Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Speech Processing for Cochlear Implants with the Discrete Wavelet Transform: Feasibility Study and Performance Evaluation

Alessia Paglialonga, Gabriella Tognola, Giuseppe Baselli, Marta Parazzini, Paolo Ravazzani, and Ferdinando Grandori

Abstract—**An innovative approach is investigated for speech processing in cochlear implants (CI). Differently from the traditional filter-bank spectral analysis strategies, the proposed method analyses the speech signal by means of the discrete wavelet transform (DWT).**

Preliminary tests were conducted in order to compare the WT and the filter-bank analysis methods. Additionally, the intelligibility of the speech processed with the proposed WT strategy was tested on normal hearing people by means of the acoustic simulations and a comparison was made with respect to traditional CI algorithms.

Results showed that the WT could be a suitable method for speech coding in CIs. The information loss was minimal and, in speech recognition tests, the WT performance was similar to traditional filter-bank strategies.

I. INTRODUCTION

The cochlear implant (CI) is a prosthetic device that allows partial restoration of hearing in profound hearing impaired people. By means of a speech processor, the CI analyses the acoustic signals that come to the implant recipient and stimulates the acoustic nerve fibers with a train of electric impulses through an array of electrodes implanted inside the cochlea. The electrical stimulation is delivered according to the tonotopic distribution of tuning frequencies along the cochlear nerve fibers [1].

It is clear that most of patient's performance is due to the speech processing strategy, whose efficiency is related to its capability to mimic the function of a healthy cochlea. So far, all speech processing strategies used in CIs are based on a so-called filter-bank approach: a bank of band-pass filters estimates the signal spectral content in all the CI stimulation channels. As reported in [2], with these filter-bank strategies, psychophysical consonant recognition tests in implant recipients give average recognition scores around 60%, and in any case well below 90% even for the best performing subjects.

From a theoretical point of view, a speech processing strategy based on the wavelet transform (WT, [3]) might improve the accuracy of the CI in coding temporal speech features. The WT has a very accurate resolution both in the frequency and in the temporal domain, thus being one of the most appropriate tools to analyze non stationary signals such as speech. Also, the ability of the WT in processing speech seems to be intrinsically related to the fact that the cochlea itself behaves as a parallel bank of WT-like filters [4], [5].

As a matter of fact, Yao and Zhang [6]-[7] developed a new adaptive WT based on a model of the active auditory system, and Cheikhrouhou and coll. [8] in their previous work adapted the wavelet shape to the speech signal. Differently from [6]-[8], in our study we focused on the feasibility of the implementation of the WT into real implant processors rather than on the development of new wavelet mathematical algorithms. In fact we took into account the real operating conditions of a CI and made a comparison between the WT method and the traditional filter-bank speech processing strategies. In particular two strategies were considered: the Continuous Interleaved Sampling (*CIS*) and the *N-of-M* strategy [1]. Also, for comparison purposes a different strategy, the *CIS-like*, was designed and evaluated.

As a first step towards the definition of a WT speech processing strategy, the aims of this study were: (i) to investigate the feasibility of using the WT as a CI speech processing strategy and (ii) to evaluate the performance of a WT-based speech processing strategy in coding speech in terms of speech intelligibility. As concerns the mathematical implementation of the WT, in this study we chose the simplest definition of the transform, i.e. the discrete wavelet transform (DWT) [3].

II. MATERIALS AND METHODS

A. The Speech Material

For the evaluation of the speech processing strategies two different sets of speech material were used.

i) A set of 12 consonants taken from the Iowa consonant test in the vowel context /aCa/ (single male speaker without noise, 16 kHz sampling rate).

ii) A list of 10 bisillabic Italian words taken from the Bocca-Pellegrini test (single male speaker without noise, 16 kHz sampling rate).

B. The Speech Processing Strategies

For each of the four speech processing strategies (i.e., the CIS, the N-of-M, the CIS-like and the DWT), the usual high pass pre-emphasis filter (1st order Butterworth cut-off 1200 Hz [9]) was applied. Then, the different strategies followed different processing paths, as described next.

A. Paglialonga is with the CNR Institute of Biomedical Engineering and the Department of Biomedical Engineering, Polytechnic of Milan, I-20133, Milan, Italy.

G. Tognola, M. Parazzini, P. Ravazzani, and F. Grandori are with the CNR Institute of Biomedical Engineering, I-20133, Milan, Italy (e-mail: gabriella.tognola@polimi.it).

G. Baselli is with the Department of Biomedical Engineering, Polytechnic of Milan, I-20133, Milan, Italy.

1-4244-0033-3/06/$20.00 ©2006 IEEE.

