PROJECT BASED REPORT

ON

SPEAKER RECOGNITION

submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

In

ELECTRICAL ENGINEERING

By

D.S. Sai Pranav (EE18B008)

Ch. Tharun (EE18B007)

Under the Esteemed Guidance of

Dr. POOJA VYAVAHARE *Ph.D*

Assistant Professor

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY TIRUPATI

(2020)

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that,we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Ch. Tharun (EE18B007)

D.S Sai Pranav (EE18B008)

BONAFIDE CERTIFICATE

This is to certify that the project based report entitled " **SPEAKER RECOGNITION**" submitted by **Ch. Tharun (EE18B007), D.S Sai Pranav (EE18B008),** to the Indian Institute of Technology, Tirupati, in partial fulfilment of the requirements for the award of the degree of the **BACHELOR OF TECHNOLOGY** in Department of Electrical Engineering, is a bonafide work done by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. POOJA VYAVAHARE
Assistant Proffesor
Department of Electrical Engineering
IIT Tirupati- 517501

ACKNOWLEDGEMENTS

The Success in this project would not have been possible without the timely help and guidance rendered by many people. Our sincere thanks to all those who has assisted us in one way or the other for the completion of our project.

Our greatest appreciation to our guide **Dr. Pooja Vyavahare,** Assistant Professor, Department of Electrical Engineering and Ms scholar, **Mr. Santhan Reddy,** which cannot be expressed in words for their tremendous support, encouragement and guidance for this project.

We thank all the members of teaching and non- teaching staff members, and also who have assisted us directly or indirectly for the completion of this project.

Finally, we sincerely thank our friends and classmates for their kind help and co operation during our work.

Ch. Tharun (EE18B007)

D.S Sai Pranav (EE18B008)

ABSTRACT

Speaker recognition is the process of automatically recognizing who is speaking using speaker-specific information in speech waves [1–3]. Many applications have been considered for speaker recognition. These include secure access control by voice, customizing services or information to individuals by voice, indexing or labelling speakers in recorded conversations or dialogues, surveillance, and criminal and forensic investigations involving recorded voice samples. Currently, the most frequently mentioned application is access control, which includes voice dialing, banking transactions over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, and remote access to computers. Speaker recognition technology, as such, is expected to create new services in smart environments and make daily life more convenient.

Speaker diarization, in which an input audio channel is automatically annotated with speakers, has been actively investigated. It is useful in speech recognition, facilitating the searching and indexing of audio archives, and increasing the richness of automatic transcriptions, making them more readable. Another important application of speaker recognition technology is as a forensics tool.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

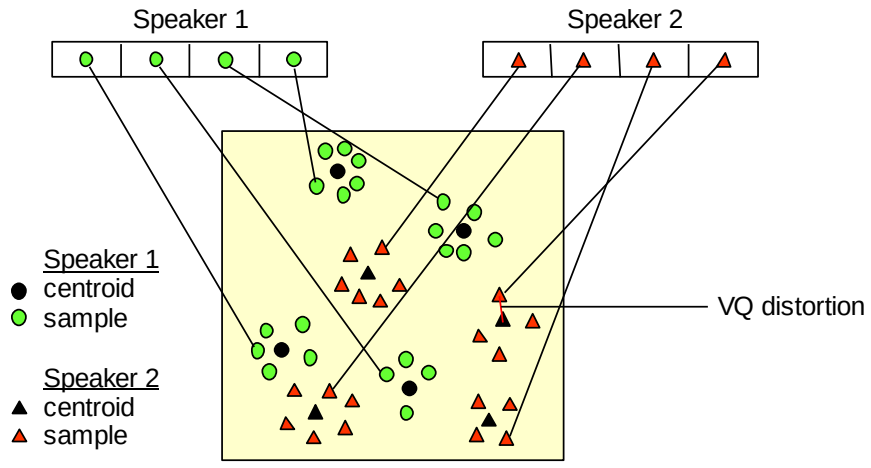
Voice can combine what people say and how they say it by two-factor authentication in a single action. Other identifications like fingerprints, handwriting, iris, retina, face scans can also help in biometrics but voice identification is needed as an authentication that is both secure and unique. Voice can combine two factors, namely, personal voice recognition and telephone recognition. Voice recognition systems are cheap and easily understood by users. In today's smart world, voice recognition plays a very critical role in many aspects. Voice based banking, home automation and voice recognition based gadgets are some of the many applications of voice recognition. The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called *pattern recognition*. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called *patterns* and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as *feature matching*.

