Robust ASR Systems using Auditory Filter in Impulsive Noise Environment

Issam Bel Haj Yahia
Research Laboratory of Signal Image and Information Technology (RLSIIT) National Engineering School of Tunis, TUNISIA

Zied Hajaiej
Research Laboratory of Signal Image and Information Technology (RLSIIT) National Engineering School of Tunis, TUNISIA

ABSTRACT
This paper is dedicated to the development of new automatic methods for recognizing of isolated words with impulsive sounds. This article presents a parameterization technique of speech signal with impulsive noise based on auditory filter modeling by the gammachirp filterbank (Gammachirp Filter Banc (GFB). This work includes two parts; the first is devoted to traditional techniques. The second deals with modern methods incorporating a model of auditory filter called gamma chirp. In this section, we will extract the characteristics of a single word with impulsive noise from the TIMIT database using parameterization technique Perceptual Linear Preduction(PLP) with the GFB.The recognition system is implemented on Hidden Markov Model Toolkit HTK platform based on HMM. For evaluation a comparative study was operated with standard PLP and Mel Frequency Cepstral Coefficient (MFCC). We propose a study of the performance of new parameterization technique GFB_PLP and GFB_MFCC proposed in the presence of different impulsive noises. Three types of impulsive noise are used (blast door, glass breaks, and explosion) Tests were carried out at different SNR levels (15dB, 10dB, 5dB, 0 dB and -3 dB) The GFB –PLP technique give the better results in different tests.

General Terms
Impulsive noise ,gammachirp ,asr recognition

Keywords
Mfcc, plp, gfb mfcc, gfb plp

1. INTRODUCTION
speech is a natural mode of communication for humans. It isvery efficient, for transmission of information conversational speaking rates can be as high as 200 words perminute. And for reception of information, has others advantages as well. Speech recognition is today a quite common element in our lives. Cellular phones, computers, telephone services and many more products use speech recognition. An important drawback affecting most of the speech processing systems is the environmental noise and its harmful effect on the system performance. The presence of noise normally degrades the performance of speech recognition; therefore it is very important that a speech recognizer in some way deals with possible noise. A large amount of work has therefore been spent in this area and there exists a lot of technique that improves the speech recognizer'smodels are generally a filterbank, none uniformly spaced in frequency and with non-uniform bandwidths, narrows at low frequencies, and broad at high frequencies, which converts the input speech signal into a performances in noisy conditions. Signal theory tools for representation of signals and systems in the time domain or in the spectral domain, their study and analysis, modeling and interpretation. Detecting the absence or presence of a signal, signal with a noise and speech recognition are treated frome problems. Indeed, the natural sounds are composed of noise, and the ear is sensitive to information related to this part [8]. With this noisy component, which is considerate for several years, we present the different characteristics of the noise part.The purpose of this article is to introduce several important concepts in signal processing and illustrate them with relatively simple examples. At first, to focus on the study and analysis of impulsive noise by incorporating a model auditory filter called gammachirp [1][3]. In this paper, we propose a techniques for parameterization speech signals based on a gammachirp filterbank (GFB) following the approach used in the technical PLP. For this we will develop a system for automatic recognition of isolated words with impulsive noise based on Hidden Markov Models HMM, the recognition system will serve as an evaluation of the impulsive signal by gammachirp filter. We propose a study of the performance of parameterization technique GFB_PLP and GFB_MFCC proposed in the presence of different impulsive noises. The sounds are added to the word with different snr (15dB, 10dB, 5dB, 0 dB and --3 dB). The recognition performance of this approach was evaluated using theTIMIT database.The obtained evaluation results are compared to thos of the standards techniques. .

2. SPEECH RECOGNITION SYSTEMS ASR
The speech recognition system has two major components, feature extraction and classification. In this work the system block diagram of isolated word recognition is shown in Figure1.

