A Comparative Study: Gammachirp Wavelets and Auditory Filter Using Prosodic Features of Speech Recognition In Noisy Environment

Hajer Rahali *Hajer.Rahali@enit.rnu.tn*
National Engineering School of Tunis (ENIT)
Laboratory of Systems and Signal Processing (LSTS)
BP 37, Le Belvédère, 1002 Tunis, Tunisie

Zied Hajaiej *Zied.hajaiej@enit.rnu.tn*
National Engineering School of Tunis (ENIT)
Laboratory of Systems and Signal Processing (LSTS)
BP 37, Le Belvédère, 1002 Tunis, Tunisie

Noureddine Ellouze *N.ellouze@enit.rnu.tn*
National Engineering School of Tunis (ENIT)
Laboratory of Systems and Signal Processing (LSTS)
BP 37, Le Belvédère, 1002 Tunis, Tunisie

Abstract

Modern automatic speech recognition (ASR) systems typically use a bank of linear filters as the first step in performing frequency analysis of speech. On the other hand, the cochlea, which is responsible for frequency analysis in the human auditory system, is known to have a compressive non-linear frequency response which depends on input stimulus level. It will be shown in this paper that it presents a new method on the use of the gammachirp auditory filter based on a continuous wavelet analysis. The essential characteristic of this model is that it proposes an analysis by wavelet packet transformation on the frequency bands that come closer the critical bands of the ear that differs from the existing model based on an analysis by a short term Fourier transformation (STFT). The prosodic features such as pitch, formant frequency, jitter and shimmer are extracted from the fundamental frequency contour and added to baseline spectral features, specifically, Mel Frequency Cepstral Coefficients (MFCC) for human speech, Gammachirp Filterbank Cepstral Coefficient (GFCC) and Gammachirp Wavelet Frequency Cepstral Coefficient (GWFCC). The results show that the gammachirp wavelet gives results that are comparable to ones obtained by MFCC and GFCC. Experimental results show the best performance of this architecture. This paper implements the GW and examines its application to a specific example of speech. Implications for noise robust speech analysis are also discussed within AURORA databases.

Keywords: Gammachirp Filter, Wavelet Packet, MFCC, Impulsive Noise.

1. INTRODUCTION

In order to understand the auditory human system, it is necessary to approach some theoretical notions of our auditory organ, in particular the behavior of the internal ear according to the frequency and according to the resonant level. The sounds arrive to the pavilion of the ear, where they are directed towards drives in its auditory external. To the extremity of this channel, they exercise a pressure on the membrane of the eardrum, which starts vibrating to the same frequency those them. The ossicles of the middle ear, interdependent of the eardrum by the hammer, also enter in vibration, assuring the transmission of the sound wave thus until the cochlea. The resonant vibration arrives to the cochlea by the oval window, separation membrane between the stirrup, last ossicle of the middle ear, and the perilymphe of the vestibular rail. The endolymphe of the cochlear channel vibrates then on its turn and drag the basilar membrane. The stenocils, agitated by the liquidize movements, transforms the acoustic vibration in potential of action (nervous messages); these last are transmitted to the brain through the intermediary of the cochlear nerve [1]. These mechanisms of displacement on any point of the basilar membrane, can

begins viewing like a signal of exit of a pass strip filter whose frequency answer has its pick of resonance to a frequency that is characteristic of its position on the basilar membrane [2]. To simulate the behavior of these filters, several models have been proposed. Thus, one tries to succeed to an analysis of the speech signals more faithful to the natural process in the progress of a signal since its source until the sound arrived to the brain. By put these models, one mentions the model gammachirp that has been proposed by Irino & Patterson. While being based on the impulsion answer of this filter type, it come the idea to implement as family of wavelet of which the function of the wavelet mother is the one of this one. In this paper, a design for modeling auditory is based on wavelet packet decomposition. The wavelet transform is an analysis method that offers more flexibility in adapting time and frequency resolution to the input signal. MFCC are used extensively in ASR. MFCC features are derived from the FFT magnitude spectrum by applying a filterbank 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. It will be shown in this paper that the performance of MFCC, based on the gammachirp filter and referred to as GFCC, are also compared to GWFCC which integrate the gammachirp wavelet. In the current paper, prosodic information is first added to a spectral system in order to improve their performance, finding and selecting appropriated characteristics related to the human speech prosody, and combining them with the spectral features. Such prosodic characteristics include parameters related to the fundamental frequency in order to capture the into-nation contour, and other parameters such as the jitter and shimmer. The implementation of gammachirp wavelet shows consistent and significant performance gains in various noise types and levels. For this we will develop a system for automatic recognition of isolated words with impulsive noise based on HMM\GMM. We propose a study of the performance of parameterization techniques MFCC, GFCC and GWFCC including the prosodic features proposed in the presence of different impulsive noises. Then, a comparison of the performance of different used features was performed in order to show that it is the most robust in noisy environment. The sounds are added to the word with different signal-to-noise SNR (20dB, 15dB and 10dB). Note that the robustness is shown in terms of correct recognition rate (CRR) accuracy. The evaluation is done on the AURORA database.

This paper is organized as follow; in the next section we briefly introduce the prosodic features and auditory filterbank. Section 3 introduces the gammachirp filter as wavelet. The processing steps of our gammachirp wavelet parameterization are described in section 4. Section 5 demonstrates simulations tested with new method. Finally, conclusions are given in section 6.

