# Speaker Recognition using Audio Spectrum Projection and Vector Quantization

Bikram Kar
kar.bikram2012@gmail.com

Avishek Dey
avishek.nit.07@gmail.com

*Department of Computer Science and Engineering*
Sanaka Educational Trust's Group of Institutions, India.

***Abstract* - Speaker Recognition is a process of automatically identifying the speaker from individual information included in speech waves. Speaker Recognition is one of the most useful biometric recognition techniques when security is paramount. Many organizations such as banks, institutions, industries etc are currently using this technology for providing greater security to their vast databases. Speaker Recognition have mainly two modules: feature extraction and feature matching. Feature extraction extracts a small amount of data from the speaker's voice signal that can later be used to represent that speaker. Feature matching involves the procedure to identify the unknown speaker by comparing the extracted features from his/her voice input with those already stored in the speech database. In this paper we propose for feature extraction to use the Audio Spectrum Projection technique of MPEG-7, which relies on basis decomposition, where three choices exist: Principle Component Analysis (PCA), Independent Component Analysis (ICA) and Non Negative Matrix Factorization (NNMF). Moreover for feature matching we find the Vector Quantization, VQ, distortion between the input utterance of an unknown speaker and the codebooks stored in the database. Finally based on this VQ distortion we decide whether to accept or reject the unknown speaker's identity.**

***Keywords - Speaker Recognition, Audio Spectrum Projection, Vector Quantization***

## I. INTRODUCTION

Speaker recognition is a process to identify a person on the basis of his speech. We all are aware about the fact that speech is a speaker dependent feature that's why we can recognize one of our friends over telephone. During the years ahead, it is hoped that speaker recognition will make it possible to verify the identity of persons accessing systems; allow automated control of services by voice, such as banking transactions; and also control the flow of private and confidential data. While fingerprints and retinal scans are more reliable means of identification, speech can be seen as a non-evasive biometric that can be collected with or without the person's knowledge or even transmitted over long distances via telephone. Unlike other forms of identification, such as passwords or keys, a person's voice cannot be stolen, forgotten or lost. Speech is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulator, and acoustic. Differences in these transformations appear as differences in the acoustic properties of the speech signal. Speaker-related differences are a result of a combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individuals. In speaker recognition, all these differences can be used to discriminate between speakers.

Speaker recognition allows for a secure method of authenticating speakers. During the enrollment phase, the speaker recognition system generates a speaker model based on the speaker's characteristics. The testing phase of the system involves making a claim on the identity of an unknown speaker using both the trained models and the characteristics of the given speech. Many speaker recognition systems exist and the following chapter will attempt to classify different types of speaker recognition systems.

## II. RELATED WORK

There is considerable speaker-recognition activity in industry, national laboratories, and universities. Among those who have researched and designed several generations of speaker-recognition systems are AT&T (and its derivatives); Bolt, Beranek, and Newman [2]; the Dali Molle Institute for Perceptual Artificial Intelligence (Switzerland); ITT; Massachusetts Institute of Technology Lincoln Labs; National Tsing Hua University (Taiwan); Nagoya University (Japan); Nippon Telegraph and Telephone (Japan); Rensselaer Polytechnic Institute; Rutgers University; and Texas Instruments (TI) [1]. The majority of ASV research is directed at verification over telephone lines. Sandia National Laboratories, the National Institute of Standards and Technology, and the National Security Agency have conducted evaluations of speaker-recognition systems. It should be noted that it is difficult to make meaningful comparisons between the text-dependent and the generally more difficult text-independent tasks. Text-independent approaches, such as Gish's segmental Gaussian model and Reynolds'

