International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 08, August-2015

Speaker Recognition using Cepstral Analysis

G. Manjuła
Associate Professor: Dept. of ECE
GSSSIETW
Mysuru, India

Dr. M. Shiva Kumar [2]
Prof & HOD: Dept. EIE
GSSSIETW
Mysuru, India

Basavanna. M[3]
Assistant Professor: Dept. of TE
GSSSIETW
Mysuru, India

***Abstract*— The aim of the paper is to verify the identity of the speaker using his voice as a metric. The system is trained using database of registered voice signals. The system extracts features from the user's voice and stores his voice profile in database along with his personal information. When the user has to be verified/ authenticated, user records the test voice signals. The system compares the input voice with the one stored in the database. If the input voice profile and the voice stored in the database are matched with each other, the user is verified/authenticated otherwise the user is rejected. The voice profile of the user is stored in the database in form of MFCC coefficients. The system uses a minimum Euclidian distance measurement algorithm to compare the two voice profiles. Graphical User Interface [GUI] is developed with the system and the performance of the system is tested on 10 users.**

Keywords: Speaker recognition, Cepstral Analysis, MFCC: Mel Frequency Cepstral Co efficient.

I. INTRODUCTION

Speaker recognition is process of automatically identifying who is speaking person based on the individual information embedded in the speech signal. Speaker recognition is an appealing research field for the last decades which still yields a number of unsolved problems such as recognition of speech disorders, whispered speech. Speaker recognition can be broadly classified into two types: Text dependent speaker recognition and Text independent speaker recognition. In the former case, the text must be same for both enrollment and verification of speaker recognition. In the later case, cooperation of the speaker required is less and hence it is the most commonly used technique for speaker recognition. In this technique, text employed for enrollment and testing\may be different [1]. Speaker recognition technology is the most potential technology to generate new services that will make our everyday lives more secured. An important application of speaker recognition technology is in forensic field.

The need to determine the identity and authority of users and customers is increasing in today's life. For any individual, it results in a growing number of PIN-codes to remember which requires more memory space. A simpler solution would be to construct biometric verification systems based on the individual's physical features such as fingerprints, retina and voice. The various application possibilities for speech based systems is in telecommunication, banking, shopping by just a phone call, database access services, informative services, security control for confidential information areas, and remote access to computers makes such systems an attractive alternative. Speech is a complicated signal produced as a result of several transformations occurring at different levels such as semantic, linguistic, articulator. Acoustic Differences in these transformations appear as differences in the acoustic properties of the speech signal.To increase security in speaker-based verification systems, a combination with more conventional methods could be used.

II. DESIGN METHODOLOGY

This work aims at developing a simple but effective speaker recognition system. The fundamental difficulty of speaker recognition is in extracting features from non-stationary signal which deals with noise, channel variations, voice changes due to health conditions, aging, mimicry and acoustic conditions. The speech feature extraction is a process of reducing the redundant information and retaining useful information which has desirable features such as high discrimination between sub-word classes, low speaker variability, Invariance to degradations in speech signal due to interference.

The general approach to automated speaker verification consists of five steps: Digital speech data acquisition, feature extraction, pattern matching, making an accept/reject decision, and enrollment to generate speaker reference models. A block diagram of this procedure is shown in Fig.2.1.

