

# Gammachirp Filter Frond-End for Automatic Speech Recognition

Zied Hajaiej, Kaïs Ouni and Nouredine Ellouze

*Laboratory of Systems and Signal Processing (LSTS)  
National Engineering School of Tunis (ENIT)  
BP 37 Le Belvédère, 1002, Tunis, TUNISIA*

**Hajaiej\_Zied@yahoo.fr**

**kais.ouni@gmail.com**

**N.ellouze@enit.rnu.tn**

**Abstract:** Feature computation models for automatic speech recognition (ASR) systems have long been modeled on the human auditory system. Most current ASR systems model the critical band response and equal loudness characteristics of the auditory system. This paper presents a new technique to extract the feature based on the human auditory system characteristics. It relies on the gammachirp filterbank to emulate the cochlea frequency resolution. Our proposed feature computation method differs of PLP and RASTA-PLP computation method only in the procedure used for modeling the human auditory perception by gammachirp filterbank to emulate Asymmetric frequency response & level dependent frequency response. In this paper, we study the signal processing of the above mentioned features computation methods, and point out to the differences between the two methods, and the effect of these differences on recognition performance.

**Key words:** ASR, gammachirp filter, HMM, speech recognition.

## INTRODUCTION

A major problem in speech recognition system is the decision of the suitable feature set which can faithfully describe in an abstract way the original highly redundant speech signal. Several techniques have been developed so far for solving this problem. It has been known that the cochlea, the main component of the inner ear, performs the filterbank based frequency analysis on the speech signal to extract the relevant features. Thus, most techniques are pivoting around the filterbank methodology in extracting the features. The difference in the design of the filterbank offers the extraction of different features from the signal.

The main parameters in the filterbank filter design are the frequency response, which defines the shape of the filters, the centre frequency and the bandwidth. These parameters can be selected based on the human auditory system. Dominant speech analysis techniques for ASR, namely Mel frequency cepstrum and perceptual linear predictive (PLP) [9], try to emulate the human auditory perception. The auditory models 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 set of sub-band signals. The most popular among the filterbank have been the gammatone filterbank based on the linear gammatone filter, and was a good fit to the roex [2] filters that were generally used to model human psychophysical data [3]. The gammatone filter is a linear filter and cannot model the level dependent properties. Hence Irino and Patterson introduced the gammachirp filter which was a modification of the gammatone filter [3]. The gammachirp was later modified into a compressive gammachirp [5]. This filter was shown to model most of the level dependency observed in basilar membrane filtering.

In this paper we develop a compromise between different front ends to design a model that is more coherent than the auditory models and having the advantages of the Mel cepstrum and the PLP front ends in being fast. Our technique is based on the Gammachirp auditory filterbank in extracting the relevant features. Gammachirp filter modelling is a physiologically based strategy followed in mimicking the structure of the peripheral auditory processing stage. It models the cochlea by emulate a bank of overlapping band pass filters with asymmetric frequency response and level dependent frequency response.

The performance of these techniques will be measured using the words recognition rate to show the classification ability of our techniques, the classical MFCC and the PLP techniques. Also the recognition rate based ASR system will be compared by using the above techniques. The system of recognition adopted for this study is The HTK toolkit originally developed at Cambridge University.

## 1. Perceptual linear prediction (PLP) and relative spectral (RASTA)

Indeed, Perceptual linear prediction analysis is a variation of original LPC analysis and was first introduced by Hermansky [9] 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 although 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 equal loudness curve and finally the resulting spectrum is compressed with cubic-root nonlinearity, simulating the intensity-loudness power law. The resulting low order all pole model is consistent with several phenomena observed in human speech perception. In this step, the IDFT is applied to obtain the dual autocorrelation function. The first  $M+1$  values are used for solving the Yule-Walker equation for obtaining the autoregressive coefficients of the all-pole model of order  $M$ . Along with PLP, Relative Spectral (RASTA) method is used to compensate for the channel effects in recognizers. RASTA uses filtering in the log domain of the power spectrum to produce robust representations of speech.