Furthermore, if there exists some set of patterns that the individual classes of which are already known, then one has a problem in *supervised pattern recognition*. These patterns comprise the *training set* and are used to derive a classification algorithm. The remaining patterns are then used to test the classification algorithm; these patterns are collectively referred to as the *test set*. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the

algorithm.

The state-of-the-art in feature matching techniques used in speaker recognition include Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). In this project, the VQ approach will be used, due to ease of implementation and high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a *cluster* and can be represented by its center called a *codeword*. The collection of all codewords is called a *codebook*.

The belowfigure shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training phase, using the clustering algorithm described in Section 4.2, a *speaker-specific* VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result codewords (centroids) are shown in Figure 5 by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the *total VQ distortion* is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified as the speaker of the input utterance.



CHAPTER 2

HISTORY

You command Siri to search for nearest pizza delivery joints and you are presented with a list of them. As can be deduced, voice recognition has come a long way; however, this technology is not something recent. In fact, it has its roots back in the 1950s. Let's delve into the past and take a look at the brief history of how voice recognition technology has evolved over time into the speech recognition software we know today.

In the 1980's

A major breakthrough was the development of the hidden Markov model which used statistics to determine the probability of a word originating from an unknown sound. It did not rely on speech patterns or fixed templates. Many of these programs made their way into industries and business applications.A doll was also made for children in 1987; it was known as 'Julie' and it could be trained by children to respond to their speech. But speech recognition systems of the 80s had one flaw: you had to take a break between each spoken word.

In the 1990's

With the introduction of faster microprocessors, speech software became feasible. In 1990, the company Dragon released 'Dragon Dictate' which was the world's first speech recognition software for consumers. In 1997, they improved it and developed 'Dragon NaturallySpeaking'; you could speak 100 words in a minute.
In 1996, the first voice activated portal (VAL) was made by BellSouth. However, this system is inaccurate and still is a nuisance for many people.

In the 2000's

By 2001, speech recognition technology development had hit a plateau, until Google came along. Google invented an application called 'Google Voice Search' for iPhones which utilized data centers to compute the enormous amount of data analysis needed for matching user queries with actual examples of human speech.
In 2010, Google introduced personalized recognition on Android devices which would record different users' voice queries to develop an enhanced speech model. It consists of 230 billion English words. Eventually, Apple's Siri was invented which relied on cloud computing as well, and you have a personal assistant who is not only intelligent, but funny and witty too.

CHAPTER 3

APPLICATIONS

In the last few years use of biometric system has become a reality. There are lots of commercial as well as personal applications where biometric is used for security purpose. Speaker verification has gained a huge acceptance in both government and financial sectors for secure authentication .Australian Government organization Centrelink, use speaker verification for authentication of recipients using telephone transactions. Possible applications of speaker recognition are forensic investigation, telephone banking, access control, user authentication etc. . Speaker recognition have more potential than other biometrics such as face recognition, finger prints, and retina scans. The main advantage of speaker recognition over other biometric is low cost, high acceptance and non-invasive character of speech acquisition. To develop a speaker recognition system, expensive equipment as well as direct participation of speakers is not required. Speaker recognition have potential to eliminate the need of carrying debit card, credit card, remembering password for bank account or any other security locks and many other online services .With the continuous improvement in reliability of speaker recognition technology, its usability has increased. Now days, use of speaker recognition has become a commercial reality and part of consumer's everyday life

The performance of speaker recognition system is vulnerable to change in speaker characteristics such as age, health problems, speaking environment etc. Another disadvantage is that it is possible to play a recorded voice instead of the actual voice of a speaker.

