WAVELET SPEECH PROCESSING FOR COCHLEAR IMPLANTS: A PRELIMINARY STUDY

G. Tognola*, G. Baselli**, M. Parazzini*, A. Paglialonga*,**, F. Grandori*

*Institute of Biomedical Engineering CNR, c/o Polytechnic of Milan, Milan, Italy
**Department of Biomedical Engineering, Polytechnic of Milan, Milan, Italy

gabriella.tognola@polimi.it

Abstract: An innovative approach is proposed for speech processing in cochlear implants (CI) in order to overcome the limitations that traditional strategies demonstrate in consonant coding. Instead of the usual filterbank spectral analysis, this new method decomposes the speech signal by means of the wavelet transform (WT). Preliminary analyses have been conducted to study the behaviour of the WT algorithm when submitted to the operative conditions imposed by the CI context. Results showed that the WT is a robust transform even in such a restrictive application as CI framework. Also it showed that a speech processing strategy based on WT is able, unlike traditional methods, to efficiently capture extremely rapid transitions in speech (i.e., plosive consonants). This WT speech processing strategy has then been optimized for the application in CIs and finally compared with traditional algorithms. Results led to the evidence that, with respect to filterbank methods, WT strategy better preserves speech features, giving a coding that retains most of the information present in the original signal. Thus it would be possible, on a real implant, to obtain a stimulation pattern that accurately reproduces the signal hence facilitating the patient perception and comprehension of speech.

Introduction

In patients affected by profound sensorineural hearing loss the cochlear implant (CI) is the only device that can restore hearing sensations. It detects acoustic signals from the surroundings and converts them (by means of the *speech processor*) in a spatio-temporal pattern of electric impulses in order to stimulate the fibers of the acoustic nerve, according to the tonotopic distribution of tuning frequencies along the cochlea [1]. It is clear that a great part of patient's performance is due to the speech processing strategy, whose efficiency is related to its capability to mimic the function of an healthy cochlea.

Speech processing strategies currently used in CIs all develop from a common root: the decomposition of the incoming signal into spectral channels by means of a filterbank. These strategies provide good levels of performance in implanted patients, even if it can be observed a great variability among subjects and, on average, the scores obtained in consonant recognition tasks are quite low [2]. This fact suggests that improving the consonant representation given by the speech processor can increase the quality of patient's speech comprehension. The present work attempts to achieve this improvement by introducing the wavelet transform (WT) into the processing path of a CI.

The difficulty of CI users to recognize consonants is related to the fundamental characteristic of traditional processing strategies, i.e. the bank of band-pass filters. These numeric filters are, because of their non ideality, inadequate to localize small patterns in time while the WT performs a multiresolution analysis providing a highly accurate decomposition of transient signals in the time-frequency joint domain. The numerous applications of WT to speech analysis (e.g., [3], [4]) confirm that this tool is actually suitable for CIs processing and can perform a precise coding of the rapid and short patterns that characterize some consonants. In particular, the suitability of the WT in processing speech seems to be intrinsically related to the fact that the cochlea itself behaves as a parallel bank of band-pass filters: the cochlear transfer function between the input sound pressure of a given frequency and the output basilar membrane displacement looks like a filter with a constant *Q-factor*. In other words, the outputs of the cochlear acoustic filters can be viewed as produced by the WT of the input stimulus.

It should be said that there haven't been made, at present, real efforts aimed at including the WT in the context of CIs. In our study, differently from previous works by Yao and Zang (who theoretically built a transform that models cochlear properties [5], [6]) the problem is considered from an innovative point of view, focusing on the applicability of WT to real implants and trying to insert the algorithm in a realistic framework. In the present work, with reference to the *Nucleus*® *24* system by *Cochlear Corporation* [7], the conditions under which the WT can be introduced in the flow diagram of a real implant are defined and the consequent WT degradation is introduced and tested. This adaptation necessarily brings the wavelet analysis far from its theoretical description and implies, as a consequence, an information loss from the original loss-less WT. It has been chosen to start this preliminary study from the simplest implementation of the transform, i.e. the discrete wavelet transform (DWT)

ISSN: 1727-1983 © 2005 IFMBE

[8], thus concentrating the investigation on the behaviour of the multiresolution analysis in the new context of CIs, disregarding at the moment any consideration about mathematical complexity. Once the new WT speech processing strategy has been designed, the aims of this study are: first to optimize its parameters for the particular application, then to test its performance on speech signals (in particular on consonants) and finally to make a comparison with the strategies currently used in CIs. For this last purpose two filterbank strategies have been considered: the Continuous Interleaved Sampling (CIS) [9], and the Advanced Combination Encoder (ACE) [10], both currently used in CI speech processors.