Gaussian Mixture Model [3], need to deal with unique problems (e.g., sounds or articulations present in the test material but not in training). It is also difficult to compare between the binary choice verification task and the generally more difficult multiple-choice identification task. The general trend shows accuracy improvements over time with larger tests (enabled by larger data bases), thus increasing confidence in the performance measurements. For high-security applications, these speaker recognition systems would need to be used in combination with other authenticators (e.g., smart card). The performance of current speaker-recognition systems, however, makes them suitable for many practical applications. There are more than a dozen commercial ASV systems, including those from ITT, Lernout & Hauspie, T-NETIX, Veritel, and Voice Control Systems. Perhaps the largest scale deployment of any biometric to date is Sprint's Voice FONCARD, which uses TI's voice verification engine. Speaker-verification applications include access control, telephone banking, and telephone credit cards. The accounting firm of Ernst and Young estimates that high-tech computer thieves in the United States steal $3–5 billion annually. Automatic speaker-recognition technology could substantially reduce this crime by reducing these fraudulent transactions.

As automatic speaker-verification systems have gained widespread use, it is imperative to understand the errors made by these systems.

## III. AUDIO SPECTRUM PROJECTION AND VECTOR QUANTIZATION

There are two types of errors: the false acceptance of an invalid user (FA or Type I) and the false rejection of a valid user (FR or Type II). It takes a pair of subjects to make a false acceptance error: an impostor and a target.

Because of this hunter and prey relationship, in this paper, the impostor is referred to as a wolf and the target as a sheep. False acceptance errors are the ultimate concern of high-security speaker-verification applications; however, they can be traded off for false rejection errors. After reviewing the methods of speaker recognition, a simple speaker-recognition system will be presented. A data base of 186 people collected over a three-month period was used in closed-set speaker identification experiments [1]. A speaker-recognition system using methods presented here is practical to implement in software on a modest personal computer. The features and measures use long-term statistics based upon an information-theoretic shape measure between line spectrum pair (LSP) frequency features. This new measure, the *divergence shape*, can be interpreted geometrically as the shape of an information-theoretic measure called divergence. The LSP's were found to be very effective features in this divergence shape means.

The design is first tested with MATLAB. A total of eight speech samples from eight different people (eight speakers, labeled S1 to S8) are used to test this project. Each speaker utters the same single digit, *zero*, once in a training session (then also in a testing session). A digit is often used for testing in speaker recognition systems because of its applicability to many security applications.

In our proposed Speaker Recognition System we generally used MPEG-7 Audio Spectrum Projection for feature extraction of sound and Vector Quantization for feature matching. The feature extraction of MPEG-7 ASP mainly consists of three techniques: Normalized Audio Spectrum Envelop (NASE), a basic decomposition algorithm, and a spectrum basis projection. Here we used Vector Quantization for 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*.

### *A. Speaker Identification*

Speaker identification is mapping a speech signal from an unknown speaker to a database of known speakers, i.e. the system has been trained with a number of speakers which the system can recognize. Speaker identification can be further divided into two branches. Open-set speaker identification decides to whom of the registered speakers' unknown speech sample belongs or makes a conclusion that the speech sample is unknown. In this work, we deal with the closed-set speaker identification, which is a decision making process of whom of the registered speakers is most likely the author of the unknown speech sample. Depending on the algorithm used for the identification, the task can also be divided into text-dependent and text-independent identification. The difference is that in the first case the system knows the text spoken by the person while in the second case the system must be able to recognize the speaker from any text .The process of speaker identification is divided into two main phases. During the first phase, speaker enrollment, speech samples are collected from the speakers, and they are used to train their models. The collection of enrolled models is also called a speaker database. In the second phase, identification phase, a test sample from an unknown speaker is compared against the speaker database. Both phases include the same first step, feature extraction, which is used to extract speaker dependent characteristics from speech. The main purpose of this step is to reduce the amount of test data while retaining speaker discriminative information. Then in the enrollment phase, these features are modeled and stored in

the speaker database. This process is represented in Figure 1.