- **Access control**: controlling access to computer networks
- **Transaction verification**: for telephone banking and account access control
- **Speech data organization**: used for voice mail
  Eg: speech skimming or audio mining applications
- **Personalization**: Device customization, store and fetch personal information based on the user verification.

CHAPTER 4

IMPLEMENTATION

In this project we will experiment with the building and testing of an automatic speaker recognition system. In order to build such a system, one have to go through the steps that were described in previous sections. We create two utility functions euc_dist and codebooks and two main functions: training and testing.

Our goal is to train a voice model (or more specific, a VQ codebook in the MFCC vector space) for each speaker S1 – S3 using the corresponding sound file in the train folder. After this training step, the system would have knowledge of the voice characteristic of each (known) speaker. Next, in the testing phase, the system will be able to identify the (assumed unknown) speaker of each sound file in the test folder

First we create a mfcc function. We take the input and do pre emphasis to it. Now cut the speech signal (a vector) into frames with overlap (refer to the frame section in the theory part). The result is a matrix where each column is a frame of *N* samples from original speech signal. Applying the steps "Windowing" and "FFT" to transform the signal into the frequency domain. This process is used in many different applications and is referred in literature as Windowed Fourier Transform (WFT) or Short-Time Fourier Transform (STFT).

The result of the last section is that we transform speech signals into vectors in an acoustic space. In this section, we will apply the VQ-based pattern recognition technique to build speaker reference models from those vectors in the training phase and then can identify any sequences of acoustic vectors uttered by unknown speakers. the utility function euc_distis used to compute the pairwise Euclidean distances between the codewords and training vectors in the iterative process.

Now is the final part! Using the training and testing functions we perform our Speaker Recognition.

CHAPTER 5

ALGORITHM EXPLANATION

The acoustic vectors extracted from input speech of each speaker provide a set of training vectors for that speaker. As described above, the next important step is to build a speaker-specific VQ codebook for each speaker using those training vectors. There is a well know algorithm, namely **LBG algorithm [Linde, Buzo and Gray],** for clustering a set of *L* training vectors into a set of *M* codebook vectors. The algorithm is formally implemented by the following recursive procedure:

1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).
2. Double the size of the codebook by splitting each current codebook $\mathbf{y}_n$ according to the rule
$$\mathbf{y}_n^+ = \mathbf{y}_n(1+\varepsilon)$$
$$\mathbf{y}_n^- = \mathbf{y}_n(1-\varepsilon)$$
where *n* varies from 1 to the current size of the codebook, and ε is a splitting parameter (we choose ε=0.01).

