WAVELET PACKET FILTERBANK FOR SPEECH PROCESSING STRATEGIES IN COCHLEAR IMPLANTS

*Waldo Nogueira, Andreas Giese, Bernd Edler, Andreas Büchner **

Laboratorium für Information Technologie, Universität Hannover
* Hörzentrum Hannover, Medizinische Hochschule Hannover

ABSTRACT

Current speech processing strategies for cochlear implants use a filterbank which decomposes the audio signals into multiple frequency bands each associated with one electrode. Pitch perception with cochlear implants is related to the number of electrodes inserted in the cochlea and to the rate of stimulation of these electrodes. The filterbank should, therefore, be able to analyze the time-frequency features of the audio signals while also exploiting the time-frequency features of the implant. This study investigates the influence on speech intelligibility in cochlear implant users when filterbanks with different time-frequency resolutions are used. Three filterbanks, based on the structure of a wavelet packet transform but using different basis functions, were designed. The filterbanks were incorporated into a commercial speech processing strategy and were tested on device users in an acute study.

1. INTRODUCTION

Cochlear implants are accepted as the most effective means of improving the auditory receptive abilities of people with profound hearing loss. Current cochlear implants consist of a microphone, a speech processor, a transmitter, a receiver and an electrode array which is positioned inside the cochlea. The speech processor is responsible for decomposing the input audio signal into different frequency bands and delivering the most appropriate stimulation pattern to the electrodes. The bandwidths of the frequency bands are approximately equal to the critical bands, where low- frequency bands have higher-frequency resolution than high- frequency bands.

Speech coding strategies play an extremely important role in maximizing the user's overall communicative potential, and different speech processing strategies developed over the past two decades aim to mimic firing patters inside the cochlea as naturally as possible [1]. "NofM" strategies such as Advanced Combinational Encoder (ACE) [2], separate speech signals into M sub-bands and derive envelope information from each band signal. N bands with the largest amplitude are then selected for stimulation (N out of M).

Studies by different authors have revealed that there are two basic cues for pitch perception in cochlear implant recipients [1], [3]. The first cue, known as temporal pitch, is related to the temporal fluctuations in the envelopes of each spectral band. The second cue, known as *place pitch*, is related to the location of excitation along the cochlea. Electrodes near the base of the cochlea represent high-frequency information, whereas those near to the apex transmit low-frequency information.

Place pitch perception is limited by the number of electrodes inserted inside the cochlea. With 22 electrodes at most, we are attempting to mimic the functionality of thousands of nerve fibers. This leads to a poor frequency resolution, as the bands associated with each electrode are relatively wide to accurately encode tonal components. The limited perception of *temporal pitch* may be related to the misalignment between the temporal resolution of the implant, determined by its rate of stimulation, and the temporal resolution of the filterbank used by the speech processor. In actual implants, the rate of stimulation in each electrode ranges from around 0.5 ms until to 2 ms. Although, this temporal resolution should be sufficient to represent the temporal features of speech signals, the simple signal processing strategies used in actual speech processors do not offer the possibility of analyzing the signal at such resolutions (the signals typically being analyzed in frames of 8 ms).

Therefore, in order to improve pitch perception and speech intelligibility, it has been speculated that the design of a new filterbank with higher temporal resolution may lead to better speech perception with cochlear implants. The new filterbank is based on the structure of a wavelet packet (WP) decomposition. In WP analysis, a signal is split into an approximation (low pass component) and detail (high pass component). Each of these components can be split further, making it possible to decompose the audio signal into different levels, so that the time-frequency features of the analysis can be adapted to the time-frequency features of the implant.

The paper is organized as follows: In section 2, a review of the ACE strategy is presented. Section 3 describes the design of the WP and its incorporation into a commercial ACE strategy. Section 4 outlines the method for the evaluation of speech intelligibility with cochlear implant recipients. Finally, in section 5 and 6 the results and conclusions are presented respectively.

2. REVIEW OF THE ADVANCED COMBINATIONAL ENCODER (ACE) STRATEGY

The ACE (Advanced Combinational Encoder)[2] (Figure 1) is an "NofM"-type strategy used with the Nucleus implant. A digital signal sampled at 16 kHz is sent through a filterbank. The filterbank is implemented with an FFT (Fast Fourier Transform). The block update rate of the FFT is adapted to the rate of stimulation on a channel (i.e the total implant rate divided by the number of bands selected, N). The FFT is performed on windowed input blocks of 128 samples (8 ms at 16 kHz) of the audio signal using a Hann window. However, the rate of stimulation on a channel is usually set to a minimum of 2 ms causing the above mentioned temporal misalignment.

