A speech analysis technique based on a modeling of the masking properties of the peripheral auditory system

KAIS OUNI & NOUREDDINE ELLOUZE
Department of Electrical Engineering
Ecole Nationale d'Ingénieurs de Tunis
BP.37, Le Belvédère 1002, Tunis
TUNISIA
Kais.ouni@enit.rnu.tn & N.ellouze@enit.rnu.tn

Abstract: - A novel approach about an auditory-speech analysis technique is presented in this paper. This analysis is based on an atomic decomposition of speech signal into a family of gammachirp functions. A peripheral auditory model is used to simulate the outer and middle ear filtering, the spectral behavior of the cochlea and the time domain processing of the auditory system. A spectrographic representation is finally given as an auditory image, which will be transmitted to the brain. The obtained spectrograms highlight at the same time, the harmonics of the pitch and the structure of formants. Furthermore, they present a good distinction between fricatives, which favors these auditory-spectrograms versus classical short-time Fourier spectrograms.

Key-Words: - Speech Analysis, Gammachirp Spectrogram, Auditory System Model, Masking Properties.

1 Introduction

The human auditory system is extremely well tuned to facilitate speech communication. Better modelling of this system will illuminate robust strategies for speech processing applications [2][3][11]. Masking describes the phenomenon of one sound stimulus affecting the perception of another stimulus [3]. In the frequency domain this phenomenon appears as a simultaneous masking. In the time domain, this phenomenon appears as a temporal masking.

When the sound pressure waves impinge upon the eardrum, in the outer ear, they cause vibrations that are transmitted via the stapes, at the oval window, to the fluids of the cochlea in the inner ear. This behaviour of the outer and middle ear can be simulated by a fixed filter [12]. The pressure waves in turn produce mechanical displacements in the basilar membrane. The amplitude and time course of these vibrations reflect directly the amplitude and frequency content of the sound stimulus [2]. These mechanical displacements, at any given place of the basilar membrane, can be viewed as the output signal of a band-pass filter whose frequency response has a resonance peak at frequency that is characteristic of the place. Filters with a so-called gammatone impulse response are more used for the modelling of the cochlear filterbank [3][4]. It is distinguished by a spectral bandwidth that depends on the central frequency of its corresponding cochlear filter, which is measured in Equivalent Rectangular Bandwidth (ERB).

The ERB is related to the psychophysical critical band assignment. The critical bandwidth is about 50-100 Hz at low frequencies and changes gradually to around 20 percent of the frequency at high frequencies. A new impulse response, named gammachirp, has been introduced recently by Irino [4]. Irino [4], and Irino & Patterson [5] have demonstrated that the gammachirp filter is an excellent candidate for asymmetric and level-dependent cochlear filter. The gammachirp filter is an extension of the gammatone filter with an additional chirp term to produce an asymmetric amplitude spectrum.

In this paper, we develop a speech analysis technique based on a multiresolution process with a gammachirp filterbank. This multiresolution process is done by a time-frequency gammachirp atomic decomposition into a family of gammachirp functions. This technique simulates the temporal and frequency processing of the peripheral auditory system.

The outer and middle ear filtering is simulated by a fixed filter. The frequency masking properties are simulated by a gammachirp filterbank. The temporal masking properties are simulated by a temporal window. A gammachirp spectrogram is then given as a simulation of the auditory image, which will be transmitted to the brain.

The first part of this paper presents a detailed description of the time-frequency gammachirp atomic decomposition. The second part presents the auditory model for the proposed speech analysis technique. Examples of gammachirp spectrograms of the obtained auditory image are given. Also a comparison of the gammachirp spectrograms to classical short time Fourier spectrograms is given and commented.

2 The Atomic Gammachirp Decomposition

In this first section we present a time- frequency study of an atomic decomposition of speech signal based on a gammachirp filterbank.

A linear time-frequency transformation correlates the signal with a family of waveforms that are concentrated in time and frequency. These waveforms are called time-frequency atoms[1][8]. In the present work we apply a gammachirp filterbank as a linear time-frequency transformation which correlates the signal with the impulse responses of gammachirp filters centred on a linear frequencies assessment in the range of audible frequencies.