## 2. Gammachirp auditory filter

Gammachirp filter is popular for auditory speech analysis [1], [3] and [4]. This function was introduced by Irino and Patterson. It has the following classical form:

$$g_c(t) = at^{n-1} \exp(-2\pi b \text{ERB}(f_r)t) \exp(j2\pi f_r t + jc \ln t + jc\varphi) \quad (1)$$

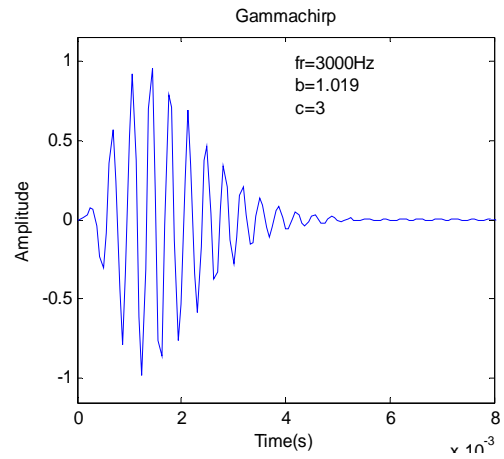


Figure 1. Example of impulse response gammachirp

Where time  $t > 0$ ,  $a$  is the amplitude,  $n$  and  $b$  are parameters defining the envelope of the gamma distribution, and  $f_r$  is the asymptotic frequency.  $c$  is a parameter for the frequency modulation or the chirp rate,  $\varphi$  is the initial phase,  $\ln t$  is a natural logarithm of time, and ERB ( $f_r$ ) is the equivalent rectangular bandwidth of an auditory filter at  $f_r$ . At moderate levels in Hz. The Fourier transform of the gammachirp in Eq. (1) is derived as follows.

$$G_c(f) = \frac{a |\Gamma(n+jc)|}{\Gamma(n)} \cdot \frac{\Gamma(n)}{|2\pi \sqrt{(b \text{ERB}(f_r))^2 + (f - f_r)^2}|^n} e^{c\theta} \quad (2)$$

$$|G_c(f)| = a_T |G_T| \cdot e^{c\theta(f)} \quad (3)$$

$$\theta(f) = \arctan\left(\frac{f - f_r}{b \text{ERB}(f_r)}\right) \quad (4)$$

$G_c(f)$  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. For a given gammatone filter, 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 asymmetric function is low-pass filter for negative values of  $c$ , a high-pass filter when  $c$  is positive, and When  $c=0$  is the complex form of the gammatone filter. The peak frequency  $f_p$  in the amplitude spectrum can be obtained analytically by setting the derivative of Eq. (2) to zero and solving the equation for the frequency. The result is by [4],[5].

$$f_p = \frac{f_r + cb \text{ERB}(f_r)}{n} \quad (5)$$

Therefore, the size of the peak shift is proportional to the chirp parameter  $c$  and the ratio of the envelope parameter  $b$ , ERB ( $f_r$ ) to  $n$ . Irino and Patterson varied the asymmetry parameter linearly with the probe level of the signal, measured in dB. In a typical case,  $c$  is negative when greater than about 30 dB.

### 3. Bandwidth and center frequency of the gammachirp filter bank

The bandwidth of each filter in the gammachirp filterbank is determined according to the auditory critical band corresponding to its centre frequency. The critical band is the bandwidth of the human auditory filter at different characteristic frequencies along the cochlea path.

The first determination of the critical band was done by Fletcher in 1938. He assumed that the auditory filters were rectangular, which greatly simplified the formulation of the signal and the noise powers within the critical band. Although the rectangular critical band concept is not realistic, it is very useful. The bandwidth of the actual auditory filters can be related to it, by suggesting an equivalent rectangular bandwidth (ERB) filter that has a unit height and a bandwidth ERB. It passes the same power as the real filters do when subjected to white noise input. This definition of ERB implies the mathematical formula:

$$ERB = \int_0^{\infty} |H(f)|^2 df \quad (6)$$

Where the maximum value of the filter transfer function,  $|H(f)|$ , is unity. Several physiologically motivated formulas have been derived for the ERB values and our preference is with that suggested by Glasberg et al. [8]. It follows the following formula at centre frequency.

$$ERB(f_r) = 24.7 + 0.108 f_r \quad (7)$$

This formula gives the highest selectivity factor, Q factor, among all the other suggested ones. The Q factor is the ratio between the centre frequency and the bandwidth of each filter. Thus to determine the bandwidth of each filter, which is now represented by the ERB value, the centre frequency of each filter has to be ready beforehand. In the human auditory system, there are around 3000 inner hair cells along the 35mm spiral path cochlea. Each hair cell could resonate to a certain frequency within a suitable critical bandwidth. This means that there are approximately 3000 band pass filters in the human auditory system. This resolution of filters can not be implemented practically using computational modeling techniques.

