International Conference on Control, Engineering & Information Technology (CEIT'14)
Proceedings - Copyright IPCO-2014
ISSN 2356-5608

Impulsive signal analysis by a gammachirp filter for the application of the speech recognition

Hajer Rahali[1], Zied Hajaiej[1], Noureddine Ellouze[1]

[1] Laboratoire des Systèmes et Traitement du Signal (LSTS)
Ecole Nationale d'Ingénieurs de Tunis
BP 37, Le Belvédère, 1002 Tunis, Tunisie
Hajer.Rahali@enit.rnu.tn
Zied.hajaiej@enit.rnu.tn
N.ellouze@enit.rnu.tn

Abstract. This article is dedicated to the development of automatic methods for recognizing of isolated words with impulsive sounds. In this paper, it focuses on the main techniques used to characterize the impulsive signals and model operation. One of the main models is the notion of auditory filters. This article includes two parts, the first is devoted to traditional techniques; Fourier transform short-term is important concepts in signal processing and is used in many fields. The second deals with modern methods incorporating a model of auditory filter called gammachirp. In this section, we will extract the characteristics of a single word with impulsive noise using parameterization technique MFCC with the gammachirp filterbank (GCFB). For this, we have developed a system for automatic recognition of isolated words with impulsive noise based on Hidden Markov Models (HMM) and Gaussian Mixtures Model (GMM). For evaluation a comparative study was operated with standard MFCC. We propose a study of the performance of parameterization technique GCFB_MFCC proposed in the presence of different impulsive noises.

Keywords: Gammachirp filterbank, MFCC, Fourier transforms FFT, impulsive noise.

1 Introduction

Speech is a natural and flexible mode of communication for humans. It is very efficient, for transmission of information; conversational speaking rates can be as high as 200 words per minute. 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's 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 from 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. In this paper, we propose two techniques for parameterization speech signals based on a gammachirp filterbank (GCFB) following the approach used in the technical MFCC. For this we will develop a system for automatic recognition of isolated words with impulsive noise based on HMM\GMM, 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 GCFB_MFCC proposed in the presence of different impulsive noises. The sounds are added to the word with different signal-to-noise (12dB, 6dB, 3dB, 0 dB and -3 dB). The evaluation is done on the TIMIT database. In this work, a new approach for speech analysis based on gammachirp filters is shown. After extracting parameters we are interested to compare their performance with standard MFCC for the application of the speech recognition, the evaluation is conducted on a database of many speakers extracted from database. This paper is organized as follow the first section is to define the traditional techniques, the second section studies the gammachirp filter and the third section shows experimental result and conclusion.

2 Characterization of speech signal

The main objective of the analysis speech signal is to extract some parameters such as voicing, pitch and formants. In this section, we provide some basic definitions and reminders that we use later in the document.

2.1 Speech production

The process of speech production is a very complex mechanism that is based on an interaction between physiological and neurological system. There are a lot of organs and muscles that are used in the production of sounds of natural language. The functioning of the human vocal device based on the interaction between three major classes of organs: lungs, larynx, and supra glottal cavities.

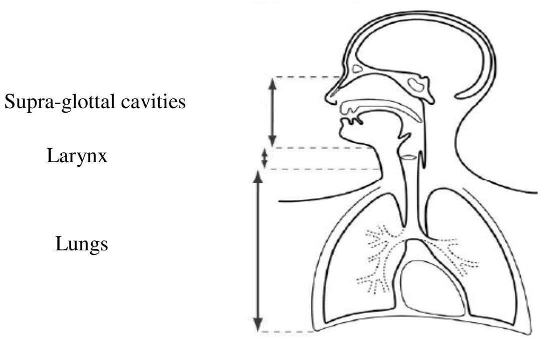


Fig.1. Diagram of the vocal apparatus

The first two classes provide what is essential for the production of any sound, whether musical or language: a source of air and noise source. The third class contains the organs that can change the sound that is produced by the joint work of the first two classes. Speech is a sequence of sound events contains voiced sounds characterized by the vibration of the vocal cords and unvoiced sounds. The spectrum of the sound emitted by the vocal cords is modulated by the resonant properties of the resonator body and lip position. Figure 2 shows the general operation of the vocal apparatus.