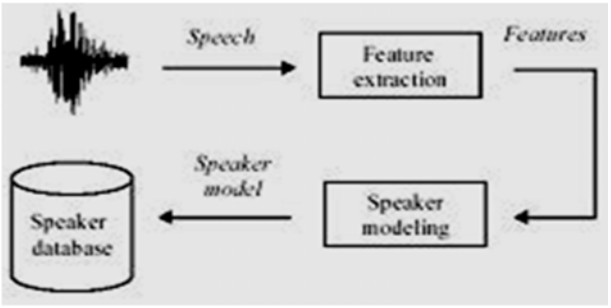



Figure 1. Enrollment Phase

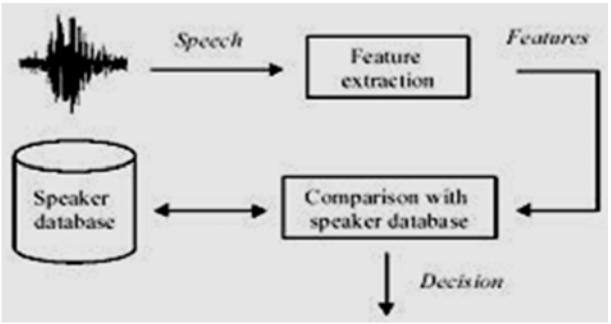



Figure 2. Identification Phase

In the identification step, the extracted features are compared against the models stored in the speaker database. Based on these comparisons the final decision about speaker identity is made. This process is represented in Figure 2.

*B. Feature Extraction*

The acoustic speech signal contains different kind of information about speaker. This includes "high-level" properties such as dialect, context, speaking style, emotional state of speaker and many others. More useful approach is based on the "low-level" properties of the speech signal such as pitch (fundamental frequency of the vocal cord vibrations), intensity, formant frequencies and their bandwidths, spectral correlations, short-time spectrum and others.

From the automatic speaker identification point of view, it is useful to think about speech signal as a sequence of features that characterize both the speaker as well as the speech. It is an important step in identification process to extract sufficient information for good discrimination in a form and size which is amenable for effective modeling. The amount of data, generated during the speech production, is quite large while the essential characteristics of the speech process change relatively slowly and therefore, they require less data. According to these matters feature extraction is a process of reducing data while retaining speaker discriminative information.

The speech wave is usually analyzed based on spectral features. There are two reasons for it. First is that the speech wave is reproducible by summing the sinusoidal waves with slowly changing amplitudes and phases. Second is that the critical features for perceiving speech by human ear are mainly included in the magnitude information and the phase information is not usually playing a key role. In our paper Vector Quantization technique is used to solve the problem of feature extraction of the sample sound.

*C. Framing and Windowing*

The speech signal is slowly varying over time (quasi-stationary) that is when the signal is examined over a short period of time (5-100msec), the signal is fairly stationary. Therefore speech signals are often analyzed in short a time segment, which is referred to as short-time spectral analysis. It works in the following way: predefined length window (usually 20-30 milliseconds) is moved along the signal with an overlapping (usually 30-50% of the window length) between the adjacent frames. Overlapping is needed to avoid losing of information. Parts of the signal formed in such a way are called frames. In order to prevent an abrupt change at the end points of the frame, it is usually multiplied by a window function. The operation of dividing signal into short intervals is called windowing and such segments are called windowed frames (or sometime just frames). The most popular window function used in speaker identification is Hamming window function, which is described by the following equation:

$$h(n) = 0.54 - 0.46 \cos(2\pi n / N-1),\ 0 \leq n \leq N-1 \qquad (1)$$

Where $N$ is the size of the window or frame. A set of features extracted from one frame is called feature vector.