2.1 The Gammachirp Filter

The gammachirp filter is a good approximation to the frequency selective behaviour of the cochlea. It is defined by the real part of the complex following expression [5]:

$$g(t) = \lambda_n t^{n-1} e^{-2\pi\alpha t} e^{j 2\pi f_0 t + j c \ln(t) + j\varphi} \quad (1)$$

with:

$$\alpha = b.ERB(f_0) \text{ and } ERB(f_0) = 24.7 + 0.108 f_0 \text{ in Hz} \quad (2)$$

Which is the equivalent rectangular bandwidth at frequency, n is the filter order, f_0 is the frequency modulation, λ_n is some normalization constant, b is a parameter defining the envelope of the gamma distribution, φ is the initial phase, ln is the natural logarithm and c is a parameter for the chirp rate. When c=0 the gammachirp function expression becomes that of the gammatone one.

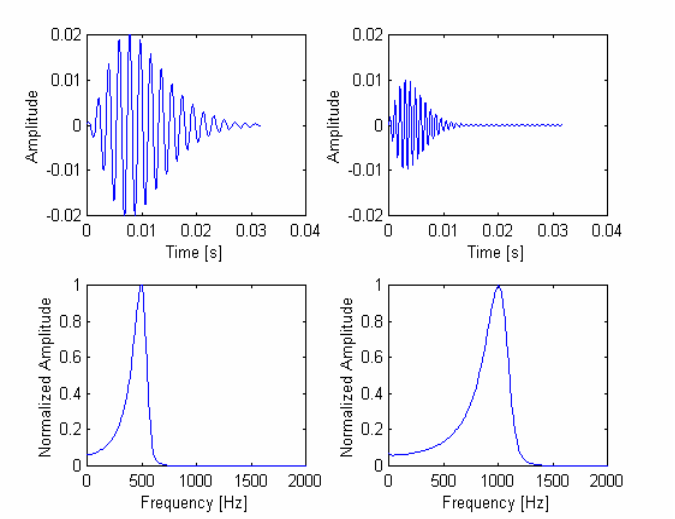

Fig.1 : Examples of impulse response and amplitude spectrum of gammachirp filters centered on 500 Hz (left side) and 1000 Hz (right side), with *b=1* and *c=-2*.

2.2 The time-frequency localization properties of the gammachirp filter

Time and frequency energy concentrations are restricted by the Heisenberg uncertainty principle. The average location of a one-dimensional normalized wave function $g \in L^2(\Re)$, (space of square - integrable functions) is [1] [8]:

$$\mu = \int_{-\infty}^{+\infty} t |g(t)|^2 dt \quad (3)$$

And the average momentum is:

$$\xi = \int_{-\infty}^{+\infty} f |G(f)|^2 df \quad (4)$$

where G is the Fourier Transform of g.

The variances around theses average values are respectively:

$$\sigma_t^2 = \int_{-\infty}^{+\infty} (t-\mu)^2 |g(t)|^2 dt \quad (5)$$

and: $$\sigma_f^2 = \int_{-\infty}^{+\infty} (f-\xi)^2 |G(f)|^2 df \quad (6)$$

In the case of the energy-normalized gammachirp function the temporal momentum and variance are respectively:

$$\mu = \frac{2n-1}{4\pi\alpha}; \sigma_t^2 = \frac{2n-1}{(4\pi\alpha)^2} \quad (7)$$

And the frequency momentum and variance are respectively:

$$\xi = f_0 + \frac{\alpha c}{n-1}; \quad \sigma_f^2 = \frac{\alpha^2}{(2n-3)}\left[1 + \frac{c^2}{(n-1)^2}\right] \quad (8)$$

The temporal and frequency variances of g satisfy the Heisenberg uncertainty principle:

$$\sigma_t \sigma_f = \frac{1}{4\pi}\sqrt{\frac{2n-1}{2n-3}\left(1 + \frac{c^2}{(n-1)^2}\right)} > \frac{1}{4\pi} \quad (9)$$

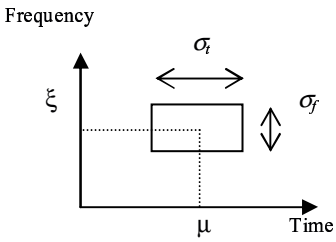

Fig.2: The time-frequency Heisenberg Box.