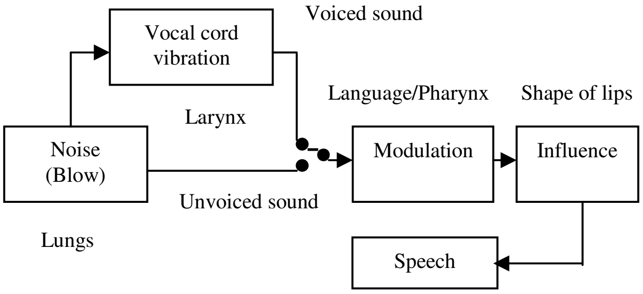


Fig.2. The general operation of the vocal apparatus

2.2 Impulsive noise

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. Include different source of impulsive noise such as door slam, explosions, phone ringtone, kick fusie…The pulses are generated by a process Y (t) and their amplitudes are defined by the sequence yi. The times of occurrence of the pulses are determined by a function N (t) called counting process. The pulses are generated from a delayed Dirac function in the form:

$$Y(t) = \sum_{i=1}^{N(t)} yi * \delta(t - ti). \qquad (1)$$

2.3 The standard MFCC

There exists in the literature a wide variety of technical parameterization of speech signals, we mention the most important of which is revolutionizing in the field of speech recognition namely MFCC. The principle of our strategy is given by fig.3.

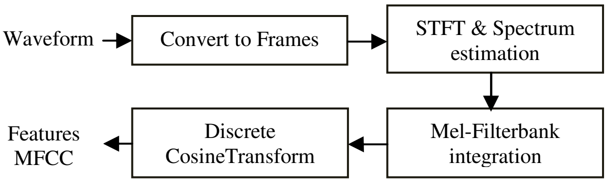


Fig.3. Example of parameterization (MFCC)

This technique consists in calculating the cepstral coefficients on a Mel scale which approximates the frequency of perception of the ear. After applying a short time Fourier transform, energy is calculated in heather critical modeled by triangular filters on the amplitude scale is expressed in decibels. The frequency scale in turn is expressed in Mel. Cepstrum is then calculated by the following expression:

$$C_n = \sqrt{\frac{2}{k}} \sum_{k=1}^{N} (\log S_k \cos (n (k - \frac{1}{2}) \frac{\pi}{k}). \qquad (2)$$

With k = 1...N and S_k representing the energy after filtering by a k triangular filter.

3 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:

$g_c(t) = a^{n-1} \exp(-2\pi b ERB(f_r)\, t) \exp(j2\pi f_r + jc\ln t + jc\varphi)$. (3)

Figure 4 shows the representation of the temporal response of the Gammachirp filter.

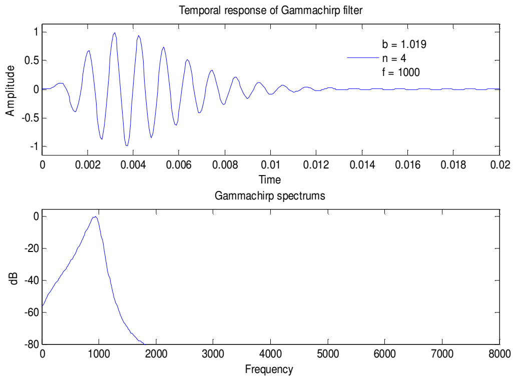


Fig. 4. Example of impulse response gammachip

Where time t>0, a is the amplitude, f_r 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 (f_r) represents the equivalent rectangular bandwidth of the filter, is given by the following relationship:

$$\text{ERB}\ (f_r) = 24.7 + 0.108 f_r. \qquad (4)$$

The Fourier transform of the gammachirp in "equation 3" is derived as follows:

$$|G_c\ (f)| = \frac{a|\sigma(n+jc)|}{\sigma(n)} * \frac{\sigma(n)}{\left|2\pi\sqrt{((bERB(f_r))^2+(f-f_r)^2}\right|^n} e^{c\theta}. \qquad (5)$$

$$|G_c\ (f)| = a_\sigma\ |G_T| * e^{c\theta(f)}. \qquad (6)$$

$$\theta\ (f) = \arctan(\frac{f-f_r}{bERB(f_r)}). \qquad (7)$$

$|G_T(f)|$ is the Fourier magnitude spectrum of the gammatone filter,$e^{c\theta(f)}$, is an asymmetric function since is anti-symmetric function centered at the asymptotic frequency. The spectral properties of the gammachirp will depend on the $e^{c\theta(f)}$, 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 behave as a high-pass filter and if "c" zero, the transfer function, behave as a gammatone filter. In addition, this parameter is connected to the signal power by the expression, [2]:

$$C = 3.38 + 0.107\ \text{Ps}. \qquad (8)$$

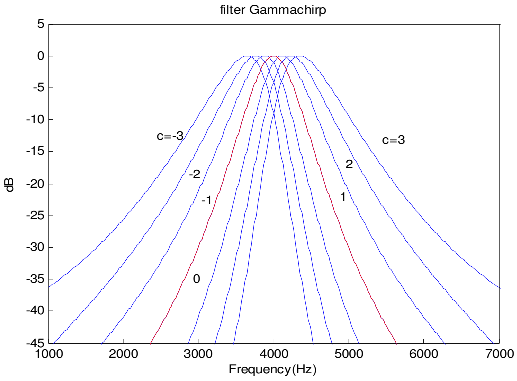


Fig.5. Example of gammachirp spectrums for different values of C

4 Characteristics of the gammachirp

The figure 6 shows a block diagram of the gammachirp filterbank. It is a cascade of three filterbanks: a gammatone filterbank, a lowpass-AC filterbank, and a highpass-AC filterbank. The output of the asymmetric compensation filterbank determines the asymmetric parameter c.

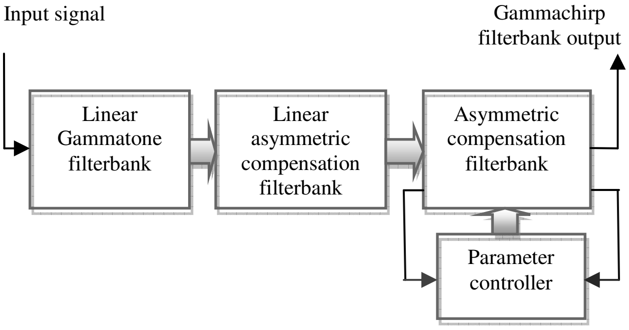

Input signal
Gammachirp filterbank output
Linear Gammatone filterbank
Linear asymmetric compensation filterbank
Asymmetric compensation filterbank
Parameter controller

Fig.6. Structure of the Gammachirp filterbank

Figure 7 shows the amplitude spectra of (a) the gammachirp and (b) the asymmetric function when the values of the chirp parameter c are integers between -3 and 3. Several characteristics are derived from this figure. Figure 7 (a) shows that the filter slope below the peak frequency is shallower than the slope it in the gammachirp when the parameter c is negative. The situation is the reverse when the parameter c is positive. The filter shape is symmetric when c is zero because it is the gammatone. The asymmetric function in fig. 7(b) is an all-pass filter when c=0. This function is a high-pass filter when c>0, and a low-pass filter when c<0. The slope and the range of the amplitude increase when the absolute values of c increases. The filter shapes of the gammachirp in fig. 7 (a) reflect these characteristics.