*D. MPEG-7 Audio Spectrum Projection*

The MPEG-7 ASP feature extraction mainly consists of a Normalized Audio Spectrum Envelope (NASE), basis decomposition algorithm such as Singular Value Decomposition (SVD) or Independent Component Analysis (ICA) and a spectrum basis projection, obtained by multiplying the NASE with a set of extracted basis functions. For the basis decomposition step, we combined a basis dimension reduction by PCA algorithm with basis information maximization by FastICA . To extract Audio Spectrum Envelope (ASE) features, the observed audio signal is analyzed using a 512-point FFT. The power spectral coefficients are grouped in logarithmic sub-bands spaced in non-overlapping 7-octave bands spanning between low boundary (62.5 Hz) and high boundary (8 kHz). The resulting 30-dimensional ASE is converted to the decibel scale. Each decibel-scale spectral vector is

normalized with the RMS (root mean square) energy envelope, thus yielding a normalized log-power version of the ASE called NASE. For each audio class, the spectral basis is extracted by computing the PCA for dimension reduction and FastICA for information maximization. The resulting spectrum projection is the product of the NASE matrix, the dimension-reduced PCA basis functions and the FastICA transformation matrix.

*E. Vector Quantization*

Speaker recognition systems are inherent of a database, which stores information used to compare the test speaker against a set of trained speaker voices. Ideally, storing as much data obtained from feature extraction techniques is advised to ensure a high degree of accuracy, but realistically this cannot be achieved. The number of feature vectors would be so large that storing and accessing this information using current technology would be unfeasible and impractical.

Vector Quantization (VQ) is a quantization technique used to compress the information and manipulate the data in such a way to maintain the most prominent characteristics. VQ is used in many applications such as data compression (i.e. image and voice compression), voice recognition, etc. VQ in its application in speaker recognition technology assists by creating a classification system for each speaker. Given the extracted feature vectors (known as code words) from each speaker, each codeword is used to construct a codebook. This process is applied to every single speaker to be trained into the system. VQ codebook algorithms are inherently difficult to implement. Although numerous VQ algorithms exist, Linde-Buzo-Gray or LBG VQ Algorithm is chosen, since it is the easiest to implement.

*E1. VQ-Linde Buzo Gray (LBG) Algorithm*

The LBG algorithm is operated on a given codebook. LBG splits the codebook into segments and performs an exhaustive analysis on each segment. The analysis compresses the training vector information creating a new codebook which is then used to compute the next segment. The Flow Diagram of VQ-LBG Algorithm is shown in figure. This code book is obtained using a splitting method. In this method an initial code vector is set as the average, and then split in two vectors. Then the iterative algorithm is run with those two vectors. Resulting two vectors are then split again into 2 vectors each. This gives us now four vectors, and the process is then repeated until the desired number of code vectors is obtained.

The process continues until all segments have been processed and the new codebook is created. The aim of this algorithm is to minimize any distortions in the data creating a codebook which is computationally optimized, while providing a sub-optimal solution. The performance of VQ analysis is highly dependent on the length of the voice file which is operated upon.

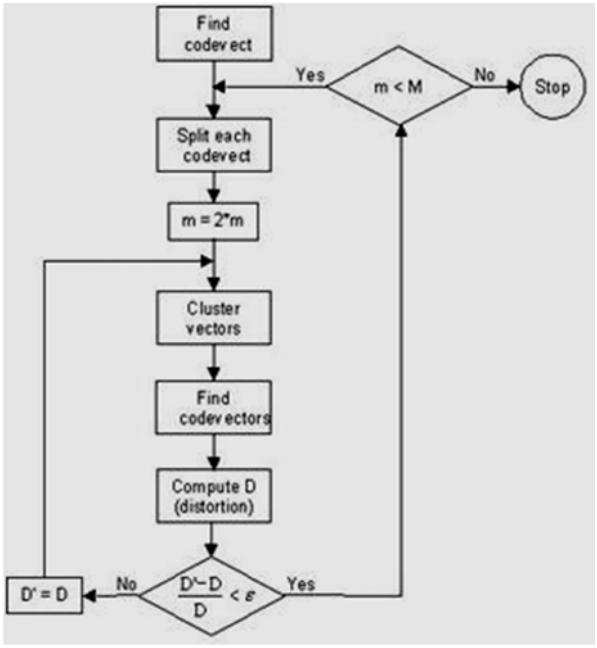
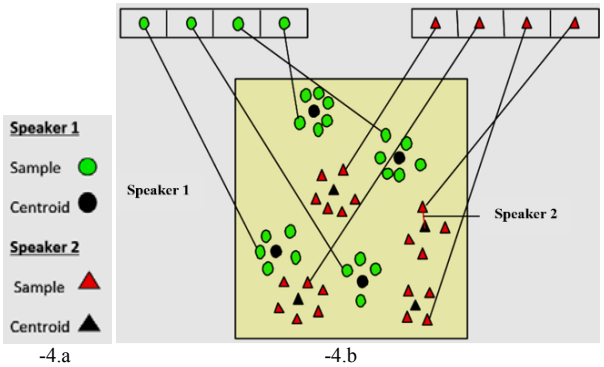



