ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 6, Issue 9, September 2019

Text Independent Speaker Modeling and Identification Based On MFCC Features

Khurrath-ul-aien M.R[1]

Department of Computer Science and Engineering, UBDTCE, Davangere, Karnataka, India[1]

ABSTRACT: In this gives an overview of automatic speaker recognition technology, with an emphasis on text-independent recognition. Speaker recognition has been studied actively for several decades. We give an overview of both the classical and the state-of-the-art methods. We start with the fundamentals of automatic speaker recognition, concerning feature extraction and speaker modeling. Here, describe a Gaussian Mixture Model Universal Background Model (GMM-UBM) speaker identification system. In this GMM-UBM system, we derive the hypothesized speaker model by adapting the parameters of UBM using the speaker's training speech and a form of Bayesian adaptation. The UBM technique is incorporated into the GMM speaker identification system to reduce the time requirement for recognition significantly.We elaborate advanced computational techniques to address robustness and session variability. In text- dependent system, the words or phrases used for verification are known beforehand and are fixed. In a text-independent system there are no constraints on the words or phrases used during verification system. The recent progress from vectors towards supervectors opens up a new area of exploration and represents a technology trend.

KEYWORDS: Discriminative models ; Feature extraction; Text-independence ;Speaker recognition; Statistical models; Supervectors

I. INTRODUCTION

Speaker recognition refers to recognizing persons from their voice. No two individuals sound identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on. State-of-the-art speaker recognition systems use a number of these features in parallel, attempting to cover these different aspects and employing them in a complementary way to achieve more accurate recognition. In addition to telephony speech data, there is a continually increasing supply of other spoken documents such as TV broadcasts, teleconference meetings, and video clips from vacations. Extracting metadata like topic of discussion or participant names and genders from these documents would enable automated information searching and indexing. Speaker diarization also known as ''who spoke when", attempts to extract speaking turns of the different participants from a spoken document, and is an extension of the ''classical" speaker recognition techniques applied to recordings with multiple speakers.In forensics and speaker diarization, the speakers can be considered non-cooperative as they do not specifically wish to be recognized. On the other hand, in telephone-based services and access control, the users are considered cooperative.

Speaker recognition systems, on the other hand, can be divided into text-dependent and text-independent ones.

- Text-dependent systems suited for cooperative users, the recognition phrases are fixed, or known beforehand. For instance, the user can be prompted to read a randomly selected sequence of numbers as described in. In text-independent systems, there are no constraints on the words which the speakers are allowed to use. Thus, the reference (what are spoken in training) and the test (what are uttered in actual use) utterances may have completely different content, and the recognition system must take this phonetic mismatch into account.
- Text-independent recognition is the much more challenging of the two tasks. In general, phonetic variability represents one adverse factor to accuracy in text-independent speaker recognition. Changes in the acoustic environment and technical factors (transducer, channel), as well as ''within-speaker" variation of the speaker him/herself (state of health, mood, aging) represent other undesirable factors. In general, any variation between two recordings of the same speaker is known as session variability.

The basic goal of our project is to recognize and classify the speeches of different persons. This classification is mainly based on extracting several key features like Mel Frequency Cepstral Coefficients (MFCC) from the speech signals of those persons by using the process of feature extraction using MATLAB. The above features may consists of pitch, amplitude, frequency etc. It can be achieved by using tools like MATLAB. Using a statistical model like Gaussian mixture model (GMM) and features extracted from those speech signals we build a unique identity for each person who enrolled for speaker recognition

Gaussian mixture models (GMM) for the modeling of speaker spectral characteristics has become the dominant approach for speaker identification systems which use untranscribed training data. A speaker model based on Gaussian Mixture Model-Universal Background Model (GMMUBM) is introduced into text-independent speaker identification. Our work focuses on applications which require high identification rates using short utterance from unconstrained (text-independent) conversational speech and robustness to degradations produced by transmission over a telephone channel.

