Computer Science Journal of Moldova, vol.**10**, no.2(29), 2002

Hints About Some Baseful but Indispensable Elements in Speech Recognition and Reconstruction

Mihaela Costin & Marius Zbancioc

Abstract

The cochlear implant (CI) is a device used to reconstruct the hearing capabilities of a person diagnosed with total cophosis. This impairment may occur after some accidents, chemotherapy etc., the person still having an intact hearing nerve. The cochlear implant has two parts: a programmable, external part, the Digital Signal Processing (DSP) device which process and transform the speech signal, and another surgically implanted part, with a certain number of electrodes (depending on brand) used to stimulate the hearing nerve.

The speech signal is fully processed in the DSP external device resulting the "coded" information on speech. This is modulated with the support of the fundamental frequency F_0 and the energy impulses are inductively sent to the hearing nerve. The correct detection of this frequency is very important, determining the manner of hearing and making the difference between a "computer" voice and a natural one.

The results are applicable not only in the medical domain, but also in the Romanian speech synthesis.

Keywords: Cochlear implant (CI), Fast Fourier Transform (FFT), Wavelet Transform (WT), Cepstrum.

1 General Considerations

As a method of first choice in the speech signal treatment, the Fast Fourier Transform for spectral analysis was used. A first attempt to determine the dominant frequencies (the formants) is made directly on the signal, the amplitude, width, and the formantic weight centre being important. The F_0 (fundamental frequency) is a characteristic of every person, and depends on the different intonations. Some persons may try to educate their voices (and search special intonations by modifying their fundamental frequencies) in order to be more eloquent in their speech. In the reconstruction of speech for cochlear implant, the signal is passed through an AGC (automatic gain control) anti-alias filter, FFT, and then it is split into 15 frequency bands (in a Mel-scale), the first being considered the fundamental frequency F_0. This is also the value used in the frequency modulation to send the signal through an antenna to the receptor placed near the ear and then, inductively, to the surgically implanted device with coils and electrodes.

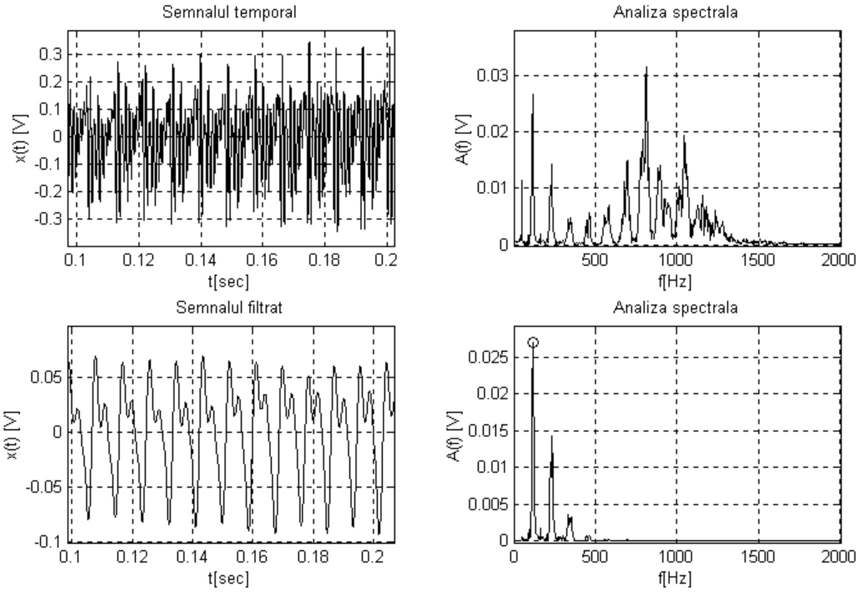


Figure 1. The fundamental frequency extraction from a filtered signal (low-pass 75-350Hz filter - the Romanian ‘a’ vowel).

The male fundamental frequency F_0 may be about 100-150 Hz, the female values are about 200-250 Hz or more and the children ones are around 300 Hz. Other formant value modifications are detected as well with respect to the fundamental frequency. For implanted patients, in order to correctly identify the voice of a speaker, this kind of details are very important. Also, in the treatment of the speech signal provided to implanted patients, it is important to select certain special features in order to correct the hearing phenomenon. For patients not performing well with the implants we will design some special neural networks to achieve this purpose [6].

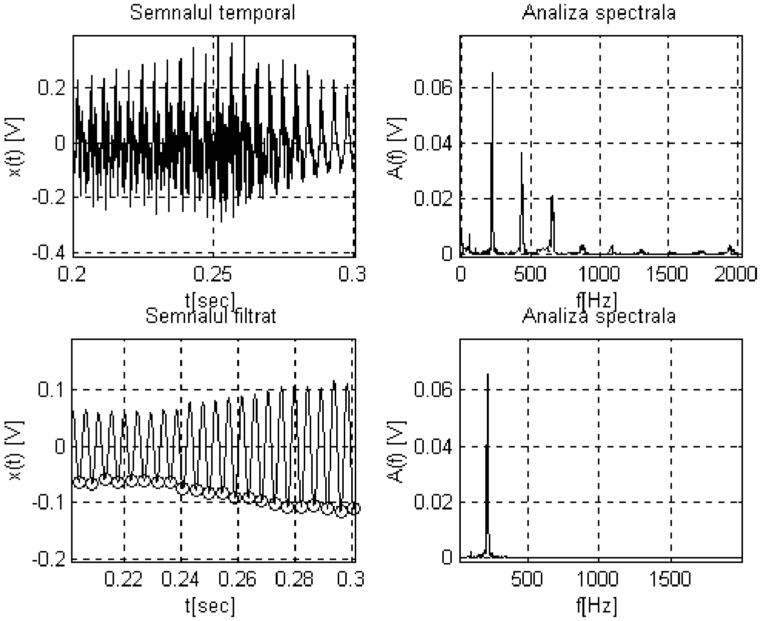


Figure 2. The fundamental frequency extraction by the mean distance computation between the local minimum vowels: Romanian ‘e’ vowel.