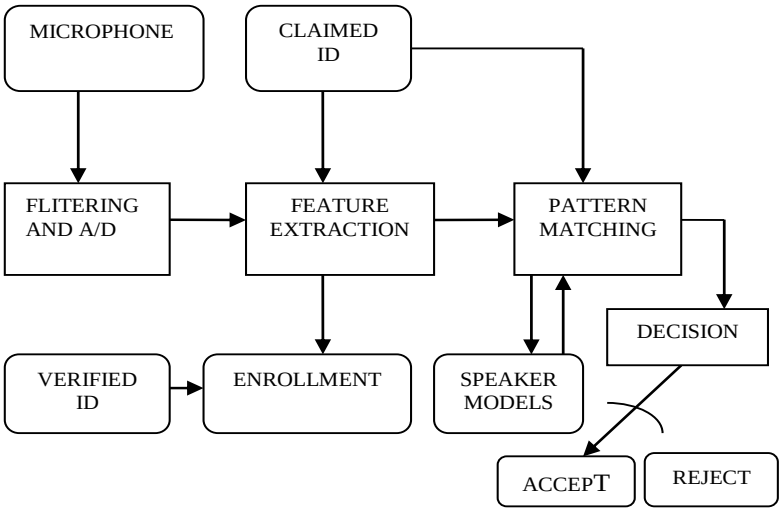

Fig. 2.1: Generic Speaker Recognition

The feature extraction maps each interval of speech to a multidimensional feature space. Then, this sequence of feature vectors is compared with speaker models by pattern matching. This results in a match score for each vector or sequence of vectors. The match score measures the similarity of the computed input feature vectors to models of the claimed speaker or feature vector patterns for the claimed speaker. A decision is made to either accept or reject the claimant according to the match score or sequence of match scores, which is a hypothesis testing problem.

For speaker recognition, features that exhibit high speaker discrimination power, high inter-speaker variability, and low intra-speaker variability are desired. Many forms of pattern matching and corresponding models are possible. Pattern-matching methods include dynamic time warping (DTW), the hidden Markov model (HMM), artificial neural networks, and vector quantization (VQ). Template models are used in DTW, statistical models are used in HMM, and codebook models are used in VQ [6].

III SYSTEM LEVEL DESIGN

In order to process a signal by a digital computer, the signal must be represented in digital form. Initially, the acoustic sound pressure wave is transformed into a digital signal suitable for voice processing. Data acquisition system comprises of three modules namely : Signal acquisition, Signal Pre-processing and Processing of speech signal. The first module consists of microphone which is used to convert speech signal to an electrical analog signal. The second module consists of anti aliasing filter. The purpose of anti aliasing filter is to limit the bandwidth of the signal to above the value of nyquist rate before sampling. The conditioned analog signal is processed by third module i.e. analog to digital converter which has a resolution of 16 bits at a sampling rate of 8000-20000 samples per second. Over-sampling is commonly used to reduce distortion introduced by practical digital to analog converters, such as zero order hold.

However, because of the large variability of the speech signal, it is better to perform some feature extraction that would reduce that variability. Particularly, eliminating various source of information, such as whether the speech signal is voiced or unvoiced and, if voiced, it eliminates the effect of the periodicity or pitch, amplitude of excitation signal and fundamental frequency etc.

Feature extraction involves analysis of speech signal. Features of the speech signal are highly variable with respect to time, feature extraction would reduce the variability of speech signal. The feature extraction techniques are classified as temporal analysis and spectral analysis techniques. In temporal analysis, the speech waveform itself is used for analysis. In spectral analysis, spectral representation of speech signal is used for analysis.

Speech signal is composed of excitation source and vocal tract system components. In order to analyze and model the excitation and system components of the speech independently and also use that in various speech processing applications, these two components have to be separated from the speech.

Cepstrum analysis is a tool for the detection of periodicity in a frequency spectrum. It provides methodology for separating the excitation e(n)*from the vocal tract shape Θ(n).

$s(n)=e(n)* \Theta (n)$ …(i)

The convolution makes it difficult to separate the two parts, therefore the cepstrum is introduced. The cepstrum is defined in the following way:

$C_s(n)= \xi^{-1}\{\log |\xi\{s(n)\}|\}$ *…(ii)*

where, ξ is the Discrete Time Fourier Transform and ξ^{-1} and is the Inverse Discrete Time Fourier Transform.

$S(w)=E(w)\ \Theta\ (w)$ *…(iii)*

$\log|S(w)|=\log|E(w)\ \Theta\ (w)|$….(iv)

$\log|S(w)|= \log|E(w)| + \log|\ \Theta\ (w)|$

$\log|S(w)|=C_e(w)+C_\Theta(w)$ …(v)

$C_s(n) = \xi^{-1}\ (C_e(w) + C_\Theta(w))$

$C_s(n) = \xi^{-1}\ \{Ce(w)\} + \xi^{-1}\ \{C\_(w)\}$

$C_s(n)=C_e(n)+C_\Theta(n)$ …(vi)

