# ANALYZING NOISE ROBUSTNESS OF PLP AND PLP-GC FEATURES IN SPEAKER IDENTIFICATION

*Amina Ben Abdallah, Zied Hajaie , Noureddine Ellouze*

Laboratory of Systems and Signal Processing National Engineering School of Tunis, TUNISIA
Amina.Benabdallah@enit.rnu.tn, Zied.Hajaiej@enit.rnu.tn, N.ellouze@enit.rnu.tn

## ABSTRACT

Several modern speaker recognition systems use a bank of linear filters as the primary step in performing frequency analysis of speech and extracting the acoustics parameters that permit characterizing the speaker identity. In this paper we point up the employ of novel feature set extracted from speech signal. The new skill for extracting these parameters is based on the human auditory system characteristics and relies on the Gammachirp Filterbank to imitate the cochlea frequency resolution with nonlinear resolution according to the equivalent rectangular bandwidth (ERB) scale. For evaluation a comparative study was operated with standard PLP, and the effect of these differences using an usual HMM/GMM for text independent speaker recognition system, for noisy environments. Performances were test database contaminated with additive noise different real-environment noises were used: car noise provided by Volvo, factory noise and white noise from Noisex92 [1]. Tests were carried out at different SNR levels (-3dB, 0dB, 3dB, 6 dB, 12 dB).

***Index Terms***— Speaker Identification, Gammachirp, PLP, PLP-GC, HMM/GMM.

## 1. INTRODUCTION

Security has turn out to be an extremely important issue due to wide use of internet technology as well as due to multi-user applications. Identifying users and yielding access only to those users who are authorized is a key to afford security. Users can be identified using a variety of approaches and their combinations. As the technology is getting advanced, more sophisticated approaches are being used to assure the need of security. Some of the most popular techniques are use of login and password, face recognition, fingerprint recognition; iris recognition etc. Use of login and password is becoming less reliable because of the ease with which hackers can usurp the password such as sophisticated electronic eavesdropping techniques [2]. Face recognition, fingerprint recognition and iris recognition also carry their own drawbacks. Users should be willing to endure the tests and should not get upset by these procedures when these techniques are used to identify them. Speaker identification permits nonintrusive monitoring and also achieves high accuracy rates which conform to most security requirements. Speaker recognition is the process of automatically recognizing who is speaking based on some unique characteristics present in speaker's voice [3]. For this recognition purpose, speaker specific characteristics present in speech signal need to be preserved. Job of Speaker recognition can be classified into two main categories, namely speaker identification and speaker verification. Speaker identification deals with distinguishing a speaker from a group of speakers. In contrast, speaker verification aims to decide if a person is the one who he/she claims to be from a speech sample. Speaker identification problem can be further classified as text dependent and text independent Speaker Identification based on relevance to speech contents [4]. Text dependent Speaker Identification involves the speaker saying exactly the enrolled or the given password/speech. Text independent Speaker Identification is a process of verifying the identity without constraint on the speech content. Compared to text dependent Speaker Identification, text independent Speaker Identification is more convenient because the user can speak freely to the system. However it requires longer training and testing utterances to achieve good performance.

## 2. SPEAKER IDENTIFICATION SYSTEM

The block diagram of a speaker identification system consists of the training phase and the testing phase as shown in fig. 1. In the training phase, the features of a speakers speech signal are stored as reference features. The feature vectors of speech are used to create a speakers model. The numbers of reference templates that are required for efficient speaker recognition depend upon the kind of features or techniques that the system uses for recognizing the speaker. In the testing phase, features similar to the ones that are used in the reference template are extracted from an input utterance of the speaker whose identity is required to be determined [5].

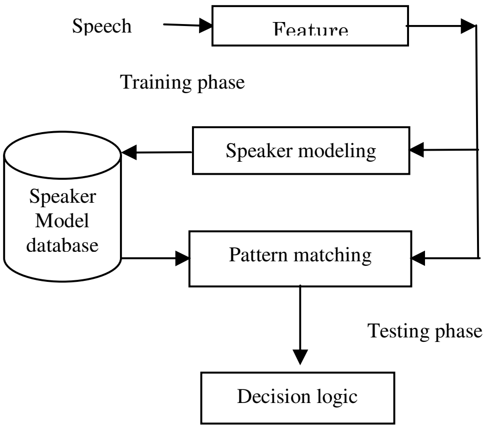

Figure 1: Training and testing modes of an automatic speaker identification system

The decision depends upon the computed distance between the reference template and the template devised from the input utterance. In speaker identification, the distance between an input utterance and all of the available reference templates is computed. The template of the registered user, whose distance with the input utterance template is the smallest, is finally selected as the speaker of the input utterance. In case of speaker verification the distance is computed only between the input utterance and the reference template of the claimed speaker. If the distance is smaller than the predetermined threshold, the speaker is accepted other the speaker is rejected as an imposter [6].