2. THEORETICAL FRAMEWORK

In this phase of feature extraction we will represent both of spectral and prosodic features which are combined for the aim of creating a robust front-end for our speech recognition system.

2.1 Prosodic features

A) Pitch (Fundamental Frequency)

The vibration of the vocal folds is measured using pitch and is nearly periodic. Pitch frequency F0 is a very important parameter using to describe the characteristic of voice excitation source. The average rate of the vibration of vocal folds measured in the frequency domain is defined as pitch. The rate of vibration is inversely proportional to the shape and size of the vocal folds. The size of vocal folds is different from speaker to speaker and hence the pitch also contains uniqueness information of a speaker. In general, the size of the vocal folds in men is larger than that in women and accordingly pitch of men is lower than that of women. The average pitch for a male speaker is about 50-300 Hz and for a female speaker it is about 100-500 Hz [3].

B) Jitter

Fundamental frequency is determined physiologically by the number of cycles that the vocal folds do in a second. Jitter refers to the variability of F0, and it is affected mainly because of the lack of control of vocal fold vibration .On the other hand, vocal intensity is related to sub glottis pressure of the air column, which, in turn, depends on other factors such as amplitude

of vibration and tension of vocal folds [3]. For our analysis, the following jitter measurements as defined in PRAAT. Mathematically, jitter is the cycle to cycle variation of the pitch period, i.e., the average of the absolute distance between consecutive periods. It is measured in μ sec. It is defined as:

$$\text{Jitter} = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i+1}|. \quad (1)$$

Where Ti is the extracted F0 period length and N is the number of extracted F0 pitch periods. Absolute jitter values, for instance, are found larger in males as compared to females.

C) Shimmer

It is the variability of the peak-to-peak amplitude in decibels. It is the ratio of amplitudes of consecutive periods. It is expressed as:

$$\text{Shimmer (dB)} = \frac{1}{N-1} \sum_{i=1}^{N-1} |20 \log(\frac{A_{i+1}}{A_i})|. \quad (2)$$

Where Ai is the peak-to-peak amplitude in the period and N is the number of extracted fundamental frequency periods. Local shimmer (dB) values are found larger in female as compared to males.

2.2 Gammachirp Auditory Filter

The gammachirp filter is a good approximation to the frequency selective behavior of the cochlea [4]. It is an auditory filter which introduces an asymmetry and level dependent characteristics of the cochlear filters and it can be considered as a generalization and improvement of the gammatone filter. The gammachirp filter is defined in temporal domain by the real part of the complex function:

$$g_c(t) = at^{n-1} e^{-2\pi Bt}\, e^{j2\pi f_r t + jc\ln t + j\varphi}. \quad (3)$$

With

$$B = b.\,\mathrm{ERB}(f_r) = b.\,(24.7 + 0.108\,(f_r)). \quad (4)$$

With: $t > 0$

- N : a whole positive defining the order of the corresponding filter.
- f_r : the modulation frequency of the gamma function.
- φ : the original phase.
- a : an amplitude normalization parameter.
- c : a parameter for the chirp rate.
- b : a parameter defining the envelope of the gamma distribution.
- $\mathrm{ERB}(f_r)$: Equivalent Rectangulaire Bandwith.

When c=0, the chirp term, c ln (t), vanishes and this equation represents the complex impulse response of the gammatone that has the envelope of a gamma distribution function and its carrier is a sinusoid at frequency f_r. Accordingly, the gammachirp is an extension of the gammatone with a frequency modulation term.

A) Energy

The energy of the impulse response $g_c(t)$ is obtained with the following expression:

$$E_{n,B} = ||g_c||^2 = < g_c, g_c > = a^2 \frac{\sigma(2n-1)}{(4\pi B)^{2n-1}}. \quad (5)$$

With σ(n) is the n-th order gamma distribution function. Thus, for energy normalization is obtained with the following expression:

$$A_{E_{n,B}} = \sqrt{\frac{4\pi B^{(2n-1)}}{\sigma(2n-1)}}. \quad (6)$$

B) Frequency response

The Fourier transform of the gammachirp in "(3)" is derived as follows [5].

$$|G_c(f)| = \frac{a|\sigma(n+jc)|}{\sigma(n)} * \frac{\sigma(n)}{\left|2\pi\sqrt{((bERB(f_r))^2 + (f-f_r)^2}\right|^n} e^{c\theta}. \quad (7)$$

$$|G_c(f)| = a_\sigma |G_T| * e^{c\theta(f)}. \quad (8)$$

$$\theta(f) = \arctan\left(\frac{f - f_r}{bERB(f_r)}\right). \quad (9)$$

$|G_c(f)|$ 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 [6]:

$$c = 3.38 + 0.107\ Ps. \quad (10)$$

C) Basic structure

Figure 1 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 [7]. The gammachirp filterbank consists of a gammatone filterbank and an asymmetric compensation filterbank controlled by a parameter controller with sound level estimation.

