Speaker Recognition in Farsi Language

Marjan. Shahchera

***Abstract*—**Speaker recognition is the process of identifying a person with his voice. Speaker recognition includes verification and identification.

***Keywords*—**Speaker Recognition, Vector Quantization (VQ), Mel Frequency Cepstrum Coefficient (MFCC) .

I. Introduction

Biometrics is defined as the method of establishing the identity of an individual based on his or her physiological and behavioral characteristics. Table 1 describes the various types of biometrics being used today. Speech is a performance biometric where the user has to perform a task in order to be authenticated. It is referred to in literature by various terms such as speaker identification, talker identification, talker authentication etc. It is a common mistake to fail to distinguish between speech recognition and speaker recognition. In speech recognition the problem is to identify spoken words irrespective of the person speaking it. In this scenario we are trying to capture the similarities of the words spoken. On the other hand in speaker recognition we are trying to distinguish between the speakers based on how they speak those words. In this scenario we are trying to enhance the differences in the way the words are spoken. This difference arises based on the shape of the vocal tracts of the person and the speech habits and accent acquired over the years. Speaker recognition is a broad problem and includes both speaker verification and identification. In speaker verification, the user claims an identity and the claim is verified by means of his voice. In speaker identification the identity of the person is not known. Given a sample of speech, it has to be matched with the speech samples already in the database, in order to come up with a possible identity. It must be noted that the user may or may not actually exist in the database. This is known as an 'open set' identification problem.

II. Procedure for Speaker Recognition

A. Overview

Section I of this report, provides an introduction to the problem of speaker recognition. This includes the problem description, an overview of the mechanism of speech production, modeling of speaker characteristics and pattern matching. This section also briefly describes the previous work in this area. Section II of the report discusses the implementation of the speaker model and the application of Cepstral analysis. The procedure for extracting the cepstrum coefficients and the derived features are explained. Section III discusses the problems of matching two speaker models using Dynamic Time Warping algorithm.

TABLE I
TYPES OF BIOMETRICS

Type	Examples
Physical	Fingerprints Iris Retinal Pattern Hand Geometry Face
Behavioral	Speech Signature Gait
Chemical	DNA Body odor

B. Problem Description

Speech is a complex and non stationary signal produced as a result of several and complex transformations and processes [1]. The mechanism of the production of speech is discussed in the next section. The speaker specific characteristics of speech arise principally due to the physiological differences in the speech production mechanisms and secondly due to the accent and speaking habits of the user. At the time of enrollment, speech sample is acquired in a controlled and supervised manner from the user. The speaker recognition system has to process the speech signal in order to extract speaker discriminatory information from it. This discriminatory information will form the **speaker model.** This model can either be stochastic, statistical or simply a template [1]. The model must have high inter speaker variability and low intra speaker variability. At the time of verification a speech sample is acquired from the user. This sample may be of short duration or taken under uncontrolled condition. The claimed identity is known in case of verification and is not known in case of identification. The recognition system has to extract the features from this sample and compare it against the models already stored before hand. This is a **pattern matching** task. There are several methods to compare and match the extracted features such as Euclidian distance, Hidden Markov Models, Gaussian Mixture Models etc. A generic speaker recognition system is shown in Fig 1.There are several variations to the process of speaker recognition. As discussed, the scenario may be that of verification or

Marjan1. Shahchera1 is the Master student in communication engineering in Islamic Azad University of Science and Technology Azarbayejansharghi , Tabriz, IRAN (corresponding author to provide phone:
00989133258307 ; e-mail: m_shahchera@yahoo.com).

recognition. The speaker may be co-operative or non co-operative with the system. The speech may be acquired in controlled condition (high quality) or
uncontrolled conditions (low quality). The recognition may be text dependent or text independent. The project will mainly discuss the process of speaker recognition in the case where the recognition is text dependent, and the speech is of high quality.

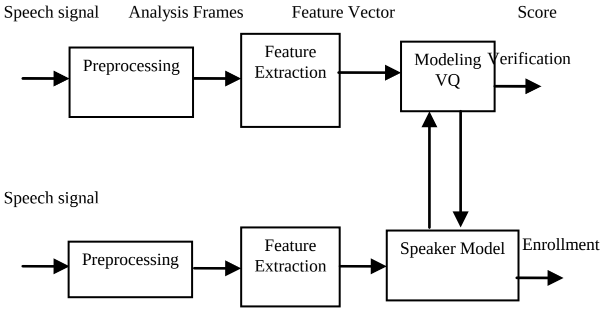

Fig. 1 Generic Speaker Recognition System

C. Speech Production

Speech is produced by the modulation of air flow by the vocal tracts. Therefore the vocal tract shape is the primary distinguishing parameter that affects the speech of a person. The vocal tract is shown in Fig 2. Refer to [1] for a more detailed description of the organs.