2 Elements in Determining the Fundamental Frequency F_0

We will detail here some methods in determining F_0 frequency that we used:

- Signal filtering with a **low-pass** filter 75-350Hz, in order to eliminate the short periods (corresponding to high frequencies) and to obtain a signal on which it is simpler to detect the T_0 period of the speech signal.

F_0 may be also extracted by the following methods:

(a) as the maximum value of the filtered speech signal;

(b) determining the T_0 period after the local *maxima* values are isolated by determining the mean fuzzy values between them;

(c) determining the T_0 period after the local *minima* values are isolated by determining the mean fuzzy values;

(d) fuzzy area computing between the zero values of the signal and the extreme value on the raising slope (or the descending one).

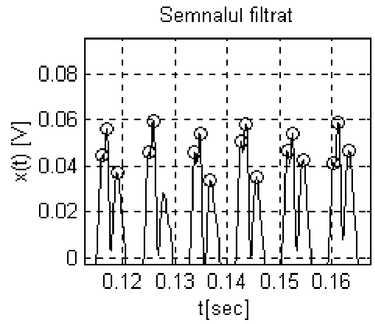


Figure 3. A particular case: F_0 detection by the help of the maximum values may produce errors.

Fig. 1 shows a male voice with $F_0 \approx 125$Hz, while in Fig. 2 a female voice was chosen with $F_0 \approx 220$Hz. In some cases, and especially for energetically pronounced vowels, the maximum values method might be difficult to implement (large amplitudes, some of them determining the unjustified neglection of some neighbours) and the minimum values method might work better.

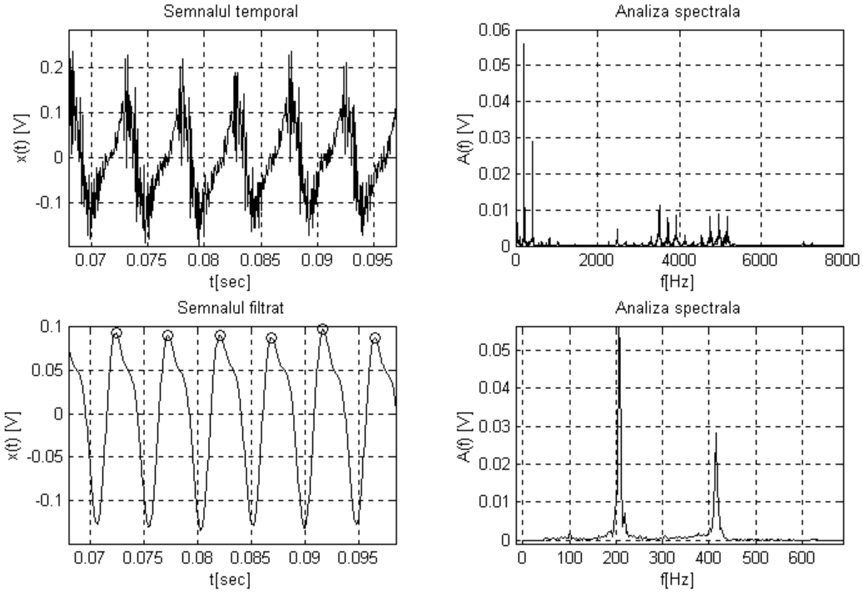


Figure 4. Fundamental frequency detection by the help of local maximum. The 'i' vowel - for a female voice.

In order to minimize errors, all the distances that overpass with a Δ a medium value A_m are eliminated. F_0 detection in the frequency domain is a more exact method if larger windows are considered. But in the time domain, T_0 can be detected in little windows as well. A little window may introduce deformations at the signal borders; we consider the important information centred. In the case of the non-vowel fones or some voiced consonants (l, m, r) it is not possible to accurately detect the T_0 period and we cannot compute F_0. The isolation of a vocalic sequence is easy to be done, due to its pseudo-periodicity, by determining the parts where the signal amplitudes sum overlaps a fixed value, while consonants are characterised by an energetic growth followed by sequences partially assimilated to noise.

- Another method to detect the fundamental frequency uses the cepstrum shape.

"**Ceps**trum" is a paraphrase from the "**spec**trum" term and "quefrency" is the term derived from the "frequency", "lifter" from "filter", etc. [1]

Cepstrum analysis is the name given to a range of techniques involving functions which can be considered as a "spectrum of the logarithmic spectrum". The first definition was in fact [1] "the power spectrum of the logarithmic power spectrum". It realises a compression of the signal passed in the "new time" or "modified time" domain - the quefrency domain. The signal is more prominent in the first part of the interval and only some maxima are emphasised in the rest of it.

