Speaker Recognition

Project Report submitted in partial fulfillment of the requirement
for degree of Bachelor of Technology

in

Electronics and Communication Engineering
under the Supervision of
Prof. Dr. Sunil V. Bhooshan

by

Ankit Gupta (081011)
Aparna Gautam (081085)
Puneet Keshav Earan (081019)

to

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
Waknaghat

Certificate

This is to certify that project report entitled Speaker Recognition, submitted by Ankit Gupta, Aparna Gautam and Puneet Keshav Earan in partial fulfillment for the award of degree of Bachelor of Technology in Electronics and Communication Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Prof. S. V. Bhooshan:____________
Head of Department
Department of Electronics and Communication

Date: 31-05-2012

"The learned is beautifully equipped to rule the world which no longer exists,

the learner is the one who rules the earth"

Acknowledgements

As we express our gratitude, we must never forget that the highest appreciation is not to utter words, but to live by them.

We express our sincere gratitude to Prof. Dr. S. V. Bhooshan, our project guide, under whose able guidance we were able to complete this project. He has been a constant source of encouragement and has inspired us to look into novice methods of research.
We would also like to thank Dr. Vinay Kumar for providing us an opportunity to work on Speaker Verification. His constant support throughout has helped us overcome various obstacles we encountered during the project.
We are very grateful to our Dean, Prof. Dr. T.S. Lamba for his valuable suggestions during the course of this project which helped us to a great extent in achieving the desired results.
To the countless people who helped us sail through this project, we will always be indebted.

Gupta Gautam PKeshav

Ankit Gupta Aparna Gautam Puneet Keshav Earan

Abstract

In todays world, security is of utmost importance and speaker recognition has gained a lot of importance in the present scenario. Our project is an attempt to carry forward the research in this field of prime importance and concern. Speaker recognition is the name given to the process of recognizing the person who is speaking based on the information included in individual speech signals. Hence it is a process of authenticating speaker identity. By virtue of this technique speaker identity is verified using the speakers voice. A speaker recognition system enables access controls of various services by voice. These systems can operate in two modes: to identify a particular person or to verify a persons claimed identity. Quite a few fields can benefit from the use of speaker recognition models. These include voice mail services, remote access to computers, telephone shopping etc. in addition it can also be employed as a forensic tool. The speaker recognition process is executed in three main phases. The first phase is the feature extraction which is done using LPC, Cepstrum methods etc. Second phase is the pattern matching in which clustering is also performed and this phase is followed by the third phase which is decision making. In our project we have used Mel Frequency Cepstrum Coefficients for feature extraction, K means method for clustering and finally the Euclidean distance for the objective of decision making. Also, since we are storing the data in separate files when the user records the training voices, our system at the time of testing turns out to be very fast. The basic objective of this project has been to understand the intricacies of the speaker recognition process, to study various methods of extracting features from voice and to finally use them successfully and develop a simple yet complete and comprehensive speaker recognition model tested on a speech database. The entire project has been realized using Matlab.

Contents

List of Figures

Chapter 1

Introduction

Speech maybe defined as the expression of or the ability to express thoughts and feelings by articulate sounds. The fundamental function of speech is communication. Moreover it may be defined as the acoustic signal used for language. Speech is and perhaps will always be the most desired means of communication between humans.

Speech production or spoken language consists of three main levels of processing. These are conceptualization, formulation and articulation. Human hearing is also adapted to speech. Humans are extremely sensitive to sounds between 1kHz and 4kHz, correspondingly this range is very important for speech. It is the approximate range of the important resonance frequencies of the human vocal tract. These resonance frequencies or formants as they are called determine the acoustic character of a speech sound. With the passage of time speech has considerably evolved and with it has evolved speech processing. Speech processing has traditionally been focuses on some problem areas which are, most of the time, overlapping. Speech processing dates back to as early as the 1960s when the first touch tone telephone was demonstrated. After that considerable work was done in the field of speech coding and compression. By the 1980s a number of speech coders had been devised. Progress in speech coding has enabled the recent adoption of low-rate algorithms for mobile telephony.

In speech the fundamental analog form is the acoustic waveform which is called as the speech signal. These signals can be converted to electrical form, processed using analog and digital signal processing, and further converted back to the acoustic form. This type of processing was first used in telephones and is still used in case of transmitting and manipulating speech signals. Speech is researched in terms of the speech production andspeech perceptionof thesoundsused inspoken language.

With the advancement of technology and availability of newer tools, speech can be analyzed more effectively and can be better utilized for betterment of humanity. Early in

1958 Dudley made a classifier that continuously evaluated spectra rather than approximations to formants. In the 1960s spectral bank estimation techniques were developed which proved to be of great significance for recognition. The 1980s focused on scaling the existing techniques (LPC, HMM etc) to tackle newer and more difficult problems. During the past two years too, considerable work has been done in this field with further evolution of techniques like Markov models, Linear predictive coding cepstrum and mel cepstrum coefficients.

Though a lot of progress has been made in the field of speech processing, specifically speaker and speech recognition, still it remains an uphill task due to various constraints. Since natural speech is continuous, without pauses, it makes identification of boundaries extremely tough to be gauged. The speech and therefore the speech spectrum may change drastically if the pronunciation or the phonemes are changed. Large or very limited vocabularies may pose another challenge to speech processing. All these factors can alter the characteristics of speech which humans can quite often compensate but is tough to be compensated by speech models.

Hence, this field provides a lot of avenues for research and is open for considerable improvement. It holds great promise for the future considering the speed at which new technologies are coming up in this research area. This project too is an endeavour to try and understand the nuances of this area of work and thereby contribute in a small but effective way towards newer developments in speaker recognition modeling techniques.

Chapter 2

Basics of speech

Spoken language is the best and most natural way used by humans to communicate information. A speech signal may provide insight into a lot many types of information. From the perspective of speech production, the speech signal conveys linguistic (e.g., message and language) and speaker information (e.g. emotional, regional, and physiological characteristics). From the speech perception point of view, it also conveys information about the environment in which the speech was produced and transmitted. Humans can easily decode most of the information, even though a lot of information is encoded in a complex form in the speech signal. Many researchers have been inspired by this exceptional ability to understand speech production and perception for developing systems that automatically extract information in speech. Speech technology has found wide applications such as automatic dictation, voice command control, audio archive indexing and retrieval etc. Lungs along with the diaphragm serve as the main energy source. The glottis, is the most important sound source in the vocal system. While

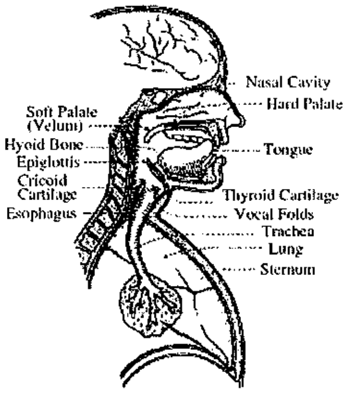


FIGURE 2.1: Human Speech Production, Speaker Recognition, Joseph P. Campbell, Jr., Department of Defense, Fort Meade, MD, j.campbell@ieee.org

speaking, the air flow is forced through the glottis between the vocal cords and the larynx to the three main cavities of the vocal tract, the pharynx and the oral and nasal cavities. The air flow from the oral cavity exits through the mouth and that through the nasal cavity, exits through the nose. From the technical point of view, the vocal system may be considered as a single acoustic tube between the glottis and mouth.
Production of speech: in producing speech, the air flow from lungs first passes the glottis followed by the throat and the mouth. Depending on which speech sound one articulates, excitation of speech signal is possible in three ways:

- **Voiced excitation:** The glottis is closed. The air pressure forces the glottis to open and close periodically thus generating a periodic pulse train (triangleshaped). This fundamental frequency usually lies in the range from 80Hz to 350Hz.
- **Unvoiced excitation:** The glottis is open and the air passes a narrow passage in the throat or mouth. This results in a turbulence which generates a noise signal. The spectral shape of the noise is determined by the location of the narrowness.
- **Transient excitation:** A closure in the throat or mouth will raise the air pressure. By suddenly opening the closure the air pressure drops down immediately. (plosive burst)

2.1 Major elements of speech production

Four major elements that contribute to speech production are:

2.1.1 Initiation

It is a function of the airstream mechanism and the direction of airflow. The airstream may be either pulmonic, glottalic, or velaric. The pulmonic airstream is initiated from the lungs. When the glottis is closed, the glottalic airstream is initiated by the vertical motion of the larynx. This produces voiceless sounds. When the tongue initiates the air pressure differential in an air-filled cavity, the airstream is called velaric.

2.1.2 Phonation

The energy that is generated by the vocal folds at the larynx is dealt with by Phonation. Unvoiced ,voiced and whisper are the three kinds of phonation.

