An Automatic Qari[1] Recognition System

Mohammed Abdullah Hussaini and Rabah W. Aldhaheri
Department of Electrical and Computer Engineering
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia
e-mail: mhussaini@kau.edu.sa

***Abstract*—In this paper, we present an automatic Qari Recognition system based on text-independent speaker recognition technique using Mel-Frequency Cepstral Coefficients. Our test database of 200 samples consisted of recordings by 20 reciters and we achieved 86.5% recognition rate for clean samples. The recognition rate is also tested for noisy samples of 25 dB to 0 dB SNR. This automatic Qari recognition system can be used by the Holy Qur'an radio and Television station anchors to help identify the reciter of the Holy Qur'an while broadcasting live and recorded prayers and recitations.**

Keywords—speaker recognition; text-independent; speech processing; Quran recitation

I. INTRODUCTION

Speaker Recognition is a biometric characterization process aimed at the identification of people by their voices. It has wide application in voice based authentication, recognition and also in forensic applications.

Speaker recognition is broadly classified into two types; text-dependent recognition and text-independent recognition. Each of them has its specific applications. In this paper, an application of a text-independent recognition system for the automatic recognition of the reciter of the Holy Qur'an is presented.

Speaker recognition was first introduced in 1960's using spectrogram of voices, or voiceprint analysis. However, this type of analysis required human interpretation and could not fulfil the goal of automatic recognition. In the 1980's various methods were proposed to extract features from voice for speaker recognition that represented features in time, frequency or in both domains. Acoustic features of speech differ amongst individuals. These acoustic features include both learned behavioural features (e.g. pitch, accent) and anatomy (e.g. shape of the vocal tract and mouth) [1]. The most commonly extracted features are the Linear Predictive Coding (LPC), Linear Predictive Cepstral Coefficient (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) which belong to the short time analysis that provides information on the vocal tract [1].

Different modelling techniques such as template matching, stochastic modeling, and neural networks were developed to model voiceprint extracted from speech. Direct template matching is time consuming when the number of feature vectors increase [2]. The number of feature vectors can be reduced by using a codebook to represent centres of the feature vectors known as Vector Quantization. The LBG (Linde, Buzo and Gray) algorithm [3] and the k-means algorithm are some of the most well-known algorithms for Vector Quantization (VQ). Neural networks and stochastic models use probability distribution such as Hidden Markov Model (HMM) and the Gaussian Mixture Model (GMM).

Although the development in the field of speech technology is moving rapidly, there are a few inherent problems that have to be solved. The reliability of the speaker recognition drops drastically when a huge user database is used or when it is used under a noisy environment. Accuracy of automatic speaker recognition system degrades severely when there is acoustic mismatch between the training and testing material [4], [5]. An acoustic mismatch can occur due to the person's health, attitude, age or due to the introduction of noise in the microphone, transmission channels and environment [6].

Methodology

In this paper, a text-independent Qari recognition system using mel-frequency cepstrum analysis and Linde-Buzo-Gray vector quantization is developed. The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency [7]. The advantage of using the mel scale is that this nonlinear representation approximates more closely to the human auditory system's response than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping allows better representation of sound in digital speaker bank. Section II explains the steps involved in extracting the MFCC. In section III we discuss the feature matching techniques used in the system. The Linde-Buzo-Gray (LBG) algorithm is used to cluster the mel-frequency cepstral coefficients and store the voice characteristics of the speaker in the database.

A training function is designed for the system to collect the information of known reciters and store it in the database. A Graphical User Interface interface is also designed