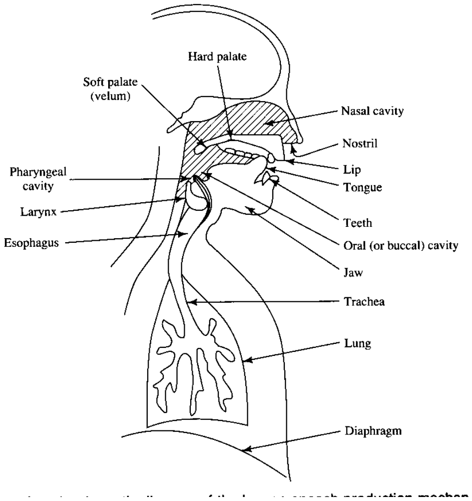

Fig 2 The Human Speech System

The vocal tract modifies the spectral contents of the speech signal as it passes through it. For each individual the vocal tract resonates at different frequencies. These frequencies are called **formants**. The shape of the vocal tract can be estimated or parameterized based on the spectral characteristics and location of the formants. The human speech system is driven by an excitation source, namely airflow from the lungs. The path of the air flow through the vocal tract modulates the signal to a significant level. The extent of modulation depends on the user. The excitation fall into the following categories:

1. *Phonation*

   This is also called as *voiced* speech. Phonation excitation occurs when the air flow is modulated by the vocal chords. E.g. /u/ in food
2. *Whispering*

   Whispering is produced when the air flows through the small triangular opening at the rear of the nearly closed vocal chords
3. *Frication*

   Frication is produced by constrictions in the vocal tract. E.g. /s/ in sleep.
4. *Compression*

   Compression is characterized by small sudden burst of air flow. E.g. /b/ in boom
5. *Vibration*

   Vibration is produced when air is forced through other closures apart from the vocal chords such as lips. E.g. /r/ in root

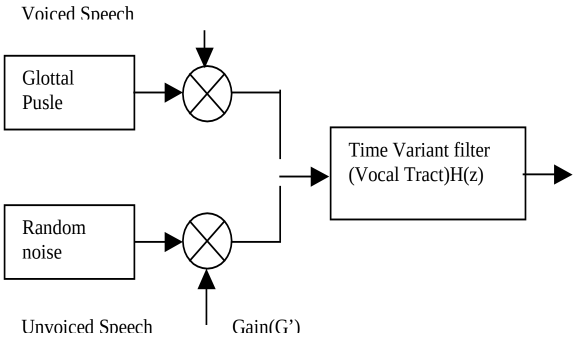

Fig. 3 Simplified Signal model for human speech

Fig 3 shows a simplified signal model for the human speech system. A more detailed model is discussed in [4]. The excitation source may be periodic as in the case of voiced speech or may be random as in the case of unvoiced speech. The vocal tract is modeled as a time variant filter that modulates the signal from either excitation source. The purpose of the speaker recognition system is to accurately model the filter characteristics.

D. Speaker Modeling

Speaker modeling involves the representation of an utterance as a sequence of feature vectors. The model must have high inter speaker variability and low intra speaker variability. Utterance spoken by the same person but at different times result in a similar yet different sequence of feature vectors. The purpose if this modeling is to capture these variations in the extracted set of features. There are two types of speech models, stochastic and template models that are used in speaker recognition. The stochastic model treats the speech production process as a parametric random process and assumes that the parameters of the underlying stochastic process can be estimated in a precise, well defined manner.

Hidden Markov Model is a very popular stochastic model. The template model attempts to model the speech production process in a non-parametric manner by retaining a number of sequences of feature vectors derived from multiple utterances of the same word by the same person[3]. The project uses the template model using normalized cepstral coefficients.