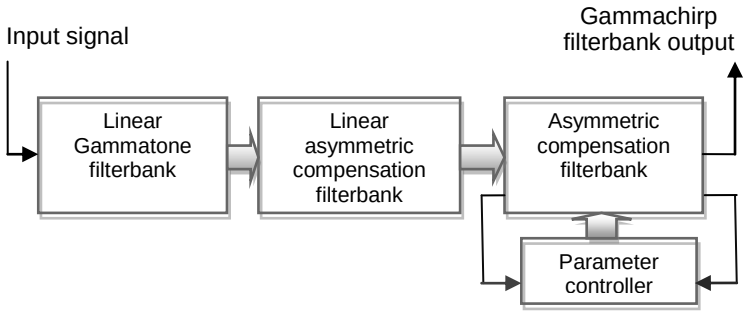

FIGURE 1: Structure of the Gammachirp Filterbank.

This decomposition, which was shown by Irino in [8], is beneficial because it allows 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.

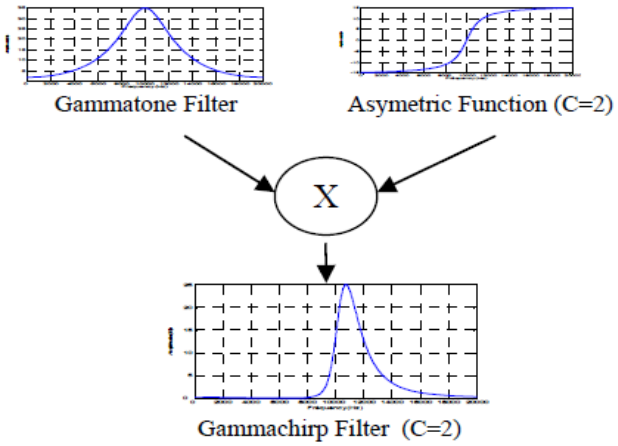

FIGURE 2: Decomposition of the Gammachirp Filter.

3. THE GAMMACHIRP FILTER AS A WAVELET

In this work, a new approach for modeling auditory based on gammachirp filters for application areas including speech recognition. The psychoacoustic model is based on the functioning of human ear. This model analyzes the input signal on several consecutive stages and determines for every pad the spectrum of the signal. The gammachirp filter underwent a good success in psychoacoustic research. Indeed, it fulfils some important requirements and complexities of the cochlear filter [5].

3.1 Wavelet Transform Analysis

The wavelet transform (WT) can be viewed as transforming the signal from the time domain to the wavelet domain. This new domain contains more complicated basis functions called wavelets, mother wavelets or analyzing wavelets. A wavelet prototype function at a scale s and a spatial displacement u is defined as: [9]

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\,\psi\left(\frac{t-u}{s}\right) \quad (u \in \mathbb{R},\ s \in \mathbb{R}_+^*). \tag{11}$$

The WT is an excellent tool for mapping the changing properties of non-stationary signals. The WT is also an ideal tool for determining whether or not a signal is stationary in a global sense. When a signal is judged non-stationary, the WT can be used to identify stationary sections of the data stream. Specifically, a Wavelet Transform function f (t) ϵ L2(**R**) (defines space of square integrals functions) can be represented as:

$$W_{u,s}(f) = \int_{-\infty}^{+\infty} f(t)\ \psi_{u,s}{}^{*}(t)\,dt. \tag{12}$$

The factor of scale includes an aspect transfer at a time in the time brought by the term u, but also an aspect dilation at a time in time and in amplitude brought by the terms s and $\sqrt{s}$.

3.2 The Gammachirp Filter As a Wavelet

The Gammachirp function which is a window modulated in amplitude by the frequency f_r and modulated in phase by the parameter c can thus be seen as wavelet roughly analytical [10] [11]. This wavelet has the following properties: it is with non compact support, it is not symmetric, it is non orthogonal and it does not present a scale function. The gammachirp function can be considered like wavelet function and constitute a basis of wavelets thus on the what be project all input signal, it is necessary that it verifies some conditions that are necessary to achieve this transformation. Indeed it must verify these two conditions:

- The wavelet function must be a finished energy "(5)":

$$||g_c||^2 = a^2\,\frac{(2n-1)!}{(4\pi B)^{2n-1}}. \tag{13}$$

$||g_c||^2=1$ if $a = \sqrt{\left(\frac{(4\pi B)^{2n-1}}{(2n-1)!}\right)}$ which define the filter of normalized energy.

- The wavelet function must verify the admissibility condition:

$$C_{g_c} = \int_0^{+\infty} \frac{|G_c(f)|^2}{f}\,df < +\infty. \tag{14}$$

If the condition "(14)" is satisfied by the function G_c, then it must satisfy two other conditions:

- The mean function g is zero: $G_c(0) = \int_{-\infty}^{+\infty} g_c(t)dt = 0$
- The function G_c (f) is continuously differentiable

To implement the gammachirp function g_c as wavelet mother, one constructs a basis of wavelets then girls $g_{p,q}$ and this as dilating by factor 'p' and while relocating it of a parameter 'q'.

$$g_{p,q}(t) = \frac{1}{\sqrt{p}}\ g_c\left(\frac{t-q}{p}\right). \tag{15}$$

Studies have been achieved on the gammachirp function [10], show that the gammachirp function that is an amplitude modulated window by the frequency f_r and modulated in phase by the c parameter, can be considered like roughly analytic wavelet. It is of finished energy and it verifies the condition of admissibility. For this family of wavelet, the frequencies of modulation are $f_m = f_r \,.\, s_0^{-m}$ and the bandwidths are $B_m = B \,.\, s_0^{-m}$, s_0 is the dilation parameter and m∈Z.

