MareText Independent Speaker Identification based on K-mean Algorithm

Allam Mousa

Electrical Engineering Department
An Najah University
allam@najah.edu

Abstract: This paper proposes a text-independent speaker identification system based on Mel Frequency Cepstral Coefficients as a feature extraction and Vector Quantization technique that would minimize the data required for processing. The correlation between the identification success rate and the various parameters of the system including the feature extraction tools and the data minimization technique will be examined. Extracted features of a speaker are quantized by a number of centroids and the K-mean algorithm has been integrated into the proposed speaker identification system. Such centroids constitute the codebook of that speaker. MFCC are calculated in both training and testing phases. To calculate these MFCC speakers uttered different words, once in a training session and once in a testing one. The speakers were identified according to the minimum quantization distance which was calculated between the centroids of each speaker in the training phase and the MFCC of individual speakers in the testing phase. Analysis was carried out to identify parameter values that could be used to improve the performance of the system. The experimental results illustrate the efficiency of the proposed method under several conditions

Keywords: Speaker Recognition, Speaker Identification, Feature Extraction, MFCC, VQ, Clustering, Text-independent

1. Introduction

Speaker recognition aims at recognizing speakers from their voices as each person has his own speech characteristics and his way of speaking. Speaker recognition is basically divided into speaker identification and speaker verification. Speaker identification is the process of determining which registered speaker provides the speech input, while verification is the task of automatically determining if a person really is the person he or she claims to be. Speaker recognition has many particular applications as a speaker's voice can be used to verify their identity and control access to services such as banking by telephone, database access services, voice dialing telephone shopping, information services and voice mail. Another important application of speaker recognition technology is for forensic purposes [1].

Speaker recognition can be classified as based on text-dependent or text-independent methods. In the text dependent method, the speaker has to say key words or sentences having the same text for both training and recognition trials. Whereas in the text independent case the system can identify the speaker regardless of what is being said [2], [3], [4].

The goal of this study is a real time text-independent speaker identification system, which consists of comparing a speech signal from an unknown speaker to a database of known speakers. The system will operate in two modes: a training mode and a recognition mode. During the training mode users will record their voices and make a feature model it. The recognition mode will use the information that the user has provided in the training mode and attempt to isolate and identify the speaker. The Mel Frequency Cepstral Coefficients (MFCC) and the Vector Quantization (VQ) algorithms are used to implement this process. The simple K-means clustering algorithm is used in this study whereas the LBG is used in other similar work [4].

Received: November 29th, 2010. Accepted: March 26th, 2011

2. Feature Extraction

The main objective of feature extraction is to extract characteristics from the speech signal that are unique to each individual and that will be used to differentiate speakers. Since the characteristic of the vocal tract is unique for each speaker, the vocal tract impulse response can be used to discriminate speakers. Therefore in order to obtain the vocal tract impulse response from the speech signal, a deconvolution algorithm like the MFCC is applied [5].

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), (MFCC), and others. MFCC is perhaps the best known and most popular. MFCC's are based on the known variation of the human ear's critical bandwidths in response to frequency. The MFCC technique makes use of two types of filters, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, a signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cords rather than the speech waveforms themselves, the MFCC are less susceptible to the said variations [6].

3. The Mel Frequency Cepstrum Coefficient

The (MFCC) is used to resolve the speech signal into a sum of two components. This computation is carried out by taking the Discrete Cosine Transform (DCT) of the logarithm of the magnitude spectrum of the speech frame. The convolution of the two components is changed to multiplication when Fourier Transform (FT) is performed. A typical flowchart of the MFCC process is shown in Figure1.