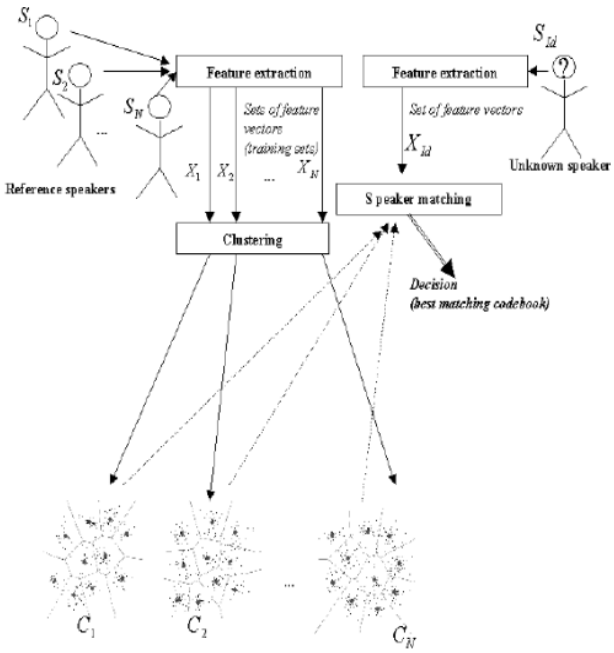

Fig 3.1 : Structure of VQ based Speaker Recognition System

In the Mel Frequency Cepstral Coefficients, the calculation of the Mel Cepstrum is same as the real Cepstrum except the Mel Cepstrum's frequency scale is warped to keep up a correspondence to the Mel scale. The Mel scale is mainly based on the study of observing the pitch or frequency perceived by the human. The scale is divided into the units mel. The mel scale is generally speaking a linear mapping below 1000 Hz and logarithmically spaced above. The mapping is usually done using an approximation (where f_{mel} is the perceived frequency in mels)[2]:

$f_{mel} = 2595 * log10(1 + f/700)$ *…(vii)*

Vector quantization (VQ) is a lossy data compression method based on the principle of block coding. In Vector Quantization, a large set of feature vectors are taken and a smaller set of measure vectors is produced which represents the centroids of the distribution. The vector distributions are defined over a high-dimensional space. In the current work, vectors are represented in 12-dimensional space. The set of

vectors stored in the rows of the feature matrix. Each row of the feature matrix represents a point in the 12- dimensional space. Each frame of the audio signal represents a point in the 12-dimensional space. The speech signal does not vary much in neighboring frames so all the similar frames in the silence are representing points in the 12-dimensional space which are very near to each other. Also the frames of the vowel part of the voice have similar characteristics as they have a continuous periodic signal. All the similar frames contain redundant information about the voice signal. Hence, all the redundant frames should be removed from the feature matrix to decrease the size of the matrix. The process of combining the nearby points in the space in to a single center is called Vector Quantization [5].

There are many algorithms available to achieve the goal of Vector Quantization. Clustering involves dividing a set of data points into non-overlapping groups or clusters. where points in a cluster are "more similar" to one another than to points in other clusters. When a dataset is clustered, every point is assigned to some cluster, and every cluster can be characterized by a single reference point, usually an average of the points in the cluster. All the N frames from the voice signal represent N points. This set of N points will be considered as the sample space for the further processing. k-Means algorithm is one of the simplest unsupervised learning algorithms. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.The k-Means algorithm aims to minimize the following error function.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Here j is the index for the clusters and i is the index for the points in a particular cluster. Cj is the center of the jth cluster. ||xi (j) – Cj||2 is the distance formula to find distance between the points xi and Cj.

The Algorithm

1. Place K points into the space represented by the vectors that are being clustered. These points represent initial group centroids.
2. Assign each vector to the group that has the closest centroid.
3. When all vectors have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the vectors into groups from which the metric to be minimized can be calculated.

The k-Means algorithm gives optimum solution in terms of minimum VQ distortion. Several iterations of k-Means algorithm are needed to converge in a optimum VQ dataset for large number of points in input dataset. k-Means algorithm implemented for two dimensions is shown in Fig 3.2. The diagrams show results during two iterations in the partitioning of 9 two-dimensional data points into two well separated clusters using the standard k-means algorithm. Points in cluster 1 are shown in red, points in cluster 2 are shown in black; data points are denoted by open circles and reference points by filled circles. Clusters are indicated by dashed lines. Note that, the iteration converges quickly to the correct clustering because the number of input data points is less in number.