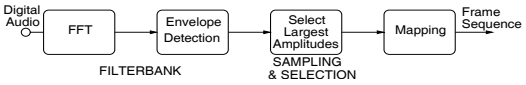


Fig. 1. ACE block diagram.

The uniformly-spaced FFT bins are combined by summing the powers to provide the required number of frequency bands M; the envelope in each spectral band is thus obtained. The frequency bounds of the spectral bands are uniformly spaced below 1000 Hz, and logarithmically spaced above 1000 Hz. Each spectral band is allocated to one electrode and represents a single channel.

In the "Sampling and Selection" block, a subset of N ($N < M$) envelopes with the largest amplitude are selected for stimulation. If N is decreased, the spectral representation of the audio signal becomes poorer, but the channel stimulation rate can be increased, resulting in improved temporal representation of the audio signal. If the channel stimulation rate is decreased, however, then N can be increased, providing an enhanced spectral representation of the audio signal.

The "Mapping" block, determines the current level from the envelope magnitude and the channel characteristics. A description of the process by which the audio signal is converted into electrical stimuli is given in [1].

3. DESIGN OF THE ACE SPEECH PROCESSING STRATEGY USING WAVELET PACKET

WP are efficient tools for speech analysis, they involve using two-band splitting of the input signal by means of filtering and downsampling at each decomposition level. An example of a WP filterbank is presented in figure 2.

Designing the WP filterbank involves choosing the decomposition tree and then selecting the filters for each decomposition level of the tree.

The decomposition tree has been chosen to mimic, as precisely as possible, the bands associated with each electrode using the commercial ACE strategy (similar to the Bark scale). The tree decomposition is presented in Figure 3. The nodes represent the decomposed signals obtained after filtering.

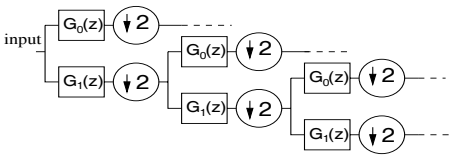


Fig. 2. Example of WP filterbank.

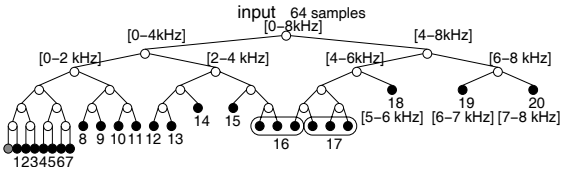


Fig. 3. Decomposition tree.

The black nodes depicted in Figure 3 represent the final node decomposition selected. Node 0, shown in grey, was not used as it chiefly contains noise information and plays no role in speech perception. For each decomposition level there is a different time-frequency resolution; in cochlear implants, however, the rate of stimulation on a channel is a fixed parameter for all the electrodes. The power in each node was, therefore, adapted to the stimulation rate. For example, if the stimulation rate on each channel is set to 2 ms, then the power in node 1 - which theoretically corresponds to a time resolution of 4 ms - was halved. For node 20, however, where the temporal resolution is 0.5 ms, the energy of the node was combined over four temporal frames in order to provide information every 2 ms. Finally, the powers in each node were weighted following a similar process to that used with the FFT in the commercial ACE strategy and the envelope in each spectral band was obtained by calculating the square root.

Once the decomposition tree has been selected, the next step involves selecting an appropriate wavelet filter for each decomposition level of the tree. The following sections present the three solutions adopted for filter selection.

3.1. Haar wavelet

The intended purpose of the new filterbank is to improve temporal resolution in order to allow enhanced perception of the *temporal pitch*. We therefore require filters with good time localization. The simplest way to achieve this may be to limit the impulse response length of the filters. However, this leads to a worse frequency resolution and therefore, at each level of the decomposition aliasing will be introduced when the subbands are sub-sampled by a factor of two. An example of such

a filter is the Haar wavelet [4]. Figures 4a and 4b present the impulse and frequency response of the Haar WP at node 1 of the tree shown in Figure 3.

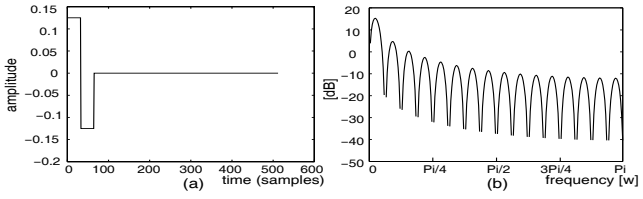


Fig. 4. (a) Haar WP impulse response at node 1.
(b) Haar WP frequency response at node 1.