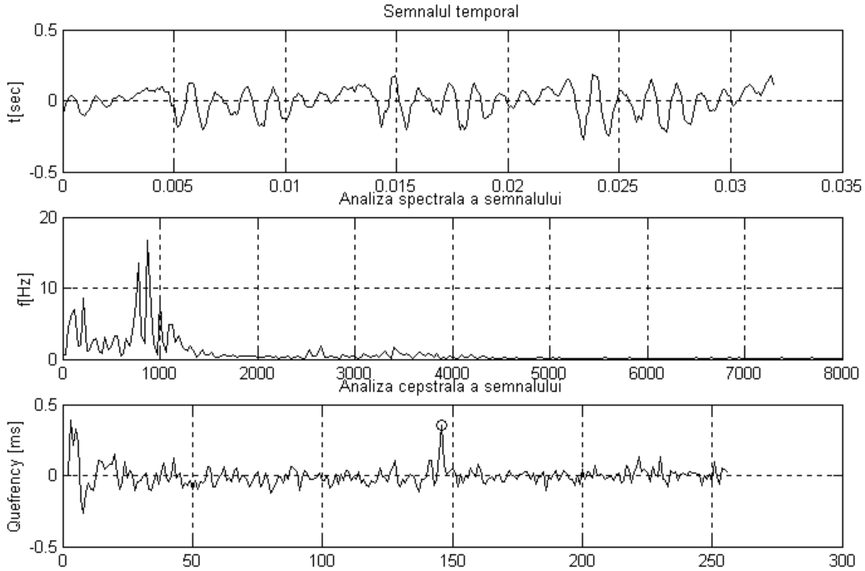


Figure 5. F_0 detection using cepstral analysis. A 32 ms window - Romanian 'a' vowel, male voice.

The distinctive feature of the cepstrum is not that it is a spectrum of a spectrum, but rather the logarithmic conversion of the original spectrum. In fact the auto-correlation function is the inverse Fourier transform of the power spectrum and can thus also be considered as a "spectrum of a spectrum" but the most common definition of the cepstrum nowadays is "the inverse Fourier transform of the logarithmic power spectrum".

The method consists of: extracting the vector, choosing the maximum value situated between the values of 50-200 Hz and then applying the rate of 16000/maximum-value to determine a fundamental frequency situated in the range of 80-320 Hz (the sampling frequency used to record the wave forms is set to f_e=16000 Hz).

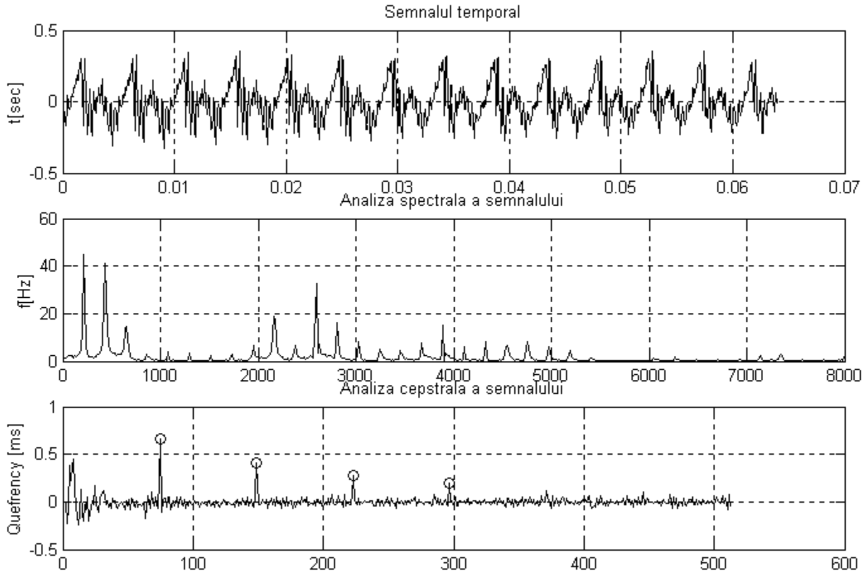


Figure 6. F_0 detection by cepstral analysis. A 64 ms window -Romanian 'e' vowel, female voice.

The advantages of using the "cepstrum" are mainly observed on small analysing windows (24-32 ms), the vector length being very small and the computing time very short as well. Neither the filter nor the T_0 period search is necessary. For large analysing windows we may observe 4-5 echoes of F_0 to almost equidistant intervals.

Another aspect to be taken into consideration is the fact that the vector obtained by the cepstral analysis has always the first values more important, imposing to ignore them when the pitch is searched.

In order to determine this maximum, a mean value has to be computed and the detected maximum has to be at least 3 - 4 times bigger than the mean value in order to be selected.

For the example in Fig. 5, $F_0 = f_e/v_{max} = 16000/146 = 109.5$Hz. Applying the other extraction methods we obtained almost the same values (99% of cases).

For the same vowel, child voice, $F_0 = f_e/v_{max} = 16000/72 = 222.2$Hz was obtained (higher than in the other cases - male, female, as expected). Fig. 6 puts into evidence the appearance in echo of the F_0 around the values of 75, 150, 225, 300Hz.

