Investigation of distance effect on Gaussian Mixture Models in Speaker Identification

Nada A.GH.Shindala
Assistant lecture
college of engineering/computer dep

Abstract

This paper investigate the effect of distance on the Gaussian Mixture Models (GMM) for text dependent speaker identification. Three stages are used for three different distances from the microphone (1m, 2m, and 3m). The set of feature extraction used here include Mel frequency cepstral coefficient (MFCC), Bark frequency cepstral coefficient (BFCC) and linear predictive cepstral coefficient (LPCC). These features are obtained from 20 speakers (10 adults and 10 children) ;all spoke five Arabic words in 5 seconds. The set of classification includes two types GMM and multilayer perceptron neural network (MLP). Total results show that MFCC has the best performance in feature extraction, and GMM has better recognition than MLP as total recognition in GMM is 93.15% and recognition in MLP is 88.06%.The results show also that the recognition rate decreases from 93.15% to 80.82% as the distance is increased from 1m to 3m.

Keywords: Speaker Identification, Gaussian Mixture Models Preceptron Neural Network

دراسة تأثير المسافة على موديلات الخليط الكاوسي في تمييز هوية المتكلم

ندى عبد الغني شندالة
مدرس مساعد/كلية الهندسة/جامعة الموصل
قسم هندسة الحاسبات

الخلاصة

يتناول هذا البحث دراسة تأثير المسافة على موديلات الخليط الكاوسي (GMM) لتمييز هوية المتكلم، استخدمت ثلاث مراحل ولثلاث مسافات مختلفة البعد عن المايكروفون (1 مترو 2 مترو3 متر)، أنواع استخلاص الصفات هي معاملات التردد الميلي (MFCC)و معاملات التردد الباركي (BFCC) ومعاملات التنبؤ الخطي (LPCC) وهذه الطرائق استخلصت من 20 متكلم (10 بالغين، 10 أطفال) وكل متكلم نطق خمس كلمات عربية ولمدة خمسة ثوان .إن طرائق التمييز المستخدمة تتضمن نوعان: الأول موديلات الخليط الكاوسي (GMM)، والثاني الشبكة العصبية متعددة الطبقات (MLP) وأثبتت النتائج إن استخدام طريقة معاملات التردد الميلي هي الأحسن في استخلاص الصفات وطريقة (GMM) هي الأحسن في التمييز، حيث كانت نسبه التمييز في (GMM) 93.15% وفي الشبكة العصبية 88.06%. كما بينت النتائج أن نسبة التمييز تقل من93.1% إلى 80.82% كلما زيدت المسا فه من 1متر إلى 3متر.

Received: 6 – 3 - 2011 Accepted: 4 – 7 - 2011

1- Introduction

The Speech signal conveys several levels of information Primarily the speech signal conveys the words or message being spoken, but on a secondary level, the signal also conveys information about the identity of the talker. While the area of speech recognition is concerned with extracting the underlying linguistic message in an utterance, the area of speaker recognition is concerned with extracting the identity of the person speaking the utterance. As speech interaction with computers becomes more pervasive in activities such as telephone financial transactions and information retrieval from speech databases, the utility of automatically recognizing a speaker based solely on vocal characteristics increases[1].

Speaker recognition uses the acoustic features of the speech signal to discriminate between individuals. These acoustic features can vary greatly from one speaker to another depending upon their anatomy and behavioral characteristics. Modeling these acoustic features is useful in speaker recognition, as they can be used to identify individuals[2].

Speaker recognition can be classified into identification and verification. *Speaker identification* is the process of determining which registered speaker provides a given utterance. *Speaker verification*, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. The system that we will describe is classified as *text-independent speaker identification* system since its task is to identify the person who speaks regardless of what is saying while text-dependent uses the same words[3].

All technologies of speaker recognition, identification and verification, text-independent and text-dependent, each has its own advantages and disadvantages and may requires different treatments and techniques. The choice of which technology to use is application-specific. At the highest level, all speaker recognition systems contain two main modules: feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers[4].

Implementation of a speaker recognition system requires the human speech content to convey meaning to a machine. The human voice consists of sounds that are characterized by the behavior and physiology of the individual. For instance, utterances produced by an individual are from the same vocal tract, and tend to have a typical pitch range, along with the characteristics associated with dialect or gender [5].

Speaker identification can be classified into two types, based on anonymity. These are open-set and closed-set speaker identification. Both sets use a database of registered speakers for identification, with the main difference being in the decision process. For open-set the decision is based upon the enrolled speakers together with the possibility that the speaker is unknown. Closed-set only considers the best match from the enrolled speakers. Figure (1) shows the categories of speaker recognition[4].

There are many algorithms and models that can be used for speaker recognition including Neural Networks, unimodal Gaussians, Vector Quantization, Radial Basis Functions, Hidden Markov Models and Gaussian Mixture Models(GMMs). These perform well under clean speech conditions, but in many cases performance degrades when test utterances are corrupted by noise, mismatched conditions or if there are small amounts of training and testing data. Among these methods GMMs are usually preferred because they offer high classification accuracy while still being robust to corruptions in the speech signal[5].