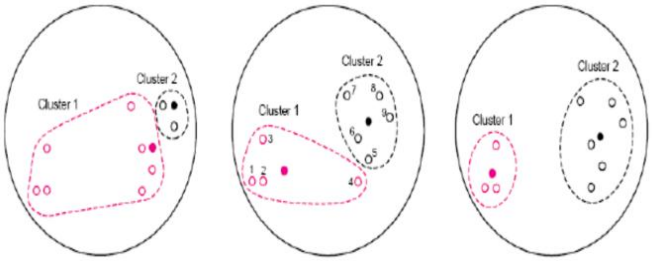

Fig 3.2 Standard k-Means algorithm in two dimensions

The K-means algorithm partitions the X feature vectors into M centroids. The algorithm first chooses M cluster centroids among the X feature vectors [7]. Then, each feature vector is assigned to the nearest centroid, and the new centroids are calculated. This procedure is continued until a stopping criterion is met, that is the mean square error between the feature vectors and the cluster-centroids is below a certain threshold or when there is no more change in the cluster-centre assignment. Typically, the value of N is 500 and the range for C is from 50 to 100. The final feature matrix is used as the voice profile of the speaker. To compare the profiles of two speakers, a Euclidian distance measurement algorithm is used.

Pattern Matching and Distance Measurement

The method used to compare two feature matrices is the measurement of the distortion distance between two vector sets by minimizing the Euclidian distance. The Euclidean distance is the distance between the two points that one would measure with a ruler, which can be proven by repeated application of the Pythagorean theorem in two dimensions [6]. The formula used to calculate the Euclidean distance in n dimensions is as follows.

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

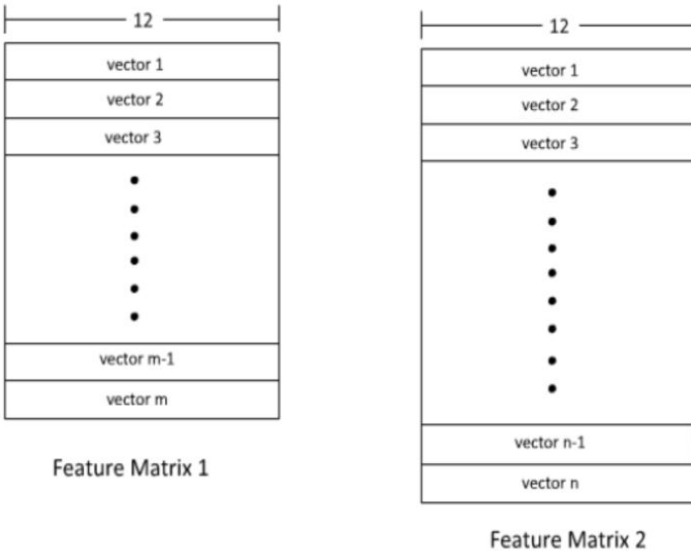

Fig 3.3 Feature Matrix Comparison

Two feature matrices to be compared using distance measurement algorithm is shown in Fig 3.3 . These matrices are the final features after application of the VQ algorithm. In a typical case, one matrix is generated by the real time recording of the voice at the time of verification and the other matrix is stored in the computer memory or in the system database. They differ in size even for the same speaker because the time of recording and the voice variations are different at registration and verification time.

The distance measurement algorithm (Euclidian distance algorithm to compare two matrices in MATLAB) is as follows :

1. Find Euclidian distance between each vector i of the matrix1 and every vector j of matrix 2 using the distance formula given above. Hence, we get a complete distance matrix of size m x n and having elements d_{ij} (d_{11}, d_{12} etc. up to d_{mn}). Here d_{ij} is the Euclidian distance between vector i of the matrix 1 and vector j of matrix 2 shown in Fig 3.3.
2. Now find a vector j in matrix 2 that is closest to vector i in matrix 1. To do this, find the minimum of each row of the distance matrix. Hence, we get a minimum distance vector of length m. Similarly for columns, we get a minimum distance vector of length n.
3. Find average of the two minimum distance vectors and calculate their mean value. The mean of the averages of the distance vectors will be the final distance between the two matrices.