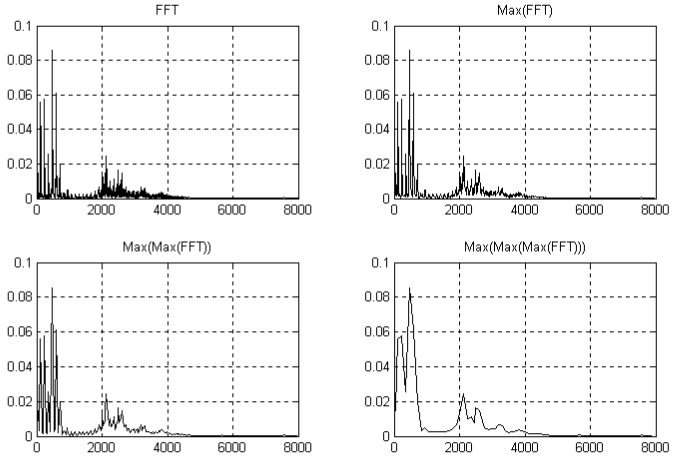


Figure 7. Cover extraction for an ‘e’ vowel - masculine voice.

A special attention have to be payed to the filtering modality of the speech signal taking into account that real filters are imperfect - there is a continuus contradiction between the two important exigences: the deep edge filters have the non-uniform band characteristic and the uniform characteristic filters have a slight descending edge (see Fig. 21).

The optimum solution of this problems consists in oversampling the signal, that is sampling with a much more important value than the Nyquist frequency:

$$f_s = (5 \ldots 10) f_N \qquad (f_N = 2 f_m, \text{ Nyquist frequency})$$

Thus, the sampling signal is containing the noisy component with a frequency much higher than the non-altered f_m. This part of the signal may be ulteriorly eliminated by digital filtering.

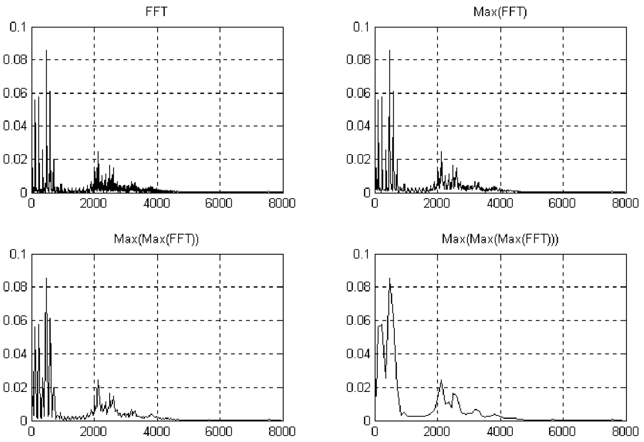


Figure 8. The extraction of the spectrum cover (envelope); 'a' vowel - masculine voice.

Taking into account these aspects that may influence the formantic values, too, [4], [5], we implemented in software a new filtering method, later detailed.

For the other formants extraction it becomes very useful the implementation of a function in order to obtain the envelope passing through the maximum values of the spectral vector. Applying this function for two or three times results in a small number of maximum values, mainly corresponding to the speech signal formants. Yet, sometimes, worthy values are eliminated and the number of function applications cannot be precisely predicted.

This is the reason to apply a domain feature extraction in order to detect these maxima values. It is preferably to establish these domains on a logarithmic scale, because the frequencies are mainly concentrated between 1000-1500 Hz. If we apply this algorithm many times so that finally the maxima number decreases under a certain value, we have to stop the algorithm in order not to loose important values.

By this way we tried to eliminate the cases where, after applying the function more times, on some intervals we have a large number of maxima and on the other hand, eliminating more essential values, a smooth and irrelevant curve is obtained.

Given the non-uniform amplitude distribution on a vocal signal spectrum, we may define application limits of this method, specific to every interval that has higher values for lower frequencies and lower values for higher frequencies.

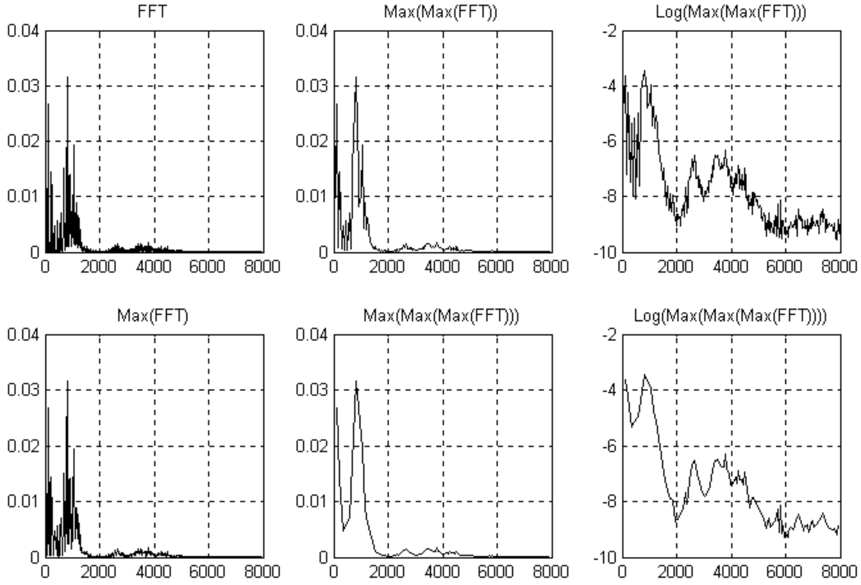


Figure 9. Applying the log function to the repeated maximum on the spectrum, amplifies the values for formants frequency detection.