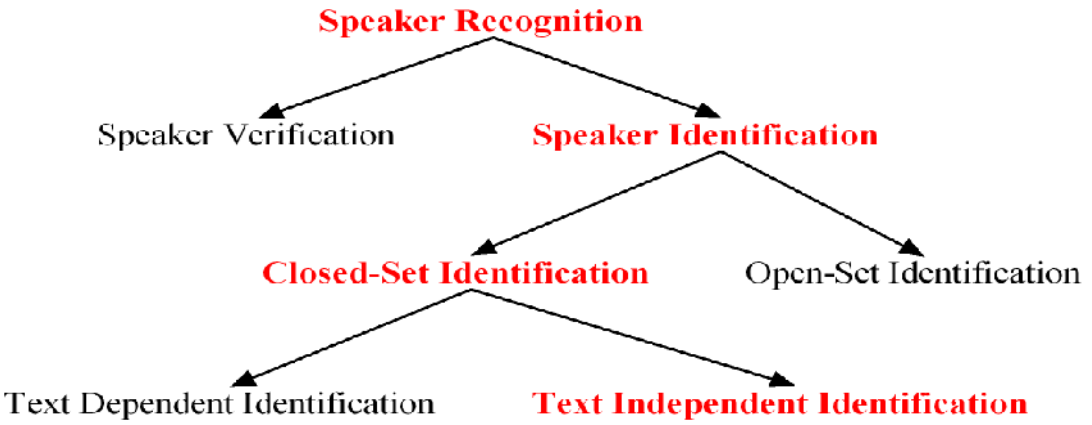

Fig 1: Categories of speaker recognition

When speech is corrupted by noise or by the limited bandwidth of telephone lines, speaker recognition accuracy degrades. The feature vectors generated from corrupted speech are no longer similar to the class distributions learned from the training data. Because of the channel effects, there is inherently more variability in the training data, and as a result, the variance of the distributions of the speaker classes increases.

This broadening of the class distributions leads to increased classification errors over the case where the training and test speech are both clean. There is also an intrinsic variation in a person's voice which is more pronounced when the voice samples are collected at widely separated times.

In this paper three stages are used for three different distances from microphone (1m,2m,3m),with three feature extraction methods ,namely; Mel frequency cepstral coefficient (MFCC), Bark frequency cepstral coefficient (BFCC) ,and Linear predictive cepstral coefficient (LPCC). MLP and GMM classifications are used for 20 speakers; all spoke Arabic words in 5 sec. The resulting features are compared at these per mentioned distances and show that MFCC has the best performance ,GMM has better recognition than MLP.

Besides this introduction, this paper contains another three sections. Section 2 describes the proposed block diagram of speaker recognition. Experimental results are given in section 3. Finally, section 4 concludes this paper.

2 Block Diagram Of Speaker Recognition

The proposed speaker recognition system uses a text-dependent speaker recognition to simplify a complex speaker recognition system which can improve the accuracy of the speaker recognition task by studying the input data system. This system consists of three main phases as show Figure 2:-

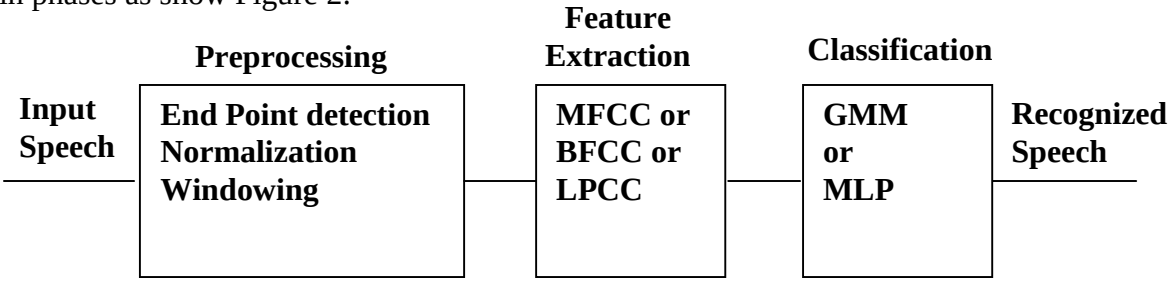

Fig.2 Recognition Block

1. Speech pre-processing phase,
2. Feature extraction phase, and
3. Recognition phase which consists of a training module and a testing module [6].

2.1Speech preprocessing:

First step of preprocessing was removed silence part ,we use the endpoint detection algorithm based short time energy and zero crossing rate to locate the beginning and ending of a speech signal. The short time energy and zero crossing are defined as:[7].

