`

Speaker Recognition Using MFCC and Vector Quantisation

Ayman M.Fahmy, Sherief S.Ahmed , Moustafa M.Okbelbab , Mohamed M.Moawad ,
Mohamed H.Zanaty , Aya F. Ahmed , Ahmed H . Kandeel

Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University,
Giza, Egypt.

ABSTRACT

Speech recognition is a popular topic in today's life.

The applications of Speech recognition can be found everywhere, which make our life more effective. Feature extraction is the first step for speaker recognition. Many algorithms are suggested/developed by the researchers for feature extraction. In this work, the Mel Frequency Cepstrum Coefficient (MFCC) feature and VQLBG (Vector Quantisation-Linde, Buzo, and Gray) algorithm have been used for designing a text independent speaker identification system. Some modifications to the existing technique of MFCC for feature extraction are also suggested to improve the speaker recognition efficiency. The MFCC algorithm is used to simulate feature extraction module. Using this algorithm the cepstral co-efficient are calculated of Mel frequency scale. VQ (Vector Quantization) method will be used for reduction of amount of data to decrease computation time. In the feature matching stage Euclidean distance is applied as similarity criterion. The proposed system is implemented in Matlab 9.0 environment and showing 85.00% results as correct acceptance and correct rejections with the error rate of 15.00%.

I. INTRODUCTION

Speech is one of the oldest and most natural means of information exchange between human beings. We as humans speak and listen to each other in human-human interface. For centuries people have tried to develop machines that can understand and produce speech as humans do so naturally .Speech recognition algorithms can be broadly divided into speaker dependent and speaker independent. Speaker dependent system focuses on developing a system to recognize unique voiceprint of individuals. Speaker independent system involves identifying the word uttered by the speaker. Mark Gales et al.[5] introduced that HMMs are generative models and although HMM-based acoustic models were developed primarily for speech recognition, it is relevant to consider how well they can actually generate speech, in 1995 the error was 50% , but now by using it the error become below 20% . Muhirwe Jackson [6] introduced that by using HMM the results revealed that 94.47% of the tested data were correctly recognized . Abeer et al.[7] introduced a text dependent speech identification system by using MFCC and K-Mean Algorithm with efficient 98% . Sarika et al.[8] introduced that by increasing the coefficients of the MFCC ,the accuracy of the system increase which reach 74.72% after using 12 coefficients . Kashyap et al.[9] introduced speech recognition using MFCC and by testing it , this give him error rate is about 13% .Vibha Tiwari [10] introduced that by increasing the Number of MFCC filters the efficiency increase which reach 80% and show that the more efficiency when using Hanning window that make the it reach 75% .Shumaila, Tahira et al.[11] introduced that with MFCC for extracting acoustic features and then used to trained HMM parameters through forward backward algorithm which lies under HMM and finally the computed log likelihood from training is stored to database. showing 86.67% results as correct acceptance .Garima et al.[12] introduced that performance of MFCC is affected by the number of filters, the shape of filters, the way that filters are spaced, and the way that the power spectrum is warped. the optimum values of above parameters are chosen to get an efficiency of 99.5 %

II. MATERIALS AND METHODS

Most of the speaker recognition systems contain two phases. In the first phase feature extraction is done. The unique features from the voice signal are extracted which are used latter for identifying the speaker. The second phase is feature matching in which we compare the extracted voice features with the database of known speakers. The overall efficiency of the system depends on how efficiently the features of the voice are extracted and the procedures used to compare the real time voice sample features with the database .

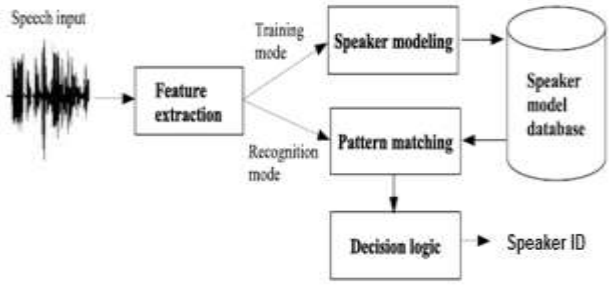


Figure 1.1 : Schematic diagram of the closed-set speaker identification system.

Referring to the diagram above, you can see that the input speech will pass through two major stages in order to get the speaker identity, they are:

1- Feature extraction.