This problem, regarding the smaller spectrum amplitudes after 2000 Hz, may be solved applying more times the log function. Thus, less evident formant values will be stressed, and it will be possible to extract them more precisely.

Obviously, by the exponential function we may return to the original function form.

The F_1 and F_2 formant extraction (the frequency where is situated the maximum amplitude value of the central frequency and its weight) may conduct to a vowel classification. (It is well known the F_1, F_2 vowel triangle.) Extracting the other formants may conduct to a more detailed phonemes classification.

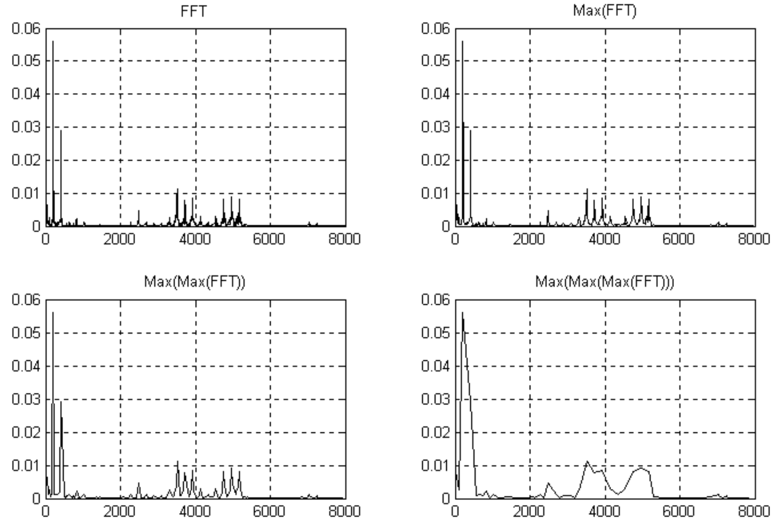


Figure 10. The extraction of the spectrum cover (envelope); ‘i’ vowel - female voice.

3 Frequency bands for cochlear implants

In this study we try to find some special methods to be applied in cochlear implants (see a block diagram in Fig. 22), due to the 15 bands in use in the studied model, so extracting formants become less important.

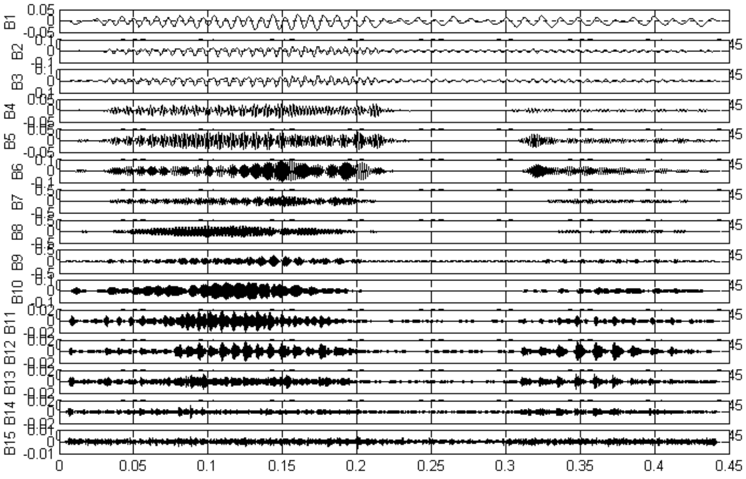


Figure 11. The speech signal is split in 15 frequency bands.

Dividing the speech signal in 15 frequencies bands using a filter, the electrodes will be or not stimulated depending on the the energy detected on the corresponding bands [6].

The 15 frequency band limits are: 0, 125, 250, 375, 500, 625, 750, 875, 1000, 1250, 1500, 1875, 2375, 3125, 4500, 6375 Hz. By automatic detection [6] we can stress some important bands, as resulted from neural network simulation. The reconstructed signal can be listened.

If a filter is used in order to extract a certain band (the Butterworth filter for example) this will modify the edge [4, 5], and differences may be observed in the signal reconstruction.

In order to eliminate this problem, instead of using a filter, which affects the edge information (very important in the CI case), we may programme a filter to extract the signal directly from the spectrum, setting to zero all the values situated outside the specified interval.

After the Inverse FFT (IFFT), we obtain by reconstruction an identical signal to the original one.