E. Pattern Matching

In speaker verification pattern matching is the task of computing scores, which is the measure of the similarity of the input with the model. The speakers are enrolled into the system and a model of their voice based on the extracted features are generated and stored (possibly on a smart card). The matching algorithm compares extracted features of the input voice with the model of the claimant based on which the user's authentication is decided. For stochastic models the pattern matching is probabilistic and results in a measure of likelihood, or conditional probability, of observation of the given model. For template model the scores is dependent on the Euclidean distance between the input and the model. Dynamic time warping which is a text-dependent template model has been used for the project.

F. Previous work

Some of the previous work using cestrum coefficients used for speaker recognition along with their error factor has been listed in Table II.

TABLE II
PREVIOUS WORK

SL	**Author**	**Feature**	**Method**	**Err**
1	Atal	Cepstrum	Pattern Matching	i:2%@0.5s v:2%@1s
2	Furui	Normalized Cepstrum	Pattern Matching	v:0.2%@3s
3	Li & Wrench	LP, Cepstrum	Pattern Matching	i:21%@3s i:4%@10s
4	Higgins & Wohlford	Cepstrum	DTW Likelihood Scoring	i:10%@2.5s i:4.5%@10s
5	Higgins	LAR, LP-Cepstrum	DTW Likelihood Scoring	v:1.7@10s

III. Preprocessing Subsystem

The input of system is the voice that recording via microphone. The output vector include sampling and filtering signals. The signal preprocessing system is shown in fig 4.

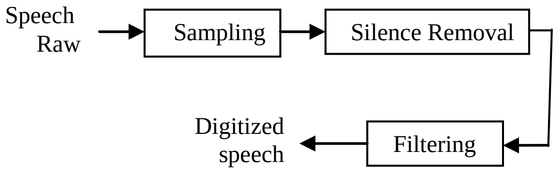

Fig. 4 Signal preprocessing system

A. Sampling subsystem

This subsystem is sampling from raw and analog speech signal with bit rate of 200kbps.

B. Silence removal subsystem

With removing the silence in the beginning and end of speech signal the process of recognition is very improved because of decreasing the volume of frames . denotation of silence normally based on features of the signals. One of the best ways is short-term analysis of power of signal with the equation (1).

$$E_n = \sum_{k=1}^{Wn} |x[k]|^2 w[n-k]$$

$$E_{avg} = \frac{1}{N}\sum_{k=1}^{N} |x[k]|^2 \qquad (1)$$

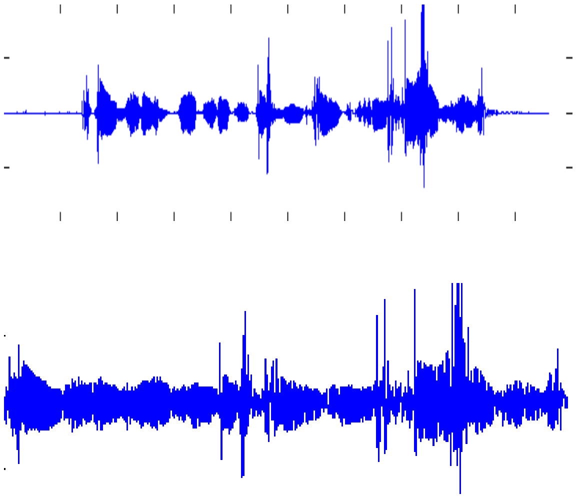

Fig. 5 Silence removal subsystem

C. Filtering Subsystem

The special filter is used for restriction of sensitivity of signal in the way of process with the equation H(z) =(1−0.95z−1).

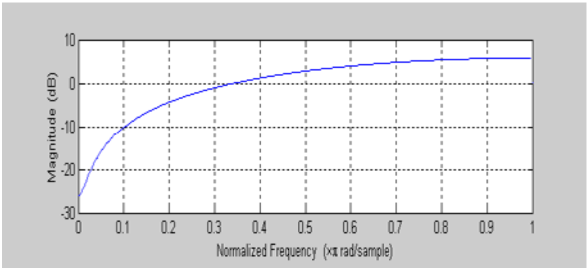

Fig 6 Filter diagram

IV. FEATURE EXTRACTION SUBSYSTEM

The input of subsystem is digitalized signal of speech that represent as vector of all sampling value of speech signal and output is Set of Mel-Cepstral Coefficients. The feature extraction subsystem is shown in fig 7.