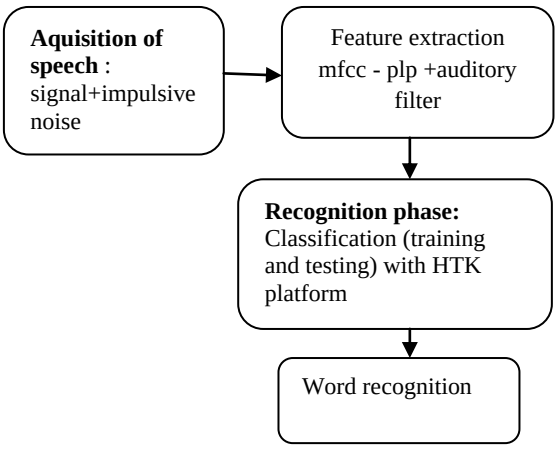

Fig 1: Structure of asr System.

3. PERCEPTUAL LINEAR PREDICTION (PLP)

Perceptual linear prediction analysis is a variation of the original LPC analysis and was first introduced by Hermansky [7] in 1990. The main idea of this technique is to take advantage of three principal characteristics derived from the psychoacoustic properties of the human ear for estimating the audible spectrum. This concept is Spectral resolution of the critical band, Equal-loudness curve and Intensity loudness power law. The audible spectrum is then approximated from an all pole autoregressive model. The PLP analysis is nearer to the behavior of the human ear than the traditional LPC technique. This last characteristic renders this method more robust in speaker-independent conditions. The PLP analysis is however computationally efficient and permits a compact representation of speech. The method considers the short term power spectrum of speech and makes a convolution of it with a simulated critical band masking pattern. Then, the critical-band is re sampled at about one Bark scale intervals. At this point, a preemphasis operation is performed with a fixed equalloudness curve and finally the resulting spectrum is compressed with cubic-root nonlinearity, simulating the intensity-loudness power law. The resulting low order all pole models is consistent with several phenomena observed in human speech perception. In this step, the IDFT (Inverse Discrete Fourier Transforms) is applied to obtain the dual autocorrelation function. The first M+ 1 values are used for solving the Yule-Walker [13] equation for obtaining the autoregressive coefficients of the all-pole model of order M [21] [2].

4. MEL FREQUENCY CEPSTRAL COEFFICIENT (MFCC)

The cepstral coefficients are derived from the output of the Mel filter bank formed from triangular filters and positioned uniformly across Mel. This technique consists in calculating the cepstral coefficients on a Mel scale that approximates the frequency perception of the ear. After applying a Fourier transform in the short term, energy is calculated in critical heath modeled by triangular filters when the amplitude scale is expressed in decibels. The frequency scale in turn is expressed in Mel. Then the Inverse Cosine Transform (IDFT) provides lesser correlated coefficients. The cepstrum is then given by the fol-lowing expression:

$$C_n = \sqrt{\frac{2}{k}} \sum_{k=1}^{N} \left(\log S_k \cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{k} \right] \right)$$

Sk : representing the energy after filtering by a k triangular filter

5. AUDITORY GAMMACHIRP FILTER

The gammachirp filter is used in the psychoacoustic research as a reliable model of cochlear filter. The gammachirp filter is defined in the time domain (impulse response function) as:

$$gc(t) = a^{n-1} \exp(-2\pi b ERB(f_r)\, t) \exp(j2\pi f_r + jc \ln t + jc\varphi)$$

a: an amplitude normalization parameter.

c: a parameter for the chirp rate.

b: a parameter defining the envelope of the gamma distribution.

ERB (fr): Equivalent Rectangulaire Bandwith.

φ : the original phase.

The gammachirp to be expressed as the cascade of a gammatone filter with an asymmetric compensation filter. Figure 2 shows the framework for this cascade approach.