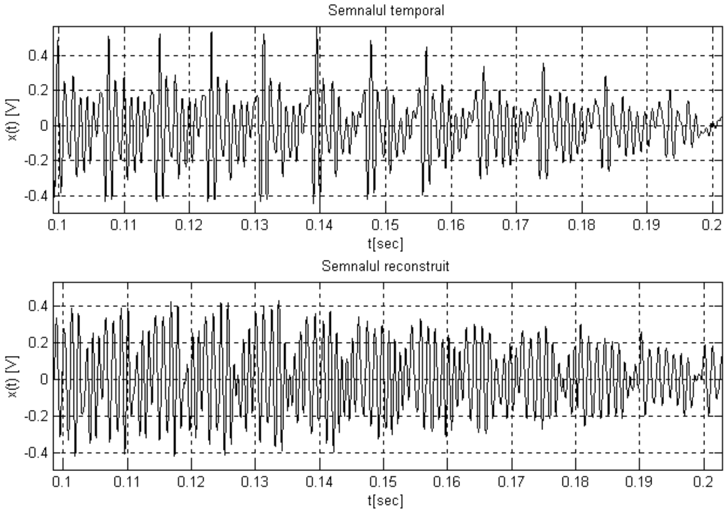


Figure 12. Differences between the original signal and the reconstructed one in the conditions of a Butterworth filter band selection - ‘aba’ phoneme.

For example, Figs. 12 and 13 show the differences between the original signal and the reconstructed one, in the conditions of a classical filter extracting the frequency bands.

The same considerations are valuable in the case of extracting the maximum values corresponding to each frequency band, and this first vector may be used in training a neural network. Next to the significant values (formants), we may have less important values. In this case a pre-emphasizing filter is necessary in order to stress the high frequencies that have small amplitudes but are, however, important.

In Fig. 15, even if the spectral values for the frequencies over 1500 Hz are almost near to zero, the clear presence of some formants in these intervals may be as important as the maximum values corresponding

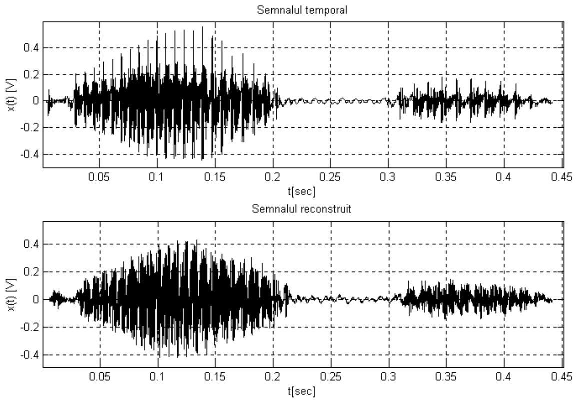


Figure 13. The amplitude modification of the reconstructed signal in the case of a filter application.

to the low frequencies, and even more.

The input training vector may be obtained apllying the cepstrum, as well, and in this case, as it has been seen before, the important values are situated in the first part of the cepstral signal.

Discrete Cosine Transform (DCT) gives good results and presents the advantage (over the FFT) of an easier extraction of the more important peaks.

Even if the vector that is obtained is almost identic with the spectral analysis one, the process of maximum values isolation is simpler.

In the two Figs. 15 and 16 we may observe the fact that the two consonants have only low frequencies, less than 1600 Hz, after this value the high frequencies entirely missing.

As observed, the graphs scale is normalised, and a multiplication by $f_e/2$ (in our case 8000) has to be applied in order to obtain the frequency value.

For the 'e' vowel (Figs. 17 and 18) uttered by the same person, but with a more energetic voice, two important modifications may be observed: F_0 (the first significant peak) is slightly shifted and F_1 grows in amplitude at least four times, the other values having almost the

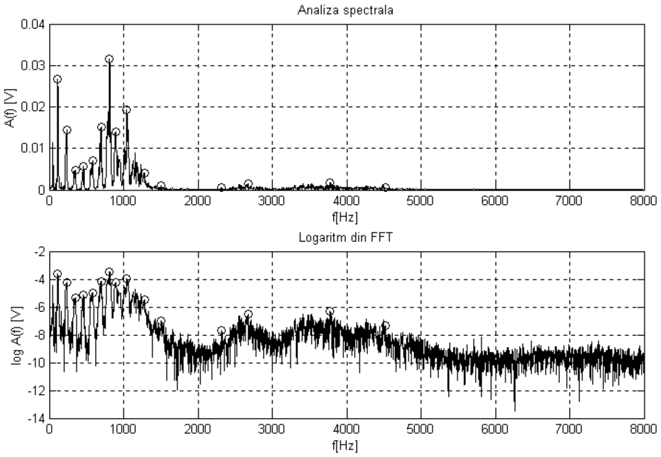


Figure 14. The maximum frequencies extraction on the 15 bands.

same size.

If, for an energetic vowel (phoneme) pronunciation, or a different intonation, we may observe such important transformations, it is obvious that this fact is very important for the language prosodia, the stress on the beginning, middle or end of a word, or a sentence being different as structure.

Compared to the 'e' vowel, that presents only 3 important formants,

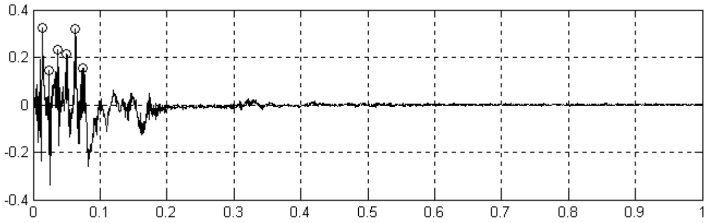


Figure 15. The Discrete Cosine Transform for the 'g' consonant.