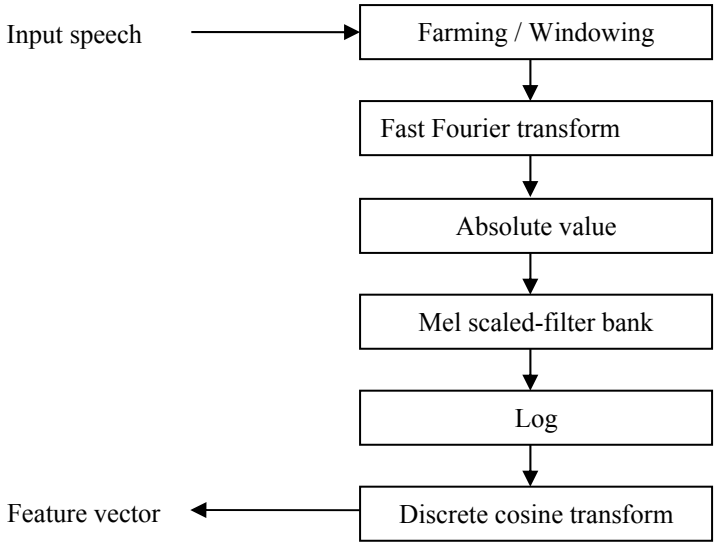

Figure1. Flow diagram of the MFCC process

A. Framing and Windowing

Framing is the process of segmenting the speech signal into small frames that range in length from ten to thirty milliseconds. In this range, the speech signal is for the most part stationary [7].

Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum. The choice of the window involves a tradeoff between several factors. In speaker recognition, the most commonly used window shape is the hamming window.

B. Fast Fourier Transform

To convert the signal from a time domain to a frequency domain in preparation for the next stage, Mel Frequency Wrapping (MFW) is applied to. The basis of performing a Fourier transform is to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain into multiplication in the frequency domain [8].

C. Mel-Scaled Filter Bank

The information carried by low frequency components of the speech signal is more important compared to the high frequency components. In order to place more emphasis on the low frequency components, mel scaling is performed. It is a unit of special measure or scale of the perceived pitch of a tone. It does not correspond linearly to the normal frequency, but behaves linearly below 1 kHz and logarithmically above 1 kHz. This is based on studies of the human perception of the frequency content of sound. The relationship between the frequency (in hertz) and the mel scaled frequency is given in Eq.1;

$$mel(f) = 2595 * \log 10(1 + \frac{f}{700}) \quad (1)$$

In order to perform mel-scaling, a number of triangular filters or filterbank are used. To implement such filterbanks, the magnitude coefficient of each Fourier transformed speech segment is bounded by correlating them with each triangular filter in the filterbank.

D. Cepstrum

The log mel spectrum has to be converted back to time producing Mel Frequency Cepstrum Coefficients (MFCCs). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. These MFCCs are real and so they may be converted to the time domain using the Discrete Cosine Transform (DCT). The MFCCs values may be calculated using Eq.2 [4];

$$C_n = \sum_{m=1}^{M} (\log S_m) \cos\left[(m - \frac{1}{2})\frac{\pi n}{M}\right] \quad (2)$$

Where n is the index of the cepstral coefficient and S_m is the output of an M-channel filterbank. The number of mel cepstrum coefficients, M, is typically chosen as (10-15). The set of coefficients calculated for each frame is called a feature vector. These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker. Therefore each input utterance is transformed into a sequence of acoustic vectors [9], [10]. The next section describes how these acoustic vectors can be used to represent and recognize the voice characteristic of a speaker.

4. Classification and Feature Matching

The decision making process to determine a speaker's identity is based on previously stored information. This step is basically divided into two modes: training and testing. Training is a process of enrolling a speaker into the identification system database by constructing a unique model for each speaker based on the features extracted from the speaker's speech sample. Testing is a process of computing a matching score, which is a measure of similarity of the features extracted from the unknown speaker and the stored speaker models in the database. The speaker with the minimum matching score is chosen to be identified as the unknown speaker.

The classification or speaker modeling techniques including Hidden Markov Modeling (HMM), Dynamic Time Warping (DTW), Gaussian Mixture Modeling (GMM), and Vector Quantization (VQ). The VQ approach has been used in this work due to its ease of implementation and high accuracy.

A. Vector Quantization

The number of feature vectors that is generated from the training mode is so large that storing every single vector is impossible. So, Vector Quantization is used to compress the information and manipulate the data in such a way as to maintain the most prominent characteristics [11]. The 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 represent the whole vector space, is called a codebook. In this work, VQ is applied on the set of feature vectors extracted from speech sample and as a result the speaker codebook is generated.