[1] A reciter of the Holy Qur'an

978-0-7695-4959-0/13 $25.00 © 2013 IEEE
DOI 10.1109/ACSAT.2012.48

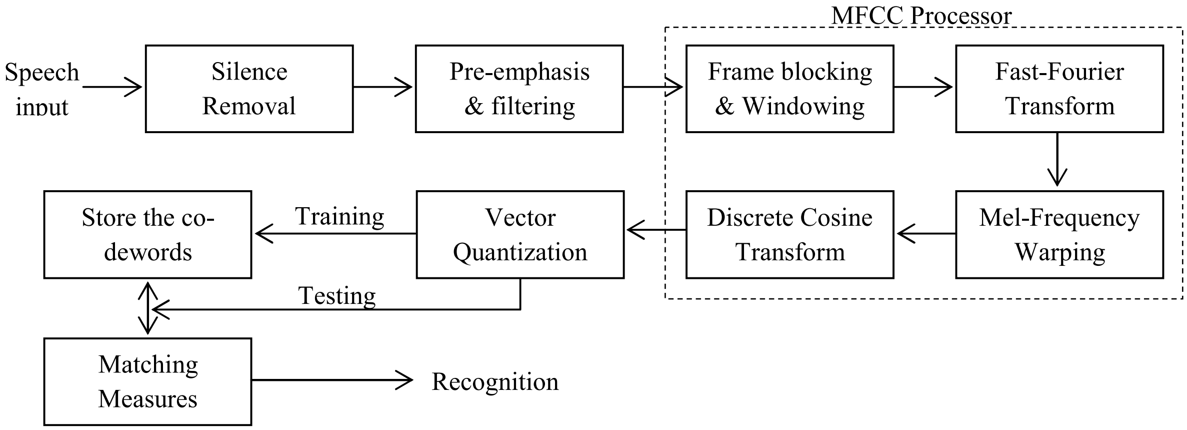

Figure 1. Block Diagram of a basic speaker recognition system

to facilitate the training and testing of the recitations and to display the name of the recognized Qari.

The designed system has the ability to read .wav and .mp3 audio files from a specified folder, and can also record any voice from the microphone input of the computer for test purpose. Section IV discusses the results and observations.

II. Speech Feature Extraction

Fig. 1 shows the block diagram of a basic speaker recognition system. The first step in an automatic speaker recognition system is Speech Feature Extraction. Speech signal is a *quasi-stationary* signal which varies slowly with time. Its characteristics are fairly stationary when examined over a short period of time, typically up to 100 ms. Therefore, the most common way to characterize a speech signal for speaker recognition is to use *short-time spectral analysis*.

A. *Silence Removal*

Most of the recorded audio files have a few milliseconds of silence at the beginning of the file. This silence is removed by finding the first sample with less than 1% magnitude of the maximum amplitude available in the file and trimming the file up to that sample.

B. *Pre-emphasis and Filtering*

Pre-emphasis reduces the noise and enhances the high frequency signals [7]. In order to pre-emphasize the speech signal $x(n)$, it is filtered through an FIR filter $(1 - \alpha z^{-1})$. α typically ranges from 0.9 to 1 [8] but 0.97 is more commonly used [9].

$$y(n) = x(n) - 0.97 \times x(n - 1) \qquad (1)$$

The speech signal is also band-limited in a passband of 60 Hz to 4 kHz with a 6th order Butterworth bandpass filter. This is done to remove the high frequency background noise and retain only the sound generated by the human reciter.

C. *Frame Blocking*

The continuous speech signal is digitized and partitioned into frames of N samples with N - M overlapping ($M < N$). N and M are calculated using the sampling rate to attain 30 ms of speech in each frame with 33% overlapping of adjacent frames.

D. *Windowing*

To minimize the spectral distortions at the beginning and end of each frame, they are passed through a Hamming window which has the form:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right),\ \ 0 \leq n \leq N-1 \qquad (2)$$

Windowing reduces the Gibbs effect by tapering the signal to zero at the beginning and end of each frame.

E. *Fast Fourier Transform (FFT)*

The frames are then converted to the frequency domain by taking their Fast Fourier Transform. For each frame of N samples, the FFT is calculated as

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{j2\pi kn}{N}},\quad k = 0,1,2,\ldots,N-1. \qquad (3)$$

The results of FFT are complex numbers but only the frequency magnitudes are considered by taking their absolute values. These values give the *spectrum* or *periodogram* of the speech signal.