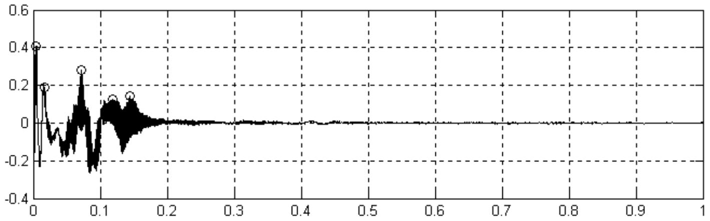


Figure 16. The Discrete Cosine Transform for the 'p' consonant.

the 'a' vowel permits almost every time the identification of 6-10 peaks, the last of them, situated between 750 Hz-1500 Hz, being more energetically rich, carrying more energy.

The maximum values characteristics, corresponding to the high frequency, amplitude, energy density, on each band, may constitute a training set for a neural network, too.

4 Discussion

Regarding classical existing methods for F_0 detection, several principal categories are prominent in the last decades. Their utilization depends on the task that have to be solved. We may mention the first attempts of physiological measurements and detection of larynges vibrations dur-

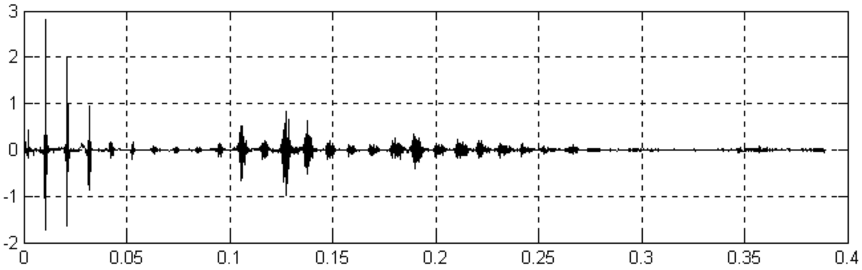


Figure 17. The DCT for the 'e' vowel - female voice.

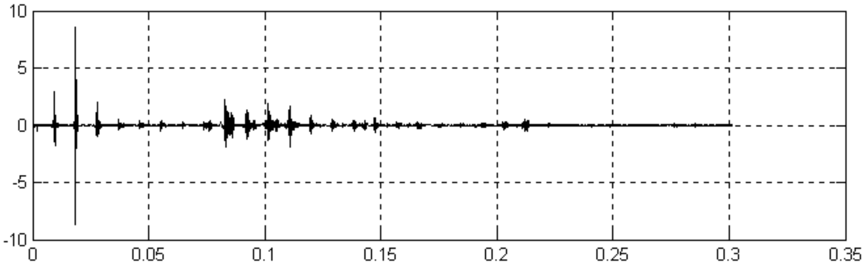


Figure 18. The DCT ‘e’ vowel, energetic pronunciation, female voice.

ing the pronunciation (4.1), measurements made in the temporal (time) domain (4.2) and in the frequency domain (4.3), etc.

4.1 Physiological attempts

Larynges movement frequency may considerably vary with phonation. Extreme variations may be sometimes observed (transitions from 100 to 400 Hz) inside only two or three cycles. The method is very imprecise (even successive cycles may present important variations reported to the mean value). Video recording presents also estimation errors and is not consequent with speech. Cycle edges detection is subject to errors, too. Another difficulty to reduce error is due to the physiological reflex phenomenon on the vocal tract.

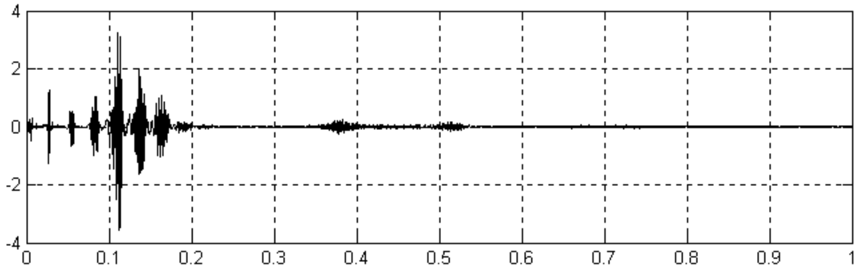


Figure 19. The DCT for the “a” vowel - child voice.

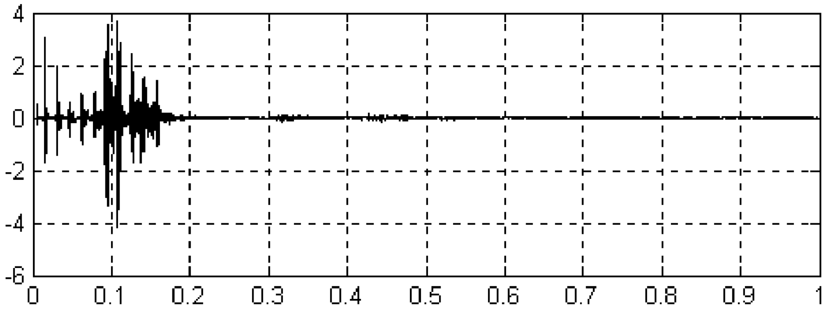


Figure 20. The DCT for the “a” vowel - masculine voice.

4.2 F_0 detection from the time domain

The zero-crossing classical method also presents a variability inconvenient. Low-pass filters introduce their own influence on the filtered signal (its characteristics have to be appropriately chosen in order to eliminate the harmonic undesirable components). Different conditions and proportions between F_0, the signal harmonics and the filter characteristics have been proposed [7].