The results show that the value 1000 Hz are the one most compatible as central frequency of the Gammachirp function. Otherwise our work will be based on the choice of a Gammachirp wavelet centered at the frequency 1000 Hz. For this frequency range, the gammachirp filter can be considered as an approximately analytical wavelet. The choice of the gammachirp filter is based on two reasons. First reason is that the gammachirp filter has a well defined

impulse response, and it is excellent for an asymmetric, level-dependent auditory filterbank in time domain models of auditory processing. Second reason is that this filter was derived by Irino as a theoretically optimal auditory filter that can achieve minimum uncertainty in a joint time-scale representation.

4. IMPLEMENTATION

With the gammachirp filter designed as described above, a frequency-time representation of the original signal, which is often referred to as a Cochleagram, can be obtained from the outputs of the filterbank. It is then straightforward to compute MFCC, GFCC and GWFCC features from the Cochleagram. The remaining of this section presents the details of our implementation. In this study, our objective is to introduce new speech features that are more robust in noisy environments. We propose a robust speech feature which is based on the gammachirp filterbank and gammachirp wavelet.

Figure 3 shows the block diagrams of the extraction of MFCC and GFCC features. Figure 4 shows the block diagrams of the extraction of GWFCC features.

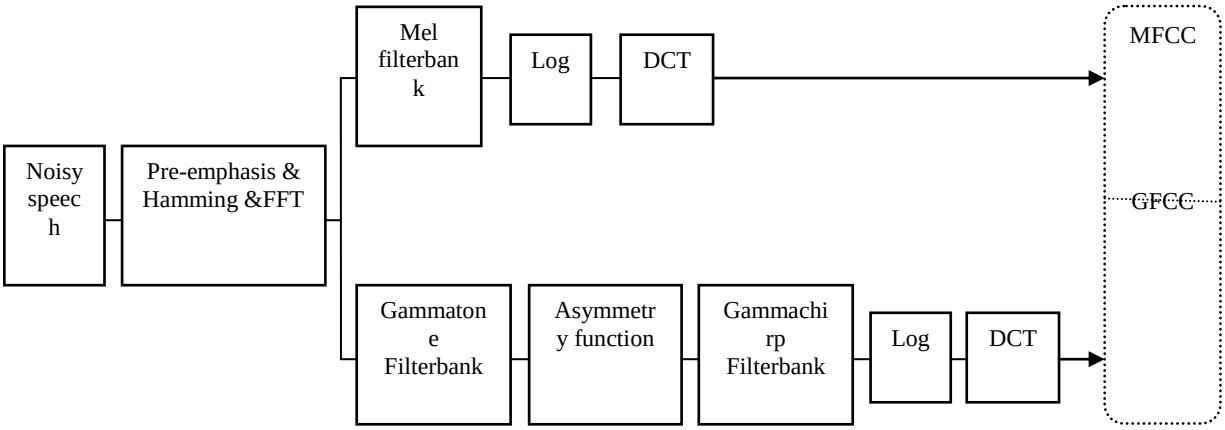

FIGURE 3: Block diagrams of the extraction of MFCC and GFCC features.