Figure 3: Flow Diagram of VQ-LBG Algorithm

## IV. RESULTS AND DISCUSSION

In this project we will experience the building and testing of an automatic speaker recognition system.




-4.a -4.b

Figure 4. Conceptual diagram illustrating Vector Quantization codebook formation.

The VQ Distortion is shown in figure 4-a where 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, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result codeword's (centroids) are shown in figure 4-b 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. VQ distortion is nothing but the Euclidian distance between the two vectors and is given by the formula:

$$d_E(x,y)=\sqrt{\sum_{i=1}^{\dim}(x_i-y_i)^2}$$

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. One speaker can be discriminated from another based of the location of centroids. In this project all these tasks are implemented in Matlab.

*B1. MATLAB Results*

Speech signals corresponding to eight speakers i.e. S1.wav, S2.wav, S3.wav, S4.wav, S5.wav, S6.wav, S7.wav, and S8.wav in the training folder are compared with the speech files of the same speakers in the testing folder. The matching results of the speakers are obtained as follows.

*B2. When All Valid Speakers are Considered*

Speaker 1 matches with speaker 1
Speaker 2 matches with speaker 2
Speaker 3 matches with speaker 3
Speaker 4 matches with speaker 4
Speaker 5 matches with speaker 5
Speaker 6 matches with speaker 6
Speaker 7 matches with speaker 7
Speaker 8 matches with speaker 8

From the plots it is clearly understand difference between the Euclidean distances of speakers are given in figure 5.1 below:

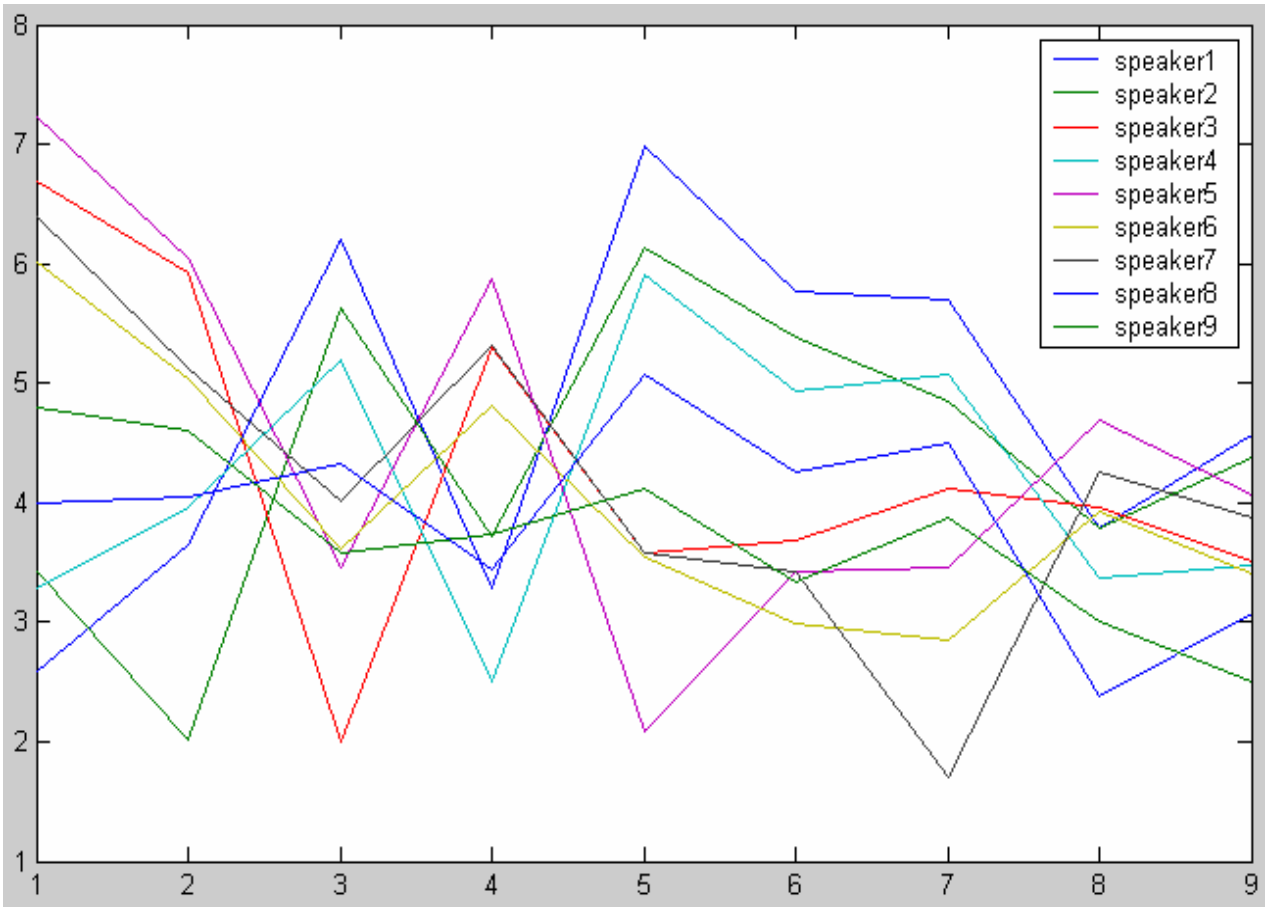



Figure 5.1: Plot for the difference between the Euclidean distances of speakers.

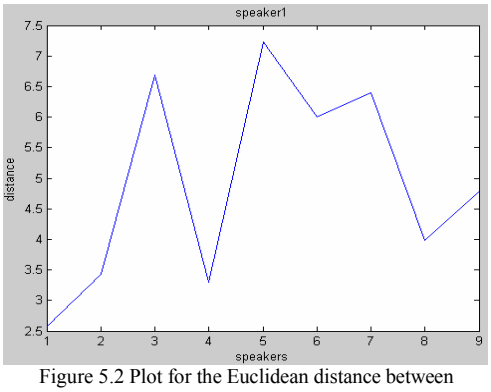



Figure 5.2 Plot for the Euclidean distance between speaker 1 and all speakers

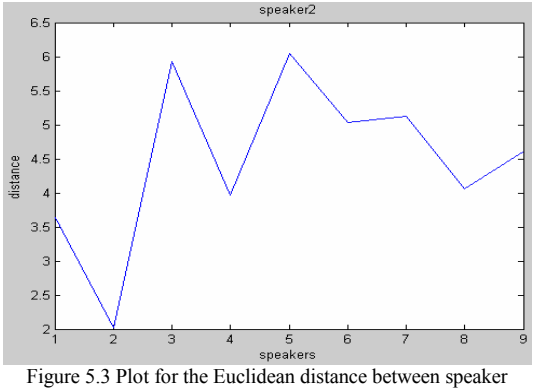



Figure 5.3 Plot for the Euclidean distance between speaker 2 and all speakers

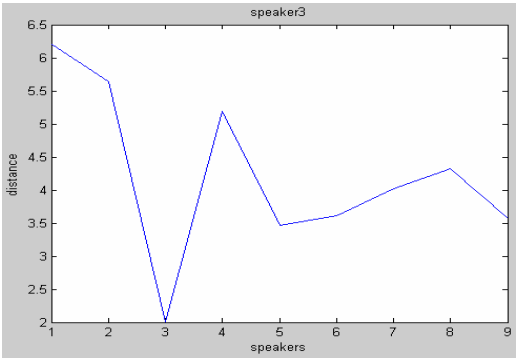



Figure 5.4 Plot for the Euclidean distance between speaker 3 and all speakers.