2.1.2.1 Voiced speech

Voiced speech consists of more or less constant frequency tones of some duration (e.g. prolonged aa sound vanilla, prolonged oo sound in book). It is produced when periodic pulses of air generated by the vibrating glottis resonate through the vocal tract, at frequencies dependent on the vocal tract shape. Because of its periodic nature, voiced speech can be identified and extracted. Majority of voiced sounds are generated through normal voiced phonation which is basically when the vocal folds are vibrating at a periodic rate and generate certain resonance in the upper chamber of the vocal tract. There are other kinds of phonations too like laryngealization and falsetto depending on the vocal tract shape.

2.1.2.2 Unvoiced Speech

Unvoiced speech is non- periodic, random-like sounds, caused by air passing through a narrow constriction of the vocal tract as (e.g. f-five, p-please). Unvoiced phonation may be either in the form of nil phonation which corresponds to zero energy or breath phonation which is based on relaxed vocal folds passing a turbulent air stream.

2.1.2.3 Whispered phonation

It happens when the speaker acts like generating a voiced phonation but the vocal folds in this are made more relaxed so that a greater flow of air can pass through and generate a more turbulent airstream. However, the vocal folds are not relaxed enough to generate an unvoiced phonation.

2.1.3 Articulation

The manner of articulation, which is mostly used for the production of consonants basically describes how the speech organs are involved in making a sound and it also alters the resonant properties of the vocal tract. This in turn brings about a change in the formants which are crucial in the identification of vowels. The different manners of articulation include:

- **Stops** are also known as plosives. The air is blocked for a moment, and then released. In English, they are p, b, t, d, k, and g. Stops maybe further classified as oral or nasal.

- **Fricatives** involve a slightly resisted flow of air. In English, these include f, v, th, dh, s, z, sh, zh, and h. Normal fricatives possess a higher air flow so that the pitch remains in the audible range. An Example is /f/ in English whereas sibilant fricatives are exceptionally high pitched sounds such as /s/ and /S/ in English.

- **Resonants:** Resonant flows are those which are produced by the passing of the air stream through a tight opening producing vocal harmonics. These are categorized into centrally and laterally resonant sounds.

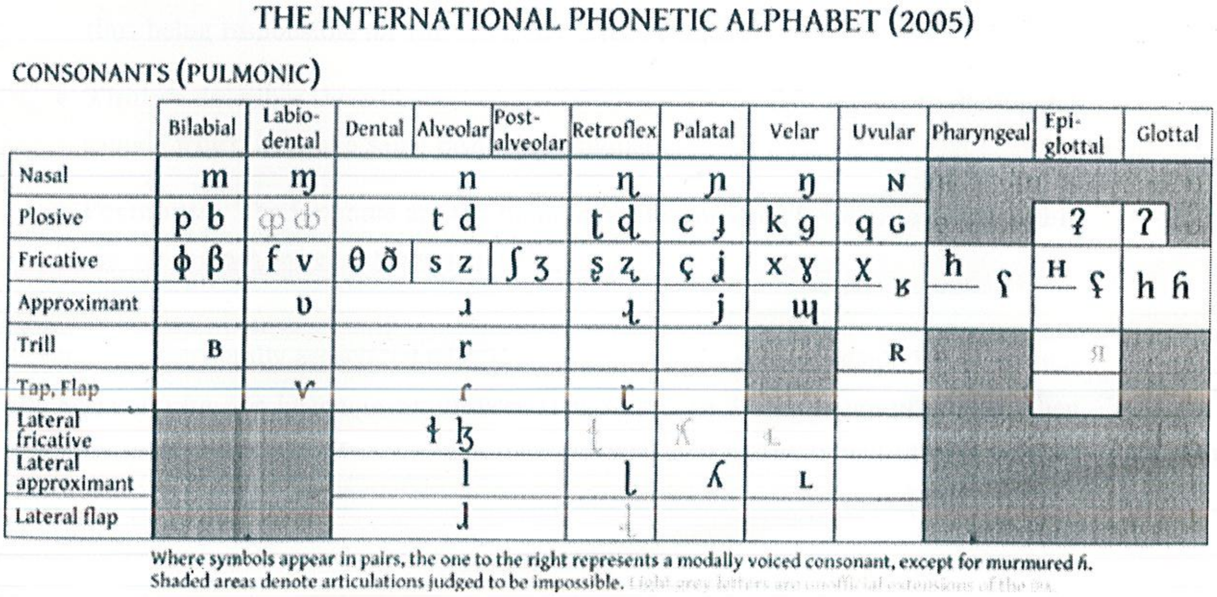
- **Affricates** are sounds that involve a plosive followed immediately by a fricative at the same location. In English, we have ch (unvoiced) and j (voiced). Many consider these as blends: t-sh and d-zh.

- **Nasals** are sounds made with air passing through the nose. In English, these are m, n, and ng.

- **Liquids** are sounds with very little air resistance. In English, we have l and r, which are both alveolar, but differ in the shape of the tongue.

- **Semivowels** are sounds that are very nearly vowels. They are also called glides, since they normally glide into or out of vowel positions (as in woo, yeah, ow, and oy).

- **Flaps and Taps:** When a quick collision of one articulator against another occurs it leads to flaps and taps. A slow collision in passing, would be called a flap and a quick almost impulsive collision would be a tap.



THE INTERNATIONAL PHONETIC ALPHABET (2005)

CONSONANTS (PULMONIC)

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		я	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ		ꞎ	𝼆	𝼄				
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		𝼈						

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ɦ. Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.

FIGURE 2.2: International Phonetic Alphabet

2.1.4 Co-ordination

The concept of co-ordination is interconnected with articulation and is inseparable. Co-ordination refers to the collaborative nature of articulatory organs to produce an advanced sound.

2.2 Characteristics of speech

- **Speech flow:** it tells about the speed at which utterances are produced as well as the number and duration of temporary breaks in speaking.
- **Intonation** is the way of producing utterances with respect to rise and fall in pitch, and leads to shifts in the tone, in either direction of the speaker's mean vocal pitch.
- Speech may be generally characterized by **pitch**, **loudness** and **timbre**.
- **Pitch**,in speech, is the relative highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the **vocal cords**.
- **Sound loudness** is a subjective term describing the strength of the ear's perception of a sound. It is intimately related to **sound intensity** but can by no means be considered identical to intensity.
- **Overtones** are the higher tones which faintly accompany a fundamental tone, thus being responsible for the tonal diversity of sounds.
- **Timbre** describes those characteristics of sound which allow the ear to distinguish sounds which have the same pitch and loudness.
- **Formants:** The formants are the fundamental resonating frequencies produced in the vocal tract, when a person speaks.

Formants are generally associated with the voiced part of a speech signal. When we plot the spectrum for the formants we observe that there are 3-5 formants produced when a person utters an alphabet. Formants maybe considered as the blueprint of a persons voice.

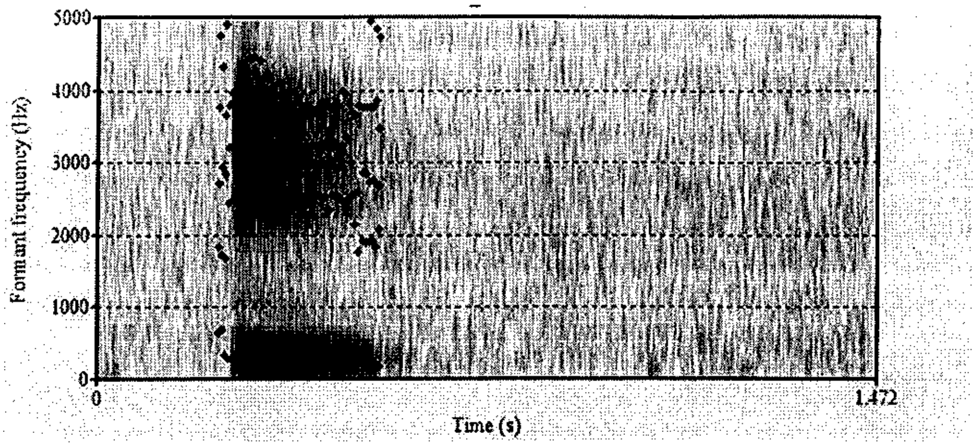


FIGURE 2.3: Formants (Ankit letter e)

