Implementation of a Speech Coding Strategy for Auditory Implants

Lamia Bouafif, Kais Ouni, and Noureddine Ellouze
Signal Processing Laboratory, Engineering Faculty of Tunis, Tunisia

Abstract: *In this paper we present the conception and the implementation of a speech processing interface for cochlea prosthesis. This module is based on a numerical speech processing algorithm which modelizes the infected ear and generates the stimulus signals for the cilia cells (brain). This interface uses a gammachirp filter bank constituted of 16 band pass filters based on IIR filters. The central frequencies and ERB bands are computed with Glasberg and Moore models; however stimulus signals are selected after a spectral energy analysis of each channel. The energy and the amplitude of the selected BFG filter bank outputs are quantized and transmitted to the cochlea electrodes. To validate our work, we tested it on vowels, consonants then on several words pronounced by different speakers. The results demonstrated a degree of discrimination and interferences between different sounds especially in multi speaker environment.*

Keywords: *Cochlea prosthesis, stimulus, gammatone filter, gammachirp, and speech coding.*

Received February 8, 2009; accepted May 27, 2009

1. Introduction

Most hearing deficiencies of the human auditory system affect the internal ear (cochlea) and then requires a specific cochlea implant [16]. It is based on the conversion of the vocal message in electric impulses to the stimulation of the nerve cells. This prosthesis is composed of on a speech processing module which models and replaces the internal ear and an electronic interface for transmission, wave generation and signal reception [1]. Many studies developed auditory models such as Flanagan model which is based on the physiological data measured by Bekesy, and a mathematical-computational model for the auditory mechanism. This model is divided into two parts, the middle ear and basilar membrane. In fact, [9] developed a model of an analogue electronic cochlea based on the ear physical functioning. This approach simulates the fluid-dynamic wave medium as a cascade of filters based on the observed properties of the medium. The action of the active outer hair cells is modelled by Automatic Gain Controls (AGC) which simulated the dynamic compression of the intensity range on the basilar membrane. In 1986, [10] developed a model of inner hair cells based on its physiology. This model was used by [12] to include a stage of neural transduction using the meddis hair cell. Finally, [4] has presented, a new model using a temporal speech analysis based on a gammatone and gammarchirp filter bank decomposition. The Equivalent-Rectangular Bandwidths (ERB) and their central frequencies are computed with glasberg and moore equations [11, 3].

2. The Speech Processing Algorithm

The principle of our speech processing strategy is given by Figure 1.

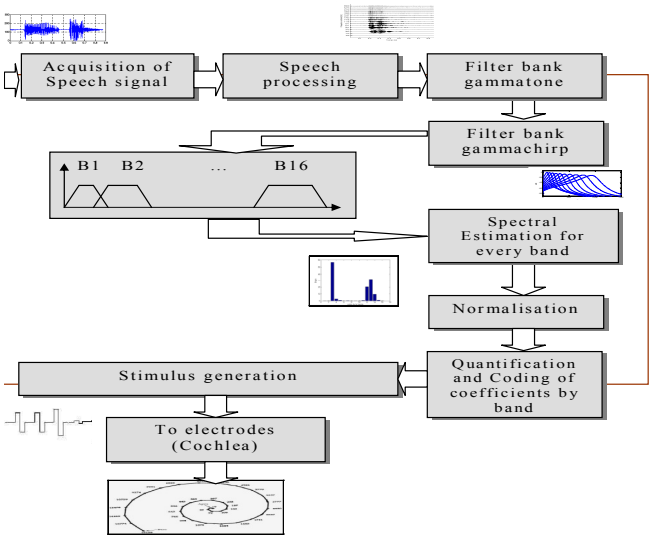


Figure 1 . The speech processing algorithm.

The voice signal processing and coding algorithms are based either on temporal or on frequency representation and modelling of the human auditory system. After signal pre-emphasis and segmentation into overlapped hamming windows, our developed algorithm uses a gammatone filter bank decomposition. Each signal of the sixteen outputs is analysed in order to compute its energy and envelope. The most significant bands (3 to 5) are selected to be coded according to CIS strategy and then transmitted to the basilar membrane electrodes.