1) The CIS and N-of-M Speech Processing Strategies: The filter-bank CIS and N-of-M strategies were implemented by means of a bank of 6th order Butterworth band-pass filters whose center frequencies reproduced those typically used in CIs. In our simulations 22 analysis filters were used to simulate the behavior of a 22-channel CI. The center frequencies of the 22 analysis filters were 250, 375, 500, 625, 750, 875, 1000, 1125, 1250, 1437.5, 1687.5, 1937.5, 2187.5, 2500, 2875, 3312.5, 3812.5, 4375, 5000, 5687.5, 6500, and 7437.5 Hz, respectively. The N-of-M strategy selects, from the entire set of available 22 channels, a subset of channels for each stimulation frame (eight in our implementation). The CIS strategy uses a smaller number of analysis filters (eight in our implementation) whose center frequencies were 312.5, 562.5, 875, 1312.5, 1937.5, 2875, 4312.5, and 6562.5 Hz. In both cases, the bank of analysis filters spans the entire signal spectral band (0-8 kHz). After spectral filtering, signal envelopes were extracted on each channel by full-wave rectification and low-pass filtering (2nd order Butterworth) with a 400 Hz cut-off frequency [10]. In the *N-of-M* strategy the eight channels with highest envelope amplitude were selected for each stimulation cycle, whereas the other channels were discarded. The signal envelopes were sub-sampled in order to obtain a sampling frequency equal to the channel stimulation rate, i.e. 500 pps.

2) The DWT Speech Processing Strategy: After the high pass pre-emphasis filter, the signal analysis was performed frame by frame using a sliding window of 128 samples (i.e., 8 ms). The signal in each frame was decomposed through the DWT algorithm into eight spectral channels. Each channel has a different number of samples: in our case (input signal sampled at *FS*=16 kHz) the sampling rates ranged from 8 kHz (for the highest frequency band) to 128 Hz (for the lowest frequency bands). Given that the final electrical stimulation pattern conveys 500 pulses per second for each electrode, the four highest bands had to be sub-sampled; one channel was already sampled at the desired rate; and the three remaining channels had to be up-sampled through linear interpolation.

3) The CIS-like Speech Processing Strategy: To implement the CIS-like strategy a filter-bank approach was used. Signal spectral analysis and envelope extraction are performed following the CIS processing, but in this case the cut-off frequencies of the eight band-pass filters were set according to the DWT spectral bands, i.e., the same cut-off frequencies as the wavelet dyadic filters are used. As described before for the CIS, in the CIS-like the signal envelopes were sub-sampled in order to obtain in each channel a sampling rate equal to 500 pps.

C. *Feasibility Study and Performance Evaluation*

To evaluate the feasibility and the performance of the WT in the CI context two different tests were performed and two different signal processing paradigms were used: (i) the *Inverse Transformation*, in order to make a comparison between the two spectral analysis approaches (i.e. the filter-bank and the wavelet), and (ii) the *Acoustic Simulation* [10], aimed at evaluating, through psychoacoustic recognition tests, the intelligibility of speech processed through each of the four strategies.

1) The Inverse Transformation: With the "Inverse Transformation" paradigm, only the spectral analysis stages were considered for comparison, i.e. the stages that sub-divide the signal in spectral channels: the filter-bank and the DWT decomposition, respectively. As concerns the filter-bank strategies, only the CIS and the CIS-like methods were considered. The signals were decomposed in eight spectral bands according to each of the three algorithms. Then, a new signal was reconstructed: the eight filtered signal components (derived with the filter-bank and the DWT approaches) were summed together. The reconstructed signals were compared with the original speech signals by means of the cross correlation index (i.e., the similarity).

2) The Acoustic Simulations and the Psychoacoustic Recognition Tests: For all the four strategies the acoustic simulations were synthesized according to the approach proposed by Shannon *et al.* [9]. On each stimulation channel, the signal envelope in the filter-bank strategies and the coefficients in the WT strategy were used to modulate a wide band noise, that was then band-limited with the same band-pass filter used in that channel in the spectral analysis stage (i.e., band-pass and wavelet filters, respectively). These modulated noise bands were then summed together and low-pass filtered at 4 kHz, thus generating an acoustic signal that could be used in psychoacoustic recognition tests.

Five normal hearing subjects (25-37 yrs) participated in the study. Stimuli were arranged in the following fashion: starting with the bisillabic words and then presenting the consonants. For the consonants, each phoneme was repeated ten times, whereas the words were played only once. Stimuli were randomized during each session. Subjects were instructed to repeat the word or consonant, and to guess if they were not sure; no feedback was given during the test. Correct answers were scored, separately for the two sets of speech material, in terms of percent recognition scores,.