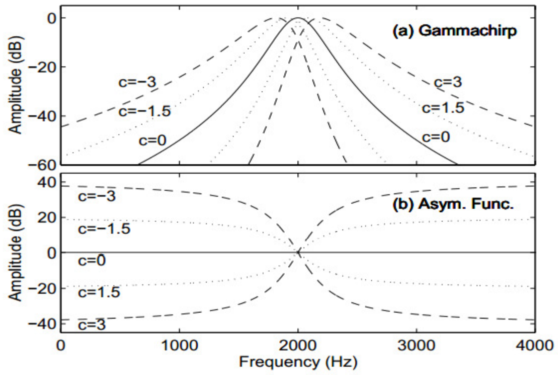


Fig.7. Amplitude spectra of (a) a gammachirp filter and (b) an asymmetric compensation filter

5 Parameterization gammachirp frequency cepstral coefficients

The Gammachirp frequency cepstral coefficients are extracted from the speech signal according to the following steps; use the gammatone filterbank with 32 filters and the bandwidth multiplying factor F = 1.5 to bandpass the speech signal. The filter spacing is linear in the ERB scale. Additional, estimate the logarithm of the short-time average of the energy operator for each one of the bandpass signals, and estimates the cepstrum coefficients using the discrete cosine transform (DCT).

These steps are the main differences between MFCC and GCFB_MFCC feature extraction. The standard MFCC uses filters with frequency response that is triangular in shape (50% filter frequency response overlap). But, GCFB_MFCC use filters that are smoother and broader than the MFCC triangular. Also, the Gammachirp MFCC filterbank is denser in frequency (controlled by the number of filters parameter). The feature extraction algorithm consists of the following steps fig. 8:

i. Filter the speech signal using a gammachirp filterbank.
ii. Estimate the energy coefficients of the framed bandpassed signals.
iii. Transform these energy coefficients into the Cepstrum domain. Only the first low-order cepstral coefficients are kept for recognition (keep the first 12 coefficients, energy).
iv. Estimate their first and second order time derivatives.

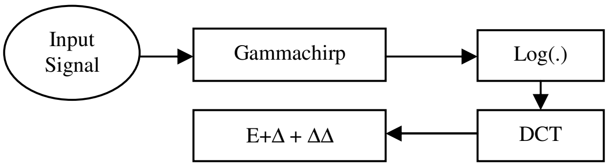


Fig.8. The feature extraction algorithm

6 HMM\GMM models parameters

Presently HMM is widely used as one of the successful speech recognition process. Using left-right HMMs [1] for the pattern recognition stage, offers the advantage that the time evolution of the signal features is taken into account. In this paper, M=3 successive states are considered for the signal features, approximately corresponding to the pulse attack, steady state, and fading phases. During the training process, the system learns the HMM characteristics of each considered signal class, by estimating a mixture Gaussian of the features, and the transition probabilities between states. This training is done with 20 iterations of the Baum-Welsh recursion [14]. During the pattern recognition process, the most probable class of signal is determined by log-likelihood estimation. Instead of the Forward-Backward Algorithm, the likelihood is evaluated using the Viterbi approximation [14], reducing the computation complexity. For each sound class, the statistical behavior of the features (Probability Density Functions) can be modeled with a mixture of Gaussians GMM. This model is characterized by the number of Gaussians, their relative weights, and their mean / covariance parameters. During a training process, the system learns the GMM parameters, by analyzing a subset of the sound database. In the recognition process, the signal to be classified is compared to the models of each class, so as to find the most probable one.

7 TIMIT database

In this study, we built several words bases extracted from the TIMIT database. This database is composed of speakers speaking 8 different dialects of the United States. We used 6132 words composed of 21 words repeated, 292 times, 36 speakers (18 males and 18 females) for training uniformly divided on 8 American dialects. For the test phase of recognition we used 2201 words, 26 speakers (13 males and 13 females) repeated 104 times uniformly divided on 8 American dialects.

These clean speech files were contaminated with additive impulsive noise, in this paper contains 464 sounds of 3 different classes: 314 door slams, 88 glass breaks and 62 explosions. Tests were carried out at different SNR levels (12dB, 6dB, 3dB, 0 dB and -3 dB). The signal to noise ratio (SNR) defined:

$$\text{SNR} = 10.log_{10}(\frac{P_{signal}}{P_{noise}}). \qquad (9)$$

Where P_{signal} and P_{noise} represent respectively the power signal and the noise.