2- Classification and Feature matching

`

III. FEATURE EXTRACTION

is a special form of dimensionality reduction, and here in our project we need to do dimensionality reduction for the input speech we will do that by extracting a specific features from the speech, these features carry the characteristics of the speech which are different from one speaker to another, so these features will play the major role in our project , as our mission is to identify the speaker and make a decision that highly depends on how much we were successful in extracting a useful information from the speech in a way enables our system to differentiate between speakers and identify them according to their features . we choose MFCC for the following reasons ; one of the most important features, which is required among various kinds of speech applications, shows high accuracy results for clean speech ,they can be regarded as the "standard" features in speaker as well as speech recognition, The most common algorithm that used for speaker recognition system. Mel frequency cepstral coefficients (MFCC) is probably the best known and most widely used for both speech and speaker recognition . A Mel is a unit of measure based on human ear's perceived frequency. The mel scale is approximately linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. The approximation of Mel from frequency can be expressed as

mel(f) = 2595*log(1+f/700) -------- (1) where f denotes the real frequency and Mel(f) denotes the perceived frequency. The block diagram showing the computation and extraction processes of MFCC is shown in Fig. 1.2

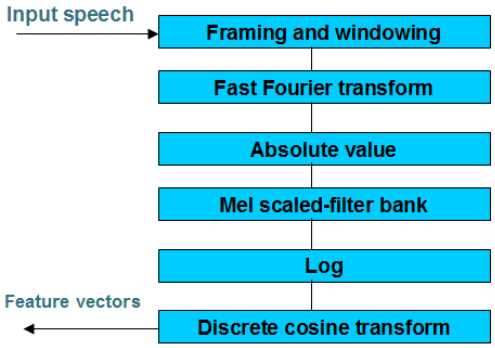


Figure 1.2 : MFCC Flow diagram

I. Preprocessing

To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. Speech signal pre-processing covers digital filtering and speech signal detection.

II. Framing

first we split the signal up into several frames such that we are analyzing each frame in the short time instead of analyzing the entire signal at once, at the range (10-30) ms the speech signal is for the most part stationary . Also an overlapping is applied to frames. Here we will have something called the Hop Size. In most cases half of the frame size is used for the hop size. Look at the figure (1.3) shown below, it represent what the frame and hop sizes are. The reason for this is because on each individual frame, we will also be applying a hamming window which will get rid of some of the information at the beginning and end of each frame. Overlapping will then reincorporate this information back into our extracted features

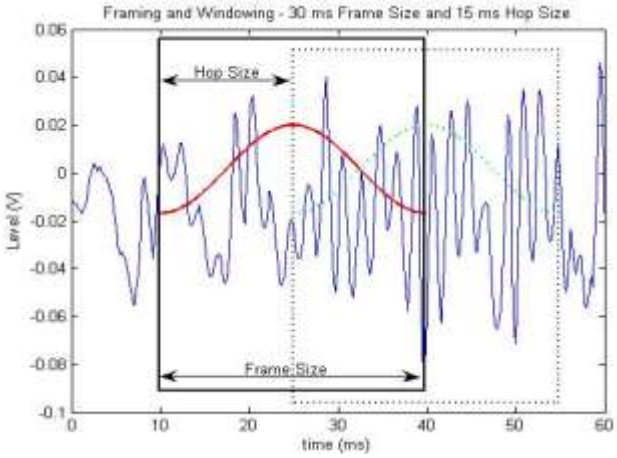


Figure 1.3: Framing and windowing.

III. Windowing

This is to select a portion of the signal that can reasonably be assumed stationary. Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum . The choice of the window is a trade off between several factors. In speaker recognition, the most commonly used window shape is the hamming window and hamming windows will decrease the possibility of high frequency components in each frame due to such abrupt slicing of the signal

.The multiplication of the speech wave by the window function has two effects, It gradually attenuates the amplitude at both ends of extraction interval to prevent an abrupt change at the endpoints and It produces the convolution for the Fourier transform of the window function and the speech spectrum .Actually there are many types of windows such as ; Rectangular window, Hamming, Hann , Cosine window... etc . The hamming window WH(n) ,defined as :-

$$W_{H(n)} = 0.54 - 0.46 \cos\left(\frac{2n\pi}{N-1}\right) \quad \ldots..(1)$$

IV. Fast Fourier Transform

The basis of performing 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 .When we perform FFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, we can still perform FFT but the discontinuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem, we have two strategies, Multiply each frame by Hamming window to increase its continuity at the first and last points and Take a frame of a variable size such that it always

contains an integer multiple number of the fundamental periods of the speech signal.

V. Mel-scaled filter bank

The mel-frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. we can use the following formula to compute the mels for a given frequency f in Hz: Mel(f)= 2595*log10(1+f/700) . one approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired Mel frequency component. The filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant Mel-frequency interval. the signal processed like that of human ear response .

$$\tilde{S}(l) = \sum_{k=0}^{N} S(K) M_i(K) \quad \text{....(2)}$$

Where :

S(l) :Mel spectrum. S(K) :Original spectrum. M(K) :Mel filterbank.