There are a number of algorithms for codebook generation such as: K-means algorithm, Generalized Lioyd algorithm (GLA) (also known as Linde-Buzo-Gray (LGB) algorithm), Self Organizing Maps (SOM) and Pairwise Nearest Neighbor (PNN). Here, we have use the K-means algorithm since it is the most popular, simplest and the easiest one to implement.

B. K-mean Clustering Algorithm

The K-means algorithm partitions the T feature vectors into M centroids. The algorithm first randomly chooses M cluster-centroids among the T feature vectors. Then each feature vector is assigned to the nearest centroid, and the new centroids are calculated for the new clusters. This procedure is continued until a stopping criterion is met, where the mean square error between the feature vectors and the cluster-centroids is below a certain threshold or there is no more change in the cluster-center assignment [5], [9], [10]. The algorithm is summarized as shown in Figure2 [12].

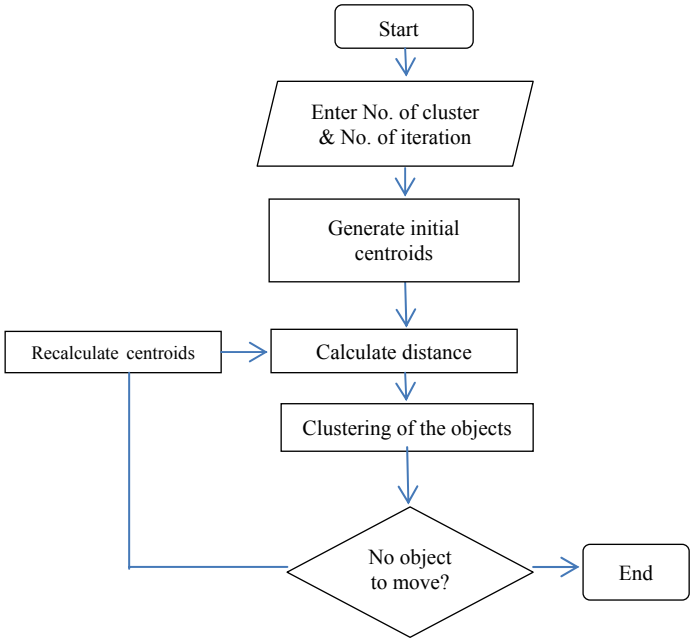

Figure 2. Flowchart of the K-means algorithm

In the recognition phase an unknown speaker, represented by a sequence of feature vectors $X= \{x_1,x_2,\ldots,x_T\}$, is compared with the reference vectors $R=\{r_1, r_2,\ldots r_K\}$ in the database. Hence, a distortion measure is computed for each codebook, and the speaker with the lowest distortion is chosen [9].

One way to define the distortion measure (D), which is the sum of squared distances between vector and its representative (centroid), is to use the average of the Euclidean distances as given by Eq.3 [4].

$$D(X,R) = \frac{1}{T}\sum_{t=1}^{T} \min d\,(x_t, r_k) \qquad \text{where } 1 \le k \le K \tag{3}$$

Thus, each feature vector in the sequence *X* is compared with all the codebooks, and the codebook with the minimized average distance is chosen.

5. Simulation Results

The following results were achieved by performing the experiments on pure speech samples from the ELSDSR database which consists of 20 speakers, 10 male and 10 female speakers. Also, another database was selected from the local environment to see how the performance of the system varies by changing the spoken text, language and the test speech length.

Here, the identification rate is defined as the ratio of the number of speakers identified to the total number of speakers tested. The number of MFCC is set to 12, the number of filter banks is 29 and the codebook size is 64. The required measured distance between the given speech and the database is illustrated in Table 1. One can see that the distance is the least when the signal is compared with itself. This indicates that each signal matches itself more than any other signal and it varies in its distance from the other signals.