As it will be described in detail on the next section, the work is composed by three different experiments, to put in evidence different particular aspects of the new strategy.

Materials and Methods

The incoming signal was processed according to each of the abovementioned strategies (the traditional CIS and ACE algorithms and the new DWT strategy), thus deriving the electrical stimulation pattern. For evaluation purposes, acoustic signals were reconstructed starting from the electrode activation map in order to simulate the patient's sensations on normal-hearing subjects and to measure the distance of these acoustic simulations from the original speech signal.

Test material. The speech material used was a subset of 16 consonants taken from the Iowa consonant test in the context /aCa/, spoken by a single male speaker without noise and sampled at 16 kHz.

Signal Processing. In each strategy a pre-emphasis filter was applied directly to the original signal. In particular a 1st order Butterworth high-pass filter was used, with a cut-off frequency of 1200 Hz [11].

The filterbank strategies (i.e. CIS and ACE) were implemented by means of a bank of 6th order Butterworth band-pass filters, whose center frequencies reproduce those of the *Nucleus® 24* system, that provides 22 electrodes for stimulation. For the ACE strategy (that selects, on each frame, only a subset of stimulation electrodes on the entire set of available channels), 22 filters were used, while for the CIS only 8 (fixed) channels were used, so only a subset of the electrodes is involved in the stimulation. On each channel, envelopes were extracted by full-wave rectification and low-pass filtering (2nd order Butterworth) with a 400 Hz cut-off frequency [10]. According to the envelope amplitude, for the ACE strategy only the 8 channels with higher magnitude were used for each stimulation cycle, discarding the others. Finally the envelope amplitudes were compressed according to the non-linear mapping function (LGF, Loudness Growth Function) to obtain a map of normalized impulse amplitudes that can be fitted to the patient's electrical parameters. A down-sampling stage must be inserted in the path in order to obtain, for the stimulation sequence, a sampling frequency equal to the channel stimulation rate (1000 pps) typically used in clinical practice. The acoustic simulations were synthesized according to the following approach: the envelope amplitudes (both before and after non-linear compression) were used to modulate a white noise, band-limited on each channel by the same filter used in the spectral analysis stage; these modulated noise-bands were finally summed and low-pass filtered at 4 kHz [12].

The DWT strategy processed the speech signal through the following steps: first of all an high-pass pre-emphasis filter was applied, as mentioned earlier, then the signal was analyzed frame by frame with a sliding window of 128 samples, and on each frame the dyadic DWT decomposition algorithm was implemented, thus generating 8 spectral channels. The wavelet time-scale decomposition structure obtained was characterized by different sampling frequencies for the different channels. Since it is necessary to generate a final stimulation map sampled at the electrical stimulation rate (i.e., 1000 pps), specific criteria were applied for interpolating frequencies lower then 1 kHz and for down-sampling the higher. From now on the path followed by the DWT strategy was the same as described before: the LGF was applied directly on the decomposition coefficients. The acoustic simulations were synthesized reconstructing the signal by applying the inverse transform on the coefficients, both before and after the LGF.

Preliminary evaluation of DWT degradation: The degradation of signal reconstructed by IDWT after parameter reduction of DWT was evaluated. The test speech signal was analyzed with temporal sliding windows of variable length and, on the set of channels obtained with the decomposition, a selection was made neglecting the smaller magnitude channels. This experiment examined the loss in performance with respect to the ideal condition, after a progressive elimination of spectral bands: the cross-correlation index between the reconstructed signal and the original was measured.