A Gaussian Mixture Model Universal Background Model (GMM-UBM) speaker identification system. In this GMM-UBM system, we derive the hypothesized speaker model by adapting the parameters of UBM using the speaker's training speech and a form of Bayesian adaptation. The UBM technique is incorporated into the GMM speaker identification system to reduce the time requirement for recognition significantly.

II. LITERATURE SURVEY

- **Alexander, A., Botti, F., Dessimoz, D., Drygajlo** are proposed An overview of text-independent speaker recognition: From features to supervectors. In this, we analyse mismatched technical conditions in training and testing phases of speaker recognition and their effect on forensic human and automatic speaker recognition. We use perceptual tests performed by non-experts and compare their performance with that of a baseline automatic speaker recognition system. The degradation of the accuracy of human recognition in mismatched recording conditions is contrasted with that of the automatic system under similar recording conditions. The conditions considered are of public switched telephone network (PSTN) and global system for mobile communications (GSM) transmission and background noise. The perceptual cues that the human subjects use to perceive differences in voices are studied along with their importance in different conditions. We discuss the possibility of increasing the accuracy of automatic systems using the perceptual cues that remain robust to mismatched conditions. We estimate the strength of evidence for both humans and automatic systems, calculating likelihood ratios using the perceptual scores for humans and the log-likelihood scores for automatic systems.

  Advantages: The accuracy of human recognition in mismatched recording conditions is contrasted with that of the automatic system under similar recording conditions.
  Disadvantages: This system is not efficient and adaptive.

- **Besacier, L., Bonastre**, are proposed the Sub band architecture for automatic speaker recognition Signal Process. Here, present an original approach for automatic speaker identification especially applicable to environments which cause partial corruption of the frequency spectrum of the signal. The general principle is to split the whole frequency domain into several sub bands on which statistical recognizers are independently applied and then recombined to yield a global score and a global recognition decision. The choice of the sub band architecture and the recombination strategies are particularly discussed. This technique had been shown to be robust for speech recognition when narrow band noise degradation occurs. We first objectively verify this robustness for the speaker identification task. We also study which information is really used to recognize speakers. For this, speaker identification experiments on independent sub bands are conducted for 630 speakers, on TIMIT and NTIMIT databases. The results show that the speaker specific information is not equally distributed among sub bands. In particular, the low-frequency sub bands (under 600Hz) and the high-frequency sub bands (over 3000Hz) are more speaker specific than middle-frequency ones. In addition, experiments on different sub band system architectures show that the correlations between frequency channels are of prime importance for speaker recognition. Some of these correlations are lost when the frequency domain is divided into sub bands. Consequently we propose a particularly redundant parallel architecture for which most of the correlations are kept. The performances obtained with this new system, using linear recombination strategies, are equivalent to those of a conventional full band recognizer on clean and telephone speech. Experiments on speech corrupted by unpredictable noise show a better adaptability of this approach in noisy environments, compared to a conventional device, especially when pruning of some recognizers is performed.

Advantages: This Experiment on speech corrupted by unpredictable noise show a better adaptability of this approach in noisy environments.
Disadvantages: This not efficient and flexible.

- **Besacier, L., Bonastre, J., Fredouille**, are proposed Localization and selection of speaker-specific information with statistical modeling**.** Statistical modelling of the speech signal has been widely used in speaker recognition. The performance obtained with this type of modelling is excellent in laboratories but decreases dramatically for telephone or noisy speech. Moreover, it is difficult to know which piece of information is taken into account by the system. In order to solve this problem and to improve the current systems, a better understanding of the nature of the information used by statistical methods is needed. This knowledge should allow to select only the relevant information or to add new sources of information. The first part of this paper presents experiments that aim at localizing the most useful acoustic events for speaker recognition. The relation between the discriminant ability and the speech's events nature is studied. Particularly, the phonetic content, the signal stability and the frequency domain are explored. Finally, the potential of dynamic information contained in the relation between a frame and its *p* neighbours is investigated. In the second part, the authors suggest a new selection procedure designed to select the pertinent features. Conventional feature selection techniques (ascendant selection, knock-out) allow only global and a posteriori knowledge about the relevance of an information source. However, some speech clusters may be very efficient to recognize a particular speaker, whereas they can be non-informative for another one. Moreover, some information classes may be corrupted or even missing for particular recording conditions. This necessity for speaker-specific processing and for adaptability to the environment (with no a priori knowledge of the degradation affecting the signal) leads the authors to propose a system that automatically selects the most discriminant parts of a speech utterance. The proposed architecture divides the signal into different time–frequency blocks. The likelihood is calculated after dynamically selecting the most useful blocks. This information selection leads to a significative error rate reduction (up to 41% of relative error rate decrease on TIMIT) for short training and test durations. Finally, experiments in the case of simulated noise degradation show that this approach is a very efficient way to deal with partially corrupted speech.

