Improved Text-independent Speaker Identification System For Real Time Applications

Nagwa M. AboElenein[1], Khalid M. Amin[2], Mina. Ibrahim[3], Mohiy M. Hadhoud[4]

[1] (Affiliation): Faculty of Computers and Information, Menofia University, Menofia, Egypt, nagwa.salim1@ menofia.edu.eg
[2] (Affiliation): Faculty of Computers and Information, Menofia University, Menofia, Egypt, k.amin@ci.menofia.edu.eg
[3] (Affiliation): Faculty of Computers and Information, Menofia University, Menofia, Egypt, mina. Ibrahim@ menofia.edu.eg
[4] (Affiliation): Faculty of Computers and Information, Menofia University, Menofia, Egypt, mmhadhoud@yahoo.com

***Abstract*— Speaker identification identifies the speaker among a set of users by matching against a set of voiceprints. In speaker identification, the identification time depends on the number of feature vectors, their dimensionality and the number of speakers. In this paper, text independent speaker identification model is developed by taking in MFCCs with VQ to obtain pressed feature vectors without losing much information, and the numbers of speakers are reduced in the test by gender detection algorithm. Gaussian Mixture Model (GMM) is used a modeling technique. Results show that proposed approach always yields better improvements in accuracy and brings almost 50% reduces in time processing**.

Keywords*— *Speaker Identification (SI); Gaussian Mixture Model (GMM); Mel Frequency Cepstral Coefficients (MFCC); Vector Quantization (VQ).

I. INRTODUCTION

The human speech carries different types of information. A lot of algorithms were proposed during the last two decades to solve different speech processing problems such as recognition of speech, speaker, speaker emotion, or language being spoken. The technology of automatic speaker recognition is employed to extract, characterize and recognize the information about speaker identity[1]. Speaker recognition can be divided into two different types: speaker verification and identification. Speaker verification task is to verify the claimed identity of person from his/ her voice. In speaker identification, the system determines who the speaking person is, which there is no identity claim for the person. For the identification, the task is usually divided into text-dependent and text-independent identification. The difference is that in the text-dependent identification, the system knows the text spoken by the person while in the text-independent identification, the system must be able to recognize the speaker from any text.

A wide variety of applications exists for speaker recognition systems. Security and forensics are the most widespread applications for speaker recognition[2].

For security, Most of applications are typically speaker verification systems intended to control access to privileged transactions or information remotely. Access control or telephone banking systems are good examples of such applications. For forensics, as voice is easy to acquire and can help to identify a perpetrator. Forensic applications are likely to be text independent speaker identification tasks, where a sample of speech from a suspect is compared to a recording of the perpetrator. They can either be text-dependent or text-independent, depending on the data available.

The process of speaker identification consists of two main phases: speaker enrollment and identification [3]. In the first phase, speech samples are gathered from the speakers, and the utterances are used to train their models. Finally, the collection of enrolled models is stored in a speaker database. In the second phase, a test utterance from an unknown speaker is compared against the speaker database. Feature extraction step is included in both phases which is used to extract speaker dependent characteristics from speech. The main purpose of feature extraction step is to retain speaker discriminative information and to reduce the amount of test data. Then these features are modeled and stored in the speaker database during the enrollment phase, or compared against the speaker database during the identification phase.

Vector Quantization (VQ) is one of the pattern classification techniques. Application of vector quantization in Speaker Recognition is found in [4]. VQ is widely applied to the Image compression [5] . The Gaussian model is the basic parametric model which is used in speaker identification [6]. Both the VQ and GMM were combined a mathematical model [7]. In the year 2006 VQ and GMM are widely applied to the speaker verification [8]. For our proposed system VQ and GMM with gender detection are used .It can improve time of processing to use in real time application.This paper is organized as follows, section II describes proposed method, In section III, the experimental results are discussed. Finally, the conclusion is reported in section IV.

II. PROPOSED METHOD

The proposed method stages are preprocessing, gender detection of the speech, feature extraction, vector quantization and feature matching. Figure. 1 represents the block diagram of the proposed method.