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)\ \omega\ (n - m)\ ]^2 \quad .....(1)$$

$$Z_n = \sum_{m=-\infty}^{\infty} |\ sgn[x(m)]\ - sgn[x(m\ -1)]\ |\ \omega\ (n - m) \quad .....(2)$$

ω(n) is the window function. The hamming window has a wide main lobe and small side lobes, making it a smooth low-pass filter with less leakage. Therefore,
the hamming window has been adopted, given as

$$\omega(n) = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi\ n}{N\ -1}\right), 0 \le n \le N\ -1 \\ 0, \end{cases} \quad .....(3)$$

The endpoint detection algorithm is implemented through the following steps:
1) Remove the dc offset in a signal to find the zero-crossing rate of the background noise.
2) Calculate the average energy and the average and standard deviation of the zero-crossing rate of the background noise, denoted by IMN, and σ_{IZC}, respectively. In addition, the zero-crossing threshold IZCT is determined as

$$IZCT\ = \min(\ IF\ , \overline{IZC}\ + 2\sigma_{\ IZC}\ ) \quad .....(4)$$

Where IF is a fixed threshold (i.e., 25 crossings/10 ms).
3) Calculate the average energy En of the entire signal and find the peak energy IMX. The lower energy threshold ITL and the upper energy threshold ITU are se set as

I1=0.03 * (IMX − IMN) + IMN …….(5a)
I2 =4 * IMN ……(5b)
ITL = min(I1, I2) …….(5c)
ITU =5 * ITL ……...(5d)

4) Find an interval of En that exceeds threshold ITU. Then, back off toward the signal beginning to find the first point at which En falls below ITL, denoted by N1, and the point N2 is searched in a similar way.
5) Move M samples backward from N1 to N1 −M, compare Zn with IZCT, and find the first point where Zn exceeds IZCT. Similarly, move M samples forward from N2 to N1 +M, compare Zn with IZCT, and find
the last point where Zn exceeds IZCT. These two points are declared as the final endpoints[8].

To illustrate the aforementioned endpoint detection algorithm, an example is given in Fig 3. The three thresholds are important in detecting the endpoints of speech signals. If the thresholds are set too small, the background noise may wrongly be regarded as a speech signal, resulting in less reliable stable speech features extracted. High threshold level will lead to the loss of speech information. The thresholds in the algorithm are experimentally set by testing over a variety of recording conditions and a large number of speakers

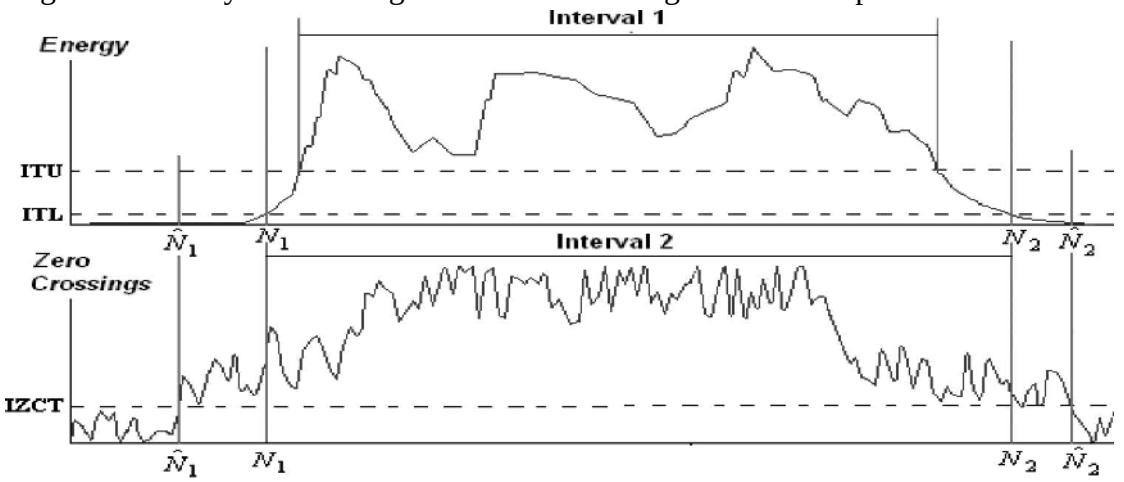

Fig.3 Example of an endpoint detection

2.2 Feature Extraction:

The most important parts of a speaker recognition system are the feature extraction which converts the properties of the important signal for the pattern recognition task to a format that simplifies the distinction of the classes. The recognition step aims to estimate the general extension of the classes within feature space from a training set [2].

An important problem in speech recognition systems is to determine a representation that is well adapted for extracting information content of speech signals .That information contains pitch ,formant frequency. In this paper we used three type of extraction :

2.2.1 Mel Frequency Cepstral Coefficient(MFCC)

MEL Frequency Cepstral Coefficients (MFCC) are used extensively in Automatic Speaker Recognition (ASR).MFCC features are derived from the FFT magnitude spectrum by applying a filter bank which has filters evenly spaced on a warped frequency scale. The logarithm of the energy in each filter is calculated and accumulated before a Discrete Cosine Transform (DCT) is applied to produce the MFCC feature vector. The frequency warping scale used for filter spacing in MFCC is the Mel (Melody) scale. The scale was devised through human perception experiments where subjects were asked to adjust a stimulus tone to perceptually half the pitch of a reference tone. The resulting scale was one in which 1Mel represents one-thousandth of the pitch of 1 kHz and a doubling of Mel's produces a perceptual doubling of pitch[9].

In sound processing, 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. MFCCs are coefficients that collectively make up an MFC.