The time frequency resolution of g is presented in the time frequency plan (t, f) by a Heisenberg box centred at (μ, ξ), whose width along time is the temporal standard deviation σ_t and whose width along frequency is the frequency standard deviation σ_f (Fig.2). In the

case of a classical Short Time Fourier Transform (STFT) analysis, the Heisenberg boxes are independent of μ and ξ (Fig.3). But in the case of the gammachirp analysis, the time frequency spread is presented in a multiresolution analysis in which the Heisenberg boxes depend on the position of μ and ξ (Fig.4).

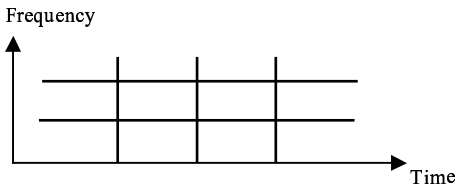

Fig.3: The time-frequency Heisenberg Boxes in the case of STFT.

2.3 The Gammachirp Spectrogram

For a speech signal x, the spectrogram $P_g x(\mu, f_0)$ measures the energy of x in the time-frequency neighbourhood of (μ, f_0) specified by the Heisenberg box of g_{μ,f_0}, which defines the time-frequency energy spread of the signal correlated by specific analysis windows around μ and f_0 in the time - frequency plane [8]. The gammachirp spectrogram $P_g x(\mu, f_0)$ has the following expression, in which g is the gammachirp impulse response centred on f_0 and translated by μ.

$$P_g x(\mu,f_0) = \left| \int_{-\infty}^{+\infty} x(t)\, g^*(t-\mu)\, e^{-j2\pi f_0 t} dt \right|^2 \qquad (10)$$

We construct then a spectrogram based on the contribution of a 512 gammachirp filters and based on the number of the analysed windows of the speech signal. Fig.6 gives the Heisenberg boxes in the case of the gammachirp spectrogram for some selected filters.

Fig.5 shows an example of gammachirp filterbank with selected filters in spectral representation.

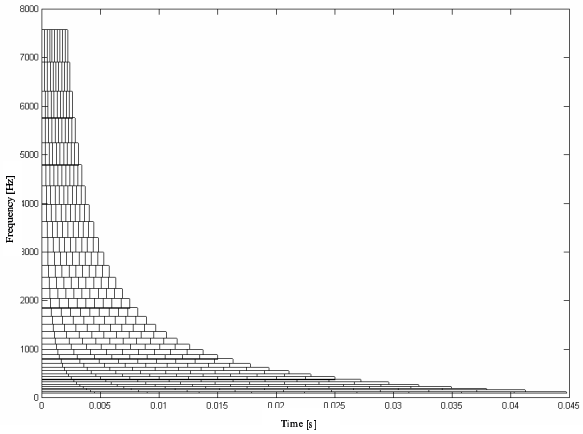

Fig. 4: Some selected Heisenberg boxes in the case of the proposed gammachirp filterbank.

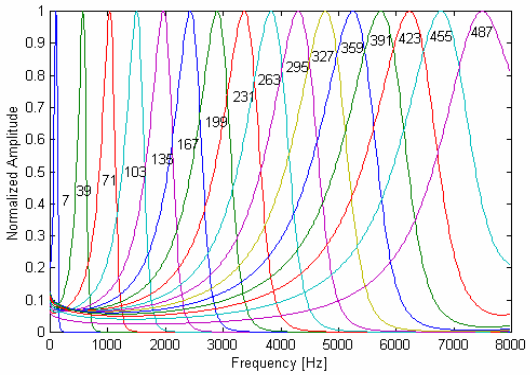

Fig.5: Example of normalized amplitude gammachirp filters. The numbers indicate the corresponding position of each filter in the filterbank.

3. The Auditory System Model

The Speech analysis technique developed in this work, is based on a simulation of the outer and middle ear filtering and the temporal and frequency masking properties of the auditory system. This proposed auditory model is based on four stages (Fig.6).

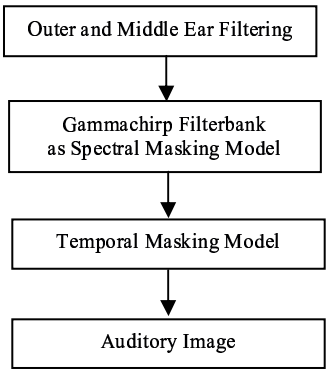

Fig. 6: The different stages of the auditory model.