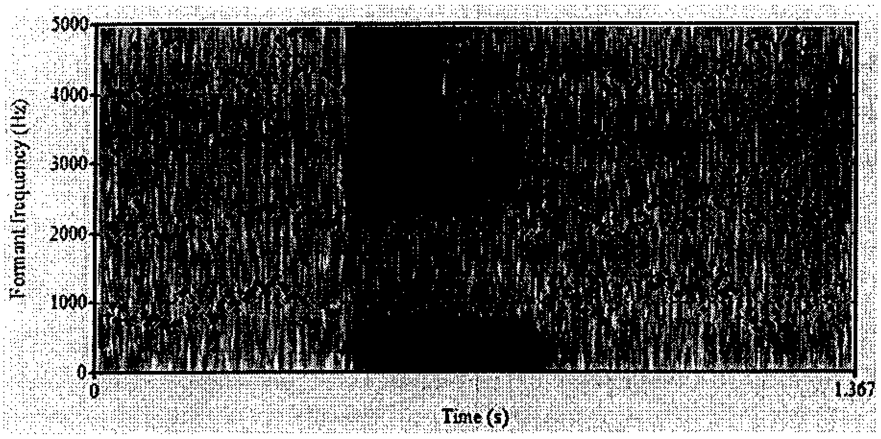


FIGURE 2.4: Formants (Aparna letter e)

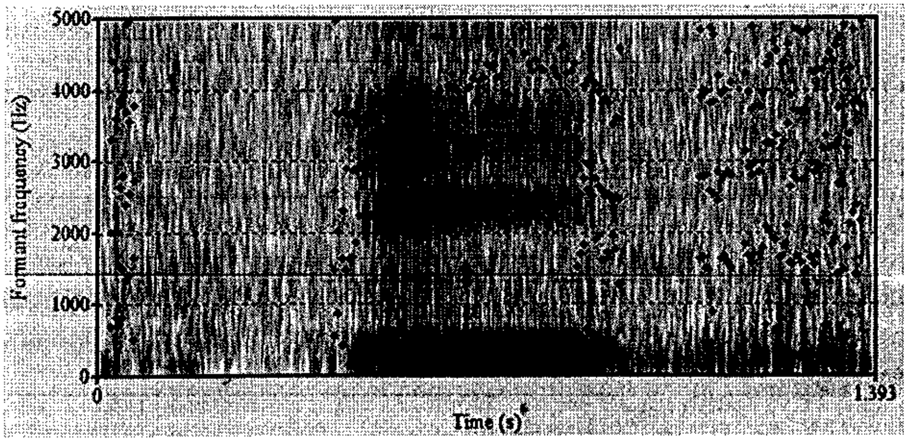


FIGURE 2.5: Formants (Puneet letter e)

2.3 Branches of speaker recognition

Speaker recognition is a generic term that refers to any procedure involving the knowledge of a person based on his/her voice. The speaker recognition discipline has many branches which are either directly or indirectly related. The branches may be classified as simple or compound. Simple branches include **speaker verification**, **speaker identification**, and **speaker classification** whereas speaker segmentation, speaker detection, and speaker tracking come under the category of compound branches.
Speech segregation can be done in two ways:

- **Based on speaker identification-** Speech identification is the process in which the algorithm or software designed itself identifies the differences in speech without using any input from user other than the samples.
- **Based on speaker verification-** In speech verification the user claims his identity and then the software or algorithm identifies the different voices in the sample. When the user claims the identity it is easier for the algorithm to compare and then differentiate the voice from the sample.

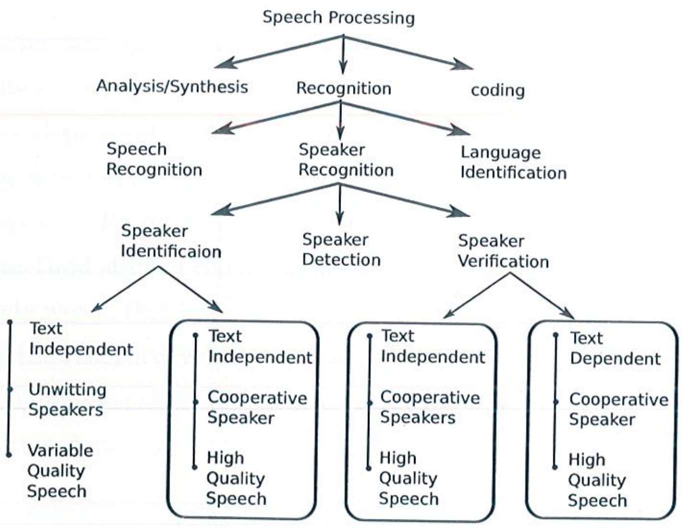


FIGURE 2.6: Speech Processing Classification, Speaker Recognition, Joseph P. Campbell, Jr., Department of Defense, Fort Meade, MD, j.campbell@ieee.org

Speaker verification is defined as deciding if a speaker is who he claims to be. This is different than the speaker identification problem, which is deciding if a speaker is a specific person or is among a group of persons. In speaker verification, a person makes an identity claim (e.g., entering an employee number or presenting his smart card).
Speaker Identification includes comparing the speech signal of an unknown person to a database of known people.

In text-dependent recognition, the phrase is known to the system and it can be fixed or not fixed and prompted (visually or orally). The claimant speaks the phrase into a microphone. This signal is analyzed by a verification system that makes the binary decision to accept or reject the users identity claim or possibly to report insufficient confidence and request additional input before making the decision.

Further, in speaker verification, the identity claim of the speaker is either accepted or rejected whereas in identification includes the determination of the registered speaker of a given speech.

Speaker classification is more generic and to an extent a little vague since all it does is pool all similar audio signals into individual bins.

The segmentation challenge is to be able to separate the speech produced by different speakers from each other. It is also desirable to separate, music and other non-speech segments.

Speaker detection encompasses segmentation as well as verification and identification. It includes the detection of one or more specific speakers in an audio stream.

Speaker Tracking is somewhat similar to speaker detection with the subtle difference that one or more of the speakers are tracked across the stream. In this case, one may envision conditions where no enrollment data is available, but not only is the audio segmented into single speaker segments, but the segments are also tagged with labels signifying the individual speakers in the stream.

Speaker recognition can also be classified as **text dependent** or **text independent** methods. In text dependent the key words or sentences are the same for training and recognition trials whereas in case of text independent method one does not rely on specific text being spoken. Formerly text independent methods were widely in use but later text dependent method started replacing it since they had an additional security feature of speaking a code word. But both these methods have a drawback

By playing back the recorded voice of registered speakers this system can be easily deceived with the highly developed electronics recording system that can repeat secret key words in a request order.

Chapter 3

Speaker Recognition System

The speaker recognition process comprises of three main phases. **Feature extraction**, **pattern matching** and **decision making**. Pattern matching and decision making together make up the **classification module**. The feature extraction module estimates a set of features from the speech signal that represent some speaker-specific information. The speaker-specific information is the result of complex transformations occurring at different levels of the speech production: semantic, phonologic, phonetic, and acoustic. The **semantic level** deals with vocabulary choice, sentence formulation etc.

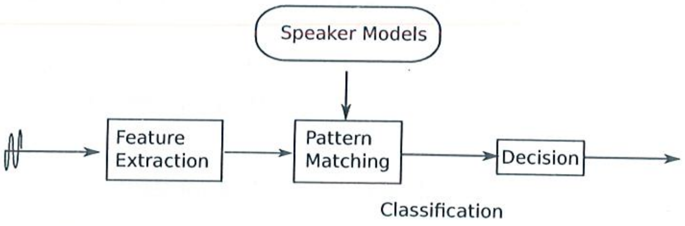


FIGURE 3.1: Speaker Recognition System

The **phonological level** deals with characteristics like duration of phonemes and intonations.

The **phonetic level** deals with the realization of the phonetic representation by the vibration of the vocal cords and the movements of articulators and the **acoustic level** deals with the spectral properties of the speech signal.

Speaker recognition and speech recognition: goal of speech recognition is to group all speakers into one and decipher the content of what is being said.

Speaker recognitions aim is to differentiate between speakers and is least interested in the content of the speech and this is why the role of vowels is crucial in case of speaker recognition. Vowels are very easy to recognize since they are all voiced and spectrally very different. Formants may be easily used for recognizing vowels.

3.1 Feature Extraction

In this project a very critical task is to extract features from the speech signal. Feature extraction basically aims at lessening the dimensionality of the input vector without losing the discriminating power of the signal. Feature extraction is required because number of test vectors required or classification increase exponentially with the dimension of the input vector. Its main aim is to reduce the input vector dimensionality without affecting the discriminating power of the signal. The number of training and test vector needed for the classification problem grows exponential with the dimension of the given input vector, so we need feature extraction. But extracted feature should meet some criteria while dealing with the speech signal, as enlisted:

- Extracted features should be easy to measure.
- Distinguish between speakers while being lenient of intra speaker variabilitys.
- Feature extraction should not be susceptible to mimicry.
- There should not be much fluctuation from one environment to another.
- Should exhibit stability with respect to time.
- It should occur frequently and naturally in speech.