3.2. db3 wavelet

The higher-order Daubechies family wavelets [4] can be used to improve frequency resolution and reduce the aliasing introduced in each decomposing level. Daubechies family wavelets are usually written by dbN, where N is the order and 2N is the impulse filter length. These wavelets are optimal in that they have minimum support length for a given number of vanishing moments. However, this family is not ideal in terms of symmetry. A disadvantage of the dbN is that it is not symmetrical. Symmetry (linear phase) is generally a desirable property for the analysis of speech signals as it means the filter has a constant group delay. Therefore, a method termed "alignment" [5] was employed in order to compensate for the fact that dbN lacks a linear phase. In this case we chose db3 in order to improve frequency resolution with respect to the Haar wavelet. It should be also mentioned that, when using the db3 mother wavelet, consideration must be given to which assumptions are made about the signal beyond the boundaries of the data, i.e. before the first and after the last sample of interest. The solution adopted to the boundary extension problem is to begin recording the signal before the region in which the decomposition will be applied, and to continue recording the signal beyond the region in which the decomposition will be applied; that is, to extend the signal with the actual signal values. Figures 5a and 5b present the impulse response of the db3 WP in node 1 of the decomposition tree shown in figure 3.

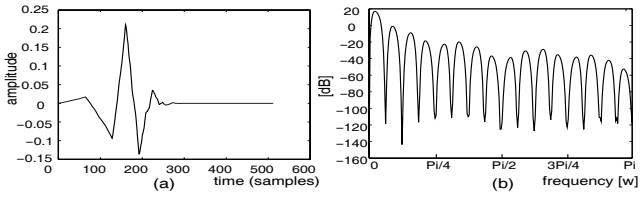


Fig. 5. (a) db3 WP impulse response at node 1.
(b) db3 WP frequency response at node 1.

3.3. Mixed wavelet

In order to find a filterbank that represents a good compromise between time and frequency resolution, a new wavelet packet filterbank was selected that uses long impulse responses at the initial stages and short impulse responses at the latter stages. The mother functions used are based on the Symmlets family. The Symmlets [4], denoted by SymN (N being the order and the impulse length being 2N), are nearly symmetrical wavelets proposed by Daubechies as modifications of the db family. The filterbank uses a Sym6 at the first level of decomposition, with its impulse response being successively halved at each subsequent level, (i.e. Symm5 is used at the second stage, and so on). Finally, at the last stage of the decomposition process it uses the Sym1, which is identical to the Haar wavelet. As different filter lengths were used at each decomposition stage, the filterbank was designated mixed WP. The same assumptions for boundary extension and "alignment" were made for the wavelet packet filterbank. Figures 6a and 6b present the impulse and frequency response of the mixed WP at the node 1 of the decomposition tree presented on Figure 3.

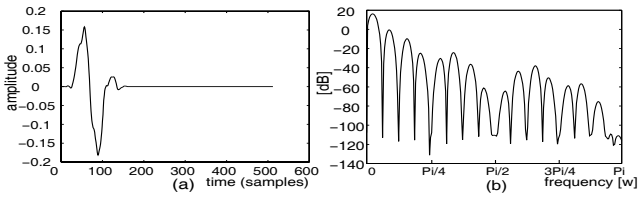


Fig. 6. (a) Mixed WP impulse response at node 1.
(b) Mixed WP frequency response at node 1.

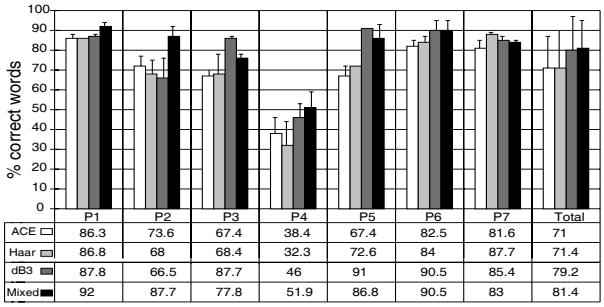
4. INTELLIGIBILITY TESTS

The Haar, db3 and mixed WP filterbanks have been incorporated into a research ACE strategy made available by Cochlear Corporation, termed NIC (Nucleus Implant Communicator). The software permits the researcher to communicate with the Nucleus implant via the standard hardware used for the fitting of patients in routine clinical practice. The NIC, processes the audio signals on a personal computer (PC). A specially initialized clinical speech processor serves as a transmitter for the instructions from the PC to the subject's implant. The three filterbanks programmed within the NIC environment were tested on subjects using the Nucleus 24 implant. The total number of electrodes for this implant is 22. However, only 20 electrodes were used by all the subjects as their everyday speech processor, the "ESPrit 3G", only supports 20 channels and the patients were familiar with this configuration.