3. Implementation

[8] Proposed a temporal model deduced from the impulse responses measured from the electric impulses of the nervous fibres of the internal ear. [5] proposed a new model of the auditory filter called gammachirp, to introduce dependence opposite the level of intensity of resonant hard working stimulus .The impulse response of the gammachirp filter is given by the following expression [6]:

$$g_c(t) = at^{n-1}\exp(-2\pi bERB(f_r)t) \times \exp(j2\pi f_r t + jc\ln t + j\phi) \quad (1)$$

where: n is a filter order, f_r is the modulation frequency of the gamma function, a is the carrier normalization parameter, c is the asymmetry coefficient of the filter, ϕ is the initial phase; BERB is the filter envelope, ERB represents the equivalent rectangular band given by [11, 14]:

$$ERB\ (fr) = 24.7 + 0.108\ fr \quad (2)$$

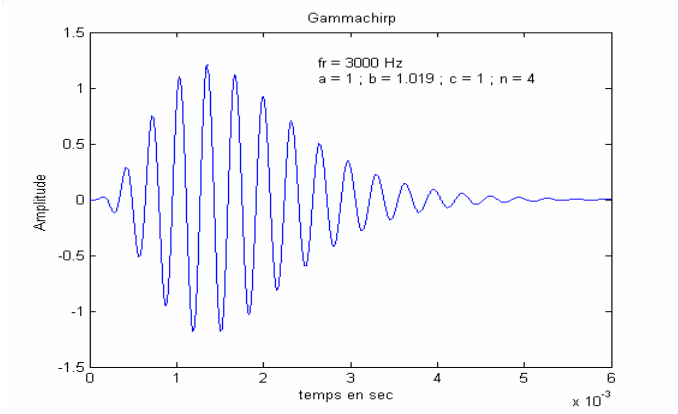


Figure 2. Temporal impulse response of the gammachirp filter (a =1, b=1.019, c=1, n=4).

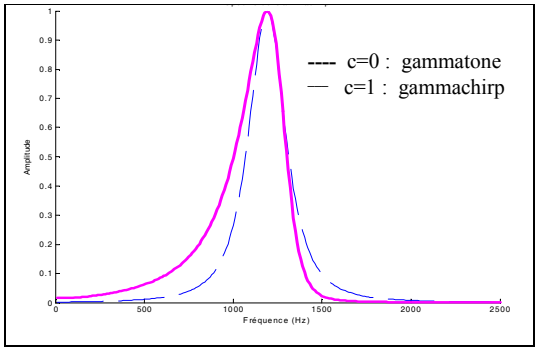


Figure 3. Frequency response of the gammachirp filter (----: gammatone :c=0 ; __: gammachirp).

The frequency response of the gammachirp filter can be expressed as:

$$G_c(f) = \frac{a|\Gamma(n+jc)|}{\Gamma(n)} \frac{\Gamma(n)}{\left|2\pi\sqrt{(bERB(f_r))^2 + (f-f_r)^2}\right|^n} e^{c\theta} \quad (3)$$

Figures 2 and 3 represent respectively the temporal impulse response and the frequency response of the gammachirp filter. The ERB is calculated in function of the central frequency (fr) according to [2]. If we use the formula of [7] and if we suppose that the signal band is between f_H and f_L with a filter recovery ratio (V) hence, the N number of filters is selected like this [15]:

$$N = \frac{9.26}{v} \ln \frac{f_H + 228\ .7}{f_L + 228\ .7} \quad (4)$$

However, the central frequencies (fr) can be premeditated by the expression [10]:

$$f_r = -\ 228\ .7 + (f_H + 228.7)e^{-\frac{vn}{9.26}} \quad (5)$$

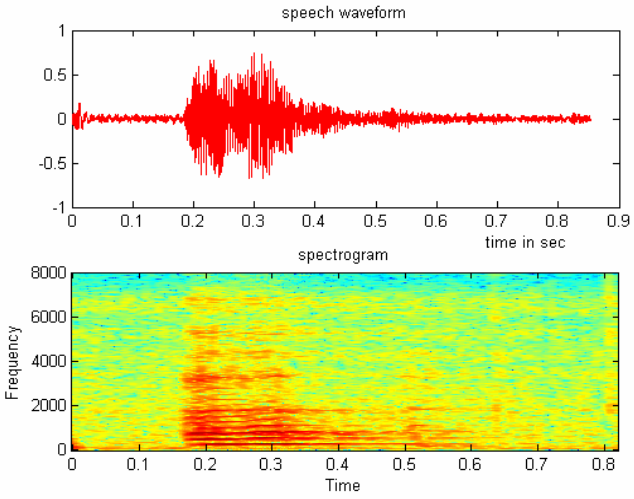


Figure 4. The speech signal and its spectrogram of the vowel “A”: female sound.

Figure 4 represents the speech signal and its spectrogram of the vowel /a/ pronounced by a female speaker. Figure 5 shows the frequency response of the gammachirp filter bank. We can observe the asymmetry of each filter resulting from the compressive behaviour of the auditory system. The waveforms of the 16 filter bank output signals are illustrated in Figure 6.