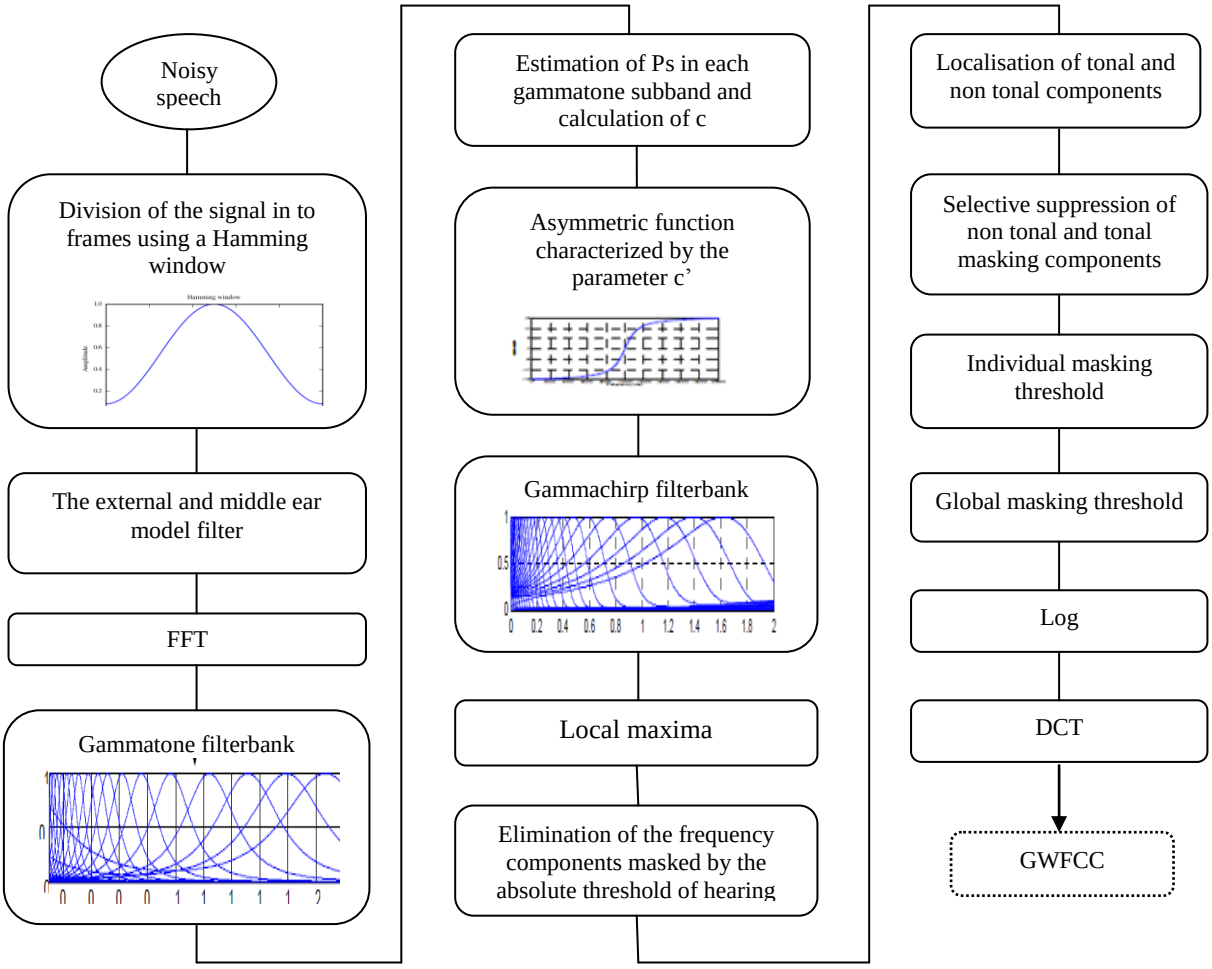

FIGURE 4: Block diagrams of the extraction of GWFCC features.

4.1 MFCC and GFCC

Generally, both methods are based on two similar processing blocks: firstly, basic short-time Fourier analysis which is the same for both methods, secondly, cepstral coefficients computation. As illustrated in figure 3, it can be seen that one of the main dissimilarity between MFCC and GFCC is the set of filters used in the extraction. In fact, triangular filter bank equally spaced in the Mel scale frequency axis is used to extract MFCC features, while in GFCC, the gammachirp filterbank are used. The Mel Cepstral features are calculated by taking the cosine transform (DCT) of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. After pre-emphasizing the speech using a first order high pass filter and windowing the speech segments using a Hamming window of 20 ms length with 10 ms overlap, the FFT is taken of these segments. The magnitude of the Fourier Transform is then passed into a filterbank comprising of 25 triangular filters. The GFCC are extracted from the speech signal according to the following steps; use the gammachirp filterbank defined in "(2)" with 32 filters and the bandwidth multiplying factor F = 1.5 to bandpass the speech signal. After, 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 DCT. These steps are the main differences between MFCC and GFCC features extraction. The standard MFCC uses filters with frequency response that is triangular in shape (50% filter frequency response overlap). But, the proposed auditory use filters that are smoother and broader than the triangular filterbank (the bandwidth of the filter is controlled by the ERB curve and the bandwidth multiplication factor F). The main differences between the proposed filterbank and the typical one used for MFCC estimation are the type of filters used and their corresponding bandwidth. In this paper, we experiment with two parameters to create a family of gammachirp filterbanks: firstly, the number of filters in the filterbank, secondly, the bandwidth of the filters ERB (f). The bandwidth of the filter is obtained

by multiplying the filter bandwidth curve ERB by the parameter F. Experimental results provided in the next section show that both parameters are important for robust speech recognition. The range of parameters we have experimented is 20 – 40 for the number of filters and 1,0 – 2,0 for the bandwidth multiplying factor F. An example of the gammachirp filterbank employing 32 filters and with F = 1.5 is shown in figure 5.

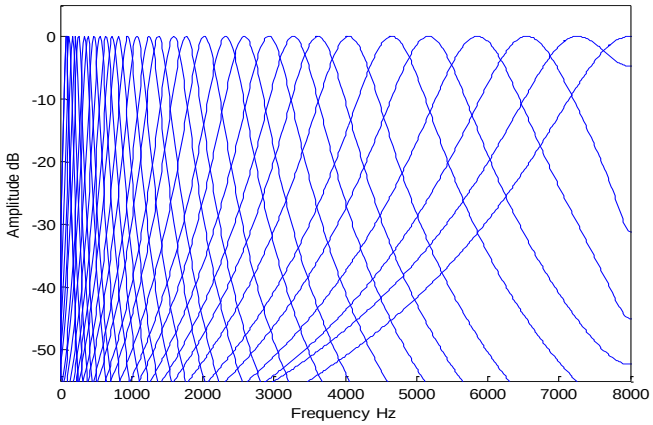

FIGURE 5: A Gammachirp Filterbank with 32 Filters.