This frequency warping can allow for better representation of sound [8]. Fig.4 illustrates extraction of MFCCs. The first step is to divide the speech signal into blocks using overlapping smooth windows such as Hamming ,Henning , etc. The next step is to take the

Discrete Time Fourier Transform (DTFT) of the windowed signal. Next, the square of the DTFT of the windowed signal is calculated. The outputs of the fourth step are the Mel-scaled filter bank energies. The fifth step involves calculating the logarithm of the Mel-scaled filter bank energies. The last step involves taking the Discrete Cosine transform (DCT) of the Mel-scaled log-filter bank energies to calculate MFCCs[10].

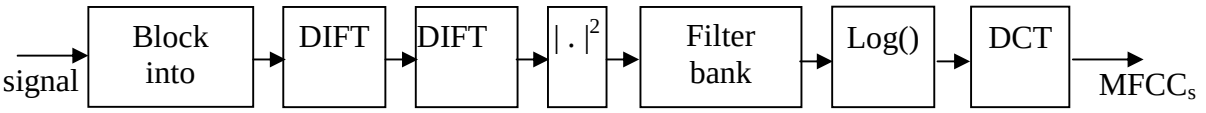

Fig .4 Extraction of the MFCCs

In the Mel frequency wrapping block, the signal is plotted(as shown in fig.5) against the Mel-spectrum to mimic human hearing. Studies have shown that human hearing does not follow the linear scale but rather the Mel-spectrum scale which is a linear spacing below 1000 Hz and logarithmic scaling above 1000 Hz. In the final step, the Mel-spectrum plot is converted back to the time domain by using the following equation:

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f/700) \quad \ldots..(6)$$

The resultant matrices are referred to as the Mel-Frequency Cepstrum Coefficients [3][1].

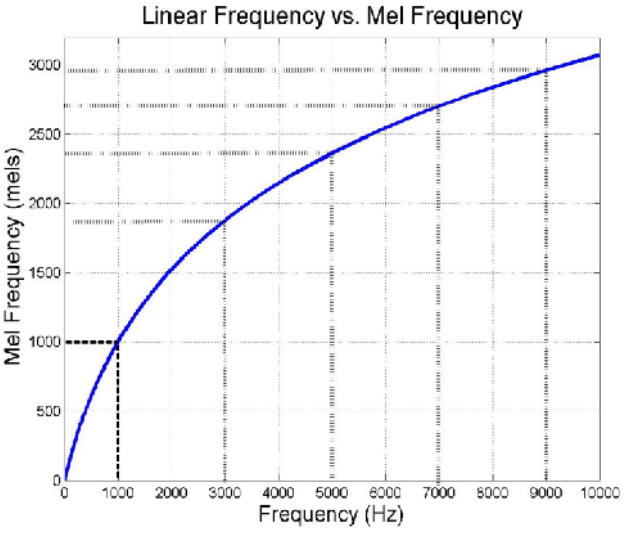

Fig.5 characteristic of Mel frequency

2.2.2 Bark Frequency Cepstrum Coefficient: *BFCC*

One of the classic approaches to analyze and process signal spectra is the Bark frequency scale(also called "critical band rate") .Based on the results of many psychoacoustic experiments, the Bark scale is defined so that the critical bands of human hearing have a width of one Bark. By representing spectral energy (in dB) over the Bark scale, a closer correspondence is obtained with spectral information processing in the ear. Based on the results of many psychoacoustic experiments, the Bark scale is defined so that the critical bands of human hearing have a width of one Bark[11].

The Bark scale ranges from 1 to 24 Barks, corresponding to the first 24 critical bands of hearing. The published Bark band edges are given in Hertz as [0, 100, 200, 300, 400, 510, 630, 770,920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000,15500]. The published band centers in Hertz are [50, 150, 250, 350, 450, 570, 700, 840, 1000,1170, 1370, 1600, 1850, 2150, 2500, 2900, 3400, 4000, 4800, 5800, 7000, 8500, 10500, 13500]. These center-frequencies and bandwidths are to be interpreted as samplings of a continuous variation in the frequency response of the ear to a sinusoid or narrow band noise process[12].

Bark-frequency cepstrum coefficients (BFCC) is used in a similar way as in the MFCC, except that the power spectrum is wrapped along its frequency axis onto the bark frequency using the following equation:

Bark(f) =26.81f/(1960+f)-0.53 ……..(7)

2.2.3Linear prediction cepstral coefficient LPCC

Linear prediction uses an all-pole model to represent the speech signal and to relate to formants [12]. The basic idea of linear-prediction coefficients (LPC) is to approximate the current speech sample as a linear combination of past speech samples, i.e.,

x(n) =a1x(n-1)+a2x(n-2)+……+apx(n-p) …..(8)

where coefficients a1, a2, . . . , a pare determined by minimizing the mean square error (MSE) between the actual and the predicted speech signals. Using autocorrelation to minimize MSE with respect to ai, (i = 1, . . . , p), we have

a = R−1r ……..(9)

where r = [r(1), r(2), . . . , r(p)] T is the autocorrelation vector. R is a p×p nonsingular Toeplitz autocorrelation matrix.