The test material was the HSM (Hochmair, Schulz, Moser) sentence test [6]. In generating the subject's program,

id	Age	Duration deafness (years)	Implant experience (years)	Rate
P1	53	1.58	2	1200
P2	53	22.58	9	900
P3	65	0	5	720
P4	40	0	5	720
P5	64	0	5	720
P6	37	15.33	5	720
P7	66	0.75	9	1080

Table 1. Subject demographics



	P1	P2	P3	P4	P5	P6	P7	Total
ACE	86.3	73.6	67.4	38.4	67.4	82.5	81.6	71
Haar	86.8	68	68.4	32.3	72.6	84	87.7	71.4
dB3	87.8	66.5	87.7	46	91	90.5	85.4	79.2
Mixed	92	87.7	77.8	51.9	86.8	90.5	83	81.4

Fig. 7. Score by patient (average and standard deviation). Scores were obtained in noise conditions (SNR=15 dB).

the same psychophysical data measured in the R126 clinical fitting software were processed using the commercial ACE and the ACE with the three new filterbanks. The signals were processed in noise, with a signal-to-noise ratio (SNR) of 15 dB. Furthermore, the test material had previously been pre-emphasized by a filter which mimics the frequency response of the microphone used in commercial cochlear implant systems. The stimulation rate was adjusted to the requirements of each test subject and the number of bands selected per frame (N) was set to 8. The test subjects (Table 1) spent some minutes listening to the processed material, using all the filterbanks, in order to become familiarized with them. For the actual testing, 2 lists of 20 sentences were presented with the ACE, the ACE with Haar WP, the ACE with db3 WP, and the ACE with mixed WP. The subjects had to repeat each sentence without knowing which strategy they were listening to. This procedure was carried out on seven patients over a period of several hours.

5. RESULTS

Figure 7 presents the averaged scores obtained by each test subject for the different filterbanks. The results were analyzed using the Wilcoxon test [7] ($p<0.05$).

The averaged results show that the mixed WP resulted in significantly better speech perception performance than achieved using the commercial ACE strategy (based on an FFT) ($p=0.016$) and the Haar WP ($p=0.047$). One reason for the improvement in recognition rates is attributed to the superior tradeoff between time and frequency resolutions achieved by the mixed filterbank, permitting a better representation of both the *temporal pitch* and the *place pitch*.

6. CONCLUSIONS

In this study, a WP filterbank was designed and incorporated into a commercial ACE strategy for speech processing in cochlear implants. Three different configurations were implemented using different mother wavelets for the WP tree: Haar based WP, db3 based WP and a WP, termed mixed WP, which uses different filter lengths at each stage of the decomposition. All these configurations were implemented in a commercial ACE strategy, and speech intelligibility tests were conducted in seven cochlear implant recipients. Averaged results of speech intelligibility tests have shown that the mixed WP filterbank leads to significantly better speech perception performance than the FFT transform (as used in the commercial ACE strategy) and the Haar WP.

7. REFERENCES

[1] P. C. Loizou, "Signal-Processing Techniques for Cochlear Implants", *IEEE Engineering in medicine and biology*, Vol.18(3), pp.34-46, May/June 2000.

[2] W. Nogueira, et al., "A Psychoacoustic NofM type speech coding strategy for cochlear implants", *Eurasip Journal on Applied Signal Processing*, Special Issue on DSP in Hearing Aids and Cochlear Implants, 2005.

[3] C. M. Mackay, "Place, temporal cues in pitch perception: are they truly independent?", *ARLO*, Vol. 1, pp. 25-30, 2000.

[4] I. Daubechies, "Ten Lectures on wavelets", *SIAM Publications, 1992*, Number 61 in CBMS-NSF Series in Applied Mathematics, Philadelphia, 1992.

[5] Jensen A., A. la Cour-Harbo. "Ripples in Mathematics", *Springer Verlag*, ISBN 3-540-41662-5, Berlin, Heidelberg, New York, 2001.

[6] I. Hochmair-Desoyer, et al., "The HSM sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users", *J. Otol*, Vol. 18, Suppl. 6, pp. 83, November 1997.

[7] W. W. Daniel, "Applied Nonparametric Statistics", *PWS-Kent Publishing Company*, 2nd Edition, ISBN: 0534381944, 1990.