III. RESULTS

Table I shows the results obtained, for the CIS and CIS-like analysis methods, in the *Inverse Transformation* test. For both sets of speech material, the cross correlation index (mean value) between the synthesized signal and the original one is shown.

TABLE I
- INVERSE TRANSFORMATION -
CROSS CORRELATION INDEX
RECONSTRUCTED SIGNAL VS. ORIGINAL SIGNAL

Strategy	Bisillabic Words	Consonants
CIS analysis	80.8 %	85.2 %
CIS-like analysis	65.4 %	80.7 %

The cross correlation with the CIS analysis is 80% and 85% for bisillabic words and consonants, respectively; with the CIS-like it is 65% and 80%.

Results from the psychoacoustic recognition tests are shown in Fig. 1 and Fig. 2 for bisillabic words and consonant phonemes, respectively. Percent recognition scores (mean value ±1s.d.) are shown for each of the four strategies: N-of-M, CIS, CIS-like, and DWT. A one-way analysis of variance (ANOVA test) revealed a significant main effect of speech strategy ($p<10^{-5}$ and p=0.04, respectively); a *post-hoc* test (Bonferroni test) showed that for bisillabic words (Fig. 1) there is no significant difference between the CIS-like and the DWT strategies, but both are significantly different from the N-of-M and CIS strategies, that give the higher recognition scores thus achieving the best performance, despite the 15%-20% information loss showed by the Inverse Transformation test. For consonant phonemes (Fig. 2) the post-hoc test showed that the DWT and the N-of-M and CIS reach an equivalent performance but the CIS-like is significantly different from each of the other three strategies and gives the worst recognition scores.

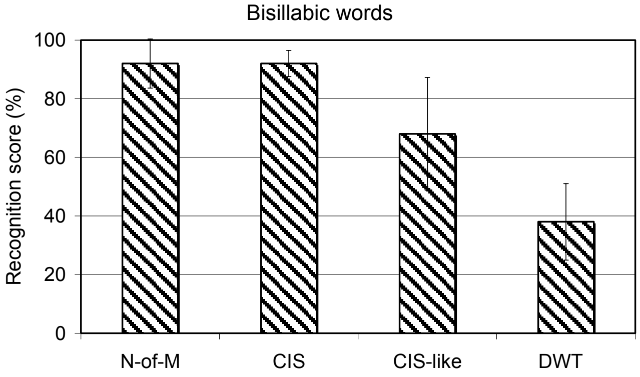

Figure 1 – Results from psychoacoustic recognition tests: bisillabic words. Percent recognition scores (mean value ±1s.d.) are shown for each of the four strategies: N-of-M, CIS, CIS-like, and DWT.

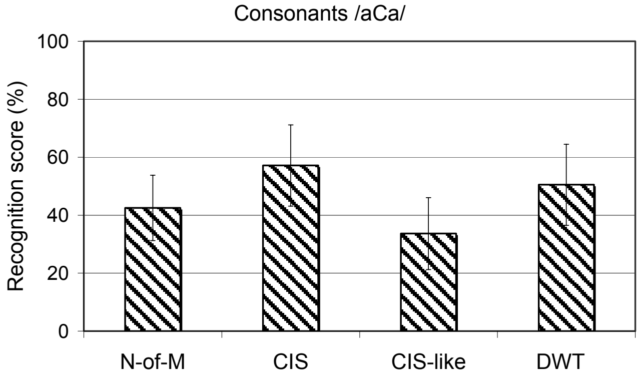

Figure 2 – Results from psychoacoustic recognition tests: consonant phonemes. Percent recognition scores (mean value ±1s.d.) are shown for each of the four strategies: N-of-M, CIS, CIS-like, and DWT.

IV. DISCUSSION

Results from the *Inverse Transformation* test (Table I) showed that the CIS and CIS-like strategies reached a cross correlation index lower than 85%, and down to 65% on bisillabic words processed with the CIS-like, differently from the WT that gives, as known, a perfect reconstruction of the original signal. This means that the traditional filter-bank strategies, splitting the signal into spectral channels, loose a part of its information content (about 20% information loss) because of the non ideal behavior of the band-pass filters. The WT, on the other hand, as far as the reconstruction is done with the *Inverse Transformation*, is able to give a reconstructed signal undistinguishable from the original one. This means that, looking at the information flow between the input signal and its spectral representation, the signal spectral components obtained with the filter-bank strategies carry only a part (average 83% for the CIS; average 73% for the CIS-like) of the information that characterizes the original speech signal, whereas the WT is able to preserve 100% of the signal features. Therefore, from a merely theoretical point of view, the WT is able to give an accurate coding of speech.