3. Nearest-Neighbor Search: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).

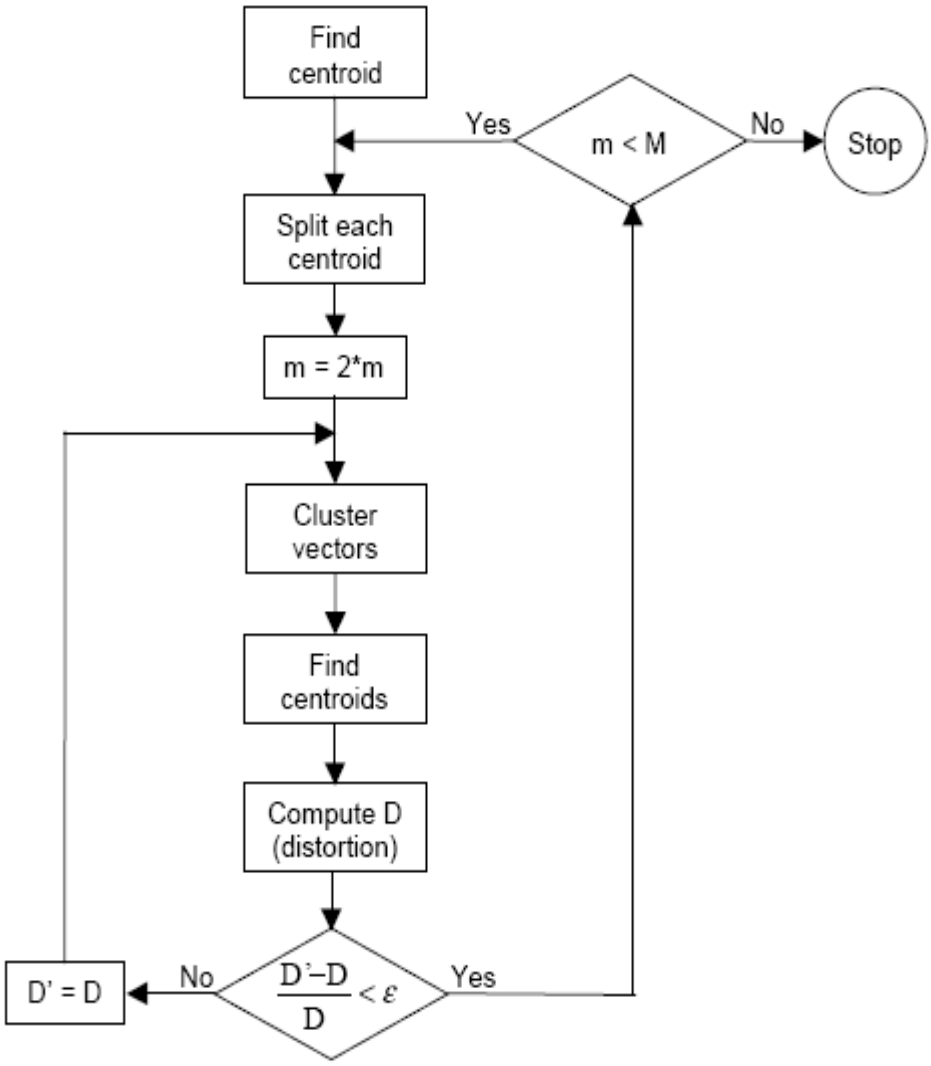
4. Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.

5. Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold

6. Iteration 2: repeat steps 2, 3 and 4 until a codebook size of *M* is designed.

Intuitively, the LBG algorithm designs an *M*-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the codewords to initialize the search for a 2-vector codebook, and continues the splitting process until the desired *M*-vector codebook is obtained.

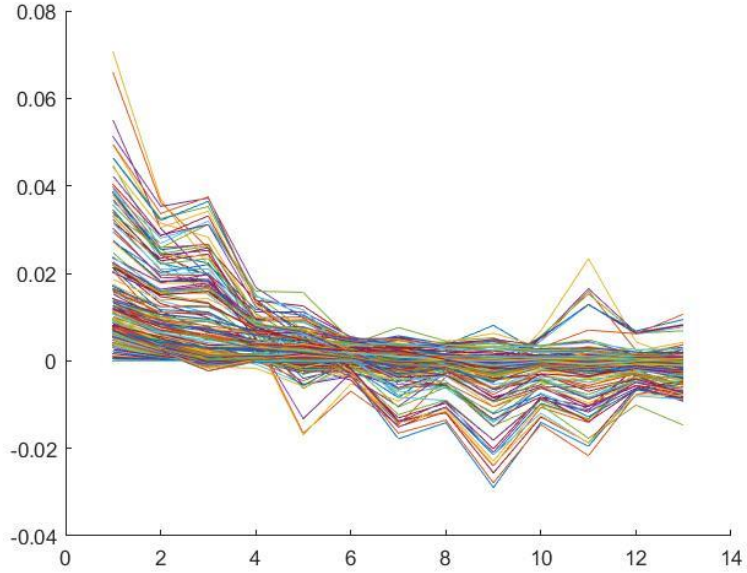
In a flow diagram, the detailed steps of the LBG algorithm. "*Cluster vectors*" is the nearest-neighbor search procedure which assigns each training vector to a cluster associated with the closest codeword. "*Find centroids*" is the centroid update procedure. "*Compute D (distortion)*" sums the distances of all training vectors in the nearest-neighbor search so as to determine whether the procedure has converged.