978-1-4673-8937-2/16/$31.00 ©2016 IEEE

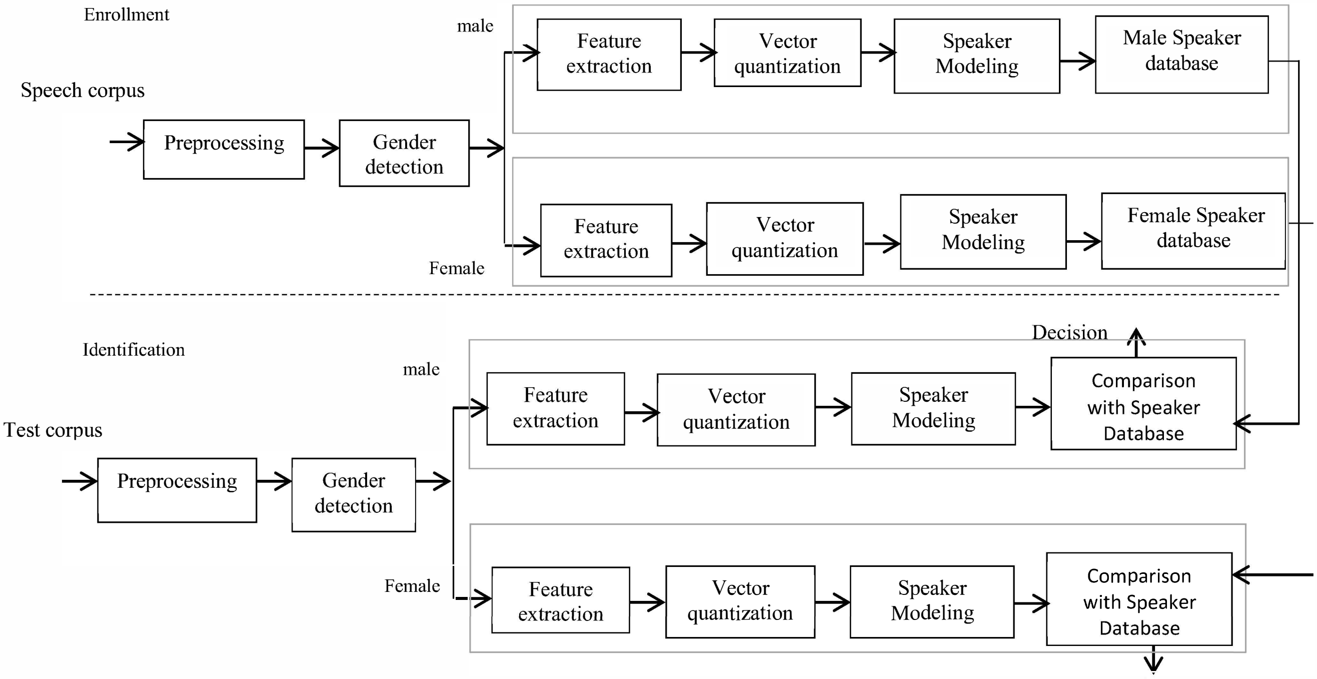

Figure. 1. Enrollment and Identification Phases

A. Preprocessing

In speaker identification system the pre-processing phase is used to increase the efficiency of feature extraction and classification stages and consequently to improve the overall speaker recognition performance. The pre-processing includes downsampling,Pre–emphasis, and silence removal.Downsampling is the process to reduce the sampling rate of a signal .The speech signal is down sampled from 44.1 kHz to 8 kHz[9] . Pre-emphasis [10] is a simple signal processing method that increases the amplitudes of high frequency bands and decreases the amplitudes of lower bands. In simple form, Pre-emphasis can be implemented as

$$Y[n] = X[n] - \alpha X[n-1]$$

Assume α= 0.95, which make 95% of any one sample is supposed to produce from the previous sample.

Speech of human does not consist only of connected speech sounds, but there are some silent regions between them. By removing these parts of speech, identification and speed are improved, because a number of frames will be reduced asimple thresholding criterion is to applied order remove the silence areas in the audio signal [11] .