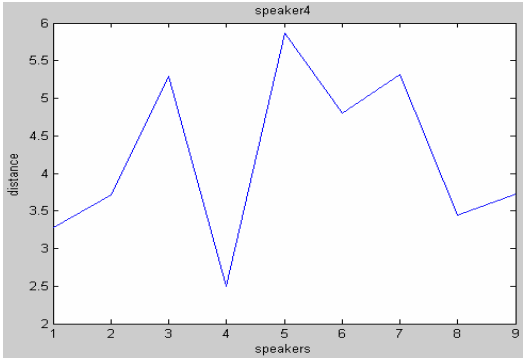



Figure 5.5 Plot for the Euclidean distance between speaker 4 and all speakers

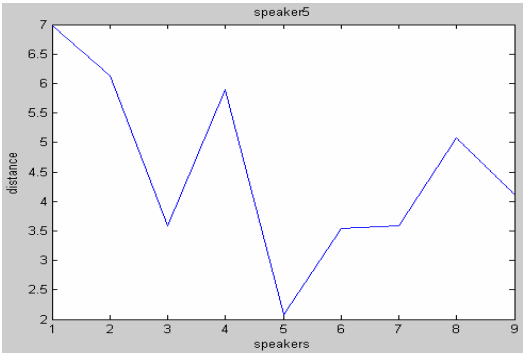



Figure 5.6 Plot for the Euclidean distance between speaker 5 and all speakers.

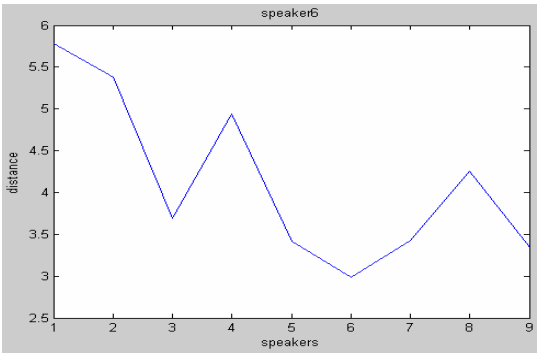



Figure 5.7 Plot for the Euclidean distance between speaker 6 and all speakers

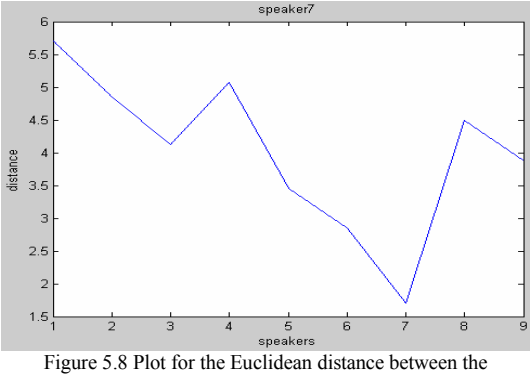



Figure 5.8 Plot for the Euclidean distance between the speaker 7 and all speakers.

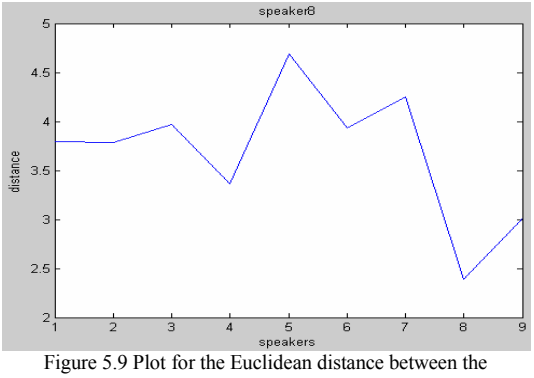



Figure 5.9 Plot for the Euclidean distance between the speaker 8 and all speakers

This system is able to recognize 7 out of 8 speakers. This is an error rate of 12.5%. The recognition rate of our system is much better than the one of a human's recognition rate. However you must be aware that this test is not really representative of the computer's efficiency to recognize voices because only 8 persons are tested, with only one training session and with only one word.