## 3. FEATURE EXTRACTION

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), Perceptual linear Predictive coefficients(PLP)[7]. Perceptual linear predictive analysis (PLP) was proposed by Hynek Hermansky in 1989 [8]. PLP analysis is similar to linear predictive coding (LPC), except that the PLP technique also uses three concepts from the psychophysics of hearing. These three concepts are the critical-band spectral resolution, equal loudness curve, and intensity loudness power law [9]. Both LPC and PLP use the autoregressive all-pole model to estimate the short-term power spectrum of speech. However, as pointed out by Hermansky, the LPC all-pole model is not consistent with human auditory perception because it does not consider the non uniform frequency resolution and intensity resolution of hearing. PLP alleviates this problem by applying the all-pole model to the auditory spectrum. The auditory spectrum is designed to be an estimate of the mean rate of firing of auditory nerve fibers [9].

### 3.1. PLP Algorithm

In the PLP technique, several well-known properties of hearing are simulated by practical engineering approximations, and the resulting auditory like spectrum of speech is approximated by an autoregressive all-pole model [10] [11].

*3.1.1. Spectral analysis*

The speech segment is weighted by the Hamming window:

$$w(n) = 0.54 + 0.46\cos[2\pi\,\mathrm{n}/(\mathrm{N}-1)] \qquad (1)$$

Where N is the length of the window. The typical length of the window is about 20ms.The discrete Fourier transform (DFT) transforms the windowed speech segment into the frequency domain. Typically, the fast Fourier transform (FFT) is used here [10]. The real and imaginary components of the short-term speech spectrum are squared and added to get the short term power spectrum [10].

$$P(w) = \mathrm{Re}[\mathrm{s(w)}]^2 + \mathrm{Im}[\mathrm{s(w)}]^2 \qquad (2)$$

*3.1.2. Critical-band spectral resolution:*

The spectrum P(w) is warped along its frequency axis w into the bark frequency Ω , by

$$\Omega(w) = 6\ln\left\{w/1200\pi + [(\mathrm{w}/1200\pi)^2 + 1]^{0.5}\right\} \qquad (3)$$

The resulting warped power spectrum is then convolved with the power spectrum of the simulated critical-band masking curve  .

This step is similar to spectral processing in Mel cepstral analysis, except for the particular shape of the critical-band curve. In PLP technique, the critical-band curve is given by:

$$\psi(\omega) = \begin{cases} 0\,for\Omega < -1.3 \\ 10^{2.5(\Omega+0.5)}\,for -1.3 \le \Omega \le -0.5, \\ 1\,for -0.5 \le \Omega \le 0.5, \\ 10^{-1.0(\Omega-0.5)}\,for 0.5 \le \Omega \le 2.5 \\ 0\,for\Omega > 2.5 \end{cases} \qquad (4)$$

The discrete convolution of  with (the even symmetric and periodic function) P(w) yields samples of the critical-band power spectrum.

$$\theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \qquad (5)$$

The convolution with the relatively broad critical-band masking curves Ψ(Ω) significantly reduces the spectral resolution of □(Ω) in comparison with the original P(w). This allows for the down-sampling of $\theta$(Ω).

3.1.3. *Equal-loudness preemphasis*

The sampled  is preemphasized by the simulated equal-loudness curve:

$$\Xi[\Omega(\mathrm{w})] = \mathrm{E(w)}[\Theta(\mathrm{w})] \qquad (6)$$

The function E(w) is an approximation to the non equal sensitivity of human hearing at different frequencies and simulates the sensitivity of hearing at about the 40-dB level. The particular approximation is adopted from Makhoul and Cosell(1976) and is given by:

$$E(w) = [(w^2 + 56.8 * 10^6)w^4] / \left[(w^2 + 6.3 * 10^6)^2 * w^2 + 0.38 * 10^9\right] \quad (7)$$

Finally, the values of the first (0bark) and the last (Nyquist frequency) samples (which are not well found) are made equal to the values of their nearest neighbors. Thus begins and ends with two equal-valued samples [10].