Moving our discussion from theory to practice, the psychoacoustic recognition tests show to what extent this suitability of the WT in preserving speech features and transmitting the whole information to the inverse reconstruction stage is maintained when the WT is put in the CI context: in this second test, the reconstruction of the signal is no longer done with the inverse transformation but the acoustic simulation are synthesized through modulated noise bands. Results obtained from these psychoacoustic recognition tests give a quantitative measure of the intelligibility of speech processed through the proposed WT strategy, thus allowing an evaluation of its performance. From Fig. 1 it can be seen that the traditional filter-bank strategies provide, as known, high recognition scores on words. There is no significant difference between the CIS and the N-of-M strategies, which are almost equivalent as concerns the recipient performance and give extremely good recognition scores (around 90%). But, if we force a filter-bank strategy (i.e., the eight-channel CIS) to work on the dyadic spectral bands, thus implementing the eight-channel CIS-like, we can see that the performance significantly worsens. This fact gives a clear indication that the major limit of applying the DWT in processing speech is using the dyadic spectral bands. It is reasonable to expect that the dyadic bands are not the best choice for speech recognition in noisy conditions: the highest dyadic band, for example, spanning the whole higher half-band of the signal, makes not possible for the listener to make a distinction between a 4 kHz and an 8 kHz speech component. So, the main reason for the observed lower WT performance could be the dyadic algorithm.

In Fig. 2 consonant recognition scores were shown. Scores are lower with respect to words recognition because of the more difficult speech material used. Even if the four distributions are very close, again according to the post-hoc test the CIS-like is significantly lower than the traditional CIS strategy, due to the dyadic spectral bands. But in this case, differently from words recognition performance, the DWT reaches a performance that is statistically equivalent to the two filter-bank strategies and performs better than the CIS-like. This difference is due to the different speech material used: consonants, differently from vowels, are mostly characterized by temporal cues and the DWT, with its fine temporal resolution for mid-to-high speech components (like consonants) is able to give a quite good consonant intelligibility, despite the dyadic spectral bands. Thus it seems that, when we speak in terms of speech recognition in noisy conditions, the temporal cues are not the only important features because if few spectral information is given (as in the CIS-like and DWT), even if we have a large amount of temporal information and minimum filter delay, as in the DWT, the recognition remains low because both time and frequency resolution should be optimized in a speech coding algorithm.

Results allowed putting in evidence that the WT can be successfully used for speech coding and that it can be easily adapted to a real CI speech processor; speech recognition tests showed that the DWT performance is close to that of filter-bank strategies and that the method can be improved trying to increase spectral resolution and formant coding.

References

[1] P. Loizou, “Introduction to cochlear implants,” *IEEE Eng. Med. Biol. Mag.*, vol. 18(1), pp. 32- 42, Jan.-Feb. 1999.

[2] P. Loizou, G. Stickney, L. Mishra, and P. Assmann, “Comparison of speech processing strategies used in the Clarion implant processor,” *Ear and Hear.*, vol. 24(1), pp. 12-19, 2003.

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

[4] G. Tognola, F. Grandori, P. Ravazzani, “Wavelet analysis of click-evoked otoacoustic emissions,” *IEEE Trans. Biomed. Eng.*, vol. 45, pp. 686-697, 1998.

[5] G. Tognola, M. Parazzini, P. de Jager, P. Brienesse, P. Ravazzani, F. Grandori, “Cochlear maturation and otoacoustic emissions in preterm infants: a time-frequency approach,” *Hearing Research*, vol. 199(1-2), pp. 71-80, 2005.

[6] J. Yao, Y. T. Zhang, “Bionic wavelet transform: A new time-frequency method based on an auditory model,” *IEEE Trans. Biomed. Eng.*, vol. 48(8), pp. 856-863, 2001.

[7] J. Yao, Y.T. Zhang, “The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations,” *IEEE Trans. Biomed. Eng.*, vol. 49(11), pp. 1299-1309, Nov. 2002.

[8] I. Cheikhrouhou, R.B. Atitallah, K. Ouni, A.B. Hamida, N. Mamoudi, and N. Ellouze, ”Speech analysis using wavelet transforms dedicated to cochlear prosthesis stimulation strategy,” 1st Intern. Symp. on Control, Communications and Signal Processing, 2004, pp. 639–642.

[9] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech recognition with primarily temporal cues,” *Science*, vol. 270, pp. 303-304, 1995.

[10] M. Dorman, P. Loizou, J. Fitzke, and Z. Tu, “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.*, vol. 104(6), pp. 3583-3585, 1998.