Linear-prediction cepstral coefficients (LPCC) is LPC in the cepstrum domain. Given a frame of speech signal x(n), let the Fourier transform of x(n) be X(ω). The cepstrum of x(n) is then defined as the inverse Fourier transform of the logarithm of the magnitude spectrum, i.e., cx(n) = ifft(log(|X(ω)|)). In practice, LPCC can be derived from LPC , We have

$$c_m = a_m + \sum_{k=1}^{n-1} \left(\frac{k}{m}\right) a_k c_{m-k}, \quad 1 \le m \le p \qquad (10)$$

LPCC is shown to be more robust and relevant for speech recognition, and hence, we include this feature in the proposed automated recognition system. However, LPCC has the disadvantage of linearly approximating speech in all frequencies, which is not the case with the perception of human hearing[14].

2.3 Classification

Two type of classification human hearing are used in this research:

2.3.1 **Gaussian Mixture Model(GMM)**

GMM classifier has grained increasing attention in pattern recognition community. GMM can be classified as a semi parametric density estimation method since it defines a very general class of functional forms for the density model. In this mixture model, a probability density function is expressed as a linear combination of basic function. Improved classification performances have been demonstrated in many pattern recognition applications. A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation,

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i g(x \mid \mu_i, \Sigma_i) \qquad (11)$$

Where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), wi, i = 1, . . . ,M, are the mixture weights, and g(x|μi,_i), i = 1, . . . ,M, are the component Gaussian densities. Each component density is a D-vitiate Gaussian function of the form,

$$g(x \mid \mu_i, \sum_i) = \frac{1}{(2\pi 2^{D/2} \mid \sum_i \mid^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu i)' \sum_i^{-1} (x-\mu_i)\right\} \quad .....(12)$$

with mean vector μi and covariance matrix Σi. The mixture weights satisfy the constraint that

$$\sum_{i=1}^{M} w_i = 1$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation,

$$\lambda = \left\{w_i, \mu_i, \sum_i\right\} \quad i=1,\ldots\ldots M \quad \ldots\ldots\ldots\ldots\ldots(13)$$

For speaker identification ,each speaker is represented by a GMM and is referred to by his/her .

Given training vectors and a GMM configuration, we wish to estimate the parameters of the GMM λ, which in some sense best matches the distribution of the training feature vectors. There are several techniques available for estimating the parameters of a GMM . By far the most popular and well-established method is maximum likelihood (ML) estimation[1].

The aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM given the training data. For a sequence of T training vectors X = {x1, . . . ,xT}, the GMM likelihood, assuming independence between the vectors, can be written as,

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(x_t \mid \lambda) \quad .....(14)$$

Unfortunately, this expression is a non-linear function of the parameters λ and direct maximization is not possible. However ,ML parameter estimates can be obtained iteratively using a special case of the expectation-maximization (EM) algorithm [15].

The EM algorithm starts with an initial guess for all the parameters to be estimated and then iterates over two steps: 1) the expectation step (E-step) and 2) the maximization step (M-step). In the E-step, a probability distribution is computed based on the current parameter estimates. In the M-step ,maximum-likelihood parameter estimates are computed based on the distributions computed in the E-step.

The basic idea of the EM algorithm is beginning with an initial model λ ,to estimate a new model $\bar{\lambda}$,such that $p(X/\bar{\lambda}) \geq P(X/\lambda)$. The new model then becomes the initial model for next iteration and the process is repeated until some convergence threshold is reached .[16][17][18].

2. 3.2 Multilayer Perceptron Neural Network(MLP):

The MLP is an artificial neural network that can model non-linear functions by using non-linear sigmoid functions in its hidden layer. In fact, it has been proven that a MLP can model any arbitrary function using only three layers, given that it has enough inputs and hidden units [19].This property makes the MLP a universal classifier/identifier and a perfect candidate for our speaker identification.

Our MLP networks contained three layers, one input with number of neurons depend upon feature vector ,one hidden layer with neurons exchange experimentally and one output layer with neurons depend upon number of speaker .The transfer function was a log sigmoid .The number of epochs that the network trained on ranged from 100 to 10000 epochs. Once again this was dependent on both the type and dimension of the input and output data used in each network[20].

3.Experimental work:

This section describe the current experimental work, the data base contains the speech data files of 20 speakers.10 adult speakers(5male,5 female).10 child speakers of different ages(6_10 years). These speech files consist of isolated Arabic words such as(close ,no ,go ,end and repeat). The words repeat three times(1meter,2m,3m) from microphone .For this type we used very high _quality microphone(impedance:150_200Ω, phantom power:12_48VDc, sensitivity: 20mv/Pascal which gives output 4dBu and -112dBu noise floor). All samples are stored in Microsoft wave format files with 11025 Hz sampling rate, 16 bit PCM and mono channels. The time for each speaker is 5 second(all was stored using MATLAB 7.4), Fig 6a shows the utterance of one word, for all speakers number of utterances are 300 half of it using in training and others for testing.

In this experiment , a continuous speech signal is first classified in to speech and non speech segments using the endpoint detection algorithm. After removing the non speech segments(as shown in Fig6b), the speech segments are divided into frames (which may have overlap) of length 10–20 ms . Feature extraction is then performed which is comprised of six stages :pre-emphasis, frame blocking, hamming window to lessen distortion, Fast Fourier Transform (FFT), triangular band pass filter, and cosine transform to get MFCC. For simplicity, we use 14 order Mel-scale cepstrum parameters14MFCC,14BFCC and 14LPCC.