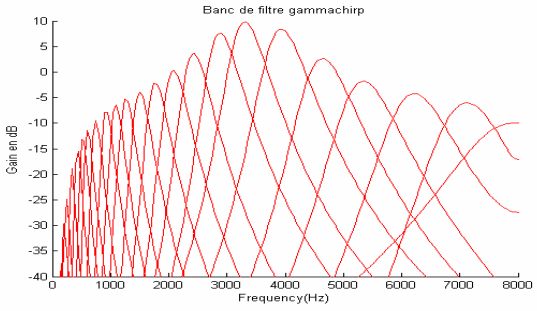


Figure 5. Frequency response of the gammachirp filterbank.

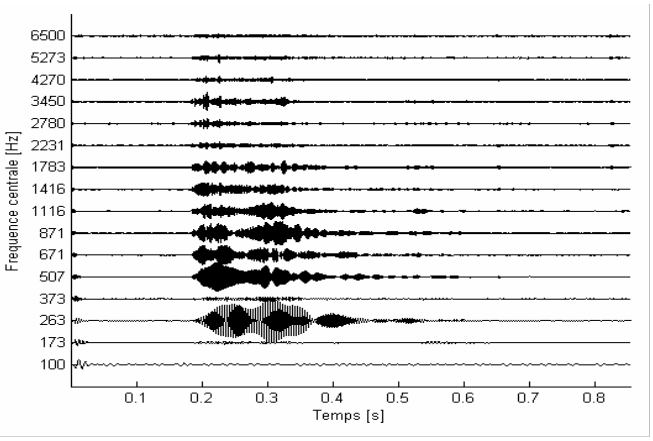


Figure 6. Temporal responses of the filterbank outputs for the vowel "a": female sound.

4. Conception of the Gammachirp Filter Bank by RII Filters

After computing the central and ERB frequencies and analysing the speech input signal through the Gammachirp FilterBank (GFB), we present in this section an approach to design the numerical filters coefficients by I.I.R (Infinite Impulse Response) filter synthesis. The advantage of this method is its simplicity of programming and the design of the equivalent analogue filter by bilinear transform. Indeed, an IIR filter is a system which can be described as:

$$y(n) = \sum_{k=1}^{N} a_k \, y(n-k) + \sum_{i=0}^{M} b_i \, x(n-i) \tag{6}$$

where *x(n)* and *y(n)* are respectively the filter input and output. The filter transfer function is:

$$H(z) = \frac{\sum_{i=0}^{M} b_i \, z^{-i}}{1 - \sum_{k=1}^{N} a_k \, z^{-k}} \tag{7}$$

4.1. Calculation of the Filters Coefficients

The main problem is therefore the calculating of the a_k and b_k coefficients of the GFB. We can use Butterworth, Chebyshev and the elliptic functions or the Bilinear transform. By using Irino model and according to relation (3), the frequency response of the GFB is [7]:

$$G_c(f) = \frac{a\Gamma(n+jc)e^{j\phi}}{\left\{2\pi bERB(f_r) + j2\pi(f - f_r)\right\}^{n+jc}} = \frac{\bar{a}}{\left\{2\pi\sqrt{\bar{b}^2 + (f - f_r)^2}.e^{j\theta}\right\}^{n+jc}} =$$

$$\bar{a} \cdot \frac{1}{\left\{2\pi\sqrt{\bar{b}^2 + (f - f_r)^2}\right\}^{n}}.e^{jn\theta} \cdot \frac{1}{\left\{2\pi\sqrt{\bar{b}^2 + (f - f_r)^2}.e^{j\theta}\right\}^{jc}}.e^{-c\theta}$$

$$= \bar{a} \cdot \left[\frac{1}{\left\{2\pi\sqrt{\bar{b}^2 + (f - f_r)^2}\right\}^{n}}.e^{-jn\theta}\right] \cdot \left[e^{c\theta}e^{-jc\ln\left\{2\pi\sqrt{\bar{b}^2 + (f - f_r)^2}\right\}}\right]$$

The last equation can be written as:

$$G_c(f) = G_T(f) \, . \, H_A(f)$$

and

$$\left|G_c(f)\right| = \left|G_T(f)\right| \, . \, \left|H_A(f)\right| = \frac{1}{\left\{2\pi\sqrt{\bar{b}^2 + (f - f_r)^2}\right\}^{n}} \, . \, e^{c\theta}$$

with

$$\theta = \arctan\frac{f - f_r}{\bar{b}}, \tag{8}$$

$$\bar{a} = a\Gamma(n + jc)e^{j\phi} \quad \text{and} \quad \bar{b} = \text{bERB(fr)}$$