F. *Mel-frequency Warping*

The human perception of the frequency contents of sounds follows a non-linear scale which has linear frequency spacing below 1 kHz and logarithmic spacing above 1 kHz. This scale is known as the *mel-frequency* scale. The following equation is used to compute the *mels* for a given frequency f in Hz [10]:

$$Mel(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right), 0 \leq f \leq F \qquad (4)$$

G. *Cepstrum*

To find the mel frequency cepstrum coefficients (MFCC), the log mel spectrum is converted back to the time domain. The local spectral properties of the speech frames are very dominant in the cepstral representation of the speech spectrum. The mel spectrum coefficients are converted to the time domain using the Discrete Cosine Transform (DCT). If the mel power spectrum coefficients are denoted as $\widetilde{S}_k, k = 0,2,...,K-1$, then the MFCC's, $\widetilde{c}_n$, are calculated as [10]:

$$\tilde{c}_n = \sum_{k=1}^{K}\left(\log \tilde{S}_k\right)\cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right], 0 \le n \le K-1 \quad (5)$$

The first component, $\widetilde{c}_0$, represents the mean value of the input signal and does not carry much speaker specific information. It is, therefore, discarded from the DCT.

The above procedure is applied to each speech frame of 30 ms with 33% overlap to compute a set of mel-frequency cepstrum coefficients. The resulting MFCCs are a set of acoustic vectors corresponding to each input utterance. These are the results of the cosine transform of the logarithm of the short-term power spectrum of the input speech signal.

III. Feature Matching

Speaker recognition systems use a number of feature matching techniques such as Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). Due to its ease of implementation and high accuracy, the VQ technique for clustering the acoustic vectors is used. Vector Quantization forms *clusters* of vectors in a finite number of regions from a large vector space. The center of each cluster is known as a *codeword* and all the codewords are stored as a *codebook*.

Fig. 2 shows a 2-dimensional plot of MFCC acoustic vectors generated in this recognition process. In the figure, three reciters and two dimensions of the acoustic space are shown. The triangles refer to the acoustic vectors from the reciter 1 (Shaikh Sa'ud Ash-Shuraim), the circles represent reciter 2 (Shaikh Ali Jaber) and the crosses are the acoustic vectors from reciter 3 (Shaikh Salah Al-Budair). In the training phase, a *reciter-specific* VQ codebook is generated for each known reciter by clustering his training acoustic vectors using the LBG clustering algorithm. In the recognition phase, the distance of the vectors of the test reciter from each trained codebook is computed. This distance is known as VQ-distortion and the reciter corresponding to the VQ codebook with the least total distortion is identified as the reciter of the test recitation.

A. *Clustering the Training Vectors*

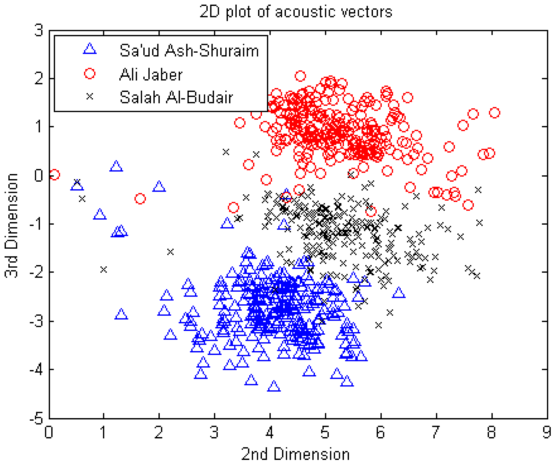

Figure 2. Two-Dimensional plot of MFCC acoustic vectors

During the enrolment session, a set of training vectors for each reciter is formed from the acoustic vectors extracted from his input recitation. In order to build a reciter-specific VQ codebook for each reciter, the LBG algorithm (Linde, Buzo and Gray, 1980 [3]) is used.

The LBG algorithm clusters the set of *L* training vectors into a set of *M* codebook vectors. It is an iterative process which starts by designing an ℓ-vector codebook, and then uses a splitting technique to further divide the codewords until the desired *M*-vector codebook is obtained.