*3.1.4. Intensity-loudness power law:*

The last operation prior to the all-pole modelling is the cubic-root amplitude compression.

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \quad (8)$$

This operation is an approximation to the power law of hearing (Stevens 1957) and simulates the nonlinear relation between the intensity of sound and its perceived loudness. Together with the psychophysical equal-loudness preemphasis, this operation also reduces the spectral amplitude variation of the critical band spectrum so that the following all-pole modelling can be done by a relatively low model order [10].

*3.1.5. Autoregressive modelling*

In the final operation of PLP analysis, Φ(Ω) is approximated by the spectrum of an all-pole model using the autocorrelation method of all-pole spectral modelling. We give here only a brief overview of its principle: the inverse DFT (IDFT) is applied to Φ(Ω) to yield the autocorrelation function dual to . The first M+1 autocorrelation values are used to solve the Yule-Walker equations for the autoregressive coefficients of the Mth-order all-pole model. The autoregressive coefficients could be further transformed into some other set of parameters of interest, such as cepstral coefficients of all-pole model [10].

*3.2 Gammachirp filterbank:*

The proposed a temporal model deduced from the impulse responses measured from the electric impulses of the nervous fibers of the internal ear. [12] proposed a new model of the auditory filter called gamma chirp, to introduce dependence opposite the level of intensity of resonant hard working stimulus .The impulse response of the gamma chirp filter is given by the following expression [13]:

$$g_c(t) = at^{n-1} e^{-2\pi bERB(f_r)t} e^{j2\pi f_r t + jc\ln t + j\varphi} \quad (9)$$

Where: n is a filter order, fr is the modulation frequency of the gamma function, as is the carrier normalization parameter, c is the asymmetry coefficient of the filter, φ is the initial phase; bERB is the filter envelope, ERB represents the equivalent rectangular band given by [14, 15]:

ERB(fr)=24.7 + 0.108 fr (10)

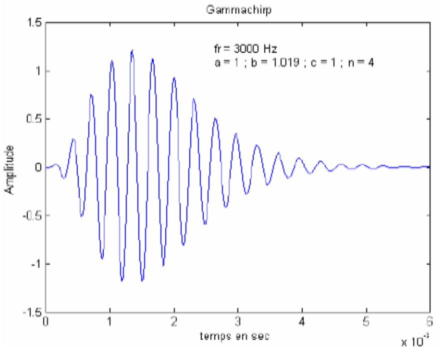

Figure 2: Temporal impulse response of the gamma chirp filter (a =1, b=1.019, c=1, n=4).

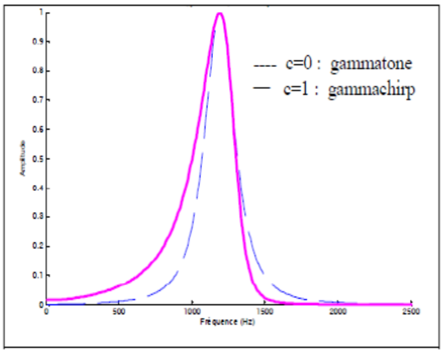

Figure 3: Frequency response of the gamma chirp filter

The Fourier magnitude spectrum of the gammachirp filter is:

$$|G_c(f)| = \frac{a\left|\Gamma(n+jc\right|e^{c\theta}}{(2\pi)\left[(bERB(\mathrm{f}_0))^2 + (f-f_0)^2\right]^{n/2}} \quad (11)$$

Where:

$$\theta = arctg(\frac{f-f_0}{bERB(f_0)}) \quad (12)$$

And Γ(n+jc) is the complex gamma distribution.