The final distance calculated using the algorithm is the minimum Euclidian distance between the two vector sets. This distance shows the similarity between the two voice profiles. It is time invariant, that means the time shift of the same phrase in two voices does not affect the distance value. Hence it is used as the comparison metric for the voice profiles.

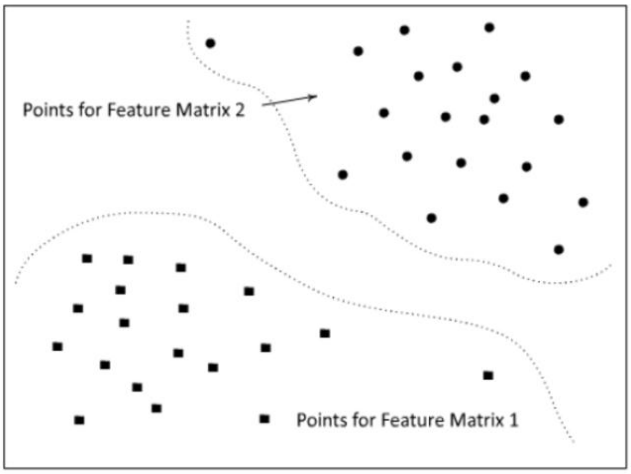

Fig : 3.4 Euclidean Distance Calculation for two feature Matrices

IV. Voice Profile comparison and Speaker Recognition

Two voice profiles can be verified by comparing the distance with a particular threshold value. The threshold value can be calculated at the training time by comparing the different sample voices of the same person.

4.1 Intra-speaker variations

At the time of registration, a system training session is implemented. During the verification of speaker, five different samples of the same person are trained with single voice password. Features are extracted and VQ algorithm is implemented on each of them. Resulting five feature matrices are compared against each other and Euclidean distances are calculated. The average of these distances is called the intra-speaker distance. This distance shows the amount of variation among the voices of the same speaker. The maximum of all the distances is the maximum detected for intra-speaker variation. This maximum distance variations among the speakers, is set as a threshold value, but there is no thumb rule to decide the threshold. There is always a tradeoff between false rejection and false acceptance mainly which depends on the threshold value. So in order to minimize false rejection and false acceptance after several iterations, maximum threshold value plus offset is set. This threshold value is saved in the system database at the time of registration along with the voice profile.

4.2 Comparison and Verification

After enrollment, the voice profile verification is performed as follows.. User enters the name and system asks the user for the password. The same quality/duration/loudness/pitch features are extracted from the submitted sample password and compared to the model of the claimed or hypothesized identity. System performs feature extraction and vector quantization operations on the real time recorded sample password speech signal. These methods all compare the similarities and differences between the input voice and the stored voice "states" to produce a recognition decision using the procedure

explained in section III. System also reads the threshold value stored in the database at the time of registration. The Euclidean distance is compared with the threshold value set. If the distance is less than the threshold, the user is accepted while if distance is more than threshold, the user is rejected.

V. SOFTWARE DESIGN

The "Speaker Recognition System" is divided into

phases:

1. Speaker Identification.

2. Speaker Verification.

i. Registration

When a biometric registration is needed for transactions between the user and speaker recognition system, the interested user need to be should enroll in the speaker recognition system. Initially, the user has to request for the service from speaker recognition system which causes its client to be redirected towards speaker recognition system for registration. The system creates a voice profile for the user by extracting features from the voice. Along with the voice profile and threshold, user's name, age, gender, id, and the date of registration are stored in the database.

ii. Verification

After registration, verification of the user is done for the probability of a fraudulent enrollment. The registered voice profile of the speaker is compared with a known voice profile stored in the database. The system temporarily records and saves the voice. The system recognizes the speaker, by comparing two speech samples, one that is used as reference which has been trained and stored in the database and the other that is collected during the test from the person who makes the claim. Now system extracts features from the voice recorded and measures the distance between the current voice profile and the saved profile using Mel Frequency Cepstral Coefficient, Vector Quantization and K-means algorithm.