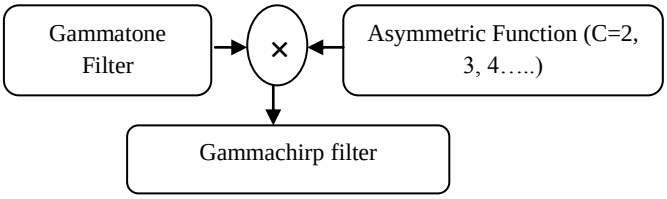

Fig 2: Decomposition of the gammachirp filter

The gammachirp filter which is derived out of the gammatone filter modification was introduced to model the asymmetry in the low frequency tail of the auditory filter response and to model level dependent proprieties such as decrease in gain, bandwidth increase and a shift in the center frequency of the filter with increase in level. Where time t>0, a is the amplitude, (fr) is the asymptotic frequency and b are parameters defining the envelope of the gamma distribution. C is a parameter for the frequency modulation or the chirp rate, φ is the initial phase, and ERB (fr) represents the Equivalent Rectangular Bandwidth of the filter, given by the following relationship:

$$ERB(f_r) = 24.7 + 0.108 f_r$$

The Fourier transform of the gammachirp in is derived as follows

$$G_c(f) = \frac{a|\Gamma(n+jc)|}{\Gamma(n)} \cdot \frac{\Gamma(n)}{\left|2\pi\sqrt{(b.ERB(f_r))^2 + (f - f_r)^2}\right|^n} e^{C\theta}$$

$$|G_c(f)| = a_\Gamma |G_T| . e^{C\theta(f)}$$

$$\theta(f) = \arctan\left(\frac{f - f_r}{bERB(f_r)} \right)$$

Gc (f) is the Fourier magnitude spectrum of the gammatone filter, e cθ (f) is an asymmetric function since it antisymmetric function centered at the asymptotic frequency. The spectral properties of the gammachirp will depend on the e cθ (f), factor; this factor has therefore been called the asymmetry factor. The degree of asymmetry depends on "c". If "c" is negative, the transfer function, considered as a low pass filter, where c is positive it behaves as a high-pass filter and if "c" zero, the transfer function, behaves as a gammatone filter. In addition, this parameter is connected to the signal power (Ps)by the expression