Figures 2 and 3 represent respectively the temporal impulse response and the frequency response of the gamma chirp filter. The ERB is calculated in function of the central frequency (fr) according to [14]. If we use the formula and if we suppose that the signal band is between fH and fL with a filter recovery ratio (V) hence, the N number of filters is selected like this [16]:

$$N = \frac{9.26}{v} \ln \frac{f_H + 228.7}{f_L + 228.7} \quad (13)$$

However, the central frequencies (fr) can be premeditated by the expression:

$$f_c = -228.7 + (f_h + 228.7)e^{-\frac{vn}{9.26}} \quad (14)$$

## 4. PERCEPTUAL LINEAR PREDICTION BASED ON GAMMACHIRP FILTERBANK (PLP-GC)

The linear predictive model is based on the mechanism of speech production. A major shortcoming of this method (LP) is that the filter spectrum uniformly distributed over all frequencies of the analysis band. Thus, it is possible that certain important details of the spectrum are not taken into account in the LP analysis. However, to take account of the collection must be based on the hearing mechanism. The purpose of the analysis is PLP-GC coefficients by estimating a model of auditory filter based on a filter bank in which every size of each GC filter varies with the input power of signal followed by a filtering operation of the external and middle ear models (equal intensity curve). The pattern of PLP-GC analysis is given in fig.4. The short term power spectrum of the speech signal is calculated. Then we multiply the spectrum of each filter of the filter bank Gammachirp by the spectrum of the speech signal in this step is to use only the model of the inner ear that is to say, This passage provides a first approximate of crudely the auditory filter, complying recovery critical bands and the masking phenomenon of the change of the template based filter Gammachirp asymmetry parameter C. Psychoacoustic experiments showed that the ear has nonlinear characteristics. Indeed, experiments conducted by Fletcher and Munson [17] showed that the intensity, when we listen to a pure constant sound intensity varies with the frequency of the pure tone. To simulate this occurrence in the PLP analysis of GC, we multiply the power spectrum resulting from the preceding step. For the model of the outer and middle ear by the given filter whose impulse response is given by the following equation:

$$w_{om}(f) = -(0.6)(3.64)f^{-0.8} + 6.5\exp\left(-0.6(f-3.3)^2\right) - 10^{-3}f^{3.6} \quad (15)$$

It is possible to estimate the model of the outer and middle ear, referring to a chart of equal loudness lines along which a pure ear gives a feeling of equal intensity. This is the origin of what we call the loudness considered subjective loudness of sounds. This curve is an approximation of the non-equal sensitivity of the human outer and middle ear for different frequencies that simulates the susceptibility of the outer and middle ear. The previous two treatments are not sufficient to be a correspondence between the measured intensity and the subjective intensity (loudness). Stevens says the law after achieving integration of critical bands and pre-emphasis, the relationship between intensity and loudness becomes:

Loudness= (intensity) 0.33 (16)

This involves applying the amplitude compression according to the cube root law (eq.8) Preemphasis with the outer and middle ear model and the application of the law of Stevens reduce the amplitude variation of the spectrum bands. The last step of the analysis PLP -GC is to approximate Φ(Ω) with the spectrum of all-pole model using the autocorrelation method. The Inverse Discrete Fourier Transform to determine the all-pole model coefficients and cepstral recursion and matrix coefficients PLP -GC.

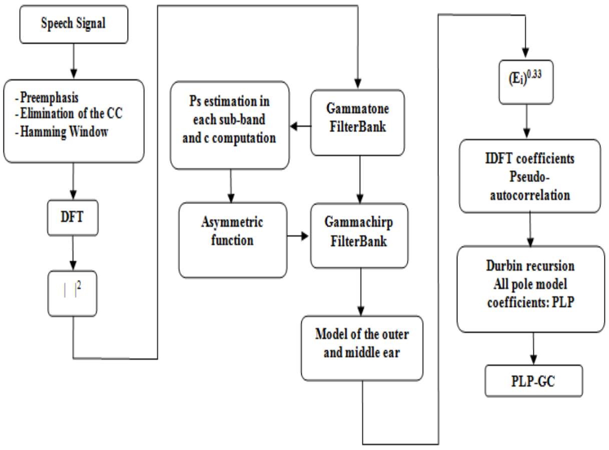
Speech Signal
- Preemphasis
- Elimination of the CC
- Hamming Window
DFT
| |$^2$
Ps estimation in each sub-band and c computation
Asymmetric function
Gammatone FilterBank
Gammachirp FilterBank
Model of the outer and middle ear
(Ei)$^{0.33}$
IDFT coefficients Pseudo-autocorrelation
Durbin recursion All pole model coefficients: PLP
PLP-GC

Figure 4: PLP-GC algorithm

## 5. EXPERIMENTAL EVALUATION

In this study, we are interested to evaluate the performance of the suggested front-end based on PLP-GC method in a text-independent speaker identification context. We consider identification task for TIMIT Speaker Database [20]. The TIMIT corpus of read speech has been designed to provide speaker data for the acquisition of acoustic phonetic knowledge and for the development and evaluation of automatic speaker recognition systems. TIMIT contains a total of 6300 sentences, is composed of speech composed of 8 American dialects.