3.1.1 Discrete Fourier Transform

We carry out spectral analysis for a signal by calculating discrete fourier transform. This is found by first calculating the discrete time fourier transform and then carrying out the quantization process for the Y-AXIS.
It is widely employed in:

- Signal Processing
- To analyse the frequencies contained in a sampled signal.
- To solve partial differential equations.
- To perform other operations such as convolutions.

Mathematical formulation for the same is:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn} \qquad k = 0, \dots, N-1 \tag{3.1}$$

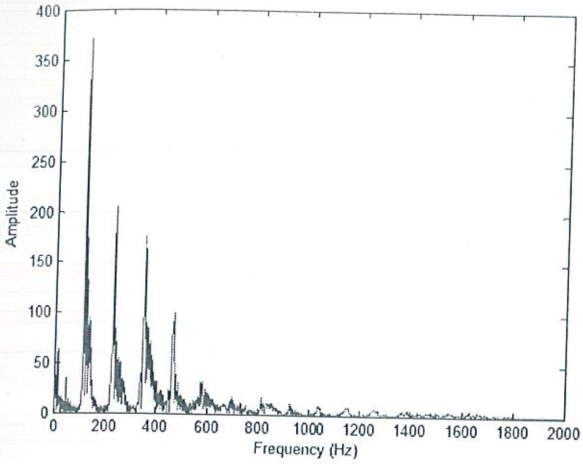


FIGURE 3.2: DFT Plot(puneet)

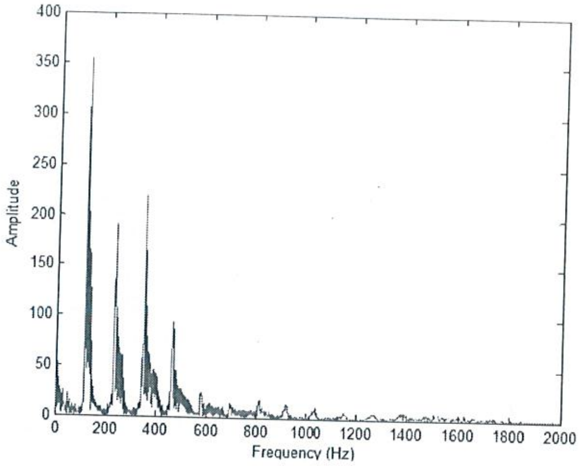


FIGURE 3.3: DFT Plot(puneet)

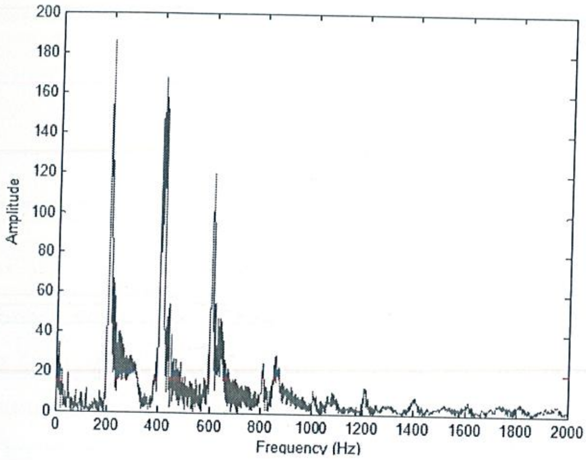


FIGURE 3.4: DFT plot(aparna)

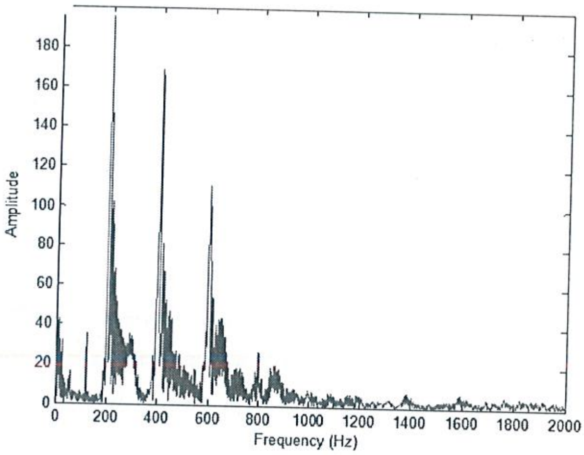


FIGURE 3.5: DFT plot(aparna)

Calculating the discrete fourier transform for a large number of voice samples ascertained the fact that voices could be discerned using this approach as the peaks in the resulting plot were consistent (see the figures below) an the frequency ranges for each individual varied to a very less extent, but speech being quasi-stationary, very accurate results cannot be obtained simply by calculation of the DFT.

3.1.2 Framing and Windowing

The figure shows a quasi stationary speech signal slowly varying with time.
It represents a word spoken by the speaker(Hello in this case). The recording was done at sampling frequency of 8KHz(mono) Time goes from left to right and amplitude is

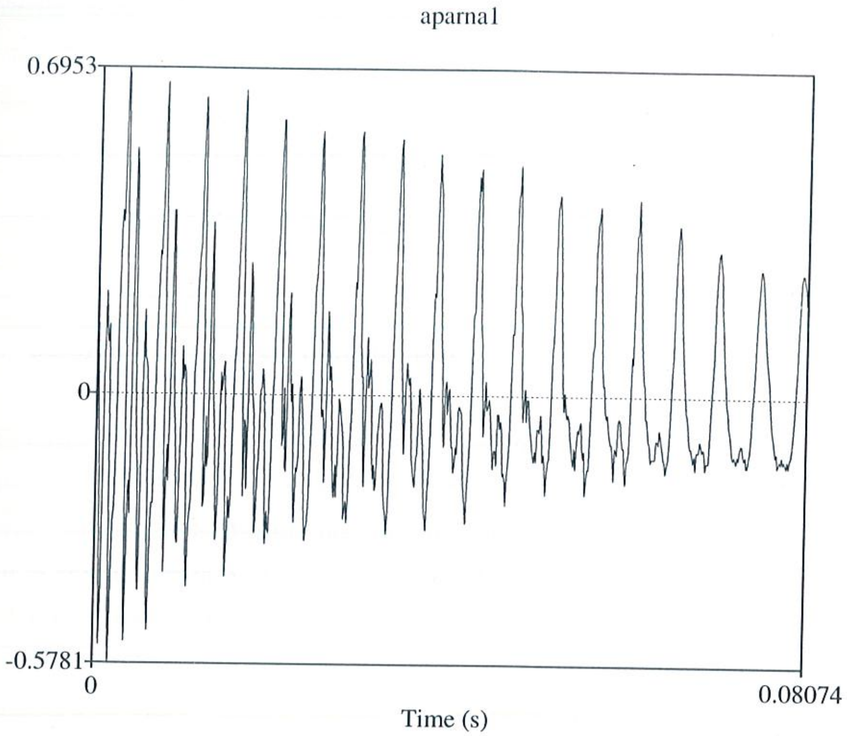


FIGURE 3.6: Quasi Stationary Nature of Voice

shown vertically. When the speech signal is examined over a short period of time such as 5 to 100 milliseconds, the signal is reasonably stationery, and therefore this signals are examine in short time segment, short time segments is referred to as a spectral analysis. This means that the signal is blocked into 20-30 milliseconds of each frame. Framing is done because speech signals are quasi stationary and hence beyond very small intervals of time the signal values keep varying.

Since the speech signal characteristics vary with time, it becomes very difficult to process the phrase since its tone, pitch etc may vary over a large range. Therefore, it is deemed necessary to split the signal into equal parts (for ease of analysis) i.e. window the signal. The most common method of windowing is splitting the signal into equal parts.

Some common windowing techniques include:

Rectangular window: It is sometimes known as a Dirichlet window. It is the simplest window, equivalent to replacing all but N values of a data sequence by zeros, making it appear as though the waveform suddenly turns on and off . It is represented as w(n)=1.

Hann window: is typically used as a window function in digital signal processing to select a subset of a series of samples in order to perform a Fourier transform or other calculations.

$$w(n) = 0.5\left(1 - cos\left(\frac{2\pi n}{N-1}\right)\right) \tag{3.2}$$

Hamming window: The "raised cosine" with these particular coefficients was proposed by Richard W. Hamming. The window is optimized to minimize the maximum (nearest) side lobe, giving it a height of about one-fifth that of the Hann window, a raised cosine with simpler coefficients

$$w(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \tag{3.3}$$

3.1.3 Linear Predictive Coefficient

Speech compression is often referred to as speech coding which is defined as a method for reducing the amount of information needed to represent a speech signal. Linear predictive coding is a digital method in which a particular value is predicted by a linear function of the past values of the speech signal. It was proposed as an encoding method by the United States Department of Defence in 1984. Human speech is produced in the vocal tract which can be approximated as a variable diameter tube. The linear predictive coding (LPC) model is based on a mathematical approximation of the vocal tract represented by this tube of a varying diameter. LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. The linear predictive filter which allows the value of the next sample to be determined by a linear combination of previous samples is the most important aspect of LPC. LPC is a lossy form of compression. It removes redundancy in signal and tries to predict next point as linear combination of previous values.