Speaker Identification:

This system works in two phases : one is registration in which database is created and the another is Verification phase which is used for verification of the speaker.

Registration phase:

(i) When the user clicks on the " Create database " button, create database window opens and the user has to enter his name in the specified space .

Fig 5.1: Creating data base for Speaker Identification

(ii) After entering the basic details, the user has to record his/her voice in order to train the system as shown in Fig 5.2

Verification Phase:

When the user wants to check his/her identity in the any registered system, he/she has to enter the "Speaker Verification system", as shown in Fig 5.2.

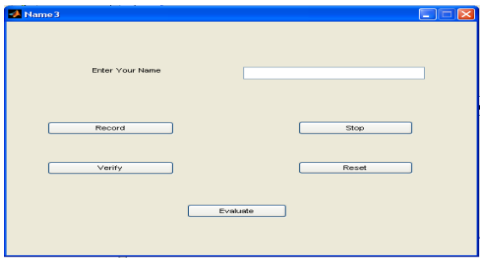

Fig 5.2: Recording and Training Window for Speaker Identification

(iii) (iii) Once the above dialog box is obtained, enter the identification name as registered during first phase, Then input the test voice signal by clicking on the "Record" button for the specified amount of time period by clicking on the "stop" button. Then click on the "verify" button to proceed for the identification process. The identity of the test voice is shown in the dialog box as shown in Fig 4.3. when the evaluation button is activated If we press the "Evaluate" button, then the performance of the speaker verification system is evaluated and is displayed as shown in Fig 5.2.

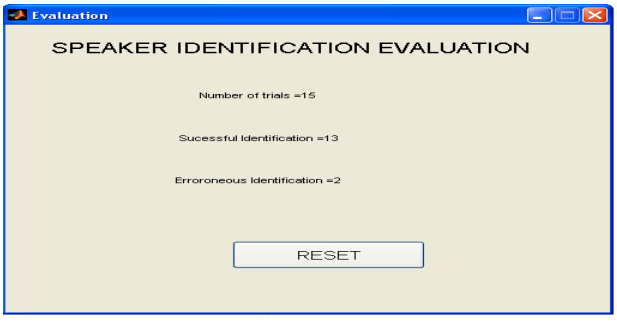

Fig 5.3: Speaker identification window

Fig 5.4: Result Window of Speaker Identification

Verification Phase:

In speaker recognition system, verification phase is the one in which the claimed identity of a speaker is verified based on the speech signal from the speaker**.**

Similar to the registration phase, Verification phase of the speaker recognition works in two phases namely :

1. Registration Phase
2. Verification Phase.

Registration phase:

Registration phase of speaker verification system is to identify a particular person . During initial configuration of registration phase is carried out during the training or enrollment person when each speaker to be verified by the system has to provide samples of speech which are then used to train the model for that speaker as shown in Fig 5.1.

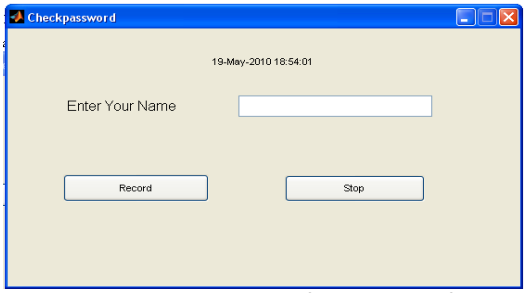

Fig 5.5 Check Password window for Speaker verification

when the individual has to make a claim as to who he/she is, and the system then proceeds to verify whether that claim is true or false. With speaker verification the speech of the unknown person is compared against both the claimed identity and against all other speakers (the imposter or background model(s)). The ratio of the two measures is then taken and compared to a threshold, if above the threshold the claim is accepted as true, if below, the claim is rejected as false as shown in Fig 5.1. Then the claimed person voice (or voice profile) is compared against the stored voice signal (or voice profile) of that person directly & and the result is displayed as shown below in Fig 5.5, if the claimed user voice does not match with the password stored in the database as shown in Fig 5.6.