In the first experiment we used a constant of mixture model(M=64) for all speaker ,then the evaluation of a speaker identification experiment processed the test speech signal by front end analysis to produce a sequence of feature vectors{x1,,x2…..xt).To evaluate different test utterance lengths ,the sequence of feature vectors were divided into overlapping fames of F feature vectors.

The first two frames from a sequence would be

First frame	Second frame
X1	X2
X2	X3
.	.
.	.
XF	XF+1

The test frame length was 5 seconds corresponds of F=500 feature vectors at 10ms frame rate .The identified speaker of each frame was compared to actual speaker of test utterance and the number of frame which were correctly identified was tabulated.

The above steps were repeated for test utterance from each speaker .The final performance was then computed as the percent of correctly identified F length frame over all test utterances.

$$\%\ \text{correct identification} = \frac{\text{number of correctly identified frames}}{\text{Total number of frames}} \qquad \text{……..(15)}$$

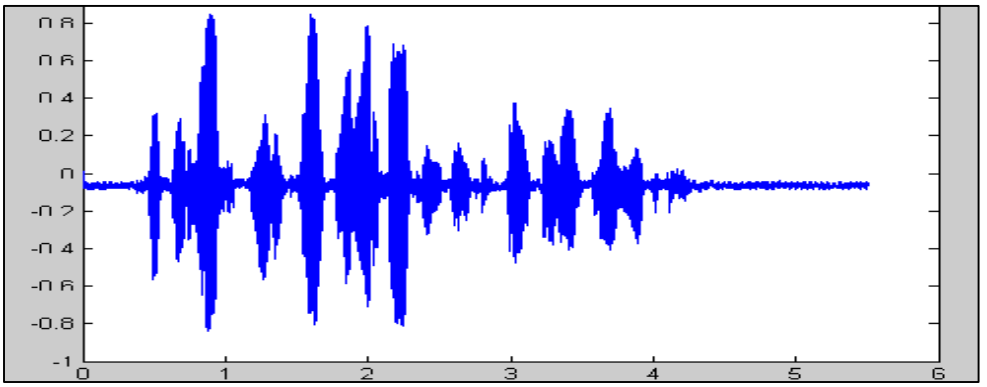

Fig 6a: Utterance of one word

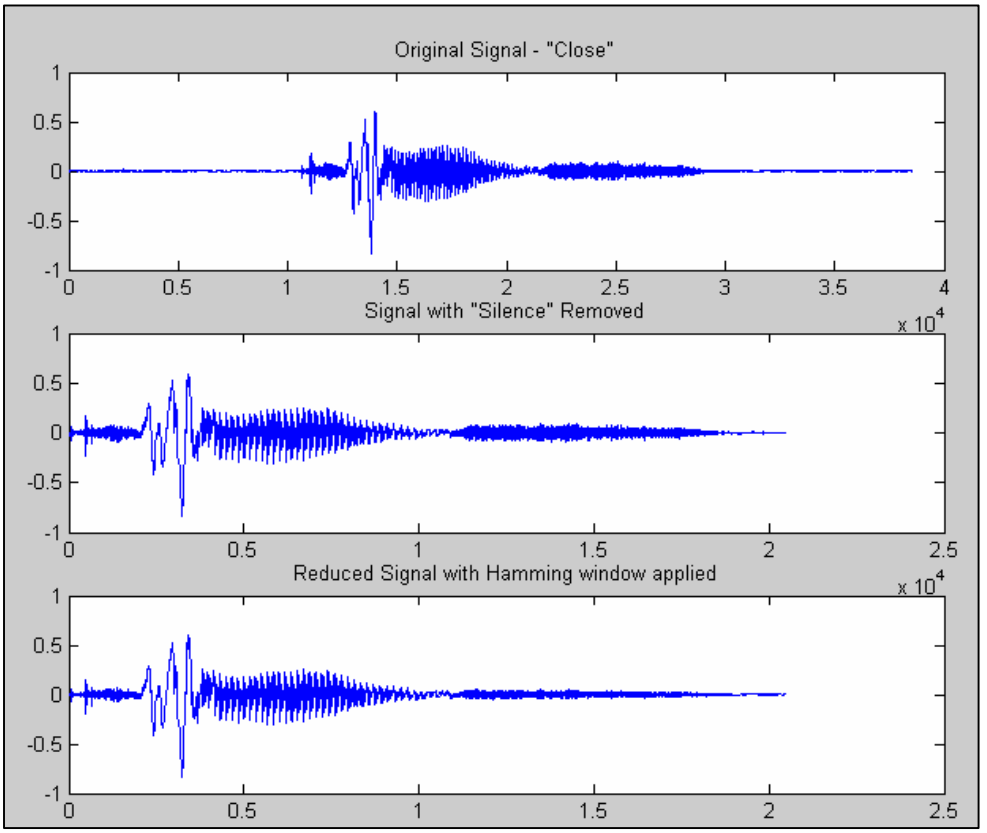

Fig 6b: removing the non speech segments

The evaluation was repeated for different value of F to evaluate performance with respect to test utterance length.