8 Word recognition in impulse noise based on a gammachirp filterbank

Speech signal processing is based either on frequency or on temporal representation and modeling of the human auditory system for improving the design of hearing devices. The analysis of speech signals is operated by using a gammachirp filterbank, in this work we use 32 gammachirp in each filterbank (of 4th order, n = 4), the filterbank is applied on the frequency band of [0 fs/2] Hz (where fs is the sampling frequency), after a pre-emphasis step and a segmentation of the speech signal into frames, and each frame is multiplied by a Hamming windows of 45ms. Each gammachirp filtering is obtained across two steps, in the first step, the speech frame is filtered by the correspondent 4th order gammatone filter, and in the second step we estimate the speech power and calculate the asymmetry parameter c as shown in the following fig. 9, the algorithm uses decomposition through a gammachirp filterbank, where the center frequency of each gammachirp filter has a bandwidth and spacing ERB that cover the 50-8000 Hz range. Each signal is analyzed in order to compute its energy and envelope. The principle of our strategy is given by fig. 9. The first step of the recognition algorithm consists in an analysis of the signal to be classified, in view of extracting some typical features. To evaluate the suggested techniques, we carried out a comparative study with different baseline parameterization technique of MFCC implemented in HTK. We tested the performance in speech signal recognition with additive impulsive noise. Figure 10 shows the difference between the recognition of the word "greasy" (with impulse noise) using traditional techniques (FFT) and gammachirp filterbank.

