SPEECH RECONSTRUCTION FOR COCHLEAR IMPLANTS

M. Vondrášek, T. Tichý and P. Sovka
Department of Circuit Theory, Czech Technical University, Prague, Czech Republic

vondram3@fel.cvut.cz, tichy@fel.cvut.cz, sovka@fel.cvut.cz

Abstract: A cochlear implant is an electronic device that bypasses a non-functional inner ear by stimulating a hearing nerves with patterns of electrical currents pulses, so that a speech can be experienced by profoundly deaf people. The perception of this imitation of speech is very individual so it is hard to say which speech strategy brings better results for a given patient. Also the new strategy development brings about the question if this strategy will be better than the old one.

This paper describes an objective measure used for the evaluation of two methods for speech reconstruction from current samples send in patient's cochlea. This measure enables the comparison of the quality of reconstructed speech with the original one and thereby allows the development of new coding strategies without demanding tests with patients. The final decision about the quality of new coding strategies has to be based on the tests with patients.

Introduction

If the hair cells in the inner ear of patient are defective, they send no information in brain and the patient hears nothing [1], [5]. The non-functionality of hair cells could be caused by a congenital defect or by labyrinthitis.

But if an electrode is inserted in cochlea, the beginning of the auditory nerve could be stimulated by current pulses. If the electrode is inserted at the beginning of the cochlea, patient hears a sound with a high frequency. If the electrode is inserted at the end of the cochlea, patient hears sound with a low frequency. The intensity of auditory asthema could be set by the amplitude, width and frequency (stimulation rate) of the current pulses.

If the array of electrodes is used, more places in the cochlea could be stimulated. In one time it is allowed to stimulate only one electrode due to cross-talk between electrodes.

We can use three modes of stimulation. First, the bipolar stimulation occurs when a potential difference is created between neighboring electrodes producing a current flow between these two electrodes. Secondly the common ground stimulation occurs when one electrode is stimulated and all the other electrodes are electronically grounded.. Finally monopolar stimulation occurs when a potential difference is created between one active electrode and distant ground outside the cochlea.

Principle of Speech Coding

The cochlear implant system consists of two parts: the implant and the speech processor [1], [2].

First, the implant (Figure 1) consist of an electronic chip in a metal box, a receiver coil with a permanent magnet and two electrodes – the intracochlear electrode array with 22 electrodes and the ball electrode, which is used in case of the monopolar stimulation. This part is responsible for stimulation of the auditory nerve and it could have another special use for measuring of the neural response of the auditory nerve. The body of this implant is situated under patient's skin and the electrode array is inserted in the cochlea.

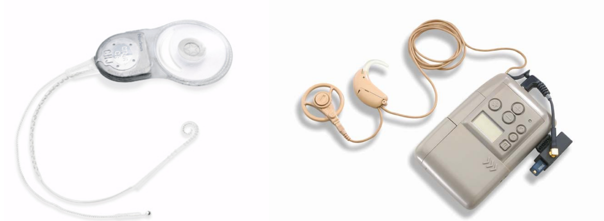


Figure 1: Nucleus implant CI24R and Sprint^{TM} bodyworn processor.

The second part of the cochlear implant system is the speech processor (Figure 1), which converts sound or speech in current pulses. The processor could be in behind-the-ear or in bodyworn configuration. The processor consists of a microphone, body with a signal processor and batteries and also a stimulating coil with a permanent magnet. Two permanent magnets (first in implant, second in stimulating coil) allow holding the stimulating coil in right position on patient's head.

The signal processing algorithm (speech strategy) specifies a set of rules which defines how the speech will be analyzed and which information will be selected and transmitted into the implant. The speech strategy has a few basic steps:

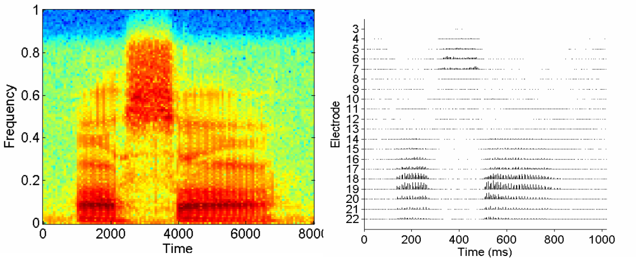
- Sound is picked up by a microphone, digitalized and divided in segments. Frequency analysis is applied on each segment and band selection is undertaken.
- The signal energy in each band is calculated and the most important information is chosen, depending on the used coding strategy and patient's setting.
- The information is coded and transmitted into the implant via radio frequency.
- The information is decoded in the implant and the nerve fiber stimulation is carried out.

ISSN: 1727-1983 © 2005 IFMBE

Speech Coding Strategies

The Cochlear implant system could use three coding strategies: Spectral Peak (SPEAK), Continuous Interleaved Sampling (CIS) and Advanced Combination Encoder (ACE) [1]. A patient uses one strategy, which is optimal for him comprehension in various environments. Each strategy has several parameters, which could be set to individualize the fit.