$$C = 3.38 + 0.107\, Ps$$

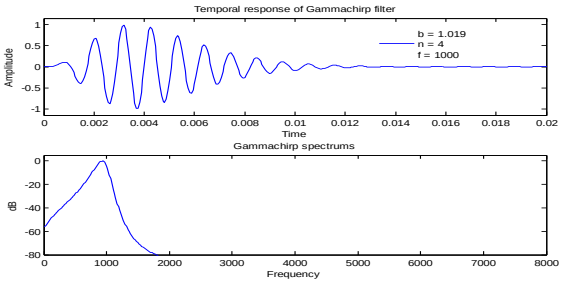

Fig 3: Example of Impulse Response Gammachirp

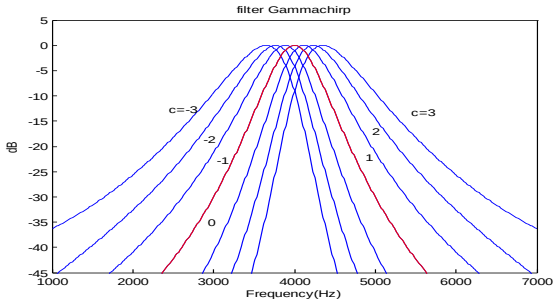

Fig 4: Gammachirp spectrums for different values of c

The figure 5 shows a block diagram of the gammachirp filterbank

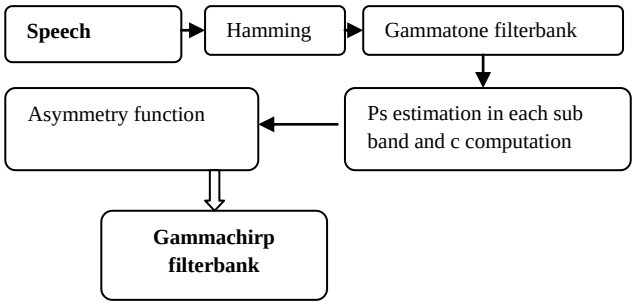

Fig 5: Block diagram of the gammachirp filterbank

The figure 6 shows the algorithm of PLP and the GFB-PLP parameterization

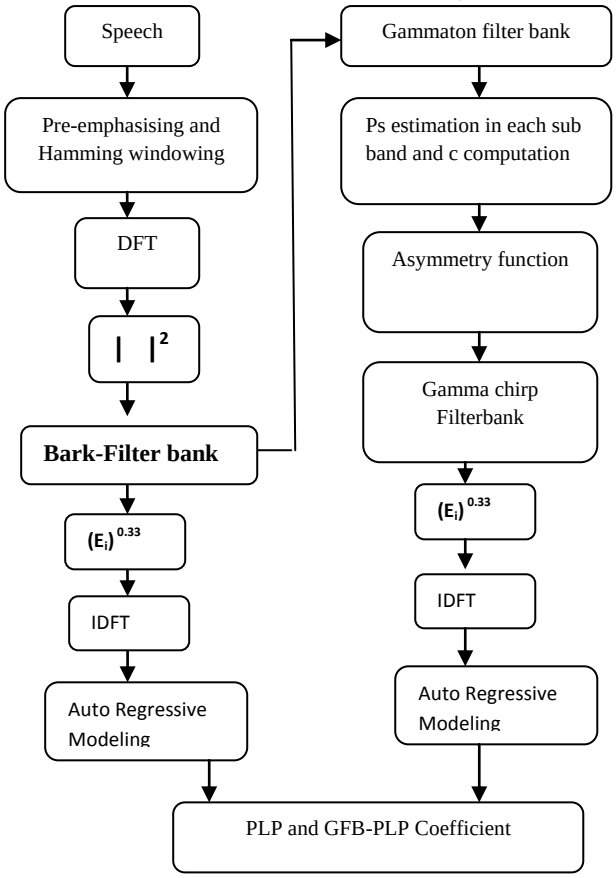

Fig 6: Block diagram of PLP and the gfb-plp parameterization

The object of the GFB-PLP analysis is to estimate coefficients by an auditory model of a filter based on filter bank whose gammachirp. The algorithm calculates GFB-PLP coefficient is based on the same step of calculated PLP coefficient in that, we only change the Bark filter bank by gammachirp filter.

The figure 7 shows the algorithm of MFCC and the GFB-MFCC parameterization

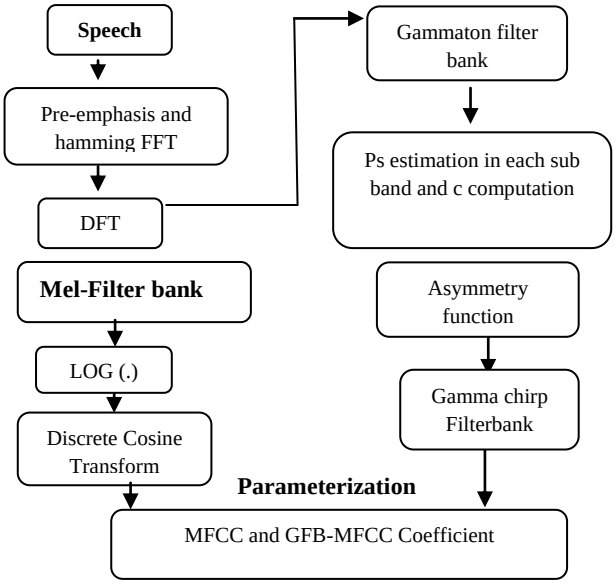

Fig 7: Block diagram of mfcc and the gfb-mfcc

The object of the GFB-MFCC analysis is to estimate coefficients by an auditory model of a filter based on filter bank whose Gamma chirp. The algorithm calculates GFB-MFCC coefficient is based on the same step of calculated MFCC coefficient in that,we only change the mel filter bank by gammachirp filter .