3.1 The first stage: The Outer and Middle Ear Model

The first stage operates the outer and middle ear filtering which is simulated by a fixed filter W_{OM}. This filter has the following frequency response equation [12]:

$$W_{OM}(f) = -0{,}6\times 3{,}64\left(\frac{f}{1000}\right)^{-0{,}8} + 6{,}5\times e^{\left(-0.6\times\left(\left(\frac{f}{1000}\right)-3{,}3\right)^2\right)} - 10^{-3}\times\left(\frac{f}{1000}\right)^{3{,}6}\ dB \qquad (11)$$

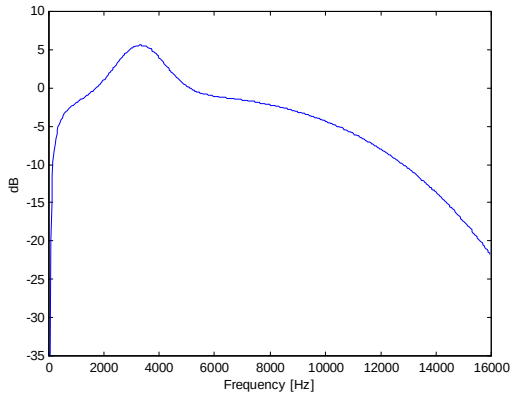

Fig.7: The frequency response of the outer and middle ear filter W_{OM}.

This equation is representative of a young listener with acute hearing. This frequency response shows that the maximum sensibility of hearing is in the range of 2000-4000 Hz (Fig.7).

3.2 The second stage: The Spectral Masking Model

The second stage operates the cochlear processing which is implemented as a gammachirp filterbank using a 512 channels. This filterbank simulates the spectral masking properties of the cochlea by the spectral characteristics of the gammachirp function centered on a linear frequencies assessment (section 2). Fig.5 gives examples of gammachirp filters centered on different frequencies.
In this example we have used the following values in Eq. (1), b=1, n=3, φ=0 and c=-2. The bandwidths of these filters depend on the location of the central frequency of the filter.

3.3 The third stage: The Temporal Masking Model

The third stage operates a temporal window model [7][9] which simulates the temporal processing of the auditory system. The models of temporal processing are of two types, functional and physiological models. The first one is based on the results given by psychoacoustic experiments and the second one is based on the actual processes observed in the periphery of the auditory pathway [7]. The functional models are usually based on some temporal weighting function. In the work of Moore et .al [9], it has been assumed that the shape of the temporal window WT is an asymmetric function. The expression of such function is :

$$W_T(t) = (1 + 2t/T_b)\ e^{(-2t/T_b)}, \quad t < 0$$
$$W_T(t) = (1 + 2t/T_a)\ e^{(-2t/T_a)}, \quad t > 0 \qquad (12)$$

Where t=0 represents the temporal centre of the window function and T_b is a parameter determining the sharpness of rising skirt of the function and T_a is a corresponding function for the descending part of the function. At moderate levels, this expression can be simplified as follow:

$$W_T = e^{(-t/T_a)} \qquad (13)$$

Recently, Plack and Oxenham [10] have proposed a revised temporal window model. It has the following expression:

$$W_T(t) = 0.975 \times e^{(t/4)} + 0.025 \times e^{(t/29)} \quad for \quad t>0$$
$$W_T(t) = e^{(-t/3,5)} \quad for \quad t<0 \qquad (14)$$

Where t is time in ms. The window function is shown in Fig.8. We have adopted this temporal window model in the present work.

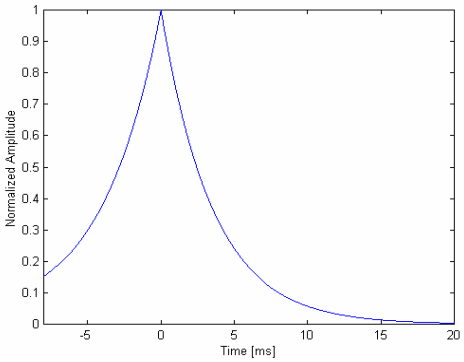

Fig.8: Temporal window function defined by Eq. (14).