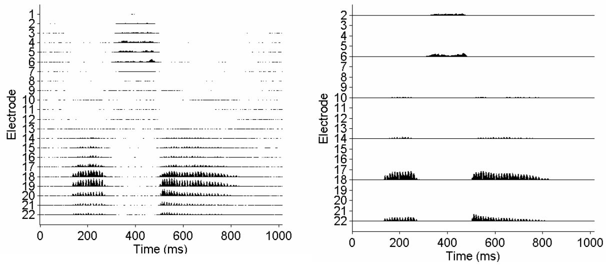
In case of the SPEAK strategy (see fig 2b)) the frequency band of the speech is divided in 20 sub bands by a filter bank. In each band the energy is calculated and some bands with the maximal energy are chosen. The number of chosen bands could be set for each patient individually from 1 to 10, typically 6 or 8 bands. In the implant, only 20 electrodes are used for the stimulation. One electrode in the implant represents one processed band. The information about selected bands with maximal energy are coded and send in the implant via radio frequency. The implant stimulate the chosen electrodes by current pulses with fixed frequency called "the stimulating rate". Figure 1b) illustrates activity of electrodes in case of the SPEAK strategy and Figure 2a) describes spectrogram of input word "asa". The SPEAK strategy on the Figure 2a) use 5 maxima.



a) b)

Figure 2: a): Spectrogram of input word. B) SPEAK strategy.

The principle of the ACE strategy is similar to the SPEAK strategy. But there are two changes. Firstly, ACE strategy uses 22 bands and also 22 electrodes. The principle of maxima selection is the same as in the SPEAK strategy. Secondly, the ACE strategy is able to stimulate with higher stimulating rate. Stimulating rate could be set with respect to the patient and the total maximal stimulating rate of the implant. The activity of the electrodes is presented in Figure 3a). The ACE strategy on this figure uses 5 maxima.



a) b)

Figure 3: a) ACE strategy. b) CIS strategy.

The CIS strategy is different. First, the fixed number of bands is selected. The numbers of bands is lower, in comparison with the ACE or SPEAK strategies. The CIS strategy uses from 4 to 12 bands. The frequency band of the processed speech is divided in selected number of sub-bands. Second, the signal energy in each band is calculated. Finally, the information about the energy of each band is coded and transmitted into the implant. The number of used electrodes is the same as the number of bands. The stimulating rate is set to a very high value when compared to the ACE strategy. Figure 3b) described the activity of electrodes In case of the CIS strategy. The CIS strategy on the Figure 3b) uses 5 maxima.

If we compare all three strategies, the SPEAK strategy could be characterized as the strategy with a high number of bands but with a small stimulating rate. The CIS strategy allows using a high stimulating rate but with small number of bands. The ACE strategy is able to offer a high number of bands and a high stimulating rate.

The discussion about what strategy works best has no practical sense because the speech perception with cochlear implants is individual. Each patient prefers different strategy with different settings.

Speech Reconstruction

Backward speech reconstruction is a process which translates the current samples mentioned above to the speech. We can do it in two ways: by a synthesis using superimposed sinusoidal signals or by a filter bank.

Synthesis with Superimposed Sinusoidal Signals (SSS)

The signal reconstructed using superimposed sinusoidal signals [6] is given by the formula

$$s(t) = \sum_{k=1}^{N} A_k(t)\sin(2pi \cdot f_k \cdot t), \qquad (1)$$

where $A_k(t)$ is the amplitude of the envelope k-th band. This amplitude is derived from the amplitude of current samples used for the stimulation. The amplitude $A_k(t)$ is non-zero only in segment where k was selected as a maximum. The frequency f_k is the central frequency of k-th band, and N is the number of the selected maxima. The spectrogram of a reconstructed word is presented in Figure 5a).

Synthesis Using Filter Bank (SFD)

The block diagram of the second method of the speech reconstruction [4] is shown in Figure 4). The current samples are represented by unit pulses. The amplitudes of these unit pulses are derived from the amplitude of the current pulses. After, the multiplexer is used to switch the unit pulses between filters in filter bank. The band pass filter bank is the same, as the one used for the speech coding in the speech processor.

Finally, all the outputs are summed up. The spectrogram of a reconstructed word is depicted in Figure 5b).

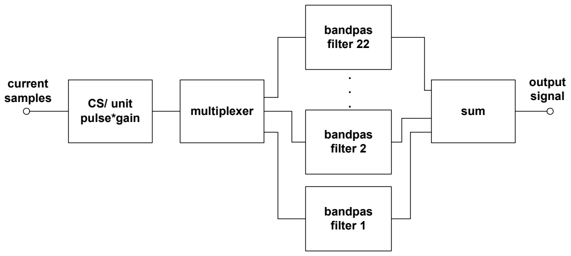
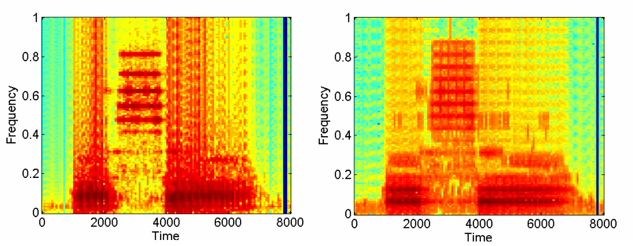


Figure 4: The block diagram of filter bank method.



a) b)

Figure 5: Spectrogram of signal reconstructed using: a) SSS and b) SFD method.

When comparing Figure 5a) with 5b) we can see differences in booth spectrograms. The question is how to evaluate these differences and their effect on hearing.