## V. CONCLUSION

The results obtained in this project using ASP and VQ are excellent. W e have computed Audio Spectrum Projections corresponding to each speaker and these were vector quantized. The VQ distortion between the resultant codebook and ASP's of an unknown speaker is taken as the basis for determining the speaker's authenticity. Here we used ASP because they follow the human ear's response to sound signals. The performance of this model is limited by a single coefficient having a very large VQ distortion with the corresponding codebook. The performance factor can be optimized by using high quality audio devices in a noise free environment. There is a possibility that the speech can be recorded and can be used in place of the original speaker .This would not be a problem in our case because the ASP's of the original speech signal and the recorded signal are different. Psychophysical studies have shown that there is a probability that human speech may vary over a period of 2-3 years. So the training sessions have to be repeated in order to update the speaker specific codebooks in the database.

## REFERENCES

[1] Campbell, J.P., Jr.; "Speaker recognition: a tutorial" Proceedings of the IEEE_Volume 85, I s s u e 9, Sept. 1997 Page(s):1437 – 1462.

[2] Roucos, S. Berouti, M. Bolt, Beranek and Newman, Inc., Cambridge, MA; "The application of probability density estimation to text-independent speaker identification" IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '82__Volume: 7, On page(s): 1649- 1652. Publication Date: May 1982.

[3] Castellano, P.J.; Slomka, S.; Sridharan, S.; "Telephone based speaker recognition using multiple binary classifier and Gaussian mixture models", IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 Volume 2, pp1075 – 1078 April 1997.

[4] D.A. Reynolds, R.C. Rose, "Robust text-independent speaker identification using Gaussian Mixture speaker models", *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, 1995, pp. 72 -83.

[5] S. Molau, M. Pitz, R. Schluter, H. Ney, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum", *Acoustics, Speech, and Signal Processing*, 2001 IEEE International Conference, Volume: 1, 2001, pp. 73-76

[6] B. S. Atal, "Automatic Recognition of Speakers from their Voices", *Proceedings of the IEEE*, vol 64, 1976, pp 460 – 475.

[7] H. Gish and M. Schmidt, "Text Independent Speaker Identification", *IEEE Signal Processing Magazine*, Vol. 11, No1994, pp. 18-32.. 4,

[8] J. R. Deller, J. H. L. Hansen, J. G. Proakis, "*Discrete-Time Processing of Speech Signals"*, Piscataway (N.J.), IEEE Press, 2000.

[9] C.-H. Lee, F.K. Soong, K.K. Paliwal: "Automatic Speech and Speaker Recognition" - advanced topics. Kluwer Academic Publishers, pp. 42-44, Norwell, Massachusetts, USA, 1996.

[10] S. Sookpotharom, S. Manas "Codebook Design Algorithm for Classified Vector Quantization" Bangkok University, Pathumtani, Thailand pp. 751-753 2002.

[11] Martin, A. and Przybocki, M., "The NIST 1999 Speaker Recognition Evaluation—An Overview", Digital Signal Processing, Vol. 10, Num. 1-3, January/April/July 2000

[12] John G. Proakis and Dimitris G. Manolakis, "Digital Signal Processing", New Delhi: Prentice Hall of India. 2002.

[13] D.A. Reynolds, R.C. Rose, "Robust text-independent speaker identification using Gaussian Mixture speaker models", *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, 1995, pp. 72 -83.

[14] J. R. Deller, J. G. Proakis and J. H. L. Hansen, Discrete-time Processing of Speech Signals, Prentice Hall, New Jersey, 1993.