As the *Z* transform of the GFB can be written as:

$$H_c(z) = \prod_{k=1}^{N} H_{Ck}(z) \, ,$$

hence:

$$H_{Ck}(z) = \frac{(1 - r_k e^{j\varphi_k} z^{-1})(1 - r_k e^{-j\varphi_k} z^{-1})}{(1 - r_k e^{j\phi_k} z^{-1})(1 - r_k e^{-j\phi_k} z^{-1})}, \tag{9}$$

where [7] $r_k = \exp\{-k.p_1 \, . \, 2\pi \, bERB(f_r)/f_s\}$, IR filter k^{th} pole module $\phi_k = 2\pi\{f_r + p_0^{k-1}.p_2 \, . \, c. \, bERB(f_r)\}/f_s$, IIR k^{th} pole argument

$$\varphi_k = 2\pi\{f_r - p_0^{k-1}.p_2 \, . \, c. \, bERB(f_r)\}/f_s \text{ , pole argument} \tag{10}$$

p_0 , p_1 *and* p_2 are of the positive coefficients,
fs : is the sampling frequency.

We can adopt the next values [7]:
$p_0 = 2$; $p_1 = 1.35 - 0.19 \, |c|$ *and* $p_2 = 0.29 - 0.004 \, |c|$. By identification of expressions 10 of the GFB with the RII expression 9, we obtain the next coefficients values: $b_0 = 1$, $a_1 = 2 \, r_1 \cos(\mu_1)$, $b_1 = -2 \, r_1 \cos \varphi_1$, $b_2 = r_1^2$ $a_2 = -r_1^2$

The r_1 , μ_1, φ_1, r_2 , μ_2, φ_2 .... values can be deduced from last expressions by puttig k =1, 2,...

5. Coding Strategy

The spectral estimation of the filter bank output signals is used to extract the stimulus parameters which are: the excited electrodes (or channels) number and their order then the stimulation speed (or spikes) deduced from the channel amplitude or envelope. These parameters once normalized, will be quantized according to an uniform quantification before coding.

5.1. The Adaptive Coding Strategy

Many coding strategies are actually used in cochlea implants such as CIS, SPEAK, ACE or ASR based either on a best temporal resolution or on a best frequency resolution of the speech signal [13]. In our study, we used an adaptive coding method based on the stimulation of a reduced number of electrodes (*N* from *M*: *N*= 3 to 5 , *M*=16) which present the more high energies as illustrated in Figure 7. The number *N* is determined by the spectral estimation of the GFB

outputs. We retain only the channels which the energy constitutes more than 90 % of the total energy. The second parameter is the stimulation speed which is adaptive with an automatic detection of the transitory speech components (consonants) and stationary components (vowels) by using a variable size analysis window. Hence, for a vowel, the analysis window is high with a slow stimulation rate, when to the consonants; the window is weak with a high stimulation rate and a small number of channels or excited electrodes (N ≤ 3). The stimulation channels orders are selected to avoid the electric interactions between adjacent bands.

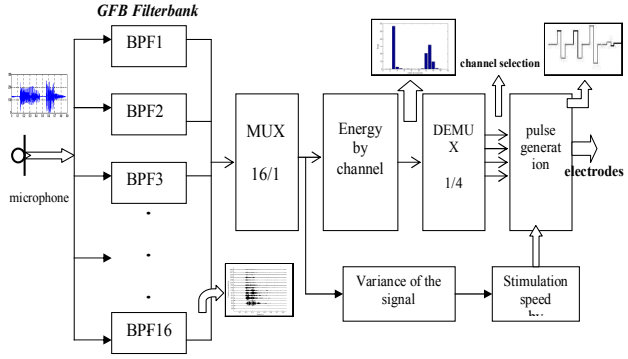


Figure 7. Coding strategy.

5.2. Stimulation Rate

The speed of the electric pulses exciting each electrode constitutes an important parameter of the cochlea implant. This cadence can be fixed or variable between the different electrodes. In our case, we use a variable stimulation rate according to two frequency bands and the variance of the analysed information. For example:

- the zone ≤ 1500 Hz which is an energetic, vocalic and melodic band, represents 40 % of the acoustical speech information and 60 % of the energy. It also contains the spectral elements generally useful for the speech perception. In this case, we use 3 to 6 stimulation channels with a rate of 400 to 500 impulses by second.
- the zone ≥ 1500 Hz where we find the fricative components, explosive and most consonant components, that represents 60 % of the acoustical speech information and 40 % of the energy. In this case, we use 3 excited channels with a stimulation rate of 1200 to 1600 impulses by second.

6. Simulation Results

We have integrated the algorithm of Figure 1 in a Matlab speech processing program for evaluating and simulation of our coding strategy. As input speech, we used sounds from TIMIT database. Figures 8 and 9 illustrate respectively the GFB filter bank temporal responses (with 16 channels) of the vowels /a/, /i/ pronounced by a female voice. Figure 10 illustrate a reconstruction of the speech signal by only three channels 3, 5 and 7. These bands are deduced from the most energetic channels as demonstrated in figure 11. We can observe that in the case of /a/ the maximum of energy is located at the 3rd channel with a variable energy distribution for the others channels. According to Figure 11, the channels 3, 5 and 7 constitute 88% of the speech energy and can therefore selected for the stimulation electrodes.