Classical filtering impose phase differences on the signal harmonics of the output signal, too. Complex microprocessor systems and software may correct some phase differences effects, dynamic variation in filters depending on frequency (problem characterising the filter banks).

A special importance has the F_0 detection by the autocorrelation method. A 10-50 ms window is sliding on the signal giving good results when the signal is not too much varying as form, from one period to the other [8].

4.3 Spectral analysis

Instead of detecting F_0 by observing cycle by cycle a filtered signal, the fundamental frequency is detected from the information found inside the harmonics of the voiced speech. A detailed description is given in [8]. Our improvements concern expecially this spectral methods.

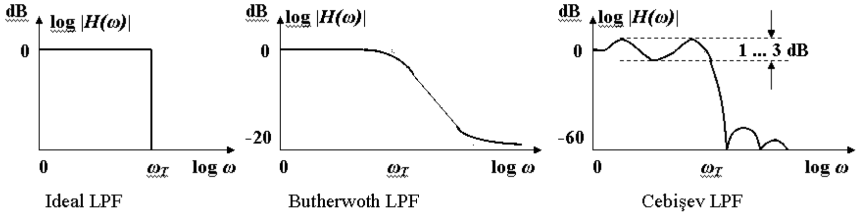
On-line results detection proves to be generally better.

5 Conclusions

After several sets of experiments and simulations we conclude that:

- the methods of determining the fundamental frequency F_0 by minimum values detection, and the cepstrum one, are the most precise, reliable and economic, in order to automatically compute it.

It is a constraint to accurately detect it because of its importance in the signal reconstruction by the help of the special cochlear implant device. This improves also patient understanding, the capacity of speaker identification, and can be applied in language prosodic aspects, too.



Ideal LPF

Butherwoth LPF

Cebişev LPF

Figure 21. Transfer characteristics for some anti-alias LP filters.

References

[1] Oppenheim, A.V. & Schafer, R.W. (1975). *Digital Signal Processing.* Prentice-Hall, N.J.

[2] Rabiner, L., Bing Huang Juang (1993). *Fundamentals of speech recognition.* Englewood Cliffs, N.J.

[3] Ghitza, O. (1991) *Auditory Nerve Representation as a Basis for Speech Processing* in "Advances in Speech Signal Processing", S. Furui and M. Sondhi, Marcel Dekker, N.Y., pp. 453-485.

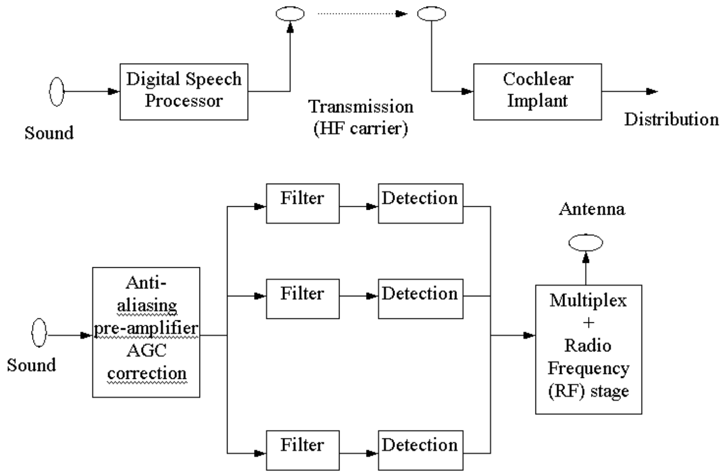


Figure 22. The block diagram of a cochlear implant device (top) and of a speech DSP (down)

[4] Strik, H. *The Effect of Low-PassFiltering on estimated voice source parameters.* University of Nijmegen, Dept. of Language and Speech, http://lands.let.kun.nl.

[5] Strik, H. and Boves, L. *Authomatic Estimation of Voice Source Parameters.* http://lands.let.kun.nl.

[6] Costin, Mihaela, Zbancioc, M., Ciobanu, A., Vachon, C.B. (2002) *Some Attempts in Improving Cochlear Implanted Patients Performances: Modeling and Automatic Methods*, Proceedings of IPMU, Annecy, France (accepted for publication).

[7] Mckinney, N.P. (1965) *Laryngeal frequency analysis for linguistic research.* Comm. Sc. Lab. Univ. of Michigan, Ann Arbor, n.14.

[8] (1989) *CALLIOPE, la parole et son traitement automatique.* Masson et CNET-ENST, Paris.

[9] Rabiner, L.R., Schafer, R.W. (1978) *Digital Processing of Speech Signal.* Prentice-Hall Inc., Englewood Clifford.

[10] Rowden, C. (1991)*Speech Processing.* McGraw - Hill Book Company.

[11] Tatham, M. (1998) *Teaching Notes* (speech.essex.ac.uk).

[12] O'Shaughnessy, D.O. (1987) *Speech Communication Human and Machine*, INRS-Telecom.

Mihaela Costin & Marius Zbancioc Received April 29, 2002

Mihaela Costin & Marius Zbancioc
Institutul de Informatica Teoretica, Academia Romana - Filiala Iasi,
B-dul Carol I nr. 8, 6600 Iasi, Romania