6. EXPERIMENTS CONDITIONS

In this paper to evaluate the GFB-PLP parameterization technique, we carried out a comparative study with different baseline parameterization techniques of MFCC and PLP implemented in HTK We tested the performance in speech signal recognition tasks wehre the training database is clean and the test database contaminated with additive impulsive noise different real-environment impulsive noises used: blast door, glass breaks, and explosion provided by Tests were carried out at different SNR levels (15dB, 10dB, 5dB, 0 dB and -3 dB).

6.1 Timit Database

The training database is built of several words extracted from the Darpa-Timit database (8). This database is composed of speech composed of 8 American dialects. We used 6132 words composed of 21 words (models) repeated 292 times from 36 speakers (18 males and 18 females) uniformly divided for American dialects. For the test phase of recognition we used 2201 words from 26 speakers (13 males and 13 females).[15][16].

6.2 Impulsive Noise

Impulsive noise consists of relatively short duration, caused by a variety of sources for example explosions and gunshots, human screams, door slams, glass breaks, dog barks, phone rings, machine noise, pieces of music or children voice sounds. In our work we used three class of impulsive noise: explosions, glass breaks, and door slams. Impulsive noise is usually non-stationary, non-Gaussian and very complex frequency behavior. It is for this reason that we are interested in the study of the noise. The duration of this noise is low in the order of second, theory feature is a Dirac. As the example in Figure 2 and 3 gives time representation and spectrogram of door slam impulsive noise.

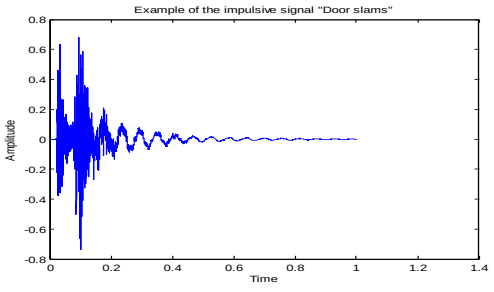

Fig 7: Time representation of door slam impulsive noise

System description

Features used in our test Mel-frequency cepstral coefficients MFCC. Another popular feature set is the set of perceptual linear prediction (PLP) coefficients. The 39-element feature vector contains 12 MFCC or PLP or GFB-PLP or GFB-MFC C(implemented in HTK platform), one energy measure, and their first and second derivatives. In our experiment, there were 21 HMM models (21Recognized words) trained using the selected feature (GFB-PLP, MFCC PLP and GFB-MFCC). The observation probabilities are modeled as a weighted sum of Gaussian probability densities. The system uses 21 models (words). Each speech model is represented by an HMM with 5 by 5 states,from left-to-right (the first and the last ones are non-emitting) and 12 mixture components per state [10]. In the Training Process parameters of HMM are estimated during a supervised process using a maximum likelihood approach with Baum-Welch re-estimation. The first step in determining the parameters of an HMM is to make a rough guess about their values. Then, the Baum-Welch algorithm is applied to these initial values to improve their accuracy in the maximum likelihood sense. Finally, The Viterbi decoding algorithm is used in the decoding process. The recognition problem is to find a state sequence of a model which is most likely to have been generated by the data. The Viterbi decoding algorithm assumes that the maximum likelihood state sequence travels through the optimal path along each state [6].

6.3 Recognition Results

In this section we present the different results obtained for the parameterization techniques, GGB-PLP, MFCC,GFB-MFCC and PLP. The performance is tested on the TIMIT databases using HTK. The features vectors consist of GFB-PLP. Each one consists of 12 cepstrals coefficients and energy (E), concatenated with the delta (D) and acceleration (A) coefficients. The features vectors are of dimension 39. The performance is compared with the MFCC, PLP, GFB-MFCC and GFB-PLP baseline as the same condition. Tables 1, 2, 3 and 4 show the average recognition accuracy.