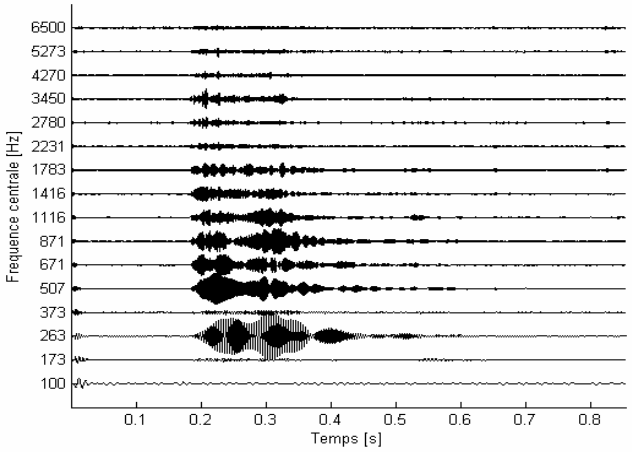


Figure 8. Filter bank outputs for the vowel /a/ (N=16 channels).

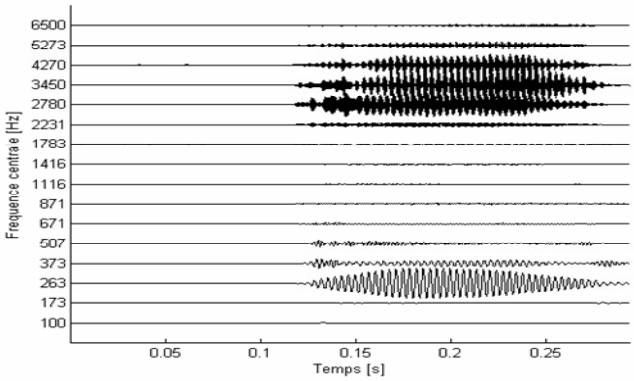


Figure 9. Filter bank outputs for the vowel /i/ (N=16).

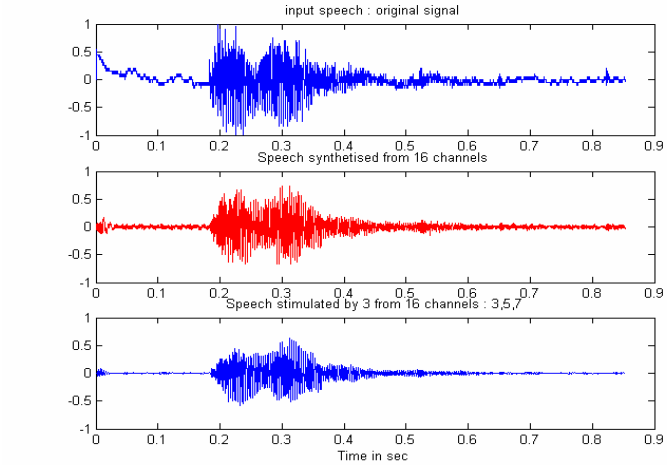


Figure 10. Speech input and reconstructed speech using 16 then 3 channels for the vowel /a/: female sound.

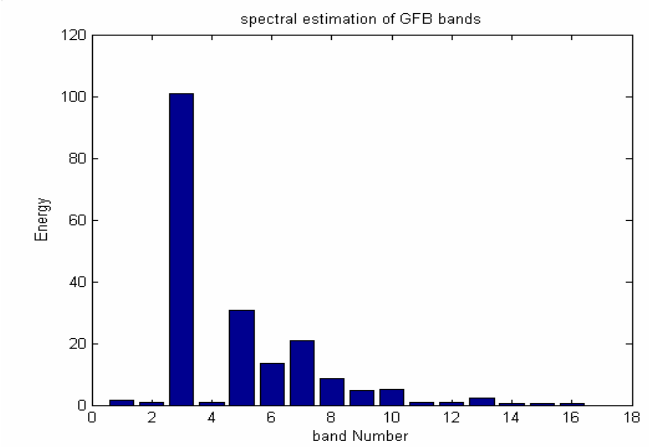


Figure 11. Spectral estimation by band for the vowel /a/: female sound.

According to our coding strategy, we can deduce that for the analysed vowel /a/, there are 3 most significant channels which are the 3rd, 5th and 7th corresponding to channels with central frequencies: 263Hz, 507 Hz, 871 Hz. As all these values are located at the energy and vocalic zone (≤ 1500 Hz), we can use 3 excited channels with stimulation rate 400 pps for the 7th electrode, 447 pps for the 5th electrode and 600 pps for the 7th electrode which was calculated in function of the three channels amplitude and variance.