B. Gender Detection

Gender detection [12]aims to predict the gender of the speaker by analyzing different parameters of the voice sample. Pitch detection algorithm PDA [13]uses autocorrelation method waveforms. The autocorrelation function is the correlation of a waveform with itself. It is based on the center-clipping method and infinite-clipping. Figure. 2 shows a block diagram of the pitch detection algorithm .The first stepof processing ,speech signal is segmented into overlapping frames 30-ms. This method requires low-pass filtering (LPF) to 900 Hz. The first stage of the process is the computation of a clipping threshold CL for the current 30-ms section of the speech. Following the clipping level is determinate, the 30-ms section of the speech is center clipped, and then infinite peak is clipped.

After that, the autocorrelation function for the 30-ms section is calculated over the range of lags from 20 to 160 samples. Additionally, the autocorrelation at 0 delay is calculated for voiced/unvoiced determination. If the maximum exceeds 55% of the autocorrelation value at 0 delay, the section is categorized as voiced and the location of the maximum is the pitch period. Otherwise, the section is categorized as unvoiced

C. *Feature Extraction*

Feature extraction is usually lossy transformation. Features play the main role in identifying the voice. For speaker recognition one of the most commonly used acoustic features are mel-scale frequency cepstral coefficient MFCC [14] . MFCC takes human perception sensitivity with respect to frequencies into consideration. MFCC is perhaps one of the best known and most popular also, it shows high accuracy results for clean speech. This is done by extracting Mel frequency cepstral coefficients (MFCC) Figure.3 shows the block diagram of extraction of Mel frequency cepstral coefficients (MFCC).

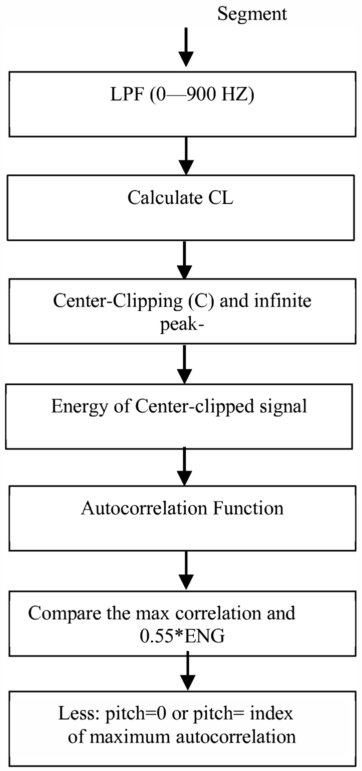
Segment
LPF (0—900 HZ)
Calculate CL
Center-Clipping (C) and infinite peak-
Energy of Center-clipped signal
Autocorrelation Function
Compare the max correlation and 0.55*ENG
Less: pitch=0 or pitch= index of maximum autocorrelation

Figure.2. Block Diagram of Pitch Detection Algorithm

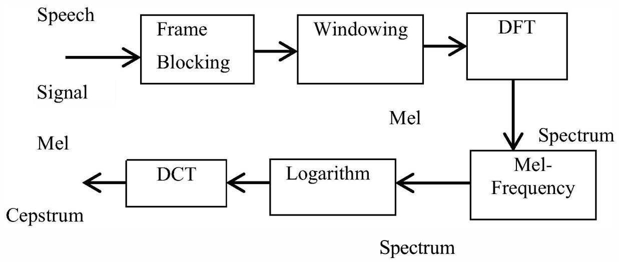

Figure. 3. MFCC Block Diagram

1) *Framing*

The framing step is the process of segmenting the speech samples into frame with N samples within the range of 20 to 30 msec. The voice signal is segmented into frames of N samples. Adjacent frames are separated by M (M<N). Typical values for *N* and *M* are *N* = 256, M=100.