Also the former step was repeated for all three distance(1m,2m,3m).. Table 1 lists the obtained results ,where Table 2 lists total recognition rate .It is obvious that distance affect dramatically on the ability of the system to recognize certain speaker and care should be paid to keep users in a suitable distance from the system.

In the second experiment the mixture model was exchanged from 8 to 128 for all three distances then the accuracy recognized. Fig 7, Fig8, and Fig9 show the accuracy for classification using GMM with different number of Gaussian components .These figures indicate that the classification performance improves as the number of Gaussian components increases from 8 to 64 and saturate at 128.

Table 1. Three Distance Recognition (F:Female,M:Male,C:Child)

		1 meter		2 meter		3 meters	
		GMM	MLP	GMM	MLP	GMM	MLP
LPCC	F	92.32	87.38	90.33	84.11	89.79	77.13
	M	91.41	84.33	90.21	82.32	88.20	80.33
	C	90.08	82.11	89.11	80.21	85.71	79.01
MFCC	F	94.01	89.10	91.31	82.20	90.21	84.12
	M	93.42	90.52	92.32	83.33	89.81	81.11
	C	91.2	86.70	88.11	81.21	87.13	80.23
BFCC	F	93.12	87.71	88.31	85.11	86.12	81.22
	M	92.47	86.50	89.88	81.13	87.12	87.02
	C	89.71	80.01	87.13	80.02	85.88	77.33

Table 2 Total Recognition

	1 meter		2 meter		3 meters	
	GMM	MLP	GMM	MLP	GMM	MLP
F	93.15	88	89.98	83.80	88.70	80.82
M	92.43	87.11	90.80	82.26	88.37	79.82
C	90.30	82.94	88.11	80.48	86.24	78.85