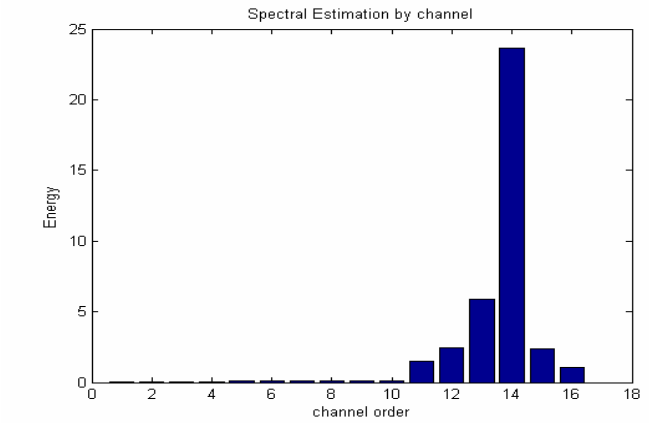


Figure 12. Spectral estimation by band for the consonant /sh/: female sound.

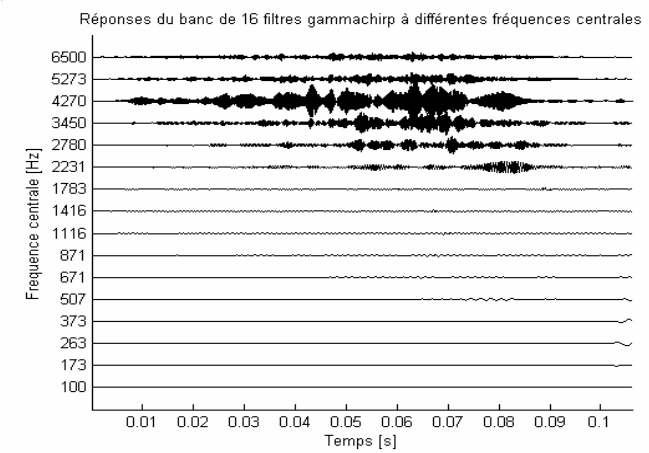
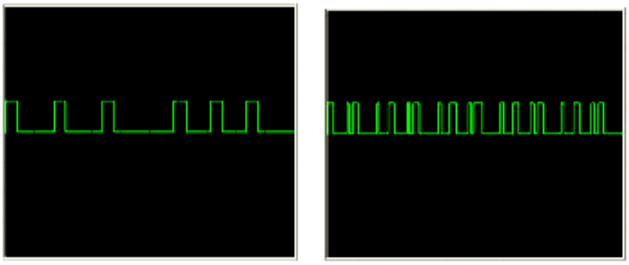


Figure 13. Temporal representation of the GFB outputs for the consonant /sh/: female sound.

According to Figures 12 and 13, we can deduce that for the analysed consonant */sh/*, there are two most significant channels which are the 13rd and 14th corresponding to channels with central frequencies: *3450Hz, 4270 Hz* . As all these values are located at the acoustic zone (≥ 1500 Hz), we can use two excited channels with stimulation rate 1200 pps for the 13th electrode, 1600 pps for the 14th electrode which was calculated in function of the channels variance.

Figure 14 illustrates two stimulation channels with different stimulation rates. The first one corresponds to the 3rd channel and corresponds to a stationary frames (vowel), however the second one refers to the 14th channel and correspond to transitory speech components (consonant, fricative , plosive,..).



Stimulation rate: 400 bps. Stimulation rate:1000 bps.

Figure 14. Stimulation rates of two channels.

7. Validation

The distances between the energy coefficients allow us to measure the separation between the stimulation parameters and to verify consequently the efficiency of our strategy. We can also observe the discrimination between the sound components in single and multi speaker environments and then compute the degree of correlation and interferences between the stimulation channels or electrodes. According to Hölder, the spectral distance is expressed by the mean quadratic norm value as :

$$d = \frac{1}{16}\left(\sum_{k=1}^{16}\left|x_k - y_k\right|^2\right) \qquad (11)$$

where x_k and y_k are the energy coefficients of the k^{th} channel.

7.1. Mean Quadratic Error

The following Figures 15 and 16 represent the result of the sixteen channel energy coefficients calculated from vowels and consonants localized in several words and pronounced by the same female speaker. We can easily observe a similar localization of the similar processed speech around 3 or 4 channels. For example, the most energy channels (which will be selected and excited) for the vowel /a/ in the words /dark/ and /had/ are the 4th,5th and 6th .