3.4 The fourth stage: The Auditory Image of Sounds

The fourth stage presents a spectrographic image as the auditory image that will be transmitted to the brain. This spectrographic image is the output result of the preceding stages, the outer and middle ear filtering, the gammachirp filterbank, and the temporal masking process. Fig.9 gives an example of this type of auditory spectrogram of the sentence "she had your dark suit", by a female speaker, depicted from the Timit database. Fig.10 and Fig.11 give the corresponding STFT spectrograms of the same sentence respectively in the case of narrowband and wideband analysis. In narrowband STFT spectrogram, the harmonics of the pitch are well localised, but on the other hand the formants are not. Conversely, in wideband STFT spectrogram, only the formants can be localized. The auditory gammachirp spectrogram gives at the same time, the harmonics of the pitch in low frequencies and the formants at high frequencies.
The following Figures present a zoom of each word in this sentence. The Fig.12 which corresponds to the

utterance /she/ shows that the fricative /sh/ is spectrally located between 1kHz and 4kHz. Note the presence of the harmonics of the pitch at low frequencies of the vowel /ae/ which begins at 80 ms and the formants at upper frequencies. Fig.12 and Fig.16 show that this auditory spectrogram presents a good distinction between the fricatives 'sh' and 's' from the word 'she' and 'suit'. Furthermore, in the words 'had', 'your' and 'dark' (Fig.13, 14 and 15), the temporal masking phenomena arise by the appearance of the formants structure of the vowels before and after their real duration.

Also, the spectral masking phenomena are shown by the harmonics of the pitch in low frequencies. This can be explained by the small bandwidths of the corresponding filters. In high frequencies, the bandwidth of the filters is increasingly broad and the masking spread is being more and more expanse, which leads to the mask of the harmonics of the pitch. Yet only the formants will be preserved.

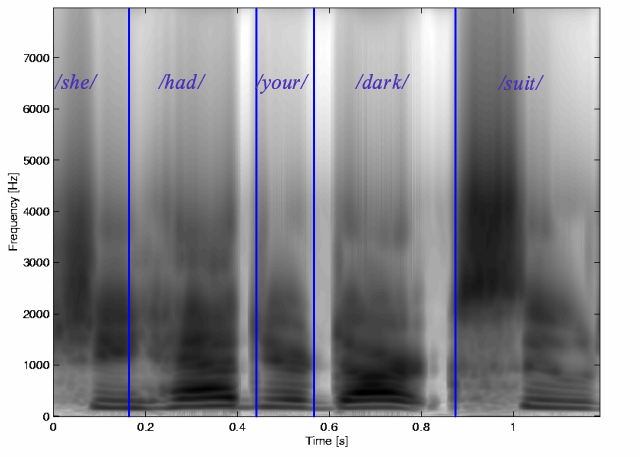

Fig.9: The Gammachirp Spectrogram of the utterance "*she had your dark suit*"

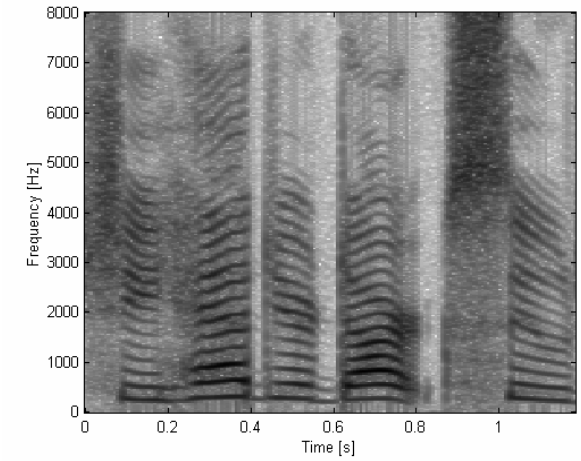

Fig.10: Narrowband STFT Spectrogram of the utterance "*she had your dark suit*"