  Advantages: It is a very efficient way to deal with partially corrupted speech. This information selection leads to a significative error rate reduction (up to 41% of relative error rate decrease on TIMIT) for short training and test durations.
  Disadvantages: It has some noise degradation and significative error.

- **Bimbot, Ivan Magrin-Chagnolleau and Luc Mathan** are proposed Second-Order Statistical Measures for Text-Independent Speaker Identication. This article presents an overview of several measures for speaker recognition. These measures relate to second-order statistical tests, and can be expressed under a common formalism. Alternate formulations of these measures are given and their mathematical properties are studied. In their basic form, these measures are asymmetric, but they can be symmetrized in various ways. All measures are tested in the framework of text-independent closed-set speaker identication, on 3 variants of the TIMIT database (630 speakers): TIMIT (high quality speech), FTIMIT (a restricted bandwidth version of TIMIT) and NTIMIT (telephone quality). Remarkable performances are obtained on TIMIT but the results naturally deteriorate with FTIMIT and NTIMIT. Symmetrization appears to be a factor of improvement, especially when little speech material is available. The use of some of the proposed measures as a reference benchmark to evaluate the intrinsic complexity of a given database under a given protocol is finally suggested as a conclusion to this work.

  Advantages: It is a high quality and the results naturally deteriorate with FTIMIT and NTIMIT.
  Disadvantages: This is difficult to implementation.

III. SCOPE OF RESEARCH

Assessing the performance of new algorithms on a common dataset is essential to enable meaningful performance comparison. In early studies, corpora consisted of a few or at the most a few dozen speakers, and data was often self collected. Recently, there has been significant effort directed towards standardizing the evaluation methodology in speaker verification. NIST evaluations include test trials under both matched conditions such as telephone only, and unmatched conditions such as language effects (matched languages vs unmatched languages), cross channel and two-speaker detection. During the evaluation, NIST releases a set of speech files as the development data to the participants. At this initial phase, the participants do not have access to the ‘‘ground truth”, that is, the speaker labels. Each participating group then runs their algorithms ‘‘blindly” on the given data and submits the recognition scores and verification decisions. NIST then evaluates the performances of the submissions and the results are discussed in a follow-up workshop. The use of ‘‘blind” evaluation data makes it possible to conduct an unbiased comparison of the various algorithms.

These activities would be difficult without a common evaluation dataset or a standard evaluation protocol. Visual inspections of the detection error trade-off (DET) curves and equal error rate (EER) are commonly used evaluation tools in the speaker verification literature. The problem with EER is that it corresponds to an arbitrary detection threshold, which is not a likely choice in a real application where it is critical to maintain the balance between user convenience and security. NIST uses a detection cost function (DCF) as the primary evaluation metric to assess speaker verification performance.

Minimum DCF (MinDCF), defined as the DCF value at the threshold for which is smallest, is the optimum cost. When the decision threshold is optimized on a development set and applied to the evaluation corpus, this produces actual DCF. Therefore, the difference between the minimum DCF and the actual DCF indicates how well the system is calibrated for a certain application and how robust is the threshold setting method. For an in-depth and thorough theoretical discussion as well as the alternative formulations of application-independent evaluation metrics. While the NIST speaker recognition benchmarking considers mostly conversational text-independent speaker verification in English, there have been a few alternative evaluations, for instance the NFI-TNO evaluation which considered authentic forensic samples (mostly in Dutch), including wiretap recordings.