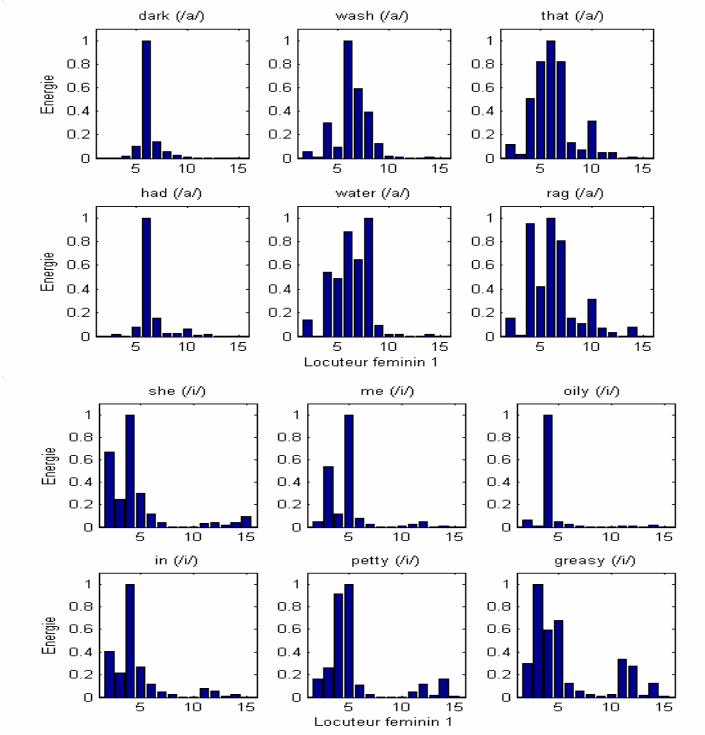


Figure 15. Representation of the 16 channels Energy parameters for the vowels /i/ and /a/ repeated by the same female speaker (in different words).

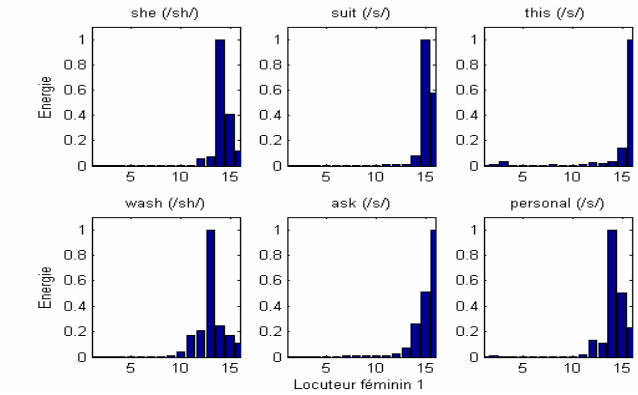


Figure 16. The 16 channels parameters of consonants (/s/, /sh/) repeated by the same female speaker.

All these results are illustrated in Table 1 on which we can observe the weak distance values between the same vowels.

with A,U or I: signify the processed vowel in the mentioned word $A_{i,j}$, $I_{i,j}$ and $U_{i,j}$: spectral energy distances of A, I and U Indice 1 : refers to the first word , 2: second word , 3: third word containing the same vowel.

7.2. Interpretation

The obtained results show that this strategy depends on the fundamental frequency. As soon as the pitch value varies from speaker to other, the distance between two identical vowels pronounced by distinct speakers is not more hopeless but remains generally the small distance in the table with the exception of some confusions between *i* and *u*. This little perturbation shows a certain overlap between certain vowels but the GFB filterbank keeps a sufficient discrimination between them in multi and inter speaker environment. The presented figures confirm these results since we see that each vowel or consonant is well localized in the selected stimulation channels.

8. Conclusions

In this study, we presented a new implementation method of speech processing and coding which is intended for cochlea implants. This strategy is based on an adaptive parameters extraction of the speech signal. These three parameters are the number of the stimulation channels (3 to 5), their order and finally their stimulation rate (number of pulses per second). The first and second parameters are chosen after a spectral energy analysis by channel of the 16 filter bank output signals. However, the last parameter is chosen in function of the envelope and amplitude signal of every stimulated channel and the vocal and acoustic information. In fact, for vocal, vowels and voiced speech, the stimulation rate is less or equal 600 pps, hence for fricatives, plosives and consonants speech frames, the stimulation rate varies between 1200 and 1600 pps in function of the signal amplitude and envelope. This technique is implemented and simulated under Matlab under several environments and speech database (TIMIT with several speakers and words).

Table 1. Spectral distances between several sounds pronounced by the same female speaker.