2) *Windowing*

The next step in the processing is to window each frame and multiplies it with a Hamming window in order to reduce the signal discontinuities of the first and the last points in the frame if the signal in a frame is denoted by s(n), n = 0,…N-1, then the signal after Hamming windowing is s(n)*w(n), where w(n) is the Hamming window defined by:

$$w(n)=0.54-0.46\cos(2\pi n/(N-1)),0\le n\le N-1$$

3) *Fast Fourier Transform*

The Fast Fourier Transform (FFT) transforms each frame of N samples from the time domain to the frequency domain. The FFT is a speedy algorithm to perform the Discrete Fourier Transform (DFT) which is defined set of N samples {x}, as follows:

4) *Mel-frequency Wrapping*

Human perception of the frequency contents is studied by psychophysical and found that sounds for speech signals does not follow a linear scale. For each tone with an actual Frequency f, measured in Hz, a subjective pitch is measured on the Mel scale. The Mel-frequency scale is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. The following formula to compute the mels for a given frequency f in Hz:

$$\text{Mel}(f) = 2595*\log 10\ (1+f/700).$$

Filter bank is used to simulate the subjective spectrum. That filter bank has a triangular band pass frequency response. The reasons for using triangular bandpass filters (i) Smooth the magnitude spectrum; (ii) Minimize the size of the features .

5) *Discrete cosine transform*

In this final step to transform the log Mel spectrum into time domain Discrete Cosine Transform (DCT) is used[15]. The result of the transformation is called Mel Frequency Cepstrum Coefficient. The set of coefficients is called acoustic vectors. So each input utterance is transformed into a sequence of the acoustic vector.

$$Cm=\Sigma k=1\ N\cos\ [m*(k-0.5)*\pi/N]*Ek,\ m=1,\ 2\ldots\ldots L.$$

Where N = number of triangular band pass filters,and L = number of mel-scale cepstral coefficients. Usually N=30 and L=20. DCT transforms the frequency domain into a time domain. These features are referred to as the mel-scale cepstral coefficients.

D. *Vector Quantization (VQ)*

Vector Quantization (VQ) is a data reduction method which means that it seeks to reduce the number of dimensions. It is a process of representing the vectors from a large vector space into a limited number of regions present in that space. The each region so obtained is called a cluster and can be determined by its centre called the code word. The collection of all code words of the vectors is called a codebook. VQ technique is implemented through Linde-Buzo-Gray algorithm[16] .Feature vectors extracted from each MFCC were applied to VQ and new feature vectors of the test recording and the trained models were fed to GMM.

E. *Speaker Modeling*

Gaussian Mixture Models (GMM) [17] is a commonly used classification system for speaker identification. The reason for using GMM for speaker recognition is due to the fact that speech features are usually assumed to be Gaussian distributed. In text-independent systems where no prior knowledge what the speaker might say is known, GMM is one of the most successful classification systems. A Gaussian mixture density is a weighted sum of M component densities, given by the equation.

$$\mathrm{P}\ (\vec{x}/\lambda) = \sum_{i=1}^{M} p_i\, b_i\ (\overrightarrow{x}),\ \text{with}\ \sum_{i}^{M} p_{i=1.}$$

Where x is a random vector of D-dimension, λ is the speaker model, p_i are the mixture weights, b_i(x) are the density. The distribution of the feature vector x is modeled clearly using a mixture of M Gaussians

$$b_i\ (\overrightarrow{x}), = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T {\Sigma_i}^{-1}(x-\mu_i)\right)$$

With mean vector $\mu_i$and covariance matrix Σ_i. The parameters are collectively represented by the notation:

$$\lambda_i = \{\ \vec{P}_i\ , \vec{\mu}_i\ , \vec{\Sigma}_i\ \}$$

Each speaker is represented by a GMM and is referred by his/her model λ.