This evaluation included open-set speaker identification and text-dependent verification tasks in addition to text-independent verification. Some of the factors affecting speaker recognition accuracy in the NIST. It is widely known that cross-channel training and testing display a much lower accuracy compared to that with same channel. Including different handsets in the training material also improves recognition accuracy. Another factor significant to performance is the duration of training and test utterances. The greater the amount of speech data used for training and/ or testing, the better the accuracy. Training utterance duration seems to be more significant than test segment duration.

III. SYSTEM DESIGN

The components of an automatic speaker recognition system. The upper is the enrollment process, while the lower panel illustrates the recognition process. The feature extraction module first transforms the raw signal into feature vectors in which speaker-specific properties are emphasized and statistical redundancies suppressed. In the enrollment mode, a speaker model is trained using the feature vectors of the target speaker. In the recognition mode, the feature vectors extracted from the unknown person’s utterance are compared against the model(s) in the system database to give a similarity score. The decision module uses this similarity score to make the final decision.

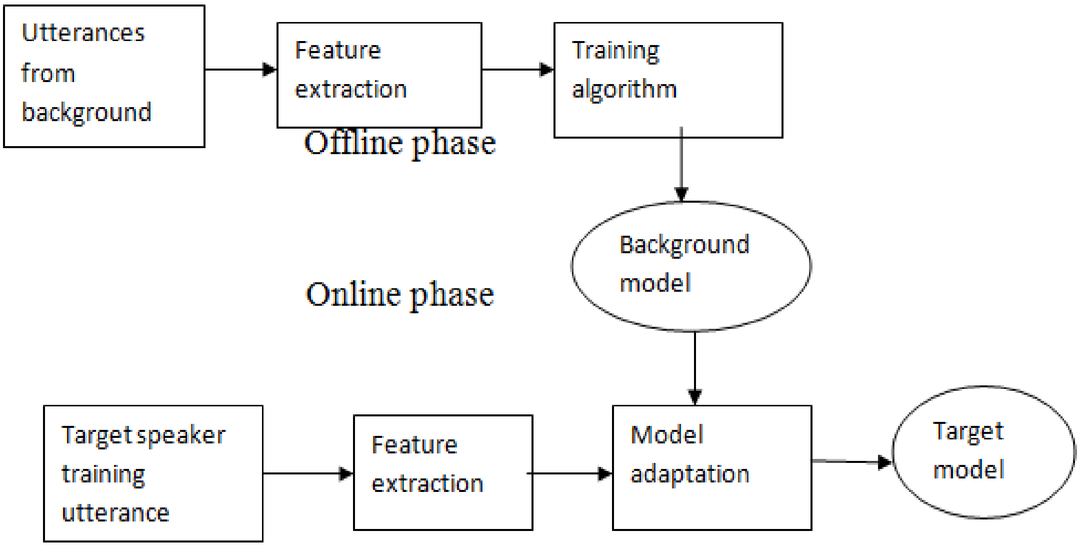


Fig.1. Speaker Enrollment

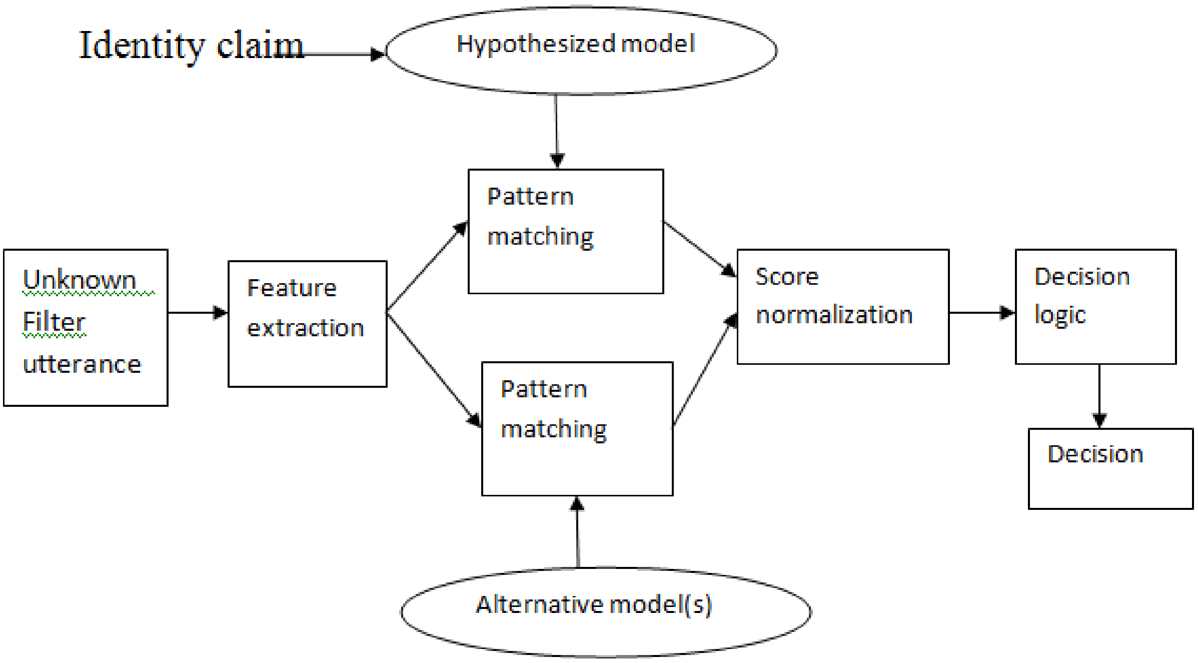


Fig.2. Speaker Verification /Identification

Virtually all state-of-the-art speaker recognition systems use a set of background speakers or cohort speakers in one form or another to enhance the robustness and computational efficiency of the recognizer. In the enrollment phase, background speakers are used as the negative examples in the training of a discriminative model. In the recognition phase, background speakers are used in the normalization of the speaker match score.

IV. FEATURE EXTRACTION