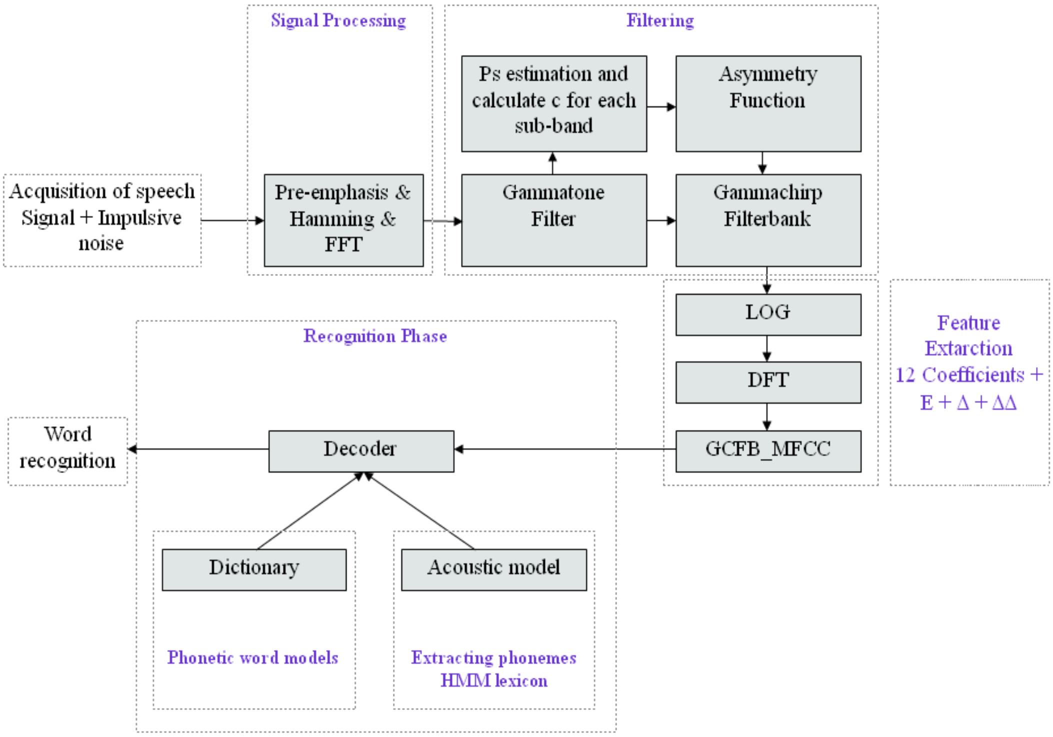
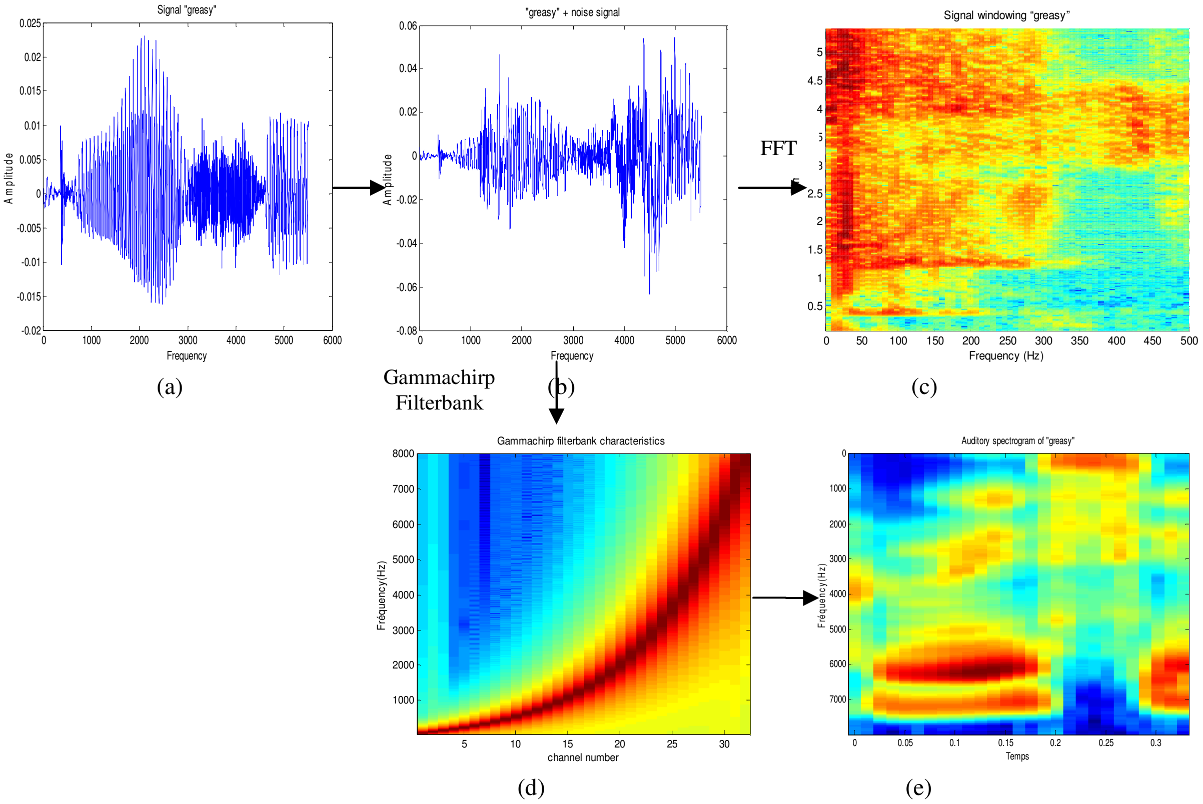

Signal Processing
Filtering
Ps estimation and calculate c for each sub-band
Asymmetry Function
Acquisition of speech Signal + Impulsive noise
Pre-emphasis & Hamming & FFT
Gammatone Filter
Gammachirp Filterbank
LOG
DFT
GCFB_MFCC
Feature Extarction 12 Coefficients + E + Δ + ΔΔ
Recognition Phase
Word recognition
Decoder
Dictionary
Acoustic model
Phonetic word models
Extracting phonemes HMM lexicon