Fig 5.6 : Result Dialog Box of Speaker Verification

VI. PERFORMANCE EVALUATION AND APPLICATIONS

All of us are aware of the fact that voices of different individuals do not sound in the same manner. Recent trends in speaker recognition have produced new tools that can be used to improve the performance and flexibility of speaker recognition. Speech is a useful parameter for identification because it is a product of the speaker's individual anatomy and linguistic background. In more specific, the speech signal produced by a given individual is affected by both the organic characteristics of the speaker (in terms of vocal tract geometry) and learned differences due to ethnic or social factors.

PERFORMANCE EVALUATION:

The human voice depends largely on many non-linear factors. Hence, the human vocal behavior can never be predicted with 100% accuracy. We tested our system on 10 different speakers for evaluating the efficiency of the system. For each person, we took five voice samples for training and 1 sample for testing. We have written a script in MATLAB to automatically test the performance of the system using the samples from these 10 people. The efficiency of the system turned out to be 95% on average. The results are as shown in Fig 6.1.

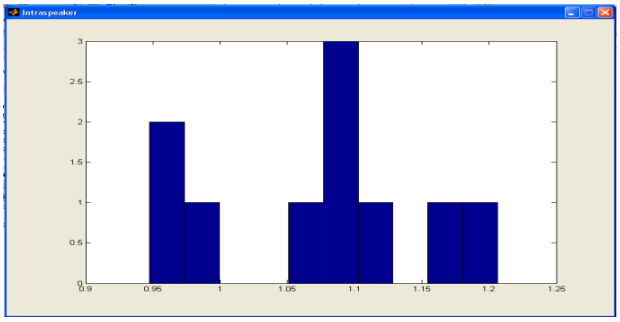

Fig 6.1: Performance Evaluation of Speakers

$$\frac{\text{num of successful recognition}}{\text{Total num of Trials}} * 100$$

These results are comparatively good. The effectiveness of this method is examined from the viewpoint of robustness against utterance variation such as differences in content, temporal variation, and changes in utterance speed

Applications:

1. The incidence of identity theft has increased rapidly in recent years, and is an issue that worries everyone. The result is that, to complete many everyday transactions, without any probability of fraudulent activities. This above function is performed by this system successfully.
2. A major aspect of our work is access control. On a security gateway, for a person to be allowed to pass through, his identity needs to be authenticated. If our system is installed there, the authentication process can be done.
3. Speaker recognition system can be used for remote authentication due to the availability of devices for collecting speech samples from microphones, telephone network etc.
4. One of the challenging and demanding areas of speaker verification application is in forensics. Usually in the cases where the crime has been committed and it has to be verified from a recorded speech signal. Speakers voice can be identified by comparing with the spectrograms. Spectrograms can be generated using Discrete Fourier Transform . But the accuracy of these methods found to be not reliable and effective. To prove the suspect to be criminal it needs to be verified beyond reasonable doubt. Hence reliable and automatic speaker verification system is desired.

VII .REFERENCES

[1] A. Subramanya, Z. Zhang, A. Surendran, P. Nguyen, M. Narasimhan, and A. Acero, "A Generative-Discriminative Framework Using Ensemble Methods For Text-Dependent Speaker Verification", in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2007), Vol.IV, pages 225-228, Honolulu, Hawaii, April 15-20, 2007

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models, "Digital Signal Processing, vol. 10(1–3), pp. 19-41, 2000.

[3] R.V Pawar, P.P.Kajave, and S.N.Mali," Speaker Identification using Neural Networks", World Academy of Science, Engineering and Technology 12 2005.

[4] Heck, L. P. and Weintraub, M., Handset-dependent background models for robust textindependent speaker recognition. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, April 1997, pp. 1071–1073.

[5] B.C. Haris, G. Pradhan, A. Misra, S. Shukla, R. Sinha and S.R.M Prasanna, "Multi-variability Speech Database for Robust Speaker Recognition", In Proc. NCC, pp. 1-5, 2011

[6] Z. Xiaojia, S. Yang and W. DeLiang, "Robust speaker identification using a CASA front-end", In Proc. ICASSP-2011, pp.5468-5471, 2011.

[7] L. Heck and N. Mirghafori, "On-Line Unsupervised Adaptation in Speaker Verification", In Proc. ICSLP-2000, vol. 2, pp. 454-457, Beijing, China, 2000