The speech signal continuously changes due to articulatory movements, and therefore, the signal must be broken down in short frames of about 20–30 ms in duration. Within this interval, the signal is assumed to remain stationary and a spectral feature vector is extracted from each frame. Usually the frame is pre-emphasized and multiplied by a smooth window function prior to further steps. Pre-emphasis boosts the higher frequencies whose intensity would be otherwise very low due to downward sloping spectrum caused by glottal voice source.

The window function (usually Hamming), on the other hand, is needed because of the finite-length effects of the discrete Fourier transform. In practice, choice of the window function is not critical. The well-known fast Fourier transform (FFT), a fast implementation of DFT, decomposes a signal into its frequency components. Alternatives to FFT-based signal decomposition such as non-harmonic bases, aperiodic functions. The DFT, however, remains to be used in practice due to its simplicity and efficiency. Usually only the magnitude spectrum is retained, based on the belief that phase has little perceptual importance. However, provides opposing evidence while described a technique which utilizes phase information. The global shape of the DFT magnitude spectrum known as spectral envelope, contains information about the resonance properties of the vocal tract and has been found out to be the most informative part of the spectrum in speaker recognition.

A simple model of spectral envelope uses a set of bandpass filters to do energy integration over neighboring frequency bands. Motivated by psycho-acoustic studies, the lower frequency range is usually represented with higher resolution by allocating more filters with narrow bandwidths.

Score Normalization:

In score normalization, the ‘‘raw” match score is normalized relative to a set of other speaker models known as cohort. The main purpose of score normalization is to transform scores from different speakers into a similar range so that a common (speaker-independent) verification threshold can be used. Score normalization can correct some speaker-dependent score offsets not compensated by the feature and model domain methods.