Same Female Speaker	A11(Dark)	A12(Wash)	A13(That)	I11(She)	I12(Me)	I13(oily)	U11(Suit)	Um12(Attitude)
A21(Had)	0.0003	0.0268	0.0843	0.1458	0.1251	0.1227	0.1498	0.1264
A22(Water)	0.1004	0.0377	0.0632	0.1595	0.1744	0.1594	0.1490	0.1657
A23(Reg)	0.0972	0.0455	0.0230	0.1150	0.1827	0.1163	0.1089	0.1228
121(in)	0.1251	0.1193	0.1351	0.0051	0.0971	0.0140	0.0287	0.0172
122(Petty)	0.1588	0.1608	0.1137	0.0484	0.0468	0.0639	0.0080	0.0697
123(greasy)	0.1702	0.1792	0.1631	0.0726	0.0483	0.1121	0.0779	0.1180
U21(Popularity)	0.1168	0.1159	0.1536	0.0286	0.1210	0.0002	0.0352	0.0006

The simulation results of the stimulation channels and their interferences in different words, demonstrated a good discrimination between these information especially for vowels, consonants, voiced and unvoiced speech frames. Besides, our adaptive strategy uses a variable stimulation channel speed with an automatic detection of the transitory components and stationary information in the processed word. This technique has the advantage to reduce the number of channels and to obtain best signal intelligibility.

References

[1] Ahmed H., Samet M., Benmessaoud M., Ghriani H., Lakhoua N., Drira M., and Mouine J., "Algorithme de Traitement de la Parole Basé sur un Banc de Filtres Numériques Programmable dédié à la Prothèse Cochléaire," *in proceedings of the International Electronic Conference Jefferson Township Education Association*, Tunis, pp. 159-163, 1998.

[2] Allen B., *Acoustical Society of America*, ASA Edition, 1995.

[3] Glasberg R. and Moore C., "Derivation of Auditory Filter Shapes from Notched-Noise Data," *Computer Journal of Hearing Research*, vol. 47, no. 2, pp. 156-160, 1990.

[4] Irino T. and Patterson D., "A Compressive Gammachirp Auditory Filter for both Physiological and Psychophysical Data," *Computer Journal of Acoustical Society of America*, vol. 109, no. 5, pp. 2008-2022, 2001.

[5] Irino T. and Patterson D., "Temporal Asymmetry in the Auditory System," *Computer Journal of Acoustical Society of America*, vol. 99, no. 4, pp. 126-129, 1997.

[6] Irino T. and Patterson D., "A Time-Domain, Level Dependent Auditory Filter: The Gammachirp," *Computer Journal of Acoustical Society of America*, vol. 101, no. 1, pp. 412-419, 1997.

[7] Irino T. and Unoki M., "An Analysis Auditory Filterbank Based on an IIR Implementation of the Gammachirp," *Computer Journal Acoustical Society of Japan,* vol. 20, no. 6, pp. 397-406, 1999.

[8] Johannesma P., "The Stimulus Response of Neurons in the Cochlear Nucleus," *Computer Journal of Symposium on Hearing Theory*, vol. 1, no. 2, pp. 58-69, 1972.

[9] Lyon F. and Mead C., "An Analog Electronic Cochlea," *Computer Journal of IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1119-1134, 1988.

[10] Meddis R., "Simulation of Mechanical to Neural Transduction in the Auditory Receiver," *Computer Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 702-711, 1986.

[11] Moore J. and Glasberg R., "Suggested Formula for Calculating Auditory-Filter Bandwidths and Excitation Patterns, *Computer Journal of Acoustical Society of America,* vol. 74, no. 3, pp. 169-173, 1983.

[12] Patterson D., Allerhand H., and Giguere C., "Time Domain Modelling of Peripheral Auditory Processing: A Modular Architecture and a Software Platform," *Computer Journal of the Acoustical Society of America*, vol. 98, no. 3, pp. 1890-1894, 1995.

[13] Roux G., *Synthèse et Réalisation d'études Cliniques sur L'implant Cochléaire*, Diplôme d'État d'Audioprothèse, Université de Rennes, 2001.

[14] Smith O. and Abel S., "Bark and ERB Bilinear Transforms," *Computer Journal of IEEE Tranactions on speech and Audio Processing*, vol. 7, no. 6, pp. 148-153, 1999.

[15] Waleed A., "Signal Processing and Acoustic Modelling of Speech Signal," *PhD Thesis*, 2002.

[16] Zwicker E. and Feldkeller R., *Psychoacoustique, l'Oreille Récepteur d'information*, Masson Press, 1981.

Lamia Bouafif obtained her engineering diploma in the field of informatics from the Science Faculty of Tunis in 1994, then her master degree in 2002 in signal processing. She is a research member of the Image and Signal Processing Laboratory of the Engineering Faculty of Tunis.

Kais Ouni is actually a professor in the field of signal processing in Engineering Faculty of Tunis ENIT.

Noureddine Ellouze is a professor at the National Engineering Institute of Tunis in signal processing. He is the director of the Image and Signal Processing Laboratory.