Applications of LPC include speech compression in standard telephone systems.Since LPC involves speech generation based on vocal tract model, it provides a perfect method for the generation of text from speech and hence is quite useful in the field of text-to-speech synthesis. Further applications of LPC and other speech compression schemes are voice mail systems, telephone answering machines, and multimedia applications since it provides smallest storage space due to its reduced bit rate.

The limitation of LPC is that it includes a trade-off between low bit rates and the quality of speech signal. It generally produces sounds that at times may sound synthetic. In order to find some degree of similarity or difference in the recorded speech signals we employed the LPC technique. Order 8 LPC coefficients were calculated with the help of Matlab (see table in Appendix) but no specific conclusions resulted and hence it prompted to employ newer techniques for speaker recognition.

3.1.4 Cepstrum

When the inverse fourier transform of the logarithm of the speech signal spectrum is taken, it results in the Cepstrum. The cepstrum can be seen as information about rate of change in the different spectrum bands. Mathematically Cepstrum of a signal maybe written as $FT(log(FT(thesignal)) + j2\_m)$ and algorithmically,

$$signal \rightarrow FT \rightarrow abs() \rightarrow log \rightarrow FT \rightarrow cepstrum \tag{3.4}$$

Cepstrum can be of two types: Real cepstrum and Complex cepstrum.

Real cepstrum uses the logarithm function and information about the spectrum magnitude.

The complex cepstrum uses the complex logarithm function and holds information of the magnitude as well as the phase of the spectrum.

This is one of the methods for feature extraction. The basic need for taking logarithmic spectrum is a better, easy and conclusive analysis of spectrum of signal which is actually convolution between the source signal and the transfer function of the filter used. In frequency domain, convolution is the multiple of frequency domain source signal and frequency domain transfer function of filter used.

A speech signal is made up of a quickly varying excitation sequence convolved with a slow varying part impulse response of the vocal system.

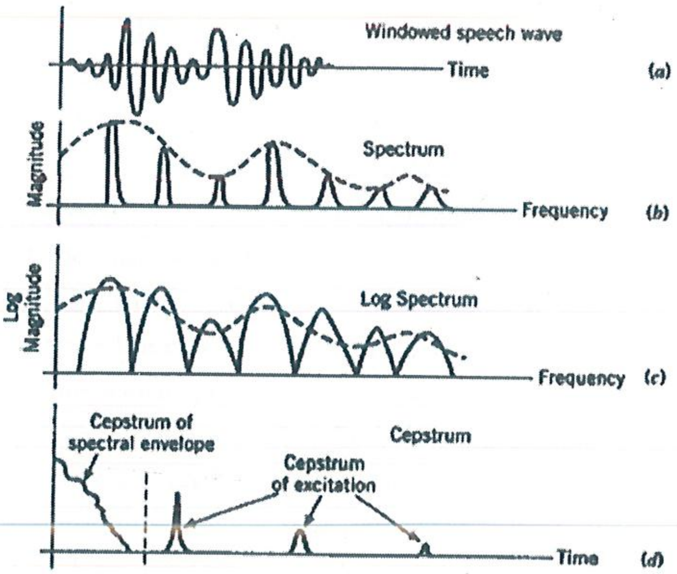


FIGURE 3.7: Source filter separation by cepstrum analysis, B. Gold and N. Morgan. Speech and audio Signal Processing. John Wiley and Sons Inc, 1999

$$s(n) = e(n) * \theta(n) \tag{3.5}$$

Due to convolution, separating these two parts is difficult and hence Cepstrum is used for the same. The equation for spectrum is:

$$C_s(n) = T^{-1}\left\{log|\mathfrak{F}\left\{s(n)\right\}|\right\} \tag{3.6}$$

The convolution changes to multiplication as the signal moves from time domain to frequency domain:

$$S(\omega) = E(\omega)\Theta(\omega) \tag{3.7}$$

By taking the logarithm of the spectrum, the multiplication gets converted to the addition.

$$log|S(\omega)| = log|E(\omega)\Theta(\omega)| = log|E(\omega)| + log|\Theta(\omega)| = C_e(\omega) + C_\theta(\omega) \tag{3.8}$$

The inverse Fourier transform, being linear, works individually on the two components.

$$C_s(n) = \mathfrak{F}^{-1}\left\{C_e(\omega) + C_\theta(\omega)\right\} = \mathfrak{F}^{-1}\left\{C_e(\omega)\right\} + \mathfrak{F}^{-1}\left\{C_\theta(\omega)\right\} = c_e(n) + c_\theta(n) \tag{3.9}$$

3.1.5 Mel Frequency Cepstral Coefficients

Mel cepstral coefficients are derived from the fourier transform of the audio signal. Speech is analyzed over short analysis window. For each short analysis window a spectrum is obtained using FFT. Spectrum is passed through Mel-Filters to obtain Mel-Spectrum. Cepstral analysis is performed on Mel-Spectrum to obtain Mel-Frequency Cepstral Coefficients. Thus speech is represented as a sequence of Cepstral vectors. It is these Cepstral vectors which are given to pattern classifiers for speech recognition purpose. The Mel-Frequency Cepstral Coefficients (MFCCs) are always real and convey information about the physical aspects of the speech signal.

The difference between thecepstrumand the mel-frequency cepstrum is the frequency bands are linearly spaced in case of normal spectrum whereas in MFC they are equally spaced on themel scale, which approximates the human auditory system's response more closely. This frequency warping can allow for better representation of sound.

As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. These coefficients allow better processing of data. MFCCs are commonly used asfeaturesinspeech recognitionsystems, such as the systems which can automatically recognize numbers spoken into a telephone. They are also common inspeaker recognition, which is the task of recognizing people from their voices. Figure shows the complete flowchart for calculation of Mel Frequency Cepstral Coefficients. The human ear does not display a frequency resolution

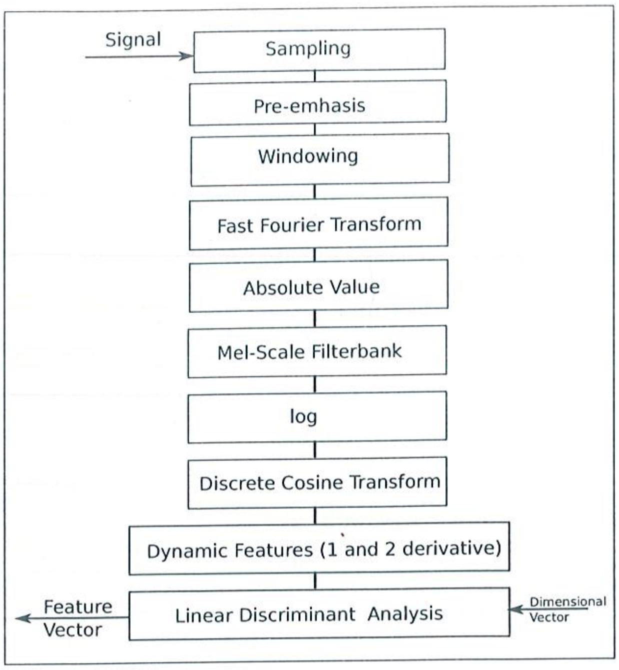


FIGURE 3.8: Speaker Recognition Process

that is linear but builds several groups of frequencies and integrates the spectral energies within a given group. Also the mid frequency and the bandwidth are non-linearly distributed. The nonlinear warping of the frequency axis is modeled by the mel-scale. The frequency groups are assumed to be linearly distributed along the mel-scale.

Mel Scale

Stevens, Volkmann and Newmann in 1937 proposed Mel as a unit of pitch. Mel means Melody. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 mels. Below about 500 Hz the mel and hertz scales coincide; above that, larger and largerintervals are judged by listeners to produce equal pitch increments. formula to convert f hertz into m mel is:

$$m = 2595log_{10}\left(1 + \frac{f}{700}\right) \tag{3.10}$$

The scale is divided into the units mel. In this test the listener or test person started out hearing a frequency of 1000 Hz, and labeled it 1000 Mel for reference. Then the listeners were asked to change the frequency till it reaches to the frequency twice the reference frequency. Then this frequency labelled 2000 Mel. The same procedure repeated for the half the frequency, then this frequency labelled as 500 Mel, and so on. On this basis the normal frequency is mapped into the Mel frequency. The Mel scale is normally a linear mapping below 1000 Hz and logarithmically spaced above 1000 Hz. Mel, an abbreviation of the word melody, is a unit of pitch. It is defined to be equal to one thousandth of the pitch of a simple tone with frequency of 1000 Hz with an amplitude of 40 dB above the auditory threshold.