Pattern Matching and Decision:

The pattern matching module deals with comparison between the estimated features to the speaker models. Some of the pattern matching methods used in speaker recognition include Hidden markov models (HMM), dynamic time warping (DTW), neural networks and vector quantization (VQ). In case of verification, this module provides an expert with a similarity score between the test sample and the claimed identity.

While, in case of identification, the module gives similarity score between the test sample and all the available samples in the database. The evaluation of these scores is done using decision module and the results are accordingly presented. The effectiveness of a speaker recognition system is measured differently for different tasks. Since the output in identification system is a speaker identity from a set of known speakers, the identification accuracy is used to measure the performance. For the verification systems, two types of error can be observed: false acceptance of an impostor and false rejection of a target speaker.

V. WORK FLOW OF VQ AND GMM

A. *Vector Quantization:*

Vector quantization (VQ) model also known as centroid model, is one of the simplest text-independent speaker models. And its roots are originally in data compression Even though VQ is often used for computational speed-up techniques and lightweight practical implementations, it also provides competitive accuracy when combined with background model adaptation. For computational reasons, however ,the number of vectors is usually reduced by a clustering method such as K-means .This gives a reduced set of vectors known as codebook .The choice of the clustering method is not as important as optimizing the codebook size.

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 6, Issue 9, September 2019

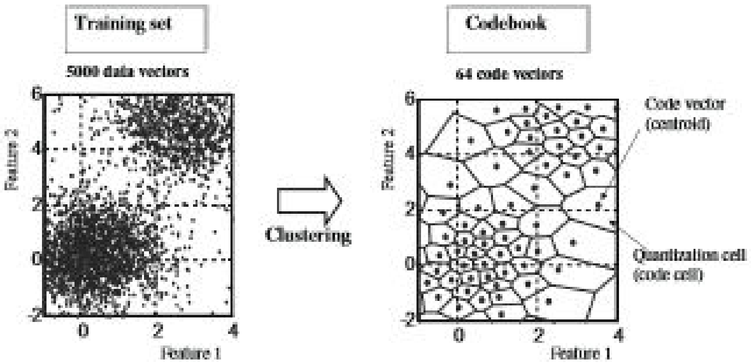


Fig.3. Codebook construction for vector quantization using the K-means algorithm. The original training set consisting of 5000 vectors is reduced to a set of K =64 code vectors (centroids)

B. Gaussian Mixture Model:

Gaussian mixture model (GMM) is a stochastic model which has become the de facto reference method in speaker recognition. The GMM can be considered as an extension of the VQ model, in which the clusters are overlapping. That is, a feature vector is not assigned to the nearest cluster as in , but it has a nonzero probability of originating from each cluster.

A GMM is composed of a finite mixture of multivariate Gaussian components.

$$p(\boldsymbol{x}|\lambda) = \sum_{k=1}^{K} P_k \, \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

In, K is the number of Gaussian components, P_k is the prior probability (mixing weight) of the kth Gaussian component, and

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}_k|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