We consider 2 male speakers and 2 female from each dialect, out of 630 speakers for speaker recognition, everyone have 5 words in the training and 5 other words for testing because here we have text independent in which we must not find the same words spoken in both of training and testing. The parameters of the system are 16 KHz sampling rate with 16 bit sample resolution. 25 millisecond Hamming window duration with a step size of 10 milliseconds. PLP coefficients with 22 as the length of cepstral liftering and 26 filter bank channels of which 12 are the number of PLP coefficients. Like an initial starting point a vector size of 39 is defined by using the 12 cepstral coefficients and the logarithmic frame energy plus the corresponding delta and acceleration coefficients. The vector size may be changed when testing with an alternative front-end that generates a different number of features.

The reference recognizer is based on the HTK software package version 2.2 from Entropic. The training and recognition parameters are defined to compare the recognition results when applying different feature extraction schemes. The task of recognition is considered without restricting the string length. The words spoken by the speakers ‘database are modeled as whole word HMMs with the following parameters:

- Single left-to-right HMM model
- mixture of 3 Gaussians per state
- Prototype model with means 0 and variances 1

The training is done in several steps by applying the embedded Baum-Welch reestimation scheme (HTK tool HERest). GMM is widely used for speaker modeling in context independent speaker identification [21]. In our research, we use HTK to design the speaker models [22].each speaker is modeled using a three-state HMM in which only one state is modeled with Gaussian mixture distributions, the other two states are dummy states. This HMM speaker model is almost the same as the GMM speaker model except that the former has state self-transition involved in the calculation of the likelihood probability. The identification rate is defined as the ratio between the number of correctly identified speech segments and total number of speech segments for each speaker. Our purpose in this study is to test whether the proposed feature has more speaker individual information, thus a three-state with two dummy states HMM is used for each speaker modeling to evaluate the proposed method.

## 6. RESULTS

The evaluation of the speaker identification performances of our systems, One Performance measures, the correct recognition rate (CORR) is adopted for comparison. They are defined as:

% CORR = no. of correct labels / no. of total labels * 100%

Tables 1, 2, 3 shows the results associated with the rate recognition of different parameterization techniques, using “energy”, “delta” and “delta delta” vectors according to the signal to noise ratio (SNR).

Please do **not** paginate your paper. Page numbers, session numbers, and conference identification will be inserted when the paper is included in the proceedings.

Table1. Speaker identification accuracy rate for white noise

| | -3 DB | 0 DB | 3 DB | 6 DB | 12 DB | AvG |
|---|---|---|---|---|---|---|
| PLP _E_D_A | 23.68 | 36.98 | 67.03 | 89.01 | 91.65 | 61.67 |
| PLP-GC_E_D_A | 25.83 | 41.56 | 72.67 | 94.35 | 97.51 | 66.38 |

Table 2. Speaker identification accuracy rate for the car noise.

| | -3DB | 0 DB | 3 DB | 6 DB | 12DB | AvG |
|---|---|---|---|---|---|---|
| PLP_e_d_a | 43.19 | 64.97 | 71.53 | 77.83 | 89.14 | 69.33 |
| PLP-GC e_d_a | 54.11 | 67.81 | 79.50 | 81.67 | 94.44 | 75.50 |

Table3. Speaker identification accuracy rate for the factory noise.

| | -3DB | 0 DB | 3 DB | 6 DB | 12DB | AvG |
|---|---|---|---|---|---|---|
| PLP _e_d_a | 41.50 | 50.65 | 58.98 | 76.82 | 92.14 | 64.01 |
| PLP-GC e_d_a | 53.39 | 57.05 | 61.30 | 82.23 | 96.93 | 70.18 |

## 7. DISCUSSION

This three tables show the identification rate of PLP _E_D_A, PLP-GC _E_D_A frontends in various SNR conditions. These results indicate clearly that the PLP-GC _E_D_A produces interesting results. However, in noisy environments, all variants of PLP-GC exceed all variants of PLP. The average identification rate of PLP-GC is about 75.50% while the average identification rate of PLP is still equal 69.33% when SNR changes from -3dB to 12dB. In others words, these results indicate that in noisy environments the PLP-GC algorithm works better than PLP and the dynamic variants of these algorithms are better suited to robust conditions.