1) *Maximum Likelihood Parameter Estimation*

The main purpose of speaker model training is to estimate the parameters of the GMM. The main aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM .The basic idea of the EM algorithm is, beginning with an initial model.The new model then becomes the initial model for the next iteration, and the process is repeated until some convergence threshold is reached. The initial model is typically derived by random equation. Consider the sequence of T training vectors X = {x1……, x_T } on each EM iteration, the following re-estimation formulas are used

Mixture weight $$Wi = \frac{1}{T}\sum_{t=1}^{T} p(i/xt,\lambda)$$

Means $$\mu i = \frac{\sum_{t=1}^{T} p(i/xt,\lambda)xt}{\sum_{t=1}^{T} p(i/xt,\lambda)}$$

Covariance $$\Sigma_i = \frac{\sum_{t=1}^{T} p(i/xt,\lambda)xt^2}{\sum_{t=1}^{T} p(i/xt,\lambda)}\ \text{-}\mu t^2$$

In testing phase, likelihood is calculated with trained w, μ,

Σ and speaker corresponds to maximum Likelihood is selected

III. RESULTS

A. *Dataset*

The experiment is performed using the CHAINS Speech Corpus, databases. The CHAINS corpus[18] is a speech database expressly designed to help to characterize speakers as individuals. The corpus contains the utterances of 36 speakers obtained in two different sessions with a time separation of about two months which includes 16 females and 20 males speakers. Speech time of 2sec is used in the experiment, solo reading is used.

B. *Experiment*

All work is implemented using Windows 7, Intel core (i-3)2.5GHz, Matlab 2012a. There are 6 sentences per speaker, so 4 of these sentences will be used for training and the 2 remaining ones will be used for testing. Each time of speech is 2sec that sums 8 sec of training data for each speaker and 2sec for testing with 2 tests per speaker. The result for gender detection is 100%. The parameters for MFCC are sampling frequency=8000Hz, frame size=256samples, frame overlap is 33%, cepstral coefficient=13, number of filter bank=32,and Gaussian component=16,32,64. The performance of biometrics system is measured using correct identification rate.

$$\mathrm{CIR}=\frac{Number\ of\ correctly\ identified\ claims}{Totalnumberofclaims}\times 100\%$$

The proposed algorithm gives good result as shown in figure 4, where VQ and GMM give accuracy 88%, proposed algorithm gives 91%. Also as shown in figure 5 ,time of testing for VQ and GMM is 0.2242sec and for our proposed algorithm is 0.1051sec .The time is reduced to almost half because instead of comparing with 36 speakers in the database it is compared with 18 speakers. results for 16 and 32 mixtures are similar. But time of testing for 16 mixture is less 32 mixture This is the reason why 16 mixtures per state will be chosen.

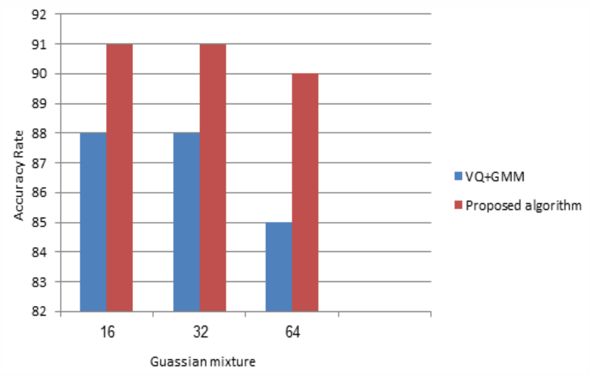

Figure 4. The performance of proposed algorithm with different number of mixtures use corpus database.

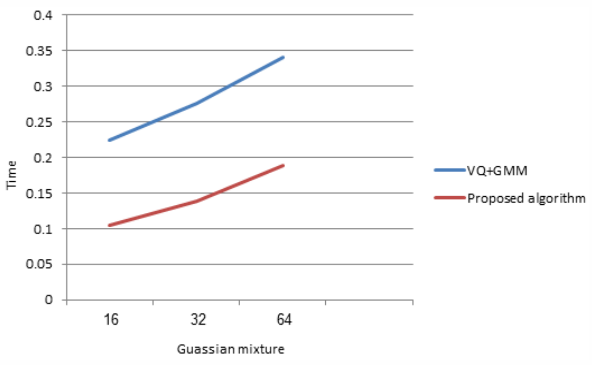

Figure 5. The Time of proposed algorithm with different number of mixtures use corpus database

With comparison between our proposed algorithm and in hybrid VQ with GMM [19], speaker identification system distinguishes between male and female speakers using VQ decision approach. Accuracy starts off highly 89.4%, and slowly decreases to approximately 77.5% and almost 20% reduce in time processing. It shown that our proposed system gives best results in accuracy and in time.