Table 1. Words recognition accuracy rate for clean speech

	Brut	**e**	**e-d**	**e-d-a**
Mfcc	91	93.14	98.14	99.05
plp	91. 46	93.68	98.15	99.35
Gfb-plp	90.12	93.43	97.56	98.93
Gfb-mfcc	93.05	49.22	98.66	99.85

Table 2. Words recognition accuracy rate for blast door impulsive noise

	-3db	**0db**	**5db**	**10db**	**15db**
Mfcc (e-d-a)	35.67	65.79	90.55	97.68	98.82
Plp(e-d-a)	32.06	65.4	91.33	97.18	98.14
Gfb-plp(e-d-a)	32.05	65.11	90.11	97.21	98.17
Gfb-mfcc(e-d-a)	40.17	68.54	91.11	98.6	98.96

Table 3. Words recognition accuracy rate for glass breaks impulsive noise

	-3db	**0db**	**5db**	**10db**	**15db**
Mfcc (e-d-a)	37.55	59.6	88.87	97.32	98.36
Plp(e-d-a)	36.44	57.7	98.23	97.73	97.96
Gfb-plp(e-d-a)	35.5	55.4	87.36	59.38	98.8
Gfb-mfcc(e-d-a)	39.40	66.98	84.55	98.77	99.14

Table 4. Words recognition accuracy rate for blast door impulsive noise

	-3db	0db	5db	10db	15db
Mfcc (e-d-a)	63.43	87.85	97.36	98.6	98.82
Plp(e-d-a)	60.11	88.37	97.77	97.27	98.14
Gfb-plp(e-d-a	59.00	86.33	97.10	97.10	98.9 6
Gfb-mfcc(e-d-a)	66.30	89.97	98.80	98.27	99.5

6.4 Discussion

From these results it can be seen that the GFB –MFCC technique give the better result (highest accuracy of recognition) 99.85% in clean speech with add energies of signal and derived parameters.

In noise speech it can be seen that the system's highest accuracy of recognitions 98.96% and 99.50% are obtained for GFB-PLP techniques in blast door and explosion, impulsive noise. With -5db SNR level. In tables 2, 3 and 4 the technique GCFB-PLPgive the better result is 40.17%, 39.40%, and 66.3% respectively with -3db SNR level.

We notice that the GFB-MFCC technique gives acceptable results in all the experience.

7. CONCLUSION

In this work, we evaluated the new technique of parameterization of the speech signals GFB- PLP (who takes account of the characteristics frequential and temporal of the ear, based on a filter bank gammachirp GFB and a classic MFCC PLP and (GFB -PLP techniques in clean and noise environment with different SNR levels (15dB, 10dB, 5dB, 0 dB and -3 dB).

The results gotten after application of this features show that this method (GFB-PLP, GFB-MFCC) give acceptable and better results by comparison at those gotten by other methods of parameterization.

We used the platform HTK toolkit, that use Hidden Markov Models and we used the TIMIT speech databases and impulsive noise in our evaluations. The results obtained after application of Gammachirp filter on the word show that this filter gives acceptable and sometimes better results.

In other issues of this work, the following strategy is applied to improve different obtained results

- Testing the state efficiency of HMM approach
- Raising the filter numbers in different techniques
- increase for example the number of coefficients in the parameters techniques

8. REFERENCES

[1] Timo Gerkmann, Richard C. Hendrikes, "Noise power estimation based on the probability of speech presence," Proc. IEEEWASPAA, pp. 145-148, New York, Oct. 2011.