## 8. CONCLUSION

An auditory motivated technique has been described to extract significant feature sets from the speech signal. It is mainly based on the Gammachirp filterbank. Gammachirp Auditory filterbank are non-uniform band pass filters, designed to imitate the frequency resolution of human hearing. The bloc diagram of PLP-GC has been implemented under Matlab and tested on the TIMIT databases using HTK. When compared to PLP, achieves better performance.

## 9. REFERENCES

[1]Varga, A., Steeneken, H,J,M., Omlison, M,T., Jones, D., “The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition”, Documentation included in the NOISEX-92 CD-ROM Set.,1992

[2] A. Jain, A. Ross, and S. Pankanti, “Biometrics: a tool for information security,” IEEE Transactions on Information Forensics and Security, vol. 1, no. 2, pp. 125–143, June 2006.

[3] U. Uludag, S. Pankanti, S. Prabhakar, and A. Jain, “Biometric cryptosystems issues and challenges,” Proceedings of the IEEE, vol. 92, no. 6, pp. 948–960, June 2004.

[4] J.P. Campbell Jr. “Speaker recognition: A tutorial,” Proc. IEEE, vol. 85, no. 9, pp. 1437-62, Sept. 1997.

[5]. R. A. Cole and colleagues, “*Survey of the State of the Art in Human Language Technology*”, National Science
Foundation European Commission, 1996.

[6] Vijender Sharma , Rakesh Garg, “Gender and Speaker Recognition Using MFCC and DTW”, IJARCSSE 2013,
Volume 3, Issue 8, pp1070-1076

[7] F. Z. Chelali, A.Djeradi, and R.Djeradi, “Speaker Identification System based on PLP
Coefficients and Artificial Neural Network ”, Proceedings of the World Congress on Engineering 2011 Vol II WCE 2011, July 6 - 8, 2011, London, U.K.

[8] M jamaati,H. Marvi and M. Lankarany, “vowels recognition using mellin transform and plp-based feature extraction”, acoustics-08 Paris.

[9] Wira Gunawan and Mark hasegawa-Johnson, “PLP coefficients can be quantizied at 400 BPS”, department of electrical and

Computer Engineering,University of Illinois at Urbana-Champaign,USA.

[10] H. hermansky,"perceptual linear predictive(PLP) analysis of speech", journal of the Acoustical Society of America, vol 87 no.4,pp 1738- 1752,1990.

[11]Sid Ahmed Selouani and Jean Caelen, "un système connexionniste modulaire pour la reconnaissance des traits phonétiques de l'Arabe ».

[12]Irino T. and Patterson D., "Temporal Asymmetry in the Auditory System," Computer Journal of Acoustical Society of America, vol. 99, no. 4, pp. 126-129, 1997.

[13]Irino T. and Patterson D., "A Time-Domain, Level Dependent Auditory Filter: The Gammachirp," Computer Journal of Acoustical Society of America, vol. 101, no. 1, pp. 412-419, 1997.

[14]Irino T. and Unoki M., "An Analysis Auditory Filterbank Based on an IIR Implementation of the Gamma chirp," Computer Journal Acoustical Society of Japan, vol. 20, no. 6, pp. 397-406, 1999.

[15]Zwicker E. and Feldkeller R., Psychoacoustiquel „Oreille Recepteur d" information, Masson Press, 1981.

[16]Waleed A., "Signal Processing and Acoustic Modeling of Speech Signal," PhD Thesis, 2002. [17]Fletcher, H., "Auditory Patterns", Review of Modern Physics, 12, pp. 47-65, 1940.

[18]L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257{286, 1989}

[19]Paliwal, K, K., "Decorrelated and Liftered Filter-Bank Energies for Robust Speech Recognition", Proc. Eurospeech, pp. 85-88. 1999.

[20] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, DARPA TIMIT: Acoustic-phonetic Continuous Speech Corpus CD-ROM. U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.

[21] D.A.Reynold,"speaker identification and verification using Gaussian mixture models", Speech Communication, vol.17, pp.91-108, 1995

[22]HTK tutorial book,http://htk.eng.cam.ac.uk/