Fig.9. Step of the recognition algorithm



(a) (b) (c)

(d) (e)

Fig.10. (a) Signal "greasy", (b) Signal "greasy" with the noise "door slams", (c) Signal windowing "greasy", (d) Characteristics of Gammachirp filterbank, (e) Auditory spectrogram of the word "greasy" with noise "door slams"; SNR = -3dB

9 Result and discussion

In this section, we illustrate the various output representations that are generated by the gammachirp filterbank and compare them to STFT representations. This article is dedicated to recognize the isolated word with impulsive sound. An extensive database with more than 500 sound samples has been built. This database is made of 3 impulsive sound classes: door slams, explosions and glass breaks.

Figure 10 shows the STFT spectrogram and the gammachirp spectrogram for the word “greasy” with the noise “door slams”. Because the values of the spectrogram are log compressed, it is difficult to observe the compressive effect of the gammachirp. However, for both the gammachirp outputs, spectral peaks for voiced segments of speech appear to be more prominent against the background in all three noises than for the STFT spectrogram. Although, the gammachirp and STFT spectrograms appear very different. First, in the segment between 0.05 and 0.2 seconds, the gammachirp output exhibits a more pronounced formant than for the STFT. On the other hand, the low frequency resonances appear to be more strongly emphasized by the gammachirp, and the bandwidths of most resonances also appear to be much narrower.

Tables 1, 2, 3, 4 and 5 shows the results associated with the rate recognition of MFCC parameterization technique, using “energy”, “delta” and “delta+delta” vectors according to the signal to noise ratio (SNR).
We define the parameters as below.
N: The total number of words to be recognized,
D: The number of words not taken,
S: The number of unrecognized words,
H: The number of recognized words,
%: The percentage rate obtained.
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%

Table 1. Recognition rates obtained by the parameterization technique combined with SNR = -3dB

	SNR=-3dB/ Explosions					SNR=-3dB/ Door slams					SNR=-3dB/ Glass breaks				
	%	N	H	S	D	%	N	H	S	D	%	N	H	S	D
MFCC_e_d_a	87.00	2201	2003	198	0	75.79	2201	2019	182	0	85.79	2201	2100	101	0
GCFB_MFCC_e_d_a	91.85	2201	2002	199	0	77.85	2201	2001	200	0	89.85	2201	2050	151	0

Table 2. Recognition rates obtained by the parameterization techniques combined with SNR = 0dB

	SNR=0dB/ Explosions					SNR=0dB/ Door slams					SNR=0dB/ Glass breaks				
	%	N	H	S	D	%	N	H	S	D	%	N	H	S	D
MFCC_e_d_a	93.14	2201	2050	151	0	83.14	2201	2052	149	0	87.14	2201	2060	141	0
GCFB_MFCC_e_d_a	95.87	2201	2112	89	0	85.20	2201	2112	89	0	90.20	2201	2100	101	0

Table 3. Recognition rates obtained by the parameterization techniques combined with SNR = 3dB

	SNR=3dB/ Explosions					SNR=3dB/ Door slams					SNR=3dB/ Glass breaks				
	%	N	H	S	D	%	N	H	S	D	%	N	H	S	D
MFCC_e_d_a	95.36	2201	2165	36	0	88.36	2201	2155	46	0	88.36	2201	2160	41	0
GCFB_MFCC_e_d_a	98.53	2201	2174	27	0	98.00	2201	2174	27	0	91.00	2201	2163	38	0

Table 4. Recognition rates obtained by the parameterization techniques combined with SNR = 6dB

	SNR=6dB/ Explosions					SNR=6dB/ Door slams					SNR=6dB/ Glass breaks				
	%	N	H	S	D	%	N	H	S	D	%	N	H	S	D
MFCC_e_d_a	98.64	2201	2171	30	0	92.64	2201	2166	35	0	98.64	2201	2170	31	0
GCFB_MFCC_e_d_a	99.10	2201	2182	19	0	96.10	2201	2170	31	0	98.72	2201	2179	22	0