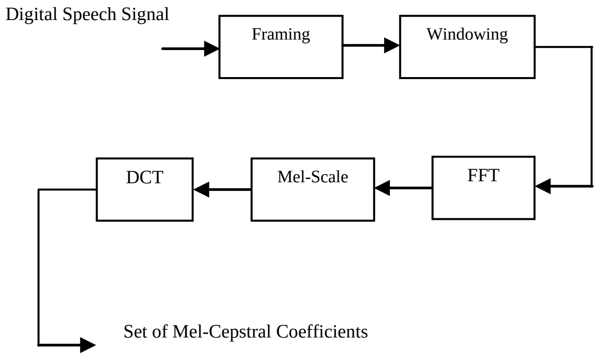

Fig 7 Feature extraction subsystem

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. A great amount of work has been already done in trying to develop identification algorithms based on the methods used by humans to identify speaker. But these efforts are mostly impractical because of their complexity and difficulty in measuring the speaker discriminative properties used by humans. 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 recognition task 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 recognition 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. Reducing data while retaining speaker discriminative information.18Based on the issues described above, we can define requirements that should be taken into account during selection of the appropriate speech signal characteristics or features:

- Discriminate between speakers while being tolerant of intra-speaker variabilities,
- Easy to measure,
- Stable over time,
- Occur naturally and frequently in speech,
- Change little from one speaking environment to another,
- Not be susceptible to mimicry.

Of course, practically, it is not possible to meet all of these criteria and there will be always a trade-off between them, based on what is more important in the particular case.

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 humans ear are mainly included in the magnitude information and the phase information is not usually playing a key role.

A. Short-Term Analysis

Because of its nature, the speech signal is a slowly varying signal or quasi-stationary. It means that when speech is examined over a sufficiently short period of time (20-30 milliseconds) it has quite stable acoustic characteristics. It leads to the useful concept of describing human speech signal, called "short-term analysis", where only a portion of the signal is used to extract signal features at one time. It works in the following way: predefined length window (usually 20-30 milliseconds) is moved along the signal with an19overlapping (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). There are several window functions used in speaker recognition area

, but the most popular is Hamming window function, which is described by the equation (2):

$$w(n) = 0.54 - 0.64\cos\left(\frac{2n\pi}{N-1}\right) \qquad (2)$$

In equation (2) N is the size of the window or frame. A set of features extracted from one frame is called feature vector. Overall overview of the short-term analysis approach is represented in fig 8.

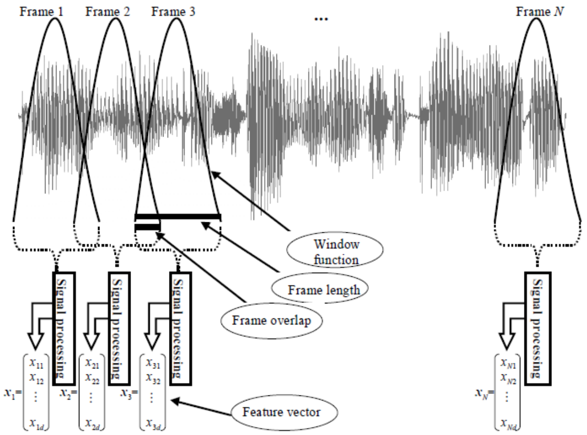

Fig 8 Short-Term Analysis

B. epstrum

We have access only to the output (speech signal) and it is often desirable to eliminate one of the components. Separation of the source and the filter parameters from the mixed output is in general difficult problem when these components are combined using not linear operation, but there are various techniques appropriate for components combined linearly. The *cepstrum* is representation of the signal where these two components are resolved into two additive parts. It is computed by taking the inverse DFT of the logarithm of the magnitude spectrum of the frame. This is represented in the equation (3).

$$Cepstrum(frame)=IDFT(log(|DFT(frame)|)) \qquad (3)$$

C. Mel-Frequency Cepstrum Coefficients

MFCC extraction is similar to the cepstrum calculation except that one special step is inserted, namely the frequency axis is warped according to the mel-scale. A “mel” is a unit of special measure or scale of perceived pitch of a tone. It does not correspond linearly to the normal frequency.

Indeed it is approximately linear below 1 kHz and logarithmic above. This approach is based on the psychophysical studies of human perception of the frequency content of sounds.

D. Filter bank

One useful way to create mel-spectrum is to use a filter bank, one filter for each desired mel-frequency component. Every filter in this bank has triangular band pass frequency response. In implement we use 20 filter bank.

Such filters compute the average spectrum around each center frequency with increasing bandwidths, as displayed in Fig6.