Table 1. The distance between various speakers

	Sp1	Sp2	Sp3	Sp4	Sp5
Sp1	10.7	13.2	17.8	14.7	13.2
Sp2	13.2	10.2	13.2	11.7	14.1
Sp3	17.5	16.1	11.9	16.2	17.7
Sp4	16.1	13.7	15.5	11.7	16.7
Sp5	14.9	15.7	17.2	17.8	12.3

From Table 1, it can be seen that the system identifies the speaker according to the theory that "the most likely speaker's voice should have the smallest Euclidean distance compared to the codebooks in the database".

The experiments conducted have shown that there are five main parameters that can greatly affect the performance of the system. These are: the number of the MFCC, the number of filters in the filter bank, the codebook size, the test speech length and the text language.

A. *Number of MFCC coefficients*

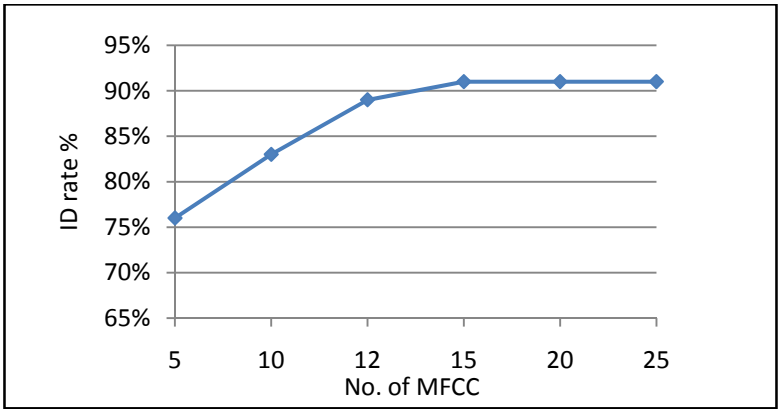

Figure 3. Identification rate (in%) Vs. the No. of MFCC

Increasing the number of mel frequency cepstral coefficients results in improving the identification rate (ID) up to a certain limit. Increasing this number has no significant improvement after a certain value as shown in Figure3. The MFCC are typically in the range 12 to 15.

B. Number of filterbanks

It is obvious that number of the filterbanks plays a major role for the purpose of improving recognition accuracy. Simulation results have shown that it is possible to obtain 100% identification rate using 40 filterbanks with a codebook size of 64 as illustrated in Figure4.

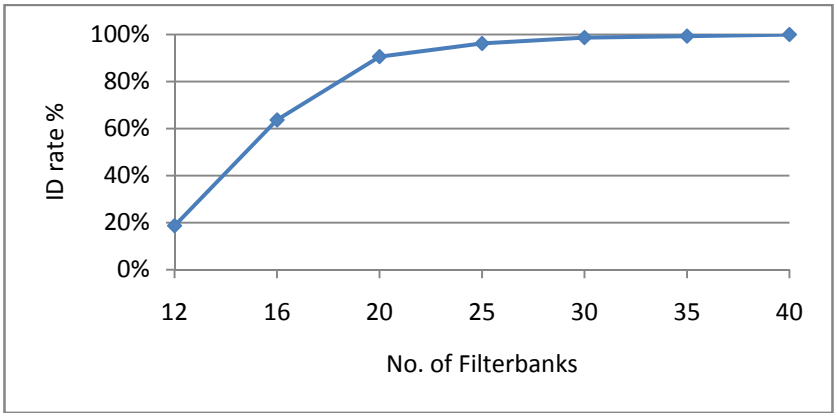

Figure 4. Identification rate (in %) for values of filterbanks

C. The codebook size

Increasing the codebook size improves the ID rate significantly such that the codebook size of 16 achieves an ID value of 98%. However, increasing the codebook size for larger values will not improve the ID rate as illustrated in Figure 5. On the other hand, the distortion measure for a speech sample text decreases as the codebook size increases as illustrated for three speakers in Figure 6.