4.2 Gammachirp Wavelet Frequency Cepstral Coefficient (GWFCC)

The operating of the new psychoacoustic model is as follows: We segmented the input signal using a Hamming window. The segmented signal is filtered using the non linear external and middle ear model. The output signal of the outer and middle ear model filter is applied to a gammatone filterbank characterized by 32 centers frequencies proposed by the wavelet transform repartition. On each sub-band we calculate the sound pressure level Ps (dB) in order to have the corresponding sub-band chirp term C. Those 32 values of chirp term “c” corresponding to 32 sub-bands of the gammatone filterbank lead to the corresponding gammachirp filterbank. On each sub-band of the dynamic gammachirp filterbank we determine tonal and non tonal components [9]. This step begins with the determination of the local maxima, followed by extracting the tonal components (sinusoidal) and non tonal components (noise) in every bandwidth of a critical band. The selective suppression of tonal and non tonal components of masking is a procedure used to reduce the number of maskers taken into account for the calculation of the global masking threshold. Individual masking threshold takes account of the masking threshold for each remaining component. Lastly, global masking threshold is calculated by the sum of tonal and non tonal components which are deduced from the spectrum to determine finally the signal to mask ratio [12]. After, estimate the logarithm and the cepstrum coefficients using the DCT. In the experiments presented here, a 12 dimensional GWFCC vector is used as the base feature, to which signal log energy is appended, after which velocity and acceleration coefficients (referred to as delta and delta-delta coefficients in the speech community) are calculated for each of the 13 original features, yielding an overall 39 element feature vector for each frame. The complete feature extraction procedure is as shown in figure 6. We note that the addition of delta-cepstral features to the static 13 dimensional GWFCC features strongly improves speech recognition accuracy, and a further (smaller) improvement is provided by the addition of double delta-cepstral features.

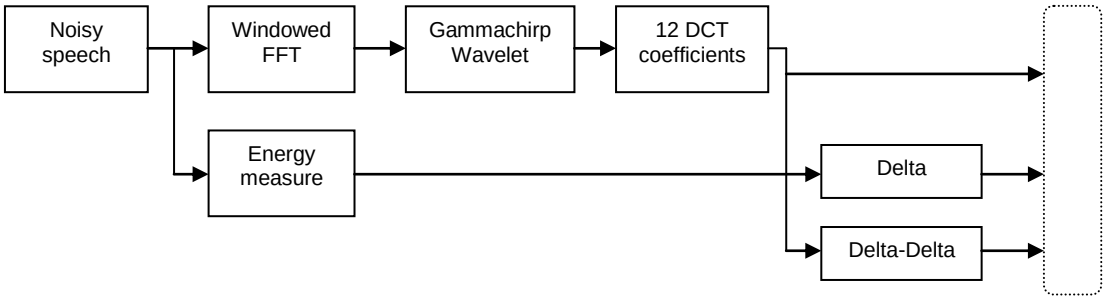

FIGURE 6: Feature extraction with temporal details.

In the next section, we investigate the robustness and compare the performance of the proposed GWFCC features to that of MFCC and GFCC with the different prosodic parameters by artificially introducing different levels of impulsive noise to the speech signal and then computing their correct recognition rate.

5. EXPERIMENTS AND RESULTS

In this section, we investigate the robustness of GWFCC in noise by artificially injecting various types of impulsive noise to the speech signal. We then present speech recognition experiments in noisy recording conditions. The results are obtained using the AURORA databases.

5.1 AURORA Task

AURORA is a noisy speech database, designed to evaluate the performance of speech recognition systems in noisy conditions. The AURORA task has been defined by the European Telecommunications Standards Institute (ETSI) as a cellular industry initiative to standardize a robust feature extraction technique for a distributed speech recognition framework. The initial ETSI task uses the TI-DIGITS database down sampled from the original sampling rate of 20 kHz to 8 kHz and normalized to the same amplitude level [13]. Three different noises (Explosion, door slams and glass breaks) have been artificially added to different portions of the database at signal-to-noise (SNR) ratios ranging from clean, 20dB to 10dB in decreasing steps of 5dB. The training set consists of 8440 different utterances split equally into 20 subsets of 422 utterances each. Each split has one of the three noises added at one of the four SNRs (Clean, 20dB, 15dB and 10dB). The test set consists of 4000 test files divided into four sets of 1000 files each. Each set is corrupted with one of the three noises resulting in a total of (3 x 1000 x 4) 12,000 test utterances. In spite of some drawbacks of the current AURORA task such as the matched test and training conditions, or the absence of natural level variations and variable linear distortions, the AURORA task is of interest since it can demonstrate the potential benefits of using noise robust feature extraction techniques towards improving the recognition performance on a task which (though with matched training and test conditions) has substantial variability due to different types of additive noise at several SNRs.

5.2 Experimental Setup

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 20ms. Generally, gammachirp filterbank and gammachirp wavelet are based on two similar processing blocks: firstly, the speech frame is filtered by the correspondent 4^{th} order gammatone filter, and in the second step we estimate the speech power and calculate the asymmetry parameter c. To evaluate the suggested techniques, we carried out a comparative study with different baseline parameterization technique of MFCC implemented in HTK. The AURORA database is used for comparing the performances of the proposed feature extractor to the MFCC and GFCC features, in the context of speech recognition. For the performance evaluation of our feature extractors, we have used the three noise of the AURORA corpus at four different SNRs (Clean, 20dB, 15dB, 10dB). The features extracted from clean and noisy database have been converted to HTK format using "VoiceBox" toolbox [14] for Matlab. In our experiment, there were 21 HMM models (isolated words) trained using the selected feature GWFCC, GFCC and MFCC. Each model had 5 by 5 states left to right. The features corresponding to each state occupation in an HMM are modeled by a mixture of 12 Gaussians. In the training process, parameters of HMM are estimated during a supervised process using a maximum likelihood approach with Baum-Welch re-estimation. The HTK toolkit was used for training and testing. In all the experiments, 12 vectors with log energy, plus delta and delta-delta coefficients, are used as the baseline feature vector. Jitter and shimmer are added to the baseline feature set both individually and in combination. Table I, II, III and VI shows the overall results. In our experiment, we tested the performance of gammachirp wavelet with additive impulsive noise and prosodic parameter, through recognition of word.