B. *Matching Measures*

There are different algorithms available to compute the VQ-distortion. In this paper, the results of three different algorithms is investigated; Euclidean distance, Mahalanobis distance and Singular Value Decomposition (SVD). Euclidean distance is the most widely used matching measure in speaker recognition systems. The Mahalanobis distance measures the dissimilarity between two vectors by finding their correlation. It was suggested in [11] that Mahalanobis distance would give a better recognition rate with equal variance scaling. SVD is an orthogonal decomposition of a matrix and it finds wide applications in rank determination and inversion of matrices, as well as in the modeling, prediction, filtering and information compression of data sequences. SVD was proposed in [12] as a better alternative to Mahalanobis distance and Euclidean distance measures in noisy environments.

IV. Results

An automatic text-independent Qari recognition system is designed and implemented. Fig. 3 shows the Graphical User Interface for the program.

The speech input is recorded at three different sampling rates for the test purpose; 8 kHz, 11 kHz and 22 kHz. All samples are passed through a 6th order Butterworth band pass filter with a pass-band of 60 Hz to 4 kHz to remove the high frequency noise.

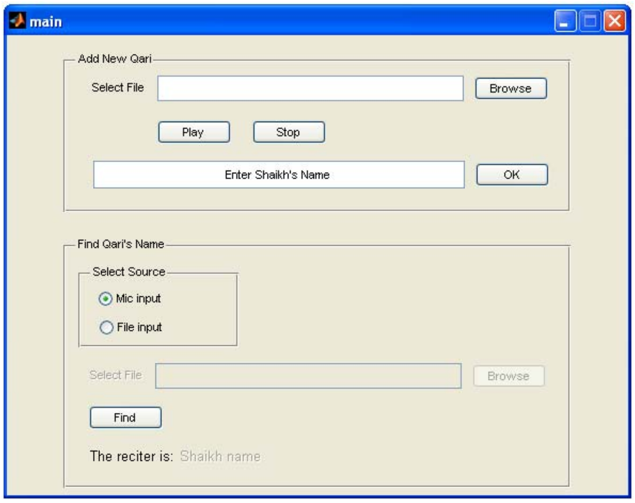

Figure 3. Graphical User Interface of the Automatic Qari Recognition System

Sample recordings from 20 different Qaris (reciters) are used to train the system with a single *Ayah* (verse) of the Holy Qur'an. Another set of ten different verses for each reciter is used to test the system. All the recordings were stored in .wav and .mp3 formats with different sampling rates. The system is implemented using MATLAB.

Table I shows the recognition rate obtained for the different distance measures. The highest recognition rate obtained for the 200 test samples is 86.5% using the Euclidean distance as the VQ-distortion measurement. The tests were conducted over a range of Signal-to-Noise Ratio (SNR) by adding Additive White Gaussian Noise to the test samples. Euclidean distance measurement yields the best result in all the tests. We observed that the use of different sampling rates for training and test files and recordings from different environments is resulting in erroneous recognition. Fig. 4 shows the effect of using different sampling rates and MFCC window size on the recognition rate. 22 kHz sampling rate with 20 ms window size and 11 kHz sampling rate with 30 ms window size yields the best results.

TABLE I

RESULTS COMPARISON

SNR (dB)	**Recognition Rate**		
	Euclidean dist.	***SVD***	***Mahalanobis dist.***
Clean Sample	86.5	56.5	59.5
25	86	54	57
20	77.5	49.5	48.5
15	59	43	38.5
10	31	34.5	23
5	11	17.5	16
0	9.5	5	11

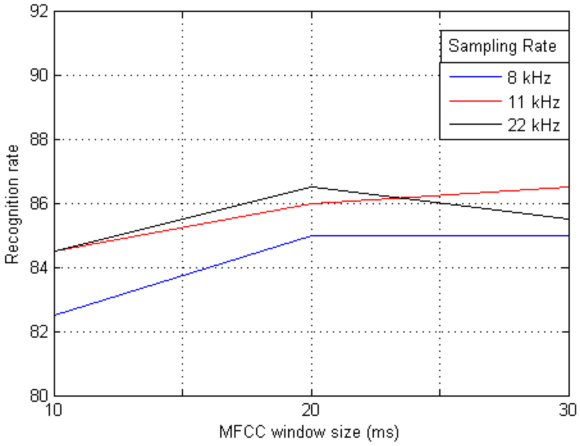

Figure 4. Effect of sampling rate and MFCC window size on the Recognition Rate

The MFCC function is adapted from Malcolm Slaney's function [13] with modifications to obtain fixed-time window size instead of fixed sample window size, and using the MATLAB's built in hamming window as well as introducing a band pass filter. The effect of removing the unvoiced speech segments from the clusters is also studied and it has been found that it improves the performance rate.