[2] Irino. T, E. Okamoto, R. Nisimura, Hideki Kawahara and Roy D. Patterson, "A Gammachirp Auditory Filterbank for Reliable Estimation of Vocal Tract Length from both Voiced and Whispered Speech," The 4th Annual Conference of the British Society of Audiology, Keele, UK, 4-6, Sept, 2013

[3] T. Irino, R. D. Patterson. ''Temporal asymmetry in the auditory system.'' J. Acoust. Soc. Am. 99(4): 2316-2331, April, 1997.

[4] T. Irino, R. D. Patterson. ''A time-domain, Level-dependent auditory filter: The gammachirp.'' J. Acoust. Soc. Am. 101(1): 412-419, January, 1997.

[5] T. Irino et M. Unoki. ''An Analysis Auditory Filterbank Based on an IIR Implementation of the Gammachirp.'' J. Acoust. Soc Japan. 20(6): 397-406, November, 1999.

[6] Patterson, R, D., Nimmo-Smith, I., ''Off-frequency listening and auditory-filter asymmetry'' J. Acoust. Soc. Am, Vol. 67, No. 1, pp. 229-245, 1980.

[7] SCHLÜTER, R., BEZRUKOV, I., WAGNER, H., NEY, H, "Gamma tone features and feature combination for large vocabulary speech recognition," In ICASSP 2007. Honolulu (HI, USA), April 2007, p. 649-652.

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

[9] Irino, T., Patterson, R, D., ''A compressive gammachirp auditory filter for both physiological and psychophysical data.'' J. Acoust Soc. Am. 109(5): 2008-2022, may 2001.

[10] Young, S., Evermann,G., Gales, M., Hain, T. D., Kershaw,X. Liu, Moore, G., Odell, J. D., Ollason, D., and Woodland. P., The HTK book (for HTK version 3.4). Cambridge University Engineering Department, Cambridge, UK, 2006.

[11] Young S. J., Woodland P. C., Byrne W. J., "HTK. Reference Manual for HTK version 3.1", December 2001.

[12] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech processing. Proceedings of IEEE, 77(2):257–286, 1989.

[13] Zied Hajaiej, Kais Ouni, Noureddine Ellouze, " Etude et évaluation d'une technique de paramétrisation perceptive des signaux de parole ", Traitements & Analyse d'Informations : Méthodes & Applications, TAIMA 05, Hammamet, Tunisie, pp.259 – 264, 1–3 octobre 2005.

[14] Smith III, J, O., Abel, J, S,. ''Bark and ERB Bilinear Transforms.'' IEEE Tran. On speech and Audio Processing, Vol. 7, No. 6, November 1999.

[15] NIST., The DARPA TIMIT Acoustic-phonetic Continuous Speech Recognition Database, 1990 .

[16] Hermansky, H., ''Perceptual Linear predictive (PLP) analysis of speech'', J. Acoust. Soc. Am. Vol. 87, No. 4, pp. 1738-1752., April 1990

[17] 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

[18] A. B. Poritz, "Hidden Markov models: A guided tour", in Proc. of the IEEE Int'l. Conf. on Acoustics, Speech and Signal Processing (ICASSP '88), May 1988, pp. 7-13.

[19] E. Loweimi and S. M. Ahadi, "A new group delay-based feature for robust speech recognition," in Proc. IEEE

Int. Conf. on Multimedia & Expo, Barcelona, pp. 1-5, July 2011.

[20] Skowronski M. D. and Harris J. G., 2002, "Increased MFCC filter bandwidth for noise-robust phoneme recognition", in Proc. ICASSP-02, Florida.

[21] L. Bréhélin, O. Gasuel. « Modèles de Markov cachés et apprentissage des séquences. Le temps, l'espace et l'évolutif en sciences du traitement de l'information », Éditions Cépaduès, pp. 407-421, 2000.

[22] Zied Hajaiej, Kaïs Ouni and Noureddine Ellouze "Gammachirp Filter Frond-End for Automatic Speech Recognition "International Conference: Sciences of Electronic, Technologies of Information and Telecommunications SETTIT, 2000