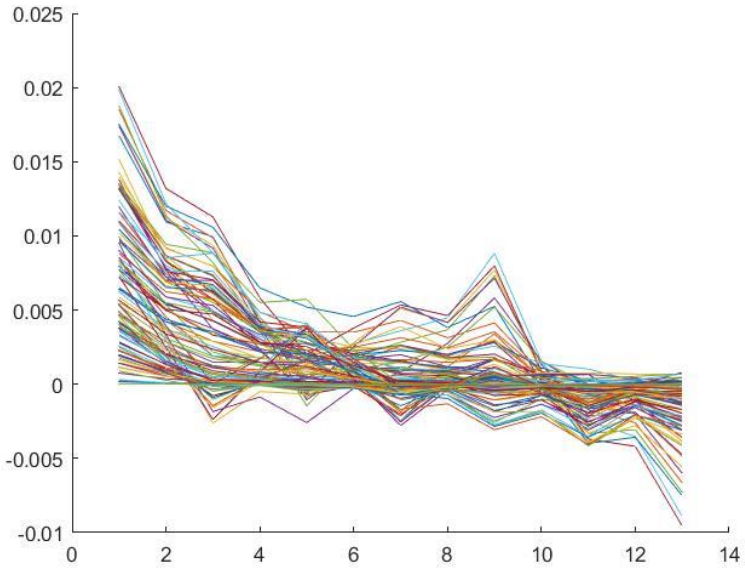


CHAPTER 6

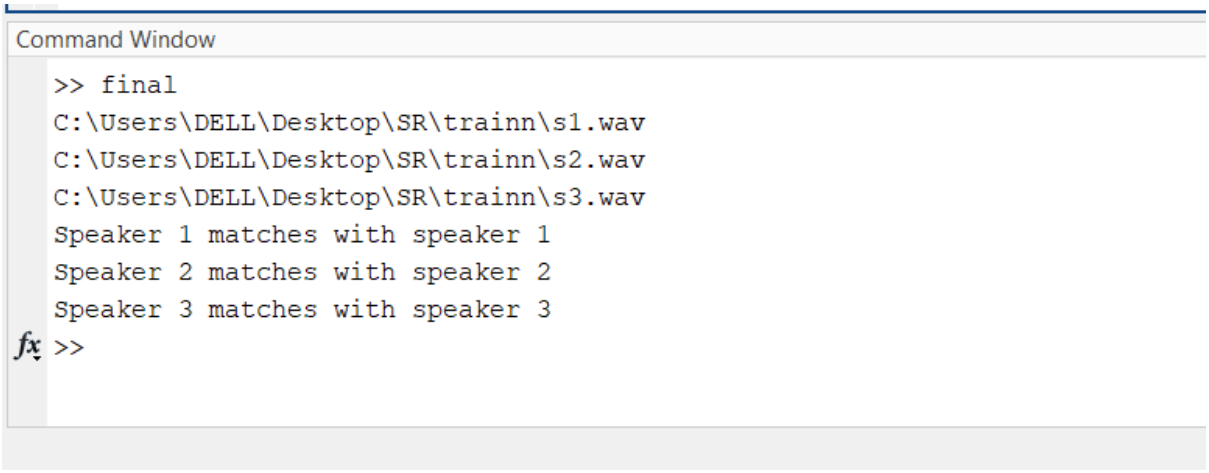
RESULTS

MFCC's for two different audio waves obtained using our code:





The output showing the success of Speaker Recognition:



CHAPTER 7

CONCLUSION AND FUTURE SCOPE

This report provides concise definition and discussion about speaker recognition technology. In addition, basic concepts of automatic speaker recognition systems, clustering technique etc. has been discussed. Speaker recognition is method of designing a system for identity of an individual through their voices. Speaker recognition has a significant prospective as it is appropriate biometric technique for security. The speaker recognition task is normally achieved by acquiring speech signal, feature extraction, pattern matching and obtaining match score.

CONTRIBUTIONS:

D.S.Sai Pranav (EE18B008) – MFCC code, Testing code , Report

Ch.Tharun (EE18B007) – Training code , codebooks code, ppt

REFERENCES:

Concepts:
http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/#steps-at-a-glance
ijirae.com/volumes/vol1/issue10/27.NVEC10086.pdf
http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/speaker_recognition.html

MATLAB:
https://in.mathworks.com/help/matlab/