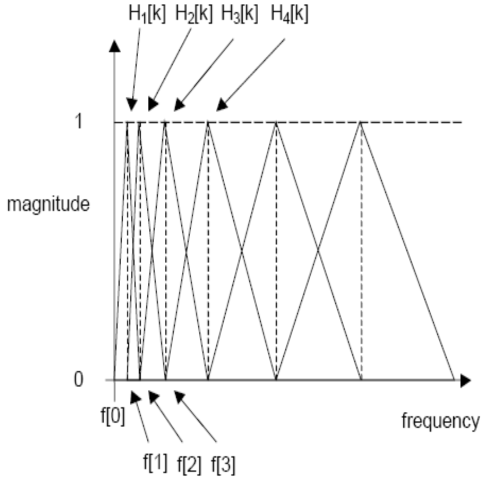

Fig 9 Triangular filters used to compute mel-cepstrum

This filter bank is applied in frequency domain and therefore, it simply amounts to taking these triangular filters on the spectrum. In practice the last step of taking inverse DFT is replaced by taking *discrete cosine transform (DCT)* for computational efficiency.

E. Vector Quantization

Vector quantization (VQ) is a process of mapping vectors from a vector space to a finite number of regions in that space. These regions are called *clusters* and represented by their central vectors or *centroids*. A set of centroids, which represents the whole vector space, is called a *codebook*. In speaker identification, VQ is applied on the set of feature vectors extracted from the speech sample and as a result, the speaker codebook is generated. Such codebook has a significantly smaller size than extracted vector set and referred as a speaker model.

All the implementations have been done with VQ-Linde Buzo Gray (LBG) algorithm as speaker modeling Paradigm.

Basically it is a method for compration the learning data of system until it is good for management and efficient .

With one code book we could decrease the initial data to the small group of indicated points.

Vector quantization could be used in both text dependent and text independent systems.

Quantization distortion is calculated with Euclidean distance Between vector and centroid that shown in equation 4 .N is the number of vectors.

$$d_E(x,y)=\sum_{i=1}^{N}(x_i-y_i)^{2} \qquad (4)$$

F. Decision subsystem

This process is depend on accommodate method and modeling method of the speaker.

Notable that accommodate method is decision make from calculated distance, and speaker model with minimum result is selected and this is the way that the speaker is recognized .

V.CONCLUSION

For examination of system the same word in Persian language with 3 different user is recorded with computer and save for speaker models resource and then ask for repeat the word with each user and recorded and for decision and comparison minimum distance is calculated and minimum distance is the maximum point for likeness .

Table III is the result of distance for each comparison. In each row section with border is the value that has minimum distance. That shown each person is successfully having the minimum distance from other speaker s and could be selected success fully from other speakers.

TABLE III PRESENTATION OF DISTANCE FOR EACH USER

	a0	a1	r0	r1	s0	s1
a0	0	0.1226	0.3664	0.3297	0.4009	0.4685
a1	0.1226	0	0.5887	0.3258	0.4086	0.4894
r0	0.3664	0.5887	0	0.0989	0.3299	0.4243
r1	0.3297	0.3258	0.0989	0	0.367	0.4287
s0	0.4009	0.4086	0.3299	0.367	0	0.1401
s1	0.4685	0.4894	0.4243	0.4287	0.1401	0

REFERENCES

[1] Rabiner and Schafer. Digital Processing of Speech Signals, Prentice Hall International, 1978.

[2] S. Furui, Digital Speech Processing, Synthesis and Recognition, New York,Marcel Dekker, 2001.

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

[4] Lau. Kelvin, Chin. Scott,leu. Lindsey, A Speaker Verification System , University of Victoria,Master's Thesis, 2002.

[5] Y. Linde, A. Buzo, and R. M. Gray, ``An Algorithm for Vector Quantizer Design," IEEE Transactions on Communications, pp. 702--710, January 1980.

[6] Prabhakar, S., Pankanti, S., and Jain, A. Biometric recognition: security and privacy concerns. IEEE Security & Privacy Magazine 1 (2003), 33–42.

[7] Li, Q., Juang, B.-H., and Lee, C.-H. Automatic verbal information verification for user authentication. IEEE Trans. on Speech and Audio Processing 8 (2000), 585–596.

[8] "Matlab VOICEBOX"
http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Marjan. Shahchera is the Master student in communication engineering in Islamic Azad University of Science and Technology Azarbayejan Sharghi, Tabriz, IRAN