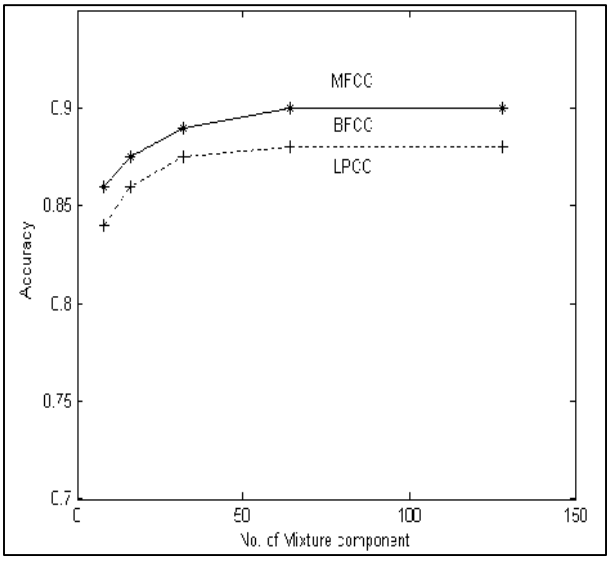

Fig 7 One Meter Accuracy for Different Mixture

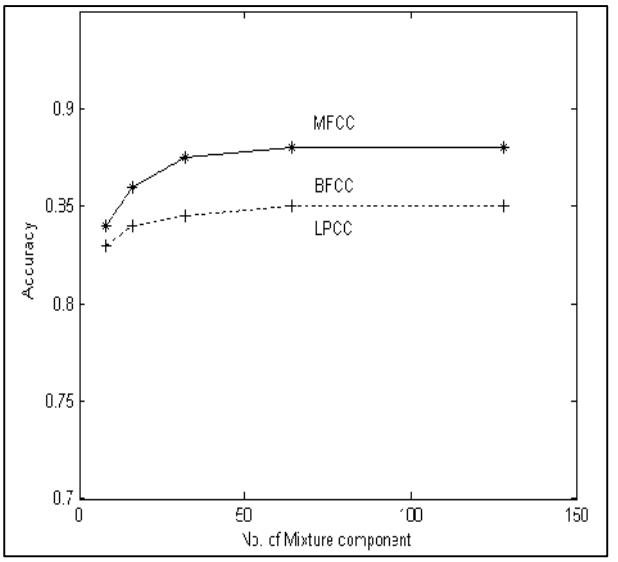

Fig 8 Two Meters Accuracy for Different Mixture

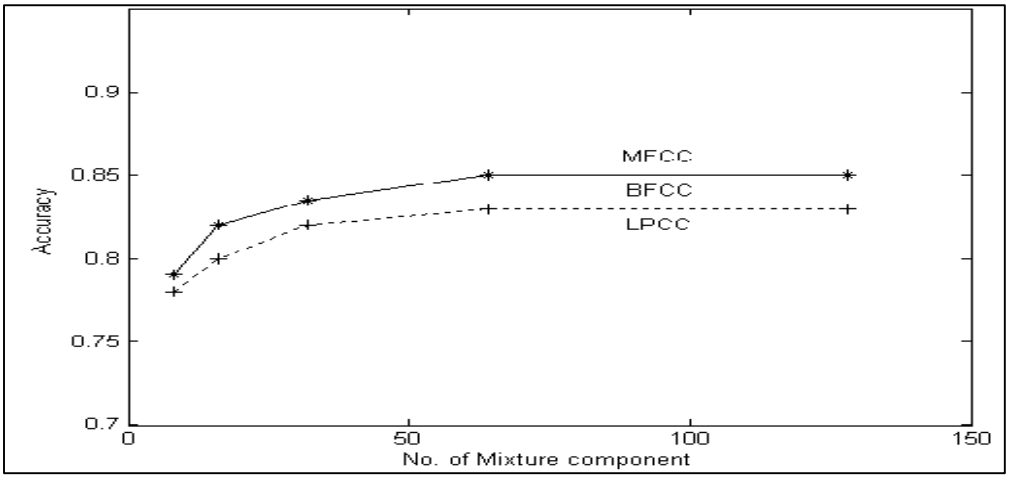

Fig 9 Three Meters Accuracy for Different Mixture

4. Conclusions:

In this paper the effect of distance between the speaker and the speaker recognition system has been studied. It has been found that speaker recognition rate decreases from 93.15 to 88.7 as the distance increases from 1m to 3m. An effective and robust feature of three types (MFCC, BFCC, LPCC) has been represented. The best one is MFCC because this type of feature captures the characteristics of speech signal and operates well in clean environment.

As a recommendation , it is suggested that the distance from microphone to the system have to (1m or less be limited) to get the better recognition.

References

[1]: D. A Reynolds and R. C. Rose, " Robust Text-Independent Speaker Identification Using Gaussian Mixture speaker Models", IEEE Transactions On Speech And Audio Processing, Vol.3, No. 1, 1995.

[2]: Campbell J.P., “Speaker Recognition: Atutorial,” Proc. IEEE, Vol.85, pp.1437-1462,Sept.1997.

[3]: G.Suvaran.kumar,K.A.Prasad,Mohan Rao,P.Saheech,"Speaker recognition using GMM",International Journal of Engineering Science and Technology,Vol.2(6),2010.

[4]: Rabiner l. juang B.H "Fundamental of speech recognition",Prentic-Hall,usa, ISBN :0-13-0151572,1993.

[5]: 5.M.I.Abdalla and H.Ali"Wavelet Based Mel-Frequency CepstralCofficient for speaker identification using Hidden Markov Models",journal of telec0munication Vol.1, ISSUE 2010.

[6]: S.Limpanakorn and ChularatTanprasert,"Voice Articulator for Thai-Speaker recognition",ThammasatIntj.sc Tech, Vol 6,no3,2001.

[7]: L.R.Rabiner and M.R.Sambur,"An algorithm for determining the end point of isolated utterances",Bell sys. Tech,Vol.54,no2,pp1399-1402,1975.

[8]: I.Lamel,L.Labine,A.Rosenlerge and J.Wilpon,"An improved End point Detector for isolated word Recognition",Vol.29,pp777-785,1981.

[9]: J.Deller.et.al"Discrete_Time processing of speech signal",MacMillan publishing Co.,ISBN:0-7803-58386-2,2000.

[10]: X.Huang,A,Acero and H.Hon,"Spoken language processing":Aguide to theory,algorithm,and system development.Prentice-Hall,Inc,2001 ,ISBN 0-13-022616-5.

[11]: Julius.O,Smith III member IEEE and Jonathan S.,Abel member IEEE"Bark and ERB Bilinear Transforms",Vol.7.No.6,IEEE 1999.

[12]: D.Dimitriadis,P.Maragos and A.Potamianos"AuditoryTeager Energy CepstrumCofficient for Robust speech Recognition".Proc. of European speech processing conference,Lisbon,Protugal,,2005.

[13]: E.Worng and S.Sridharan,"Comparison of linear predicfioncepstrum coefficient and mel-frequency cepstrum coefficient for language identification",in Proc. Intell Multimedia Video-Speech Process pp95-98,2001.

[14]: L.Shafaran,M.Riley and M.Mohri"Voicesignature",inProc.IEEE Workshop Autom speech recog. pp31-36,2003.

[15]: Minghua Shi and Amine Bermak,"An Efficient Digital VLSI Implementation of Gaussian Mixture Models-Based Classifier", IEEE,Vol.14,No.9,2006.

[16]: Y.Zeng,Z,Wu,T.Falk and W.Chan "Robust GMM based gender classification using pitch and RASTA-PLP parmaters of speech".inProc.Int.Couf.Mach.Learn Cybern,2006,pp3376-3379.

[17]: P.Mayorga,J.Martin,A.Hernands,J.I.Fiores"Gaussian Components Optimization for a Robot Controlled by Speech Commands in Mexican Spanish,IEEE 2007.

[18]: B.L.Pellon,J.H.L.Hansen"An Efficient Scoring Algorith for Gaussian Mixture Model Speaker Identification",IEEE,signal processing letters Vol.5,No.11,pp281-284,1998.

[19]: Y.Konig and N.Morgan"GDNN:Agender dependent neural network for continuous speech recognition".in Proc.IJCNN,1992,Vol.2,pp332-337.

[20]: Fan Sun,Bibo Li and HuishengChi"Some Key in speaker recognition using Neural Network approach",ch3065-0\91\0000,IEEE senior member,2005,National laboratory of machine perception.

The work was carried out at the college of Engineering. University of Mosul