Table 5. Recognition rates obtained by the parameterization techniques combined with SNR = 12dB

	SNR=12dB/ Explosions					SNR=12dB/ Door slams					SNR=12dB/ Glass breaks				
	%	N	H	S	D	%	N	H	S	D	%	N	H	S	D
MFCC_e_d_a	98.84	2201	2181	20	0	94.64	2201	2176	25	0	99.64	2201	2190	11	0
GCFB_MFCC_e_d_a	99.88	2201	2192	9	0	99.10	2201	2189	12	0	99.95	2201	2199	2	0

In the previous section, we present the results of the gammachirp parameterization and the traditional method MFCC. We can see the comparison between the MFCC and GCFB_MFCC, these MFCC Gammachirp give better results in generalization and the better performance with add energy, the delta, acceleration of signal and the SNR.

The feature-based gammachirp filterbank reduces the relative word error rate by 5-10% for the different sound. The improvement is benefited from using a gammachirp filterbank instead of the triangular mel filterbank, table 1, 2, 3, 4 and 5 detailed results of word recognition accuracy rate are shown. In table I the recognition accuracy of the GCFB-MFCC, is 91.85%, but the results change the noise of another, and we see an improvement of recognition rates with energy, the delta and acceleration coefficients.

10 Conclusion

This paper reviewed the background and theory of the gammachirp auditory filter proposed by Irino and Patterson. The motivation for studying this auditory filter is to improve the signal processing strategies employed by automatic speech recognition systems. We have presented an approach of time-frequency analysis “auditory spectrogram” for speech. This takes account of characteristics of the ear. Were analyzed the impulsive noise based on a gammachirp filterbank, in word recognition. The gammachirp was compared to the short time Fourier transforms. Parameterizations implemented showed their performance with recognition system Automatic HTK speech based on Hidden Markov Models given word recognition. We observe that the worst results are those obtained with the basic modeling and best are those obtained with the model with gammachirp filterbank. Concerning this article, we presented the implementation of the gammachirp model of the cochlear filter. We validated this implementation by its use in analysis of some word with impulsive noise. The results gotten after application of this filter on the word show that this filter gives acceptable and sometimes better results by comparison at those gotten by other methods of parameterization such MFCC and PLP.

11 References

[1] 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.

[2] T. Irino, R. D. Patterson. Temporal asymmetry in the auditory system.J. Acoust. Soc. Am. 99(4): 2316-2331, April, 1997.

[3] T. Irino, R. D. Patterson. A time-domain, Level-dependent auditory filter: The gammachirp. J. Acoust.Soc. Am. 101(1): 412-419, January, 1997.

[4] T. Irinoet M. Unoki. An Analysis Auditory Filterbank Based on an IIR Implementation of the Gammachirp. J. Acoust. SocJapan. 20(6): 397-406, November, 1999.

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

[6] T. Irino, R. D. Patterson. A compressive gammachirp auditory filter for both physiological and psychophysical data.J. AcoustSoc . Am. 109(5) : 2008-2022, may 2001.

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

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

[9] B.R. Glasberg, B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data, HearingResearch, 47, 103-198, 1990.

[10] 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.

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

[12] T.Irino, T. and Unoki, M. (1998). "A time-varying, analysis/synthesis auditory filterbank using the gammachirp,” IEEE Int. Conf. Acoust., Speech Signal Processing (ICASSP-98), 3653-3656.

[13] University of Pennsylvania Linguistic Data Consortium. Darpa-timit: a multi speakers data base.

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

[15] J. W. Pitton, K. Wang, and B. H. Juang, “Time-frequency analysis and auditory modeling for automatic recognition of speech,” Proc. IEEE, vol. 84, pp. 1199–1214, Sept. 1996.

[16] O. Ghitza, “Auditory models and human performance in tasks related to speech coding and speech recognition,” IEEE Trans. Speech Audio Processing, vol. 2, pp. 115–132, Jan. 1994.

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