5.3 Results and Discussion

The performance of the suggested parameterization methods GWFCC and GFCC is tested on the AURORA databases using HTK. We use the percentage of word accuracy as a performance evaluation measure for comparing the recognition performances of the feature extractors considered in this paper. %: The percentage rate obtained. Tables I, II and III present the average word accuracy (in %), averaged over all noise scenarios. One Performance measures, the correct recognition rate (CORR) is adopted for comparison. They are defined as:

$$\% \text{ CRR} = \text{no. of correct labels/no. of total labels} * 100\%. \quad (16)$$

Features / SNR	Explosions				Door slams				Glass breaks			
	Clean	*20 dB*	*15 dB*	*5 dB*	*Clean*	*20 dB*	*15 dB*	*5 dB*	*Clean*	*20 dB*	*15 dB*	*5 dB*
MFCC (Baseline)	85.45	82.25	78.29	78.44	84.34	80.56	78.98	77.76	87.76	85.43	77.76	76.10
MFCC+Jitter	88.05	84.85	80.89	80.04	86.94	83.16	81.58	80.36	90.36	88.03	80.06	78.70
MFCC+Shimmer	88.45	85.25	81.29	81.44	87.34	83.56	81.98	80.76	90.76	88.43	80.76	79.10
MFCC+Jitter+Shimmer	89.55	86.35	82.39	82.54	88.44	84.66	82.76	81.86	91.86	89.53	81.86	80.20

TABLE 1: Word accuracy (%) of MFCC.

Features / SNR	Explosions				Door slams				Glass breaks			
	Clean	*20 dB*	*15 dB*	*5 dB*	*Clean*	*20 dB*	*15 dB*	*5 dB*	*Clean∞*	*20 dB*	*15 dB*	*5 dB*
GFCC (Baseline)	89.85	85.23	80.27	80.34	88.24	85.56	81.98	81.76	88.76	87.53	86.76	86.32
GFCC+Jitter	92.45	87.85	82.89	82.94	90.84	88.16	84.58	84.36	91.36	90.13	89.36	88.90
GFCC+Shimmer	92.85	88.23	83.27	83.34	91.24	88.56	84.98	84.76	91.76	90.53	89.76	89.32
GFCC+Jitter+Shimmer	93.95	89.33	84.37	84.44	92.34	89.66	86.06	85.86	92.86	91.63	90.86	90.42

TABLE 2: Word accuracy (%) of GFCC.

Features / SNR	Explosions				Door slams				Glass breaks			
	Clean∞dB	*20 dB*	*15 dB*	*5 dB*	*Clean∞ dB*	*20 dB*	*15 dB*	*5 dB*	*Clean∞ dB*	*20 dB*	*15 dB*	*5 dB*
GWFCC (Baseline)	92.43	90.17	88.20	85.74	90.35	89.96	88.98	87.70	92.76	91.43	90.86	90.54
GWFCC+Jitter	95.05	92.77	90.80	88.34	92.95	92.56	91.58	90.30	95.36	94.03	93.46	93.14
GWFCC+Shimmer	95.43	93.17	91.20	88.74	93.35	92.96	91.98	91.70	95.76	94.43	93.86	93.54
GWFCC+Jitter+Shimmer	96.53	94.27	92.30	89.84	94.45	94.06	93.08	92.80	96.86	95.53	94.96	94.64

TABLE 3: Word accuracy (%) of GWFCC.

The recognition accuracy for GWFCC, ΔGWFCC and ΔΔGWFCC are obtained and presented in the table VI by different noise. The results are considered for 39 features (GWFCC+ΔGWFCC+ΔΔGWFCC).

Features / SNR	Explosions				Door slams				Glass breaks			
	Clean∞dB	*20 dB*	*15 dB*	*5 dB*	*Clean∞ dB*	*20 dB*	*15 dB*	*5 dB*	*Clean∞ dB*	*20 dB*	*15 dB*	*5 dB*
GWFCC (13)	83	82.34	80.98	75.74	82.54	80.67	80.54	79.70	85.76	85.47	84.06	83.54
GWFCC+ΔGWFCC (26)	84.23	83	82.54	80.76	90.95	89.56	88.98	85.30	91.56	90.83	90.42	89.86
GWFCC+ΔGWFCC+ ΔΔGWFCC (39)	93.64	91.17	91.20	90.09	94.21	92.87	91.98	90.10	97.76	95.43	92.06	90.87

TABLE 4: Recognition rate (%) of GWFCC, ΔGWFCC and ΔΔGWFCC.