Mel spectrum is used to reflect the perception characteristics of the human ear. In analogy to computing the cepstrum, we now take the logarithm of the mel power spectrum (instead of the power spectrum itself) and transform it into the quefrency domain to compute the socalled mel cepstrum.

Mel frequency warping is made more convenient using the bank filter, the filters are centered according to the Mel frequency. Inverse DFT is used for cepstral coefficient calculation. Insignal processing, afilter bankis an array ofband-passfiltersthat separates the input signal into multiple components, each one carrying a singlefrequencysub-bandof the original signal. The process of decomposition performed by the filter bank is calledanalysis(meaning analysis of the signal in terms of its components in each sub-band); the output of analysis is referred to as a sub-band signal with as many sub-bands as there are filters in the filter bank.

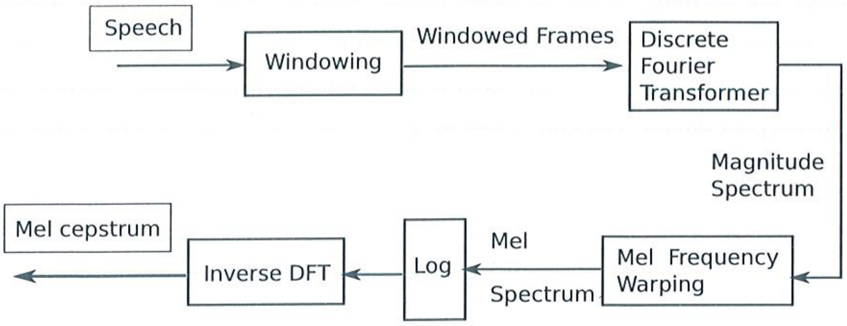


FIGURE 3.9: Process of generating Mel Cepstrum

3.2 Pattern Matching

3.2.1 Vector quantization

Quantizing feature vectors to smaller template vectors makes analysis easier and this is done using vector quantization. VQ is a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the distribution. Vector quantization is used in processes where small deviations from the mean value do not matter much in the decision process.

Vector quantization is a quantization technique for signal processing and is used for pattern matching. It is a method generally employed to compress data. It clubs together a large set of data vector which are closest and represent them as one vector. Each group is represented by a respective centroid points.

The density matching property of vector quantization is powerful, especially for identifying the density of large and high-dimensioned data. As the data is compressed so this method can be stated as lossy compression scheme.

Vector quantization technique is efficiently used in various areas of biometric modalities like finger print pattern recognition, face recognition and speech recognition. The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features. By means of VQ, storing every single vector that we generate from the training is impossible.

VQ is a is used as a modeling technique for speaker recognition.In this the entire speech signal is divided into clusters and each cluster has a centroid which represents the mean of that cluster.

3.2.1.1 K-Means Algorithm

The K-means algorithm is a way to cluster the training vectors to get feature vectors. In this algorithm clustered the vectors based on attributes into k partitions. It use the k means of data generated from gaussian distributions to cluster the vectors. The objective of the k-means is to minimize total intra-cluster variance, V.

$$V = \sum_{i=1}^{k} \sum_{j \epsilon S_i} |x_j - \mu_i|^2 \tag{3.11}$$

where there are k clusters Si, i = 1,2,...,k and i is the centroid or mean point of all the points $x_j \epsilon S_i$

The process of k-means algorithm used least-squares partitioning method to devide the input vectors into k initial sets. It then calculates the mean point, or centroid, of each set. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters, and algorithm repeated until when the vectors no longer switch clusters or alternatively centroids are no longer changed.

3.2.2 Measuring Similarity

3.2.2.1 Mahalanobis Distance

The "Mahalanobis distance" is a rule for calculating the distance between two points, which is better adapted than the usual "Euclidian distance".
Two usual cases where the Mahalanobis distance plays an important role :

1. Distance of a point to the mean of a distribution.
2. Distance between the means of two distributions.

It is based on correlations between variables by which different patterns can be identified and analyzed. It is a useful way of determining similarity of an unknown sample set to a known one.

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \tag{3.12}$$

Mahalanobis distance (or "generalized squared inter point distance" for its squared value) can also be defined as a dissimilarity measure between two random vectors and of the same distribution with the covariance matrix S.

$$d(\overrightarrow{x},\overrightarrow{y}) = \sqrt{(\overrightarrow{x}-\overrightarrow{y})^T S^{-1}(\overrightarrow{x}-\overrightarrow{y})} \tag{3.13}$$

Some important characteristics of the mahalanobis distance:

- It takes into consideration the fact that the variances in each direction are different.
- It also accounts for the covariance between variables.
- It reduces to the common Euclidean distance for uncorrelated variables with unit variances

The Mahalanobis distance accounts for the variance of each variable and the covariance between variables. Geometrically, it does this by transforming the data into standardized uncorrelated data and computing the ordinary Euclidean distance for the transformed data
Mahalanobis is mainly used in classification problems, where there are several groups and the investigation concerns the affinities between groups. The object of the study may be to form clusters of members who are similar to each other, perhaps in a hierarchical scheme
We calculated cepstrum coefficients and then Mahalanobis Distance was calculated between two samples of voices, one being the input from speaker and other from the

database. We observed that the distance between same voice sample was coming more as compared to the distance between two different voice sample. This contradicted theoretical concepts of speaker recognition. So we neglected this speech constraint.

3.2.2.2 Bhattacharaya Distance

In statistics, the Bhattacharya distance measures the similarity of two discrete or continuous probability distributions. It is closely related to the Bhattacharyya coefficient which is a measure of the amount of overlap between two statistical samples. It is given by:

$$D_B(p,q) = -ln(BC(p,q))\text{where}BC(p,q) = \sum_{x \epsilon X} \sqrt{p(x)q(x)} \tag{3.14}$$

Chapter 4

Approach and Final Model

The main objective of this project was to be able to perform speaker verification and identification and thereafter move onto the segregation of speech samples. Speaker verification has been a field that has been extensively researched and a lot many techniques have been devised to execute the same. The segregation of speech samples was a relatively new concept and hence provides scope for research.

In the initial phase of the project, focus was on speaker verification and identification. For the same, initially we calculated the Fourier transform of a spoken word by the same person taken at different points of time and tried comparing the resultant transforms of the voices. Analysis suggested that Fourier transform was not enough and some other features need to be employed in order to get tangible results. This was because the values though for voices recorded at a particular time were similar but they changed when we checked the same after a few days. The voice samples were decided to be recorded at 8 kHz so that the processing is faster and efficient at the same time. Also the optimum range a human ear can hear and distinguish is below 4 kHz.

Next, we employed Linear Predictive Coding Coefficients which is an efficeient technique for data compression and also helps predict future values as a consequence of initial linear values. Again the results were not sufficient in providing any conclusive results since the values obtained for LPC filter of order eight were very close and did not vary in a particular pattern.
With no proper results the next logical step was to employ calculation of cepstrum for speech sample. The results implied some success but still were insufficient to reach any comprehensive conclusion. On reading the literature available in this field and consulting our mentors, we found that we were not using the concepts of windowing and framing

which are essential in processing speech signals. Therefore the aforementioned technique was repeated after performing the windowing and the framing of the signal.

In addition to this we tried the resonant frequency approach to decipher if normalised DFT at all could be used for differentiating speeh signals. It was found, in case of all the speech samples, that the resonant frequencies were more or less consistent and varied very less for the consecutive samples that were recorded for each person.

Achieving this foresight we decided to model it and present the results in a concrete form.

Firstly a code segment was developed to perform framing and windowing of the speech signal. The sample was divided into 30 ms frames with an overlap of 50% of the frame duration. The windowing was performed to avoid signal truncation. The most suited windowing technique for speech processing is the hamming window which was employed. The hamming window function is given by:

$$w(i) = 0.54 + 0.46 * cos\left(\frac{2\pi i}{N}\right) \tag{4.1}$$

Algorithm for calculating the Mel-Cepstrum was implemented. The process for calculating melcepstrum employs the following:

Fourier Transform: The fourier transform was calculated with the help of the built in matlab function fft, which calculates it by the fast fourier transform method.

Absolute Value is taken using the abs() function in Matlab.