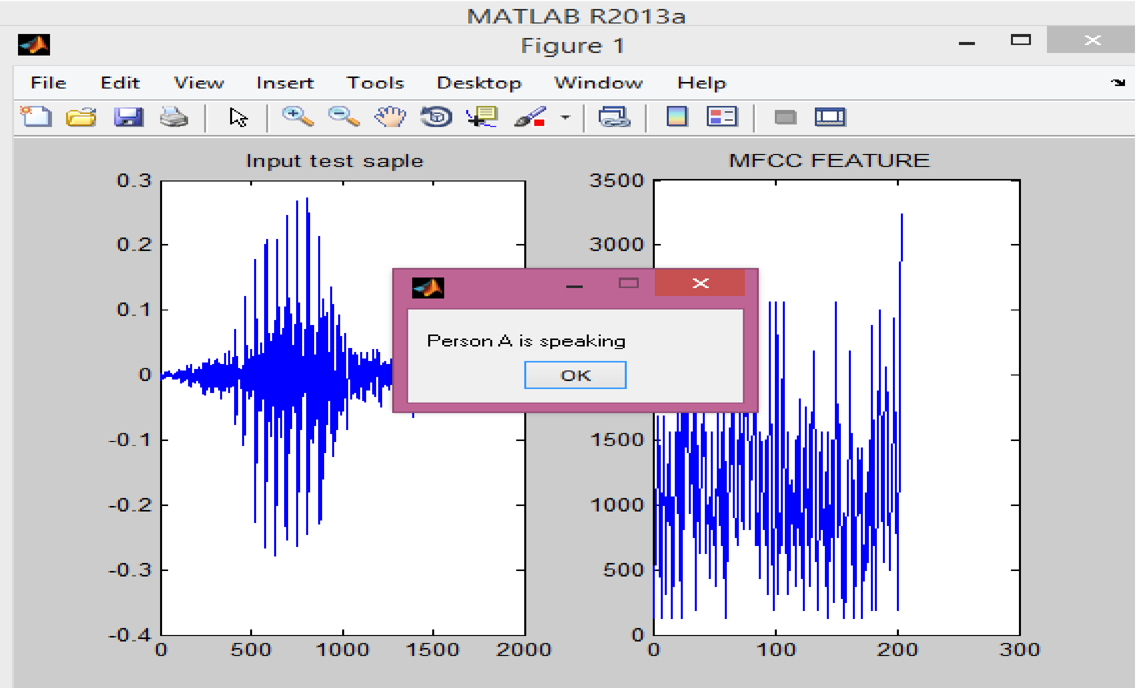
is the d-variate Gaussian density function with mean vector μ_k and covariance matrix $\sum_k$. The prior probabilities P_k**Error! Reference source not found.**0 constrained as $\sum_{k=1}^{K} P_k = 1$.

Training a GMM consists of estimating the parameters $\lambda = \{P_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ from a training sample.The basic approach is maximum likelihood (ML) estimation. The average log-likelihood of x with respect to model λ is defined as,

$$\mathrm{LL}_{\mathrm{avg}}(\mathcal{X}, \lambda) = \frac{1}{T}\sum_{t=1}^{T}\log\sum_{k=1}^{K} P_k \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

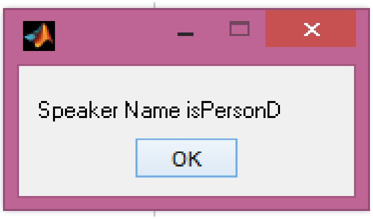
The higher the value, the higher the indication that the unknown vectors originate from the model λ. The popular expectation–maximization (EM) algorithm can be used for maximizing the likelihood with respect to a given data. Note that K-means can be used as an initialization method for EM algorithm

VI. EXPERIMENTAL RESULTS



Person A is speaking

Fig.4. Waveforms Generated By Using VQ



Speaker Name isPersonD

Fig.5. Identifying Person Using GMM

REFERENCES

1. Alexander, A., Botti, F., Dessimoz, D., Drygajlo, A., 2004. The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. Forensic Science International 146S, December 2004, pp. 95–99.
2. Besacier, L., Bonastre, J.-F., 2000. Subband architecture for automatic speaker recognition. Signal Process. 80, 1245–1259. Besacier, L., Bonastre, J., Fredouille, C., 2000. Localization and selection of speaker-specific information with statistical modeling. Speech Comm. 31, 89–106.
3. Besacier, L., Bonastre, J., Fredouille, C., 2000. Localization and selection of speaker-specific information with statistical modeling. Speech Comm. 31, 89–106.
4. Bimbot, F., Magrin-Chagnolleau, I., Mathan, L., 1995. Second-order statistical measures for text-independent speaker identification. Speech Comm. 17, 177–192.
5. Deller, J., Hansen, J., Proakis, J., 2000. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York.
6. Reynolds, D., Rose, R., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3, 72–83.

7. Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted gaussian mixture models. Digital Signal Process. 10 (1), 19–41

8. Soong, F.K., Rosenberg, A.E., Juang, B.-H., Rabiner, L.R., 1987. A vector quantization approach to speaker recognition. AT & T Technical J. 66, 14–26.