Table I, II and III presents the performance of three voice features in presence of various levels of additive noise. We note that the GWFCC features that are extracted using the gammachirp wavelet exhibit the best CRR. Also, it is observable that the performance of the three features decreases when the SNR decreases too, that is, when the speech signal becoming more noisy. Similarly, the performance of GFCC shows a decrease, but it is a relatively small decrease, whereas the GWFCC features have the overall highest recognition rate throughout all SNR levels. These results assert well the major interest of the gammachirp wavelet and of the auditory filterbank analysis. In additive noise conditions the proposed method provides comparable results to that of the MFCC and GFCC. In convolutive noise

conditions, the proposed method provides consistently better word accuracy than all other methods. Jitter and shimmer are added to the baseline feature set both individually and in combination. The absolute accuracy increase is 2.6% and 3.0% after appending jitter and shimmer individually, while there is 4.1% increase when used together. As we can see in the tables, the identification rate increases with speech quality, for higher SNR we have higher identification rate, the gammachirp wavelet based parameters are slightly more efficiencies than standard GFCC for noisy speech (94.27% vs 89.33% for 20 dB of SNR with jitter and shimmer) but the results change the noise of another. We can see the comparison between the two methods parameterization, these GWFCC give better results in generalization and the better performance. The improvement is benefited from using a gammachirp wavelet instead of the auditory filterbank. From the above table VI, it can be seen that the recognition rates are above 90%, this is recognition rates are due to the consideration of using 39 GWFCC features.

From all the experiments, it was concluded that GWFCC has shown best recognition performance compared to other feature extraction techniques because it incorporates gammachirp wavelet features extraction method.

6. 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. In this paper, we concentrated on the implementation of automatic speech recognition in noisy environments. This system uses gammachirp wavelet cepstral features extracted from an audio signal after analysis by gammachirp filterbank. The proposed features (GWFCC) have been shown to be more robust than MFCC and GFCC in noise environments for different SNR values.

Several works have demonstrated that the use of prosodic information helps to improve recognition systems based solely on spectral parameters. Jitter and shimmer features have been evaluated as important features for analysis for speech recognition. Adding jitter and shimmer to baseline spectral and energy features in an HMM-based classification model resulted in increased word accuracy across all experimental conditions. The results gotten after application of this features show that this methods gives acceptable and sometimes better results by comparison at those gotten by other methods of parameterization such MFCC and GFCC.

7. REFERENCES

[1] Miller A., Nicely P. E. (1955). Analyse de confusions perceptives entre consonnes anglaises. J. Acous. Soc. Am, 27, 2, (trad Française,Mouton, 1974 in Melher & Noizet, textes pour une psycholinguistique).

[2] Greenwood, D.D. A cochlear frequency-position function for several species – 29 years later. J.Acous. Soc. Am, Vol. 87, No. 6, Juin 1990.

[3] R.E. Slyh, W.T. Nelson, E.G. Hansen. Analysis of m rate, shimmer, jitter, and F0 contour features across stress and speaking style in the SUSAS database. vol. 4. in Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, pp. 2091-4, Mar. 1999.

[4] 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.

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

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

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

[8] Irino, T., Patterson R. D. A compressive gammachirp auditory filter for both physiological and psychophysical data. J. Acoust. Soc. Am. Vol. 109, N° 5, Pt. 1, May 2001. pp. 2008-2022.

[9] S. Mallat. A Theory for multiresolution signal decomposition: Wavelet representation. IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. 11. No. 7 pp 674-693 July 1989.

[10] Alex Park. Using the gammachirp filter for auditory analysis of speech. May 14, 2003. 18.327: Wavelets and Filter banks.

[11] Stephan Mallat. Une exploitation des signaux en ondelettes. Les éditions de l'école polytechnique.

[12] H.G. Musmann. Genesis of the MP3 audio coding standard. IEEE Trans. on Consumer Electronics, Vol. 52, pp. 1043 – 1049, Aug. 2006.

[13] H. G. Hirsch, D. Pearce. The AURORA Experiment Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition. ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium, France, 2000.

[14] M. Brookes. VOICEBOX: Speech Processing Toolbox for MATLAB. Software, available [Mar, 2011] from, www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[15] E. Ambikairajah, J. Epps, L. Lin. Wideband speech and audio coding using gammatone filter banks. Proc. ICASSP'01, Salt Lake City, USA, May 2001, vol.2, pp.773-776.

[16] M. N. Viera, F.R. McInnes, M.A. Jack. Robust F0 and Jitter estimation in the Pathological voices. Proceedings of ICSLP96, Philadelphia, pp.745–748, 1996.

[17] Salhi.L. Design and implementation of the cochlear filter model based on a wavelet transform as part of speech signals analysis. Research Journal of Applied Sciences2 (4): 512-521, 2007 Medwell-Journal 2007.

[18] P. Rajmic, J. Vlach. Real-time Audio Processing Via Segmented wavelet Transform. 10th International Conference on Digital Audio Effect , Bordeaux, France, Sept. 2007.

[19] P.R. Deshmukh. Multi-wavelet Decomposition for Audio Compression. IE (I) Journal –ET, Vol 87, July 2006.

[20] WEBER F., MANGANARO L., PESKIN B. SHRIBERG E. Using prosodic and lexical information for speaker identification. Proc. ICASSP, Orlando, FL, May 2002.