Then the fourier transform values are converted to **cepstrum** which in turn are con-

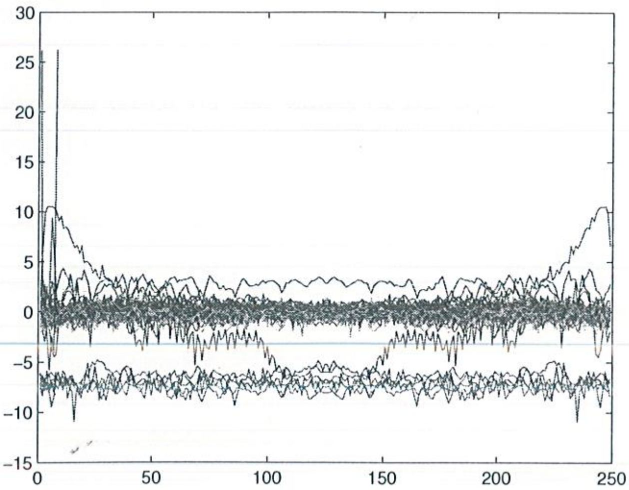


FIGURE 4.1: Plot of Cepstrum of a Voice Sample

verted into **mel cepstrum** with the help of mel scale with the help of the function melbank. The melbank function is a function developed Mike Brookes and is provided in the matlab toolbox Voicebox.

Then the **log** of the signal was taken and the **Discrete Cosine transform** was applied on the signal. The matlab function dct was used in the process.

Then the **kmeans function** was executed on the obtained mel frequency cepstrums. The kmeans function clusters the coefficients into different groups.

After successful implementation of the code segments mentioned above, the euclidean distance was calculated for the test data and the individual samples of each speaker. The euclidian is basically the distance calculated between the centroids of the two speech samples.

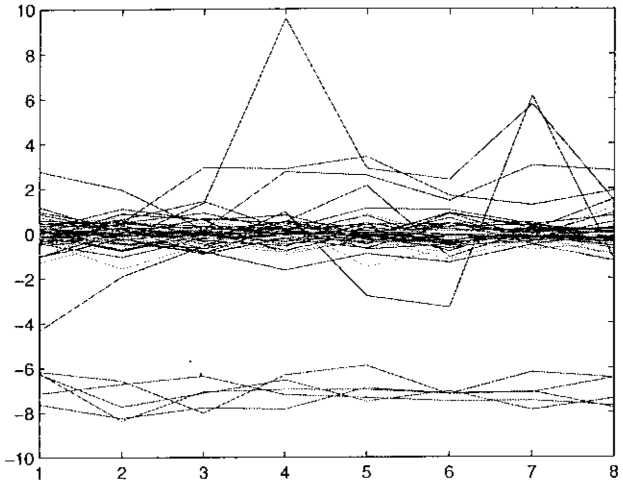


FIGURE 4.2: Plot of kmeans of a voice sample

The decision making was based on the values of the distance obtained. A minimum distance value indicated that the two speech samples matched and hence spoken by the same person.

Appendix A

Matlab Codes

A.1 Speaker.m

10

```
clear
prwd='\\172.16.73.3\081011\Project\speech data\Final\';
disp('pwd is ');
disp(prwd);
yn=input('do you want to change the present working directory y
   /n','s');
if(yn=='y'||yn=='Y')
    prwd=input('enter the correct directory address','s');
end
yn=input('do you want to record training samples y/n  ','s');
if(yn=='y'||yn=='Y')
    createnew(prwd);
else
    yn=input('Would you like to check for matching data y/n','s
   ');
    if(yn=='y'||yn=='Y')
        check(prwd);
    end

end
```

A.2 createnew.m

10

```
function []=createnew(prwd)
pwdtemp=prwd;
name=input('enter your name ','s');
fn=cat(2,prwd,name);
p=exist(fn,'dir');
if (p==0)
mkdir(prwd,name);
else
    p=input('the name already exists do you want to retrain y/n ','s')
    if(p=='y'||p=='Y')
        mkdir(prwd,name);
    else
        return
    end
end

prwd=cat(2,prwd,name,'\');
Fs=8000;
recobj1=audiorecorder();
disp('the computer will now ask you to record code word ')
disp('when ready to speak the code(single word), press a key say the code word ')
pause on;
pause
disp('start ')
recordblocking(recobj1, 2)
disp('end of recording');
first = getaudiodata(recobj1);
firstname=cat(2,prwd,name,'1.wav');
wavwrite(first ,Fs,8,firstname);

max=0;
for i=1:length(first)
    if(first(i)>max)
```

```
        max=first(i);
    end
end
first=first-mean(first);                        % to remove dc component
first=first/max;
Nsamps = length(first);
t = (1/Fs)*(1:Nsamps);
%Do Fourier Transform
first_fft = abs(fft(first));                    %Retain Magnitude
first_fft = first_fft(1:Nsamps/2);              %Discard Half of Points
f = Fs*(0:Nsamps/2-1)/Nsamps;                   %Prepare freq data for plot

%Plot Sound File in Time Domain
figure;
plot(t, first);
xlabel('Time (s)');
ylabel('Amplitude');
title(firstname);
save2word(cat(2,prwd,'figures.doc'));
%Plot Sound File in Frequency Domain
plot(f, first_fft);
xlim([0 2000]);
xlabel('Frequency (Hz)');
ylabel('Amplitude');
title(firstname);
save2word(cat(2,prwd,'figures.doc'));

disp('Loading data...')
[train.data]=Load_data(pwdtemp)
disp('Calculating mel-frequency cepstral coefficients for training set...')
[train.cc] = mfcc(train.data,Fs,pwdtemp)
disp('Performing K-means...')
C=8                      % number of centroids
[train.kmeans] = kmean(train.cc,C,pwdtemp)
```

A.3 Load_data.m

10

```
function [data] = Load_data(prwd)
% Training mode - Load all the wave files to database (codebooks) %

di=dir(prwd);
count=0;
for d=3:length(di)
    if di(d).isdir==1
        count=count+1;
        a=(di(d).name);
        c=cat(2,prwd,a,'\',a,'1.wav');
        data{count}=wavread(c)';
        name{count}=(di(d).name);
    end
end
c=cat(2,prwd,'data');
save(c,'data');
c=cat(2,prwd,'name');
save(c,'name');
```

A.4 mfcc.m

10

```
function [cepstral] = mfcc(x,fs,prwd)
% Calculate mfcc's with a frequency(fs) and store in ceptral cell.
cepstral = cell(size(x,2),1);
for i = 1:size(x,2)
    disp('loop')
cepstral{i} = melcepst(x{i},fs,'x');
plot(cepstral{i})
hold on
end
c=cat(2,prwd,'cepstral');
```

```
save(c,'cepstral');
```

A.5 melcepst.m

10

```
function c=melcepst(s,fs,w)
nc=12;                          %no. of cesrral coefficients
p=floor(3*log(fs));             %no. of filters in filterbank 2.1
   per octave
n=pow2(floor(log2(0.03*fs)));    %length of frames in power of 2
    for 30 ms
fh=0.5;
fl=0;
inc=floor(n/2);                  %amount of overlap n/2=128
z=enframe(s,hamming(n),inc);     %framed matrix of the sound
f=rfft(z.');                     %dft of the signal is
   calculated
[m,a,b]=melbankm(p,n,fs,fl,fh,w);   %Use melbank to convert to
   mel scale
pw=f(a:b,:).*conj(f(a:b,:));
pth=max(pw(:))*1E-20;
ath=sqrt(pth);
y=log(max(m*abs(f(a:b,:)),ath));
c=dct(y).';                 %calculates the dct of the melspectrum
   coefficients
                            %converts to mfcc
nf=size(c,1);
nc=nc+1;
if p>nc
   c(:,nc+1:end)=[];
elseif p<nc
   c=[c zeros(nf,nc-p)];
end
end
```

A.6 enframe.m

10

```
function f=enframe(x,win,inc)
% x is the voice sample
% win is the window which is hamming for us
% inc is the increment amount which is n/2 for us
a=size(x);
if(a(1)~=1)
    x=x';
end
nx=length(x(:));%length of voice
lw=length(win); %length of window
w = win(:)';
nli=nx-lw+inc;
nf = fix((nli)/inc);
                    %total number of frames that fit in the voice sample
na=nli-inc*nf;
f=zeros(nf,lw); %matrix of zeros for output
for i=1:nf
    for j=1:lw
        f(i,j)=x(1,((i-1)*inc+j));
    end
end
w=w(ones(nf,1),:);
f = f.* w;
end
```

A.7 rfft.m

10

```
function y=rfft(x)
s=size(x);
d=2 %the dimension along which to do fft
n=s(d);
y=fft(x,n,d);
```

```
s(d)=1+fix(n/2);
y(s(d)+1:end,:,:)=[];
end
```

A.8 kmean.m

10

```
function [datakmean] = kmean(x,C,prwd)
% Calculate k-means for x with C number of centroids
train.kmean.x = cell(size(x,1),1);
train.kmean.esql = cell(size(x,1),1);
train.kmean.j = cell(size(x,1),1);
for i = 1:size(x,1)
[train.kmean.j{i} train.kmean.x{i}] = kmeans(x{i}(:,1:12),C);
end
datakmean = train.kmean.j;
c=cat(2,prwd,'kmeans');
save(c,'datakmean');
```