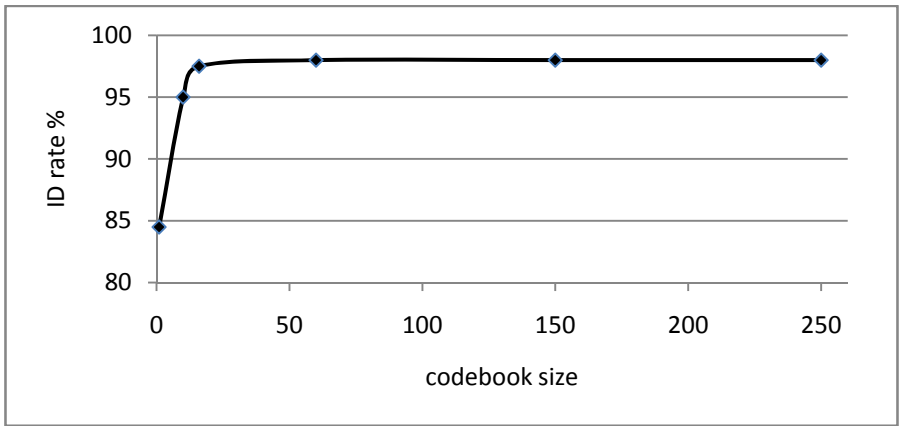

Figure 5. Codebook size Vs. Identification rate

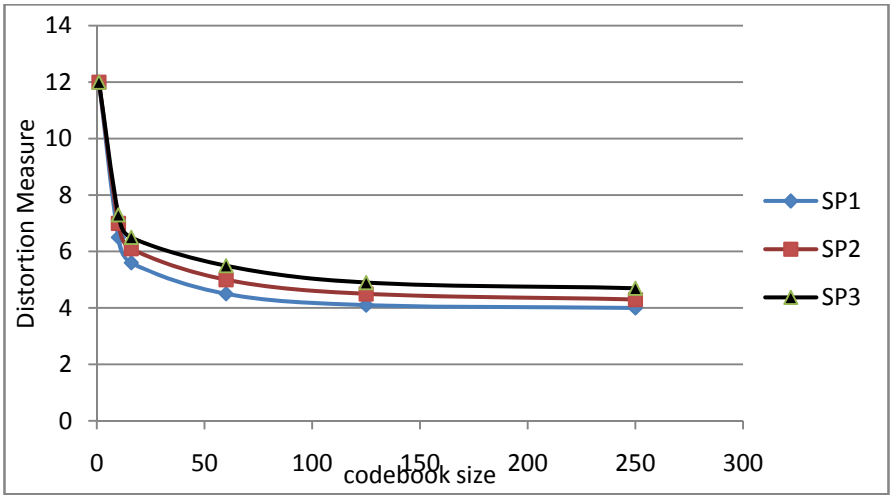

Figure 6. Distortion measure Vs. codebook size

From the obtained results, it is obvious that increasing the number of centroids results in increasing the identification rate, but the computational time will also increase. On the other hand, the matching scores (Euclidean distance) for the same speaker is decreased as the codebook size increases.

D. Test Speech Duration

To study the performance of different test shot duration, three tests were conducted using all test speakers uttering the same test speech sample with three different duration values. The ID rate accuracy is increased when the speech text duration is increased as illustrated in Table 2.

Table 2. Identification rate (in %) for different test shot duration

Test speech duration	ID (%)
0.5 sec.	60%
2 sec.	85%
Full test shots	95%

The best identification can be achieved using the whole test shots from which we can conclude that the performance of the VQ analysis is highly dependent on the duration of the speech data that is being processed.

E. The Effect of Text and Language

During the experiments, the speakers uttered different phrases in two different languages, Arabic and English, and the results showed that the system is able to identify the correct speaker regardless of the spoken text and language. This indicates that the mel frequency cepstral features extracted from the speech sample are sensitive to the speaker's voice characteristics but not to the language or text. Moreover, VQ is used to cluster the feature vectors based on their sound classes and not according to the spoken text.