IV. CONCLUSIONS

Speaker identification has been recognized as an important biometric security tool in today’s world. This paper improves the time for text independent speaker identification system the proposed algorithm use gender detection, MFCC, VQ in training and GMM in testing. The results show that proposed algorithm achieved high recognition rate 91% and testing time is 0.1051 second.

References

[1] Rohit Singh, Harish Kumar, and Samreen Zehra." *Real Time Speaker Recognition*". A project report. Babu Banarsi Das Institute of Technology .Ghaziabad .2012.

[2] Christoph Kozielski, Martin Rothbucher, and Klaus Diepold. *Online Speaker Recognition for Teleconferencing Systems*. Technical Report, Institute for Data Processing.2014.

[3] Evgeny Karpov.."Real Time Speaker Identification". Master`s thesis, Department of Computer Science, University of Joensuu, 2003.

[4] Soong, Frank K., et al..A vector quantization approach to speaker recognition. AT&T technical journal 66.2. 14-26.1987.

[5] M. Mittal, R. Lamba. "Image Compression Using Vector Quantization Algorithms": A Review," International Journal of Advanced Research in Computer Science and Software Engineering(vol. 3, no. 6, June). 2013.

[6] Reynolds, Douglas, and Richard C. Rose."Robust text-independent speaker identification using Gaussian mixture speaker models"., in Speech and Audio Processing, IEEE Transactions on 3.1. 1995

[7] Gersho, Allen. "Asymptotically optimal block quantization". in Information Theory, IEEE Transactions 1979.

[8] Kinnunen, Tomi, Evgeny Karpov, and Pasi Franti. "Real-time speaker identification and verification". in Audio, Speech, and Language Processing IEEE Transactions on 14.1.2006 .

[9] B.H.lee S.M.kub, Ed., "Real time digital signal processing implementation and application". England: John wiley.

[10] Vergin, R., and Douglas O'Shaughnessy."Pre-emphasis and speech recognition. in Electrical and Computer Engineering", Canadian Conference on.(Vol. 2. IEEE). 1995

[11] T.Giannakopoulos. "A method for silence removal and segmentation of speech signals, implemented in Matlab", Insititute of Informatics and Telecommunications (IIT).

[12] Pawan.k, Nitika.J, Anirban.B, and Mahesh.C,.2011. "Gender Classification Using Pitch and Formants". in International Conference on Communication, Computing & Security, ICCCS , india.

[13] Tan, Li, and Montri Karnjanadecha."PITCH DETECTION ALGORITHM. AUTOCORRELATION METHOD AND AMDF" . 2003

[14] Koustav Chakraborty,Asmita Talele,Savitha Upadhya. 2014. "Voice Recognition Using MFCC Algorithm". International Journal of Innovative Research in Advanced Engineering (IJIRAE), vol. 1, no. 10.

[15] A.H.M. Ashfak Habib and S. Biswas. 2010-2011. "AUTOMATIC SPEECH RECOGNITION".Computer Science and Research Journal, vol. 07.

[16] SHIKHA.G, MOHD. S. 2015.Speech Recognition using MFCC & VQ. International Journal of Scientific Engineering and Technology Research, vol. 04, (january).

[17] G.S. KUMAR,K.A.PRASAD RAJU,M. Rao CPVNJ , P.Satheesh . 2010. "SPEAKER RECOGNITION USING GMM". International Journal of Engineering Science and Technology, vol. 2(6).

[18] [Online].file:///H:/database/solo.tar.7z/doc/html/Chains_overview2.html

[19] Piyush Lotia, M.R. Khan, 2011. "Multistage VQ Based GMM For Text". International Journal of Soft Computing and Engineering (IJSCE), vol. 1, no. 2.