A.9 check.m

10

```
function []=check(prwd)
Fs=8000;
recobj1=audiorecorder();
disp('the computer will now ask you to record code word ')
disp('when ready to speak the code(single word), press a key
    say the code word ')
pause on;
pause
disp('start ')
recordblocking(recobj1, 2)
disp('end of recording');
test.data = getaudiodata(recobj1)'
```

```
disp('Calculating mel-frequency cepstral coefficients for test data...')
test.cc = melcepst(test.data,Fs,'x')
c=cat(2,prwd,'kmeans.mat');
train.kmeans=load(c);


disp('Compute a distortion measure for each codebook...')
[result index] = distmeasure(train.kmeans.datakmean,test.cc);
c=cat(2,prwd,'name.mat');
name=load(c);
name
result
disp('results with least distance is most probable')
```

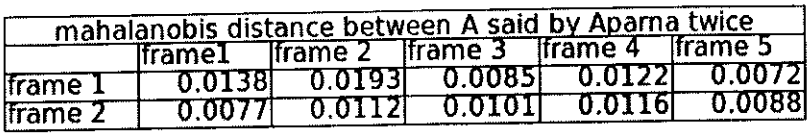
A.10 distmeasure.m

10

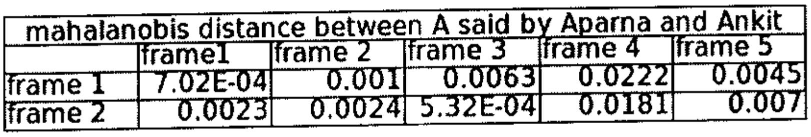
```
function [result,index] = distmeasure(x,y)
result = cell(size(x,2),1);
dist = cell(size(x,2),1);
mins = inf;
for i = 1:size(x,1)
dist{i} = disteusq(x{i}(:,1:12),y(:,1:12),'x');
temp = sum(min(dist{i}))/size(dist{i},2);
result{i} = temp;
if temp < mins
mins = temp;
index = i;
end
end
```

Appendix B

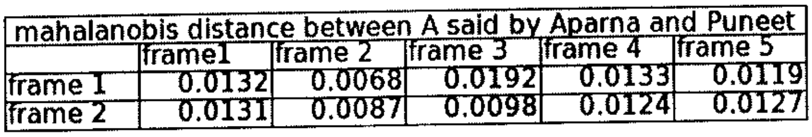
Mahalanobis Distance



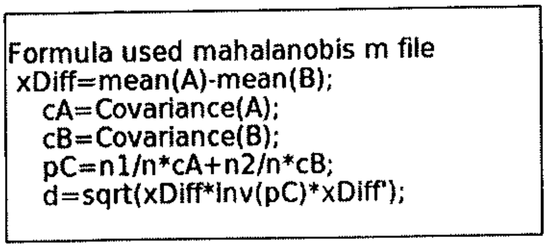
mahalanobis distance between A said by Aparna twice					
	frame1	frame 2	frame 3	frame 4	frame 5
frame 1	0.0138	0.0193	0.0085	0.0122	0.0072
frame 2	0.0077	0.0112	0.0101	0.0116	0.0088



mahalanobis distance between A said by Aparna and Ankit					
	frame1	frame 2	frame 3	frame 4	frame 5
frame 1	7.02E-04	0.001	0.0063	0.0222	0.0045
frame 2	0.0023	0.0024	5.32E-04	0.0181	0.007



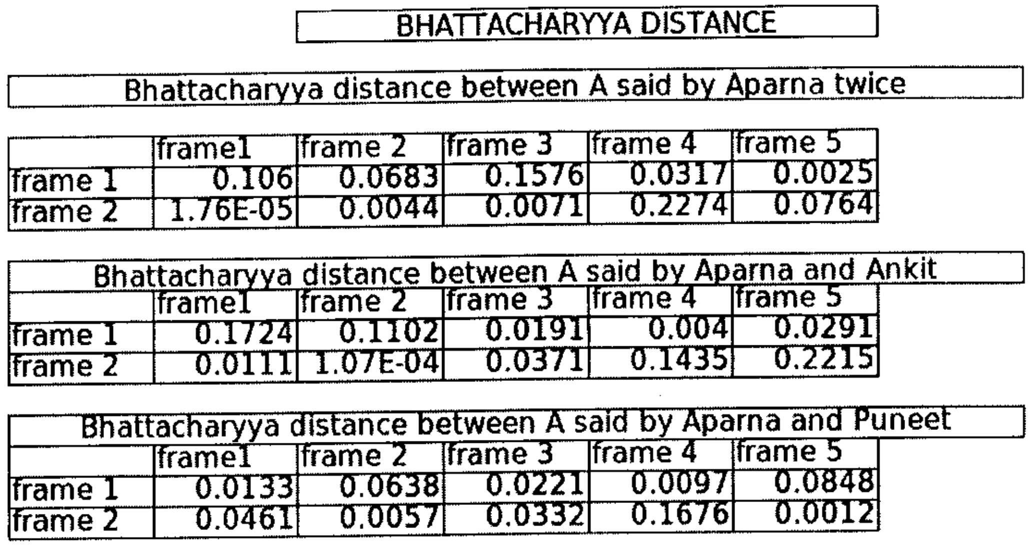
mahalanobis distance between A said by Aparna and Puneet					
	frame1	frame 2	frame 3	frame 4	frame 5
frame 1	0.0132	0.0068	0.0192	0.0133	0.0119
frame 2	0.0131	0.0087	0.0098	0.0124	0.0127



```
Formula used mahalanobis m file
 xDiff=mean(A)-mean(B);
   cA=Covariance(A);
   cB=Covariance(B);
   pC=n1/n*cA+n2/n*cB;
   d=sqrt(xDiff*Inv(pC)*xDiff');
```

Appendix C

Bhattacharaya Distance



BHATTACHARYYA DISTANCE

Bhattacharyya distance between A said by Aparna twice

	frame1	frame 2	frame 3	frame 4	frame 5
frame 1	0.106	0.0683	0.1576	0.0317	0.0025
frame 2	1.76E-05	0.0044	0.0071	0.2274	0.0764

Bhattacharyya distance between A said by Aparna and Ankit

	frame1	frame 2	frame 3	frame 4	frame 5
frame 1	0.1724	0.1102	0.0191	0.004	0.0291
frame 2	0.0111	1.07E-04	0.0371	0.1435	0.2215

Bhattacharyya distance between A said by Aparna and Puneet

	frame1	frame 2	frame 3	frame 4	frame 5
frame 1	0.0133	0.0638	0.0221	0.0097	0.0848
frame 2	0.0461	0.0057	0.0332	0.1676	0.0012

Appendix D

Execution of program

10

```
pwd is
\\172.16.73.3\081011\Project\speech data\Final\
do you want to change the present working directory y/nn
do you want to record training samples y/n  n
Would you like to check for matching data y/ny
the computer will now ask you to record code word
when ready to speak the code(single word), press a key say the code word
start
end of recording

test =

    data: [1x16000 double]

Calculating mel-frequency cepstral coefficients for test data ...

d =

     2


test =

    data: [1x16000 double]
      cc: [249x13 double]

Compute a distortion measure for each codebook...

name =

    name: {'ankit'  'aparna'  'parth'  'puneet'  'sanskar'}



result =

    [17.6023]    [24.5787]    [17.3521]    [5.4994]
   [27.0981]

results with least distance is most probable
```

Bibliography

[1] Matlab-documentation. Technical report, www.mathworks.com/help/techdoc/.

[2] Voicebox: Speech processing toolbox for matlab. Technical report, Exhibition Road, London SW7 2BT, UK.

[3] Homayoon Beigi. *Fundamentals of Speaker Recognition.* Springer Science+Business Media, Yorktown Heights, NY, USA, 2011.

[4] P. Boersma and D. Weenink. Praat-5.3.02.

[5] Jr. Campbell, J.P. Speaker recognition.

[6] W.J. Chen and W.T. Huang. Vector quantization. Technical report, http://en.wikipedia.org/wiki/Vectorquantization; accessed October 2011.

[7] L. Deng and Shanghnessy. *Speech Processing : A Dynamic And Optimization Oriented Approach.* 1st edition, 2003.

[8] E. Deza and M.M. Deza. Euclidean distance. Technical report, http://en.wikipedia.org/wiki/Euclideandistance; accesed Februaury 2012.

[9] B. Gold and N. Morgan. *Speech and audio Signal Processing.* John Wiley and Sons Inc, 1999.

[10] L. Rabiner and B. Hwang-Juang. *Fundamentals Of Speech Recognition.* Prentice Hall, 1993.