Signal processing of standard transients: The second experiment focused on the capability of the strategies to localize rapid and short temporal patterns. The three speech processing methods were tested on particular input signals, the step signal and the ideal impulse, in order to compare the speed of response of the different algorithms to rapid transitions. The I/O functions were directly compared and their duration was measured in each case.

Speech processing on test material: The last experiment made a realistic comparison between the new, optimized, DWT strategy and the traditional ones. The algorithms were tested on the same, standard, speech material and the reconstructed signals were compared to the original waveforms, even in this case by means of the cross correlation index.

Results

Table 1 shows the results of the first experiment for two different frame lengths: 1 ms and 5 ms. The ideal condition of perfect reconstruction of the original signal is indicated by a cross correlation index equal to 100%, and the two different frame lengths considered lead to two different total numbers of channels, five and seven respectively. By observing the trends of the cross-correlation values obtained with a decreasing number of selected channels, the quantitative effect of the gradual elimination of spectral channels can be measured and considerations about the DWT robustness can be inferred.

	Number of channels						
	7	6	5	4	3	2	1
L=5 ms	100	99.2	96.9	93.4	86.3	75.6	57.0
L=1 ms	-	-	100	99.1	96.2	89.0	65.7

Table 1. Results from the first experiment: cross-correlation values (%) between the reconstructed signal and the original one, against the number of spectral channels selected (by an energy criterion) for the reconstruction stage.

Table 2 shows the durations of the time-domain responses of the speech processing systems for two significant input signals: the step function and the ideal impulse. It can be immediately seen that, differently from traditional strategies, the DWT response is in both cases almost instantaneous.

	ACE	CIS	DWT
step	52	30	0.73
impulse	11	14	0.98

Table 2. Results from the second xperiment: durations (in ms) of the time-domain I/O characteristics of the speech processing strategies, obtained with a step and an impulse input.

The last experiment takes into consideration realistic operative conditions for the three strategies and tests them on standardized speech material, in order to compare the proposed WT strategy with the traditional ones. Results are given in Figures 1 and 2, according to two different points of view: in the first case, a single phoneme has been taken into consideration and the waveforms of the synthesized signals (from the compressed coefficients) with each of the three methods can be directly compared with the original signal, in addition for each case the cross correlation is given as an objective measure of distance between signals.

Figure 2 shows the final results of the comparison between strategies: the mean cross correlation values and the standard deviations, computed on the 16 phonemes, are indicated. Results on the left refer to synthesis of the acoustic simulations from the un-compressed envelopes, while as concerns results on the right, reconstruction is made after the application of the LGF.

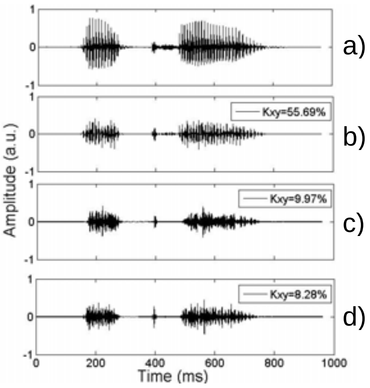

Figure 1: Reconstruction of speech signal /ata/ by means of the DWT (b), the ACE (c), and the CIS (d) strategies. The original signal is plotted on panel (a). The legends show the cross-correlation value between the original and the reconstructed signal.

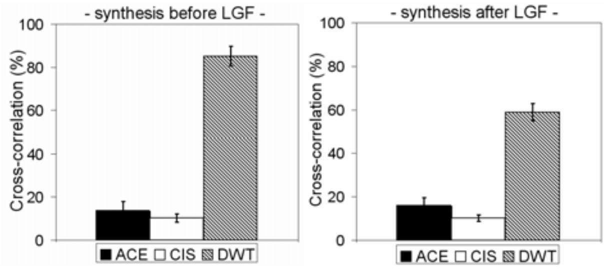

Figure 2. Mean cross correlation values (± 1 s.d.) between the original signals and the signals reconstructed with the ACE, CIS and DWT strategies. Synthesis before the LGF on the left, after LGF on the right. Means were computed on the whole data set of 16 phonemes.