However we can approximate this high resolution into some possibly implemented one. This can be achieved by specifying certain overlapping between the contiguous filters. The percentage overlapping factor  $v$  will specify the number of channels, filters, required to cover the useful frequency band. This band is decided according to the requirements of the application. In our speech recognition system this band is in the range of 50-8000Hz, as this is the useful information distribution band. If we depend on Glasberg and Moore [8] recommendation and if we suppose that the information carrying band is bounded by high frequency  $f_h$  and low frequency with  $v$

overlapping spacing the number of filters will be:

$$N = \frac{9.26}{v} \ln \frac{f_h + 228.7}{f_l + 228.7} \quad (8)$$

Then the centre frequency can be calculated by:

$$f_c = -228.7 + (f_h + 228.7) e^{\frac{vn}{9.26}} \quad (9)$$

Where  $1 \leq n \leq N$ . Having decided the locations of the centre frequency of each filter the bandwidth can be calculated from (7) and we can now proceed to the implementation stage.

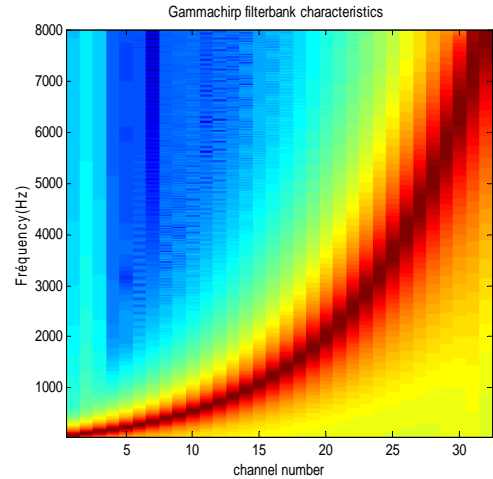


Figure 2. Gammachirp filterbank characteristics

### 4. Gammachirp auditory motivation

The physiologically motivation gammachirp filter can be used to weighting coefficients for speech signals. The speech signal, sampled at 16 kHz, is first preemphasized and cut into a number of overlapping segments. A Hamming window is multiplied and Fourier transform (DFT) is computed for each frame. The input frame is decomposed into a set of sub band signals by the filtering action of the gammatone filterbank, where the center frequency of each gammatone filter have bandwidths and spacing ERB and cover the 50-8000 Hz range.

Each section of the filterbank Gammachirp basically consists of 2 paths, one the filtering path and the other being the signal strength estimation path. The filtering path is 4th order gammatone filterbank followed by an asymmetry function whose one parameter is to be controlled to achieve level dependent filtering action. In the other path estimate  $P_s$  in each sub band and  $c$  computation using the filters outputs are subjected to equal loudness preemphasis filter. From this stage we experimented with two options. The first option, GammaChirp PLP (GC-PLP) is to augment similar steps used in preparing the PLP coefficients. The second option, GammaChirp-PLP RASTA (GC-PLP RASTA), is subjected RASTA filtering from Eq 10.

The protocol of comparison consists in listening to the compressed original sound file. Then, the possibility is offered to the tester to listen to it as much time as he wishes. The test on the following sound file is accessible when a note was validated.

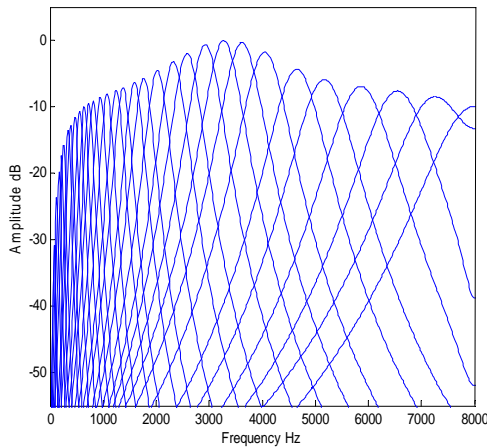
$$H(z) = 0.1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \quad (10)$$

We continue similar steps used in the PLP coefficients. Figure 3 shows the frequency response of Gammachirp filterbank covering 50-8000 Hz band after the preemphasis by the equal loudness curve,