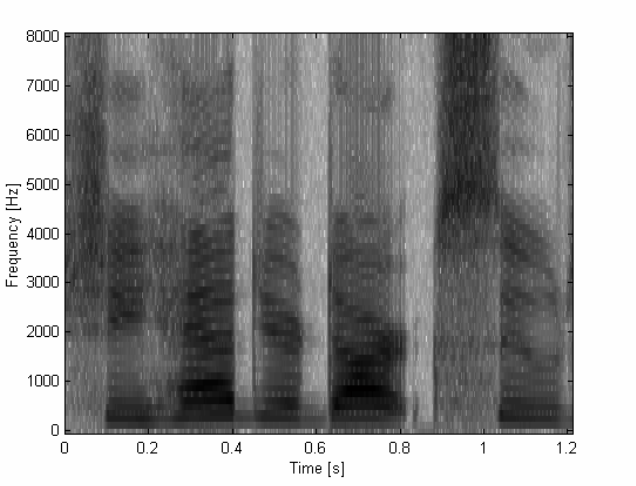

Fig.11: Wide band STFT Spectrogram of the utterance "*she had your dark suit*".

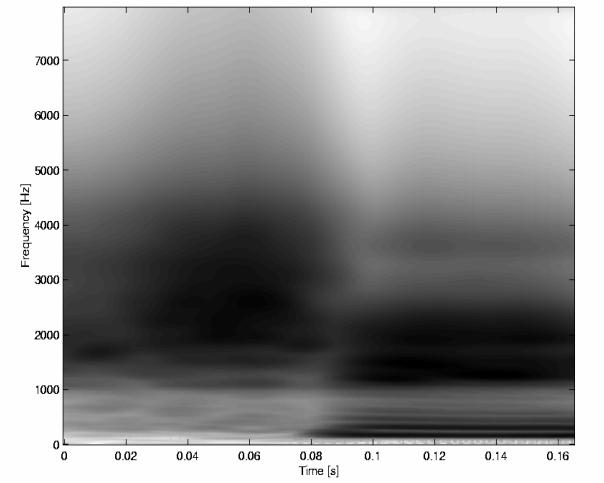

Fig.12: A zooming of the word '*she*' from the same sentence.

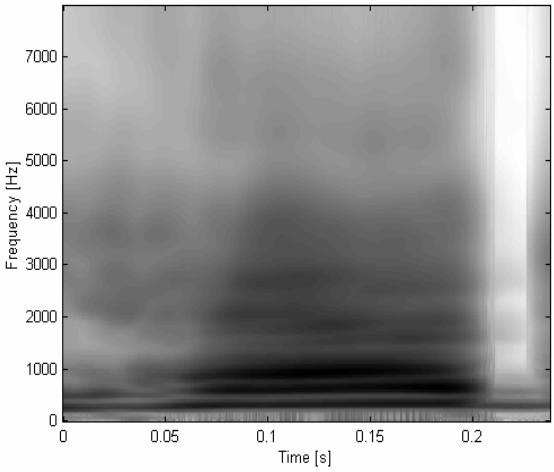

Fig.13: A zooming of the word '*had*' from the same sentence.

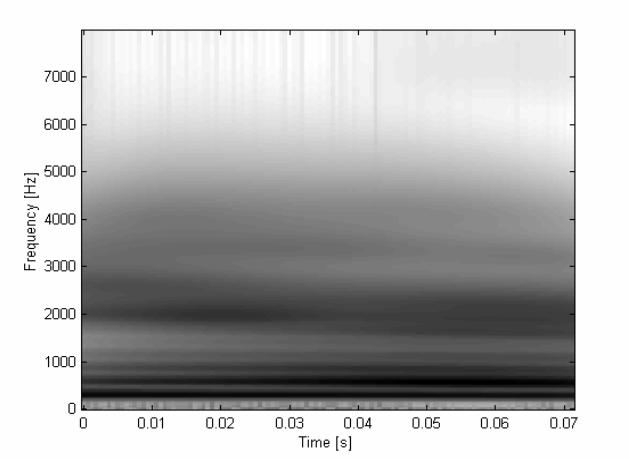

Fig.14: A zooming of the word '*your*' from the same sentence.

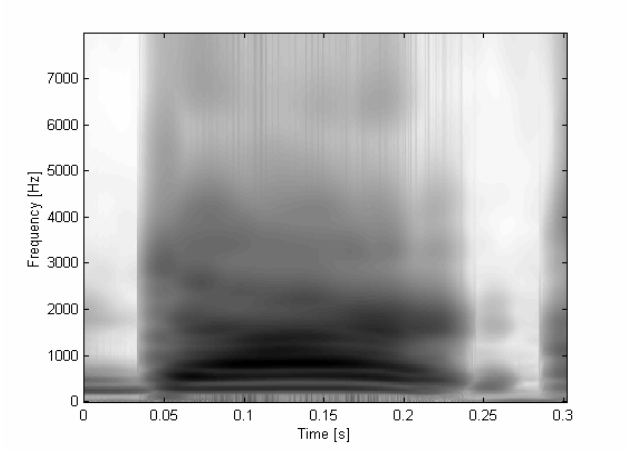

Fig.15: A zooming of the word '*dark*' from the same sentence.

