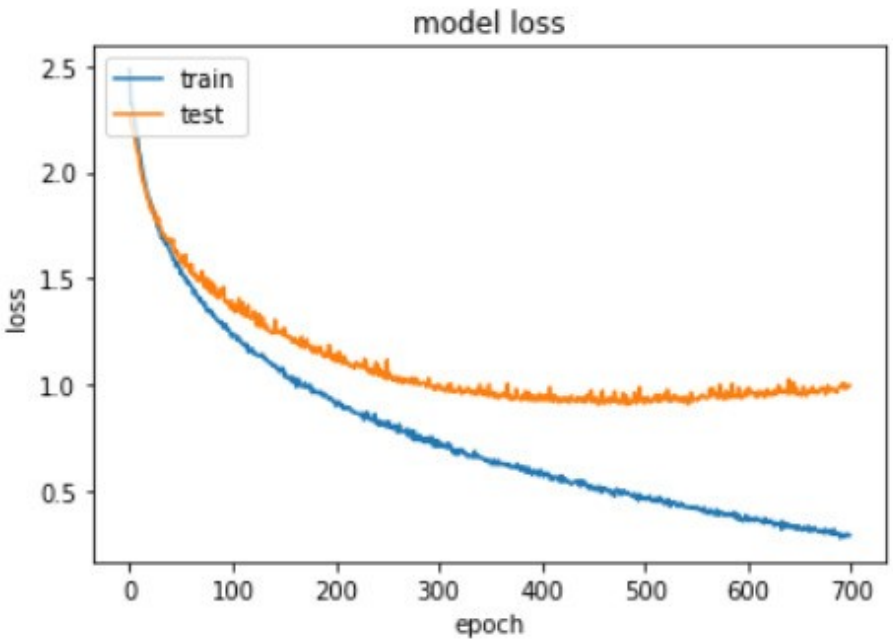


Fig.7: Result Training V/s Testing accuracy

We got the accuracy of the training model about 92% but the accuracy of cross validation about 47%. Our dataset has overfat the mode. Dataset could be modified to get better accuracy.

CHAPTER 5

Conclusions

Hence our project presents a new way to give the ability to machine to determine the emotion with the help of the human voice. It will give the machine the ability to have a

better approach towards having a better conversation and seamless conversation like human does.

CHAPTER 6

Future Scope

Our project aims to determine the emotion with the speech of a human. Our project can be extended to integrate with the robot to help it to have a better understanding of the mood the corresponding human is in, which will help it to have a better conversation.

Any e-commerce sites which have an AI based chat bot which recommends the customer to have a good experience our project will determine the customer/s mood and accordingly it can help the chat bot to have a better recommendation.

References

Research Papers and Materials Referred:

[1]Liberman, Mark, et al. Emotional Prosody Speech and Transcripts LDC2002S28. Web Download. Philadelphia: Linguistic Data Consortium, 2002.

[2]IEEE Research Paper

https://ieeexplore.ieee.org/document/7344793/?reload=true

[3]M.M.H.E. Ayadi, M.S. Kamel, F. Karray, "Survey on speech emotion recognition:
Features classification schemes and databases", *Pattern Recognition*, pp. 572-587, 2011.

[4]Stanford Project Paper
http://cs229.stanford.edu/proj2007/ShahHewlett%20-%20Emotion%20Detection%20from%20Speech.pdf