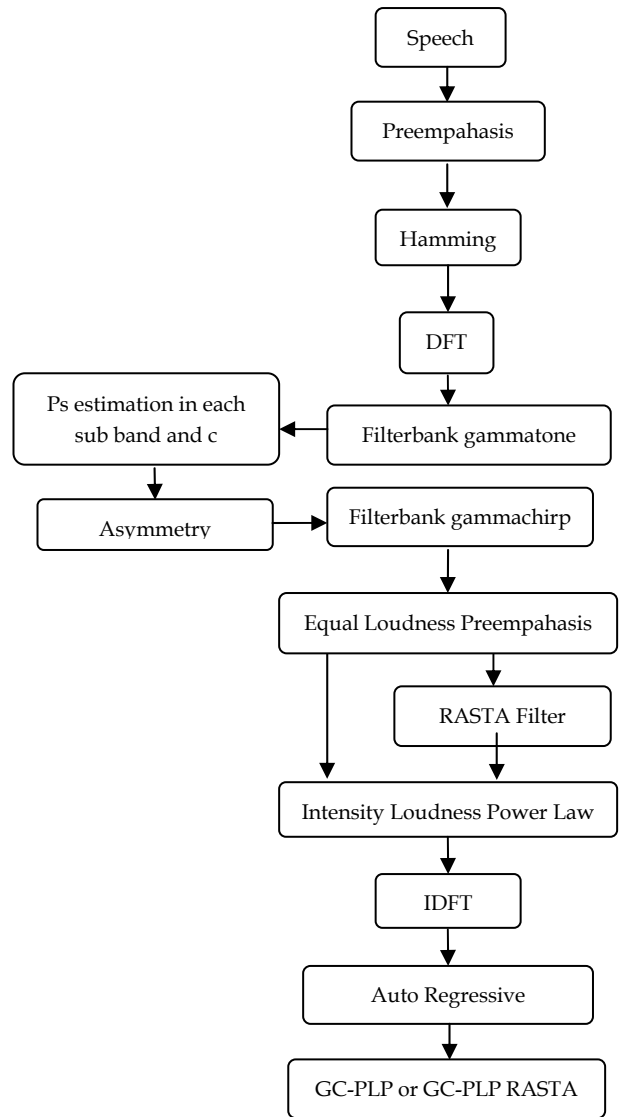
Figure 4 shows the block representation of basic signal processing model used to perform the human auditory system.

### 5. Experimental conditions and results

To evaluate the suggested technique, we carried out a comparative study with different traditional techniques of parameterization MFCC and PLP. Performance we tested it in a word recognition task. We will describe in this part the experimental conditions.



**Figure 3.** Example of gammachirp filterbank after the preemphasis by the equal loudness curve filter



**Figure 4.** Block diagram of the gammachirp parameterization

#### 5.1. The database

In this study, we built several words bases extracted from the Darpa-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.

#### 5.2. Traditional parametrisation methods

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, One energy measure, and their first and second derivatives.

### 5.3. Hidden markov models parameters

In our experiment, there were 21 HMM models trained using the selected feature (GC-PLP, GC-PLP RASTA, MFCC and PLP). With Gaussian mixture for observations corresponding to each state of an HMM are used to characterize a model. Each model had 5 by 5 states left-to-right. The features corresponding to each state occupation in an HMM are modelled by a mixture of 12 Gaussians. Although the use of 12 mixtures give slightly better recognition results.

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.

**Table 1.** Words recognition accuracy obtained by the different techniques of parameterization and different combined of energy, delta and delta-delta

	Rough state	Energy	Energy, Delta	Energy, Delta and Delta-Delta
MFCC	91.00	93.14	98.36	98.64
PLP	91.78	94.14	98.41	99.05
GC-PLP	91.94	94.32	98.86	99.42
GC-PLP RASTA	90.83	92.96	98.54	99.52

### 5.4. Recognition results

In this paragraph, we present the results of the suggested parameterization and the traditional methods MFCC, PLP. We measure the generalization score. We can see comparisons recognition rates obtained by the suggested parameterization and the traditional methods, recognition rates have been obtained after train and test iteration. These GC-PLP give better results in generalization and the better performance with add energies of signal. The results are consistent with the GC-PLP RASTA in function of the derived parameters, than spectral method like MFCC or perceptual methods PLP.

One performance measures, the correct recognition rate (CORR) is adopted for comparison. They are defined as:

$$\% \text{ CORR} = \text{no. of correct labels} / \text{no. of total labels} * 100\%$$

## 6. 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 in Figure 4 have been implemented and tested. They outperform their classical counterparts, MFCC and PLP techniques.

## REFERENCES

- [1] Ouni Kais," Contribution to the Vocal Signal Analysis Using Knowledges on the Auditory Perception and Multiresolution Time Frequency Representation of the Speech Signals" (in french), PhD Thesis on Electrical ENIT Tunis. February 2003.
- [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. 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.
- [5] T. Irino, R. D. Patterson. "A compressive gammachirp auditory filter for both physiological and psychophysical data." *J. Acoust. Soc. Am.* 109(5): 2008-2022, may 2001.
- [6] 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.
- [7] 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.
- [8] B.R. Glasberg, B. C. J. Moore. "Derivation of auditory filter shapes from notched-noise data", *Hearing Research*, 47, 103-198, 1990.
- [9] H. Hermansky. "Perceptual Linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.* Vol. 87, No. 4, pp. 1738-1752., April 1990.
- [10] 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)*, pp3653-3656.