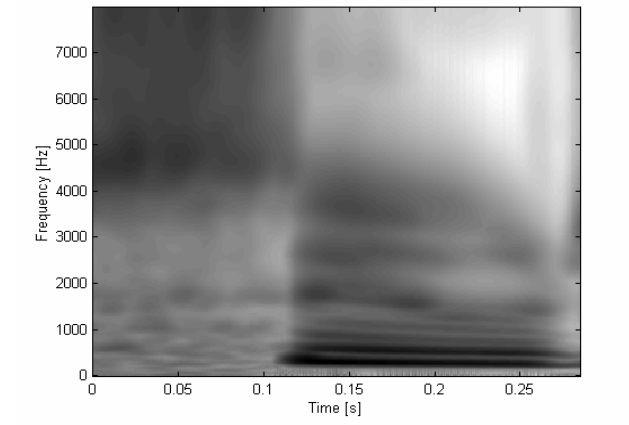

Fig.16: A zooming of the word '*suit*' from the same sentence.

4. Conclusion

In this work, we have presented a time-frequency analysis and representation of speech signal based on a modeling of the outer and middle ear filtering and temporal and frequency masking properties of the peripheral auditory system. These properties are simulated by a bank of 512 gammachirp filters. This filterbank is preceded by an outer and middle ear filtering model and followed by a time processing stage that simulated the temporal masking phenomena. The spectrograms that are based on this type of analysis highlight at the same time, the harmonics of the pitch and the structure of formants for voiced sounds and a good spectral localization of fricatives. This approach advantages this auditory spectrogram versus classical short-time Fourier spectrogram. It also presents a model of the auditory image of speech transmitted to the brain.

References:

[1] I. Daubechies, "The Wavelet Transform, Time-Frequency Localization and Signal Analysis", *IEEE Transaction on Information Theory*, Vol. 36, No. 5, September 1990, pp.961-1005.

[2] J. L. Flanagan, *Speech analysis, synthesis and perception*, second edition, Springer - Verlag, Berlin, 1972.

[3] O. Ghitza, Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition, *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 1, Part II, January 1994, pp. 115-131.

[4] T. Irino, A Gammachirp function as an optimal auditory filter with mellin transform, *IEEE ICASSP 96*, Atlanta, May 7-10, 1996, pp. 981-984.

[5] T. Irino, Patterson R. D., A compressive gammachirp auditory filter for both physiological and psychophysical data", J. Acoust. Soc. Am., Vol. 109, No. 5. Pt. 1, May 2001, pp.2008-2022.

[6] P. I. M. Johannesma, The pre-response stimulus ensemble of neurons in the cochlear nucleus, Symposium of Hearing Theory, 1972, pp. 58-69.

[7] A. Härma, Temporal masking effects: single incidents, FAMbac Technical Report, 1999.

[8] S. Mallat, *A Wavelet Tour of Signal Processing*, second edition, Academic Press, 1999.

[9] B. J. C Moore, B. R. Glasgerg, C. J. Plack and A. K. Biswas, The shape of the ear's temporal window, J. Acoust. Soc. Am., Vol. 83, pp.1102:1116.

[10] C. J. Plack, A. J. Oxenham, Basilar membrane nonlinearity and the growth of forward masking, J. Acoust. Soc. Am., Vol. 103, part.3, pp.1598-1608.

[11] S. Shamma, The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives, *Journal of Phonetics*, 1988, Vol.16, pp. 77-91.

[12] UIT-Recommendation BS.1387, *Method for Objective Measurements of Perceived Audio Quality*, December, 1998.

[13] X. Yang, K. Wang, and S. A. Shamma, *Auditory Representations of Acoustic Signals*, Technical Research Report, TR 91-16r1, Univ. of Maryland, College Park.