Discussion

Results derived from all the described experiments highlight important and relevant properties of the DWT strategy as it has been defined here, and these features are particularly valuable for the application to a cochlear implant. In particular we can stress three main aspects: (1) robustness of the algorithm, (2) localization of temporal transitions and (3) accurate coding of speech.

(1) The robustness of the DWT algorithm, even after a strong reduction of the set of coefficients, can be verified on all the experiments that have been made: in fact in each test there exists a processing stage aimed at reducing the set of available data in order to respect the final stimulation rate, which is far lower then the original signal sampling rate. In particular it is the first experiment that gives an objective measure of the effect of suppressing spectral channels: from Table 1 it can be seen how the subtraction of 3 channels from a set of 6 or 7 channels (and so a strong reduction of data) doesn’t affect the performance in a dramatic way, in fact the cross correlation remains higher then 90%. Also, even with a single selected channel, the index is about 60%. It is evident that even a selection of a very small set of the most significant coefficients does not bring, for the DWT, a really strong decrease of the information transmitted to the reconstruction stage (IDWT): the

multiresolution decomposition algorithm concentrates the energy and the information of the analyzed signal in a compact set of coefficients instead that uniformly among them, so the signal can be easily reconstructed even after a great degradation of the available data set. This consideration means that the DWT is particularly suitable for an application to CIs, where the electrical stimulation pattern is, as known, an extremely synthetic version of the spectro-temporal profile of the incoming signal: the wavelet transform is able to optimize the stimulation sequence, by conveying on the electric impulses as much information as possible, thus probably being able, in perspective, to facilitate the patient's comprehension of speech.

(2) Another primary aspect, highlighted especially in Experiment 2, is the precision of the DWT in coding the temporal localization of rapid and short components even once that the algorithm has been extrapolated from its definition and inserted in the binding processing path of a CI. On one hand, the traditional filterbank strategies, despite they use really short elementary temporal analysis windows (about 3 ms), give a rough coding of rapid transitions, altering their temporal localization, thus making difficult for the implanted subject to recognize rapid speech components. On the other hand the wavelet transform demonstrates to be able to optimize temporal resolution, showing impulse (and step) responses that last less then 1 ms; so it can clearly code (as confirmed by the fourth experiment on consonant speech material) the rapid transitions present in everyday speech, particularly in plosive consonants. From a realistic point of view this fact implies a possible improvement of the patient's performance in consonant recognition tasks and in speech recognition.

(3) The accurate coding of speech obtained with the DWT strategy is evident in Figure 1 and confirmed by Figure 2: the most important aspects and features of the original signal are preserved, better then for the traditional strategies, even after the degradation of the sampling rate and also after the application of the LGF compression function. The new DWT strategy clearly respects the temporal profile of the analyzed signal, for example it can be seen (Figure 1) how the filterbank strategies alter the effective duration of the plosive consonant /*t*/ while the wavelet method gives a synthesized signal more similar to the original phoneme, as it is reflected by the cross correlation values: about 56% against the 8-10% reached by traditional methods. This similarity comes not only from the appropriate coding of the consonant, but also from other important aspects: first of all, with the DWT the signal envelope and in general all the low frequency information is well reproduced in the reconstructed signal, and this means that this spectral information is also present (and well coded) in the electrode activation map. Besides, there is another aspect that can be highlighted: the wavelet strategy accurately encodes the speaker's fundamental frequency F0 and it represents, with good precision, the *pitch*, i.e. the impulsive vibration of the vocal cords, thus it could give to the implanted subject, in perspective, the capability to distinguish the timbre of the speaker and easily identify him.

Together with the identification of the fundamental frequency, there is another aspect concerning speech comprehension that is facilitated by wavelet transform: the extraction of useful information of the signal from background noise. The suppression of noise is one of the characteristic wavelet features: with a very low computational cost the transform can extract signal from noise, even by means of a simple selection of the coefficients based on an energy threshold, so the wavelet strategy can be exploitable, in perspective, for efficiently de-noising the stimuli coming from the real environment. These considerations imply the possibility to have a clear representation of speech and also that the implanted subject has the capability of focusing his attention on the fundamental speaker, even if he is surrounded by numerous sources of competing voices and noise.