6. Conclusion

A text-independent speaker identification system has been implemented using such that Mel Frequency Cepstral Coefficients were used for feature extraction and K-mean Vector Quantization technique was used to model the speakers. Using the extracted features, a

codebook from each speaker has clustered the feature vectors. Codebooks from all the speakers were collected in a database. A distortion measure, based on minimizing the Euclidean distance, was used when matching the unknown speaker with the speaker database. As the number of centroids increases, the identification rate of the system increases. Also, the number of centroids has to be increased as the number of speakers increase. In addition, as the number of filters in the filter-bank increases, the identification rate increases. The experiments conducted have shown that it was possible to achieve an almost 100% identification rate when using 40 filters with full training sets and full test shots. Reducing the test shot duration reduced the recognition accuracy. For real time application, the test data usually needs to be few seconds long. It has been shown that VQ based clustering is an efficient and simple way to perform text and language independent speaker identification. This system has more than 97% accuracy in identifying the correct speaker when using long enough training-sessions and testing sessions.

Acknowledgements

The author would like to thank all those who helped in improving the clarity and quality of this paper particularly the anonymous reviewers for their valuable comments.

References

[1] Frederic Bimbot, Jean-Francois Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, SylvainMeignier, Teva Merlin, Javier Ortega-Garcia, Dijana Petrovska-Delacretaz, and Douglas A. Reynolds, "A Tutorial on Text-Independent Speaker Verification", *EURASIP Journal on Applied Signal Processing*:4, pp. 430–451, 2009

[2] H. S. Jayanna, S.R. Mahadeva Prasanna, "Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition" *IETE Technical Review*, Vol.26, Issue 3, pp. 181-190, 2009

[3] M. R. Hasan, M.Jamil, M. G. Rabbani, M. S. Rahman, "Speaker Identification using Mel Fdrequency Cepstral Coefficients", *3rd Int. Conf. On Elec. & Computer Eng., ICECE 2004, Dhaka, Bangladesh*, pp. 565-568, 2004.

[4] Tomi Kinnunen, Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors", *Speech Communication*, 52:pp. 12–40, 2010

[5] Zhang Yan, Tang Zhenmin, Li Yanping "Combining Speech Enhancement and Discriminative Feature Extraction for Robust Speaker Recognition", *2009 WRI World Congress on Computer Science and Information Engineering*, Vol.5, pp. 274-278, 2009

[6] Shi-Huang Chen and Yu-Ren Luo "Speaker Verification Using MFCC and Support Vector Machine", *Proceedings of the International MultiConference of Engineers and Computer Scientists* 2009 Vol.1, IMECS 2009, Hong Kong, March 18 - 20, 2009

[7] Kondoz A. M., "Digital speech, coding for low bit rate communication systems," *Wiley*, 2004

[8] Gold and N. Morgan, Speech and Audio Signal Processing, John Wiley and Sons, New York, NY, 2000.

[9] A. Revathi1, R. Ganapathy and Y.Venkataraman "Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach" *Int. J. of Computer Science and Information Technology*, Vol. 1, No 2, pp: 30-42, 2009

[10] Wael Al-Sawalmeh, Khaled Daqrouq, Omar Daoud, Abdel-Rahman Al-Qawasmi, "Speaker Identification System-based Mel Frequency and Wavelet Transform using Neural Network Classifier" *European Journal of Scientific Research*, Vol.41 No.4, pp.515-525, 2010

[11] R. M.Gray, "Vector Quantization", *IEEE ASSP Magazine*, Vol.1, pp. 4-29, April 1984

[12] Dost Muhammad Khan, Nawaz Mohamudally, "An Agent Oriented Approach for Implementation of the Range Method of Initial Centroids in K-Means Clustering Data Mining Algorithm", *International Journal of Information Processing and Management*, Vol. 1, No. 1, pp. 104-113, 2010

Allam Mousa received his BSc (with high honor) MSc and Phd in Electrical and Electronics Engineering all from Eastern Mediterranean University in 1990, 1992 and 1996 respectively. He is an Associate Professor at the department of Electrical Engineering An Najah University. He was the chairman of the same department in 2002-2009 and was chairman of the Electronics Engineering Department Al Quds University in 1999/2000. His current research interests include Speech and Image Processing and Telecommunications. He is reviewer of several international journals and conferences. Dr. Mousa is also interested in higher education quality assurance and he is currently the Presidents Assistant for Planning Development and Quality. He is also SMIEEE and member of Engineering Association.