V. CONCLUSION

A simple speaker recognition system is designed and implemented to use it for identifying the reciters of the Holy Qur'an. The reciter characterization is based on the Mel-Frequency Cepstrum Coefficients analysis and LBG Vector Quantization. Experimental results show a very good recognition rate with pre-recorded files with same sampling rates for the training and test samples. Euclidean distance measurement for computing the VQ distortion yielded the best recognition rate. The system can be deployed for use at the Qur'an radio and TV stations for automatically identifying the reciter when a recording is played. The algorithms can be further developed to account for pitch frequency variations [7] and include Gaussian Mixed Weights [14] to achieve better recognition rates.

REFERENCES

[1] H. S. Jayanna and S. R. M. Prasanna, "Analysis, feature extraction, modeling and testing techniques for speaker recognition," IETE Tech Review, vol. 26, no. 3, pp. 181-190, 2009.

[2] R. Makhijani, U. Shrawankar and V. M. Thakare, "Speech enhancement using pitch detection approach for noisy environment," International Journal of Engineering Science and Technology, vol. 3, no. 2, pp. 1764-1769, 2011.

[3] Y. Linde, A. Buzo and R. Gray, "An algorithm for vector quantizer design," IEEE Transactions on Communications, vol. 28, no. 1, pp. 84 - 95, 1980.

[4] L. Lamel and J. Gauvain, "Speaker verification over the telephone," Speech Communication, pp. 141-154, 2000.

[5] J. Ortega-Garcia and J. Gonz´alez-Rodriguez, "A large speech corpus in Spanish for speaker identification and verification," IEEE

International Conference on Acoustic Speech and Signal Processing, p. 773–776, 1998.

[6] S. Singh and E. Rajan, “MFCC VQ based speaker recognition and its accuracy affecting factors,” International Journal of Computer Applications, vol. 21, no. 6, 2011.

[7] W. Yutai, L. Bo, J. Xiaoqing, L. Feng and W. Liha, “Speaker recognition based on dynamic MFCC parameters,” International Conference on Image Analysis and Signal Processing, pp. 406-409, 2009.

[8] M. R. Fallahzadeh, F. Farokhi, M. Izadian and A. A. Berangi, “A hybrid reliable algorithm for speaker recognition based on improved DTW and VQ by genetic algorithm in noisy environment,” International Conference on Multimedia and Signal Processing, vol. 2, pp. 269-273, 2011.

[9] J. Saastamoinen, E. Karpov, V. Hautamaki and P. Franti, “Accuracy of MFCC-based speaker recognition in series 60 device,” Journal on Applied Signal Processing, vol. 17, p. 2816–2827, 2005.

[10] B. Gold and N. Morgan, Speech and Audio Signal Processing, John Willy & Sons, 2002, pp. 189-203.

[11] V. Gupta, J. Gowdy and J. Bryan, “Evaluation of some distance measures for speaker independent isolated word recognition,” IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 460-463, 1977.

[12] R. W. Aldhaheri and F. E. Al-Saadi, “Robust text-independent speaker recognition with short utterance in noisy environment using SVD as a matching measure,” Journal of King Saud University, Computer & Information Sciences, vol. 17, pp. 23-41, 2004.

[13] M. Slaney, Auditory Toolbox, Version 2, Interval Research Corporation, 1998.

[14] Z. Weng, L. Li and D. Guo, “Speaker recognition using weighted dynamic MFCC based on GMM,” International Conference on Anti-Counterfeiting Security and Identification in Communication, Chengdu, 2010.