Conclusions

Results that emerge from this study demonstrate the appropriateness of wavelet transform for the proposed application: such an analysis tool, by concentrating energy and information in a small set of data, reveals a great potential for use in cochlear implants (where the data reduction is considerable). Moreover, it is evident how the transform is suitable for coding some crucial elements of speech, as the rapid transitions of consonants, or the low-frequency spectral content, or also the speaker's fundamental frequency and timbre, thus confirming that the speech processing for cochlear implants could get great and important benefits from the introduction of the WT. Essentially, results found here can be interpreted as a better speech coding and also a more adequate stimulation sequence that could be achieved with the new method and could be rigorously demonstrated by further studies involving patients. It would undoubtedly be useful, in fact, at this stage of the study, to experiment the strategy directly on implanted subjects delivering the stimulation sequence generated with PC simulations of the WT strategy, through a suitable interface, to the real implanted stimulator (*streaming mode*) in order to measure the effective intelligibility of the electrical stimuli derived by the simulated implant. As a future improvement of the strategy, it has already be mentioned the opportunity to insert appropriate de-noising algorithms in order to further increase the quality of the information transmitted to the electrodes and to obtain, at last, a speech processing tool able to efficiently code signals coming from a noisy context.

In brief we can conclude that it would be possible, by using the wavelet transform for speech processing in cochlear implants, to send to the electrode array a stimulation pattern able to convey the information about speech signals in an extremely synthetic way (not redundant), and deprived by the noisy components that naturally come from the sound environment around us.

References

[1] WILSON B.S., FINLEY C.C., LAWSON D.T., WOLFORD R.D. (1988): 'Speech processors for cochlear prostheses', *Proc. IEEE*, **76**, pp. 1143-1154

[2] Loizou P., STICKNEY G., MISHRA L., ASSMANN P. (2003): 'Comparison of speech processing strategies used in the Clarion implant processor', *Ear and Hear.*, **24**(1), pp. 12-19

[3] LONG C.J., DATTA S. (1996): 'Wavelet based feature extraction for phoneme recognition', Proc. of 4th Intern. Conf. on Spoken Language Processing, Philadelphia, USA (October, 1996) Vol. 1, pp. 264-266

[4] TZANETAKIS G., ESSL G., COOK P. (2001): 'Audio analysis using the discrete wavelet transform', Proc. of WSES Intern. Conf. on Acoust. and Music: Theory and Applications, 2001

[5] YAO J., ZANG Y.T. (2001): 'Bionic wavelet transform: A new time-frequency method based on an auditory model', *IEEE Trans. Biomed. Eng.*, **48**(8), pp. 856-863

[6] YAO J., ZANG Y.T. (2002): 'The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations', *IEEE Trans. Biomed. Eng.*, **49**(11), pp. 1299-1309

[7] VANDALI A.E., WHITFORD L.A., PLANT K.L., CLARK G.M. (2000): 'Speech perception as a function of electrical stimulation rate: using the Nucleus 24 cochlear implant system', *Ear and Hear.*, **21**(6), pp. 608-624

[8] MALLAT S. (1989): 'A theory for multiresolution signal decomposition: the wavelet representation', *IEEE Trans. on Patt. Anal. Machine Intell.*, **11**(2), pp. 674-694

[9] WILSON B.S., FINLEY C.C., LAWSON D.T. (1991): 'Better speech recognition with cochlear implants', *Nature*, **352**, no. 18, pp. 236-238

[10] DORMAN M., LOIZOU P., FITZKE J., TU Z. (1998): 'The recognition of sentences in noise by normal hearing listeners using simulations of cochlear implant signal processors with 6-20 channels', *J. Acoust. Soc. Am.*, **104**(6), pp. 3583-3585

[11] FU Q.J., SHANNON R.V. (1998): 'Recognition of spectrally degraded and frequency shifted vowels in acoustic and electric hearing', *J. Acoust. Soc. Am.*, **105**(3), pp. 1889-1900

[12] SHANNON R., ZENG F., KAMATH V., WYGONSKI J., EKELID M. (1995): 'Speech recognition with primarily temporal cues', *Science*, **270**, pp. 303-304