L=0 ,1 , ..., L-1 , Where L is the total number of mel filterbanks

N/2 = Half FFT size.

***Now**, we will move to the next stage to have the cpestrum or the mel frequency cpestrum coefficient .*

After the previous discussion we set the following to implement the MFCC algorithm which are pre-emphasis, framing with frame size=256 sample, Windowing by multiplying each frame with hamming window., Using 40 Mel filterbanks and Extracting 12 MFCC coefficients for each frame

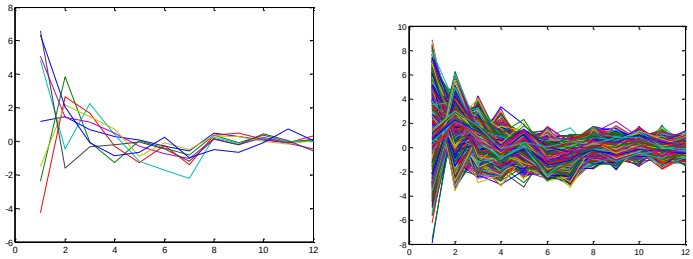
VI. CLASSIFICATION AND FEATURE MATCHING

1. Classification Methods

Stochastic models provide more flexibility and better results. It includes : Gaussian mixture model (GMM), Hidden Markov Model (HMM) and Artificial Neural Network (ANN), also linear classifier **.** For template models ,This approach makes minimal assumptions about the distribution of the features. Template models are considered to be the simplest ones. It includes : Dynamic Time Warping (DTW) and Vector Quantization (VQ) models . in this project, we chose to use vector quantization approach due to it's ease of implementation

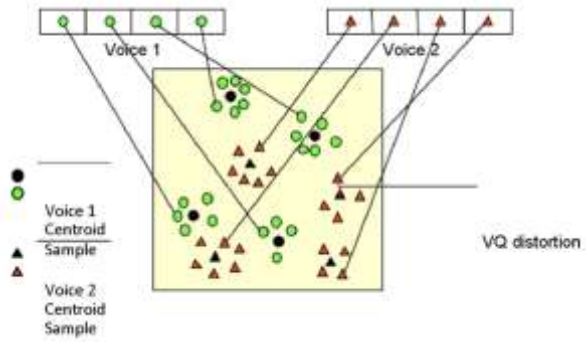
2. Vector Quantization

The quantization technique used to compress the information and manipulate the data such in a way to maintain the most prominent characteristics. VQ is used in many applications such as data compression (i.e. image and voice compression), voice recognition, etc. in its application in speaker recognition technology assists by creating a classification system for each speaker. It is a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the distribution. The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features . These Figures shows that The vectors generated from training before and after VQ.



3. Speaker Database

The first step is to build a speaker-database Cdatabase = { C1 , C2 , …. , CN} consisting of N codebooks, one for each speaker in the database. This is done by first converting the raw input signal into a sequence of feature vectors X={x1, x2, …. , xT}. These feature vectors are clustered into a set of M codewords C={c1, c2, …. , cM}.There are number of algorithms for codebook generation such as :K-means algorithm or Generalized Lioyd algorithm (GLA) which is also known as Linde-Buzo-Gray algorithm ,etc . In this project, we used the VQLBG algorithm ,We use Vector Quantization by the Linde-Buzo-Gray algorithm (VQLBG) that takes two parameters (MFCC transformed signal vector of saved signal vector in dataset, number of centroides) that delivers us output we will use below. Our Matlab code asks person to record his/her sound for a number of seconds in order to identify his/her sound. In order to recognize the person's voice, we load the recorded sound. The speaker is identified according to the minimum quantization distance which is calculated between the centroids of each speaker in training phase and the MFCC's of individual speaker in testing phase .



4. Feature Matching

In the recognition phase an unknown speaker, represented by a sequence of feature vectors {x1,x2,….,xT}, is compared with the codebooks in the database. For each codebook a distortion measure is computed, and the speaker with the lowest distortion is chosen ,

$$C_{best} = \underset{1 \le i \le N}{argmin} \{ s(X, C_i) \} \quad \text{.....(3)}$$

One way to define the distortion measure, which is the sum of

squared distances between vector and its representative (centroid) is to use the average of the Euclidean distances

$$s(X, C_i) = \frac{1}{T}\sum_{t=1}^{T} d(x_t, c_{min}^{it}) \quad .....(4)$$

The well known distance measures are Euclidean, city distance, weighted Euclidean and Mahalanobis. We use Euclidean distance in our work. Where C min denotes the nearest codeword x(t) in the codebook Ci and d(.) is the Euclidean distance. Thus, each feature vector in the sequence X is compared with all the codebooks, and the codebook with the minimized average distance is chosen

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

to be the best. The formula used to calculate the Euclidean distance can be defined as following: .....(5)
The Euclidean distance between two points P= (p1, p2…pn) and Q = (q1, q2...qn)

IV. RESULTS AND DISCUSSION

In order to evaluate the performance of the system as a real time application ,another database is selected from our environment. It consists of 10 speakers , their ages are about 22 years old. The training and testing sessions were made in our communication lab environment, by using PC with 2.4 GHZ Processor , 8 GB RAM , 1 GB VIGA and Processor CORE I7 . Training set consists of 4 seconds speech utterance . This table show a comparison between different methods used before and our proposed methods this show that the proposed method that used give us an accuracy with 85% .

Method	***Algorithm***	***Accuracy***
Mark Gales et al.	HMM	About 80%
Muhirwe Jackson	HMM	94.47%
Abeer et al.	MFCC and K-Means	98%
Sarika et al.	MFCC	74.72%
Kashyap et al.	MFCC	87%
Vibha Tiwari	MFCC	75%
Tahira et al.	MFCC	86.67%
Garima et al.	MFCC	99.5%
Proposed Method	MFCC	85%

VI. CONCLUSION

The goal of this project was to implement a text-independent speaker identification system. The feature extraction is done using mel frequency cepstral coefficients [MFCC] and the speakers was modeled using vector quantization technique. Using the extracted features a codebook from each speaker was build clustering the feature vectors using the VQLBG algorithm. Codebooks from all the speakers was collected in a database. A distortion measure based on minimizing the euclidean distance was used when matching the unknown speaker with the speaker database. The study reveals that 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 increases. The study shows that as the number of filters in the filter-bank increases, the identification rate increases. The experiments conducted using speaker recognition database, showed that it was possible to achieve 100% identification rate when using 40 filters with full training sets and full test shots. Our experiments in the communication lab environment showed that , in order to obtain satisfactory result for real time application, the test data usually needs to be more than ten seconds long. All in all, during this project we have found that VQ is an efficient and simple way to do speaker identification. Our system is 85 % accurate in identifying the correct speaker when using 4 seconds for training session .

REFERENCES

[1] sarika hegde, k. k. achary and surendra shetty, "feature selection using fisher's ratio technique for automatic speech recognition," *international journal on cybernetics & informatics (ijci)*, vol. 4, no. 2, april 201.

[2] mathur, a., saxena, t. and krishnamurthi, r., "generating subtitles automatically using audio extraction and speech recognition," in *computational intelligence & communication technology (cict), 2015 ieee international conference*, pp. 621 – 626.

[3] geeta nijhawan, and dr. m.k soni, "speaker recognition using mfcc and vector quantisation," *int. j. on recent trends in engineering and technology*, vol. 11, no. 1, july 2014.

[4] dipmoy gupta, radha mounima c., navya manjunath and manoj pb, "isolated word speech recognition using vector quantization (vq)," *international journal of advanced research in computer science and software engineering*, vol. 2, issue 5, may 2012.

[5] Mark Gales and Steve Young., "The Application of Hidden Markov Models in Speech Recognition, " *Cambridge University Engineering Department, Trumpington Street, Cambridge* , vol. 1, no. 3, 2007.

[6] Muhirwe Jackson., "automatic speech recognition: human computer interface for kinyarwanda languag, " *Computer Science of Makerere University* , August 2005.

[7] Abeer M.Abu-Hantash and Ala'a Tayseer Spaih , "Text Independent Speaker Identification System " *An-Najah National University* , May 2010.

[8] Sarika Hegde , K. K. Achary and Surendra Shetty , "feature selection using fisher 's ratio technique for automatic speech recognition" *International Journal on Cybernetics & Informatics (IJCI)*, Vol. 4, No. 2, April 2015 .

[9] Kashyap Patel, R.K. Prasad , "Speech Recognition and Verification Using MFCC & VQ " *International Journal of Emerging Science and Engineering (IJESE)* ISSN: 2319–6378, Volume-1, Issue-7, May 2013.

[10] Vibha Tiwari., "MFCC and its applications in speaker recognition " *International Journal on Emerging Technologies* 1(1): 19-22(2010) .

[11] Shumaila Iqbal, Tahira Mahboob and Malik Sikandar Hayat Khiyal., "Voice Recognition using HMM with MFCC for Secure ATM " *IJCSI International Journal of Computer Science* Issues, Vol. 8, Issue 6, No 3, November 2011 .

[12] Garima Vyas and Barkha Kumari ., "Speaker Recognition System Based On MFCC and DCT " *International Journal of Engineering and Advanced Technology (IJEAT)* ISSN: 2249 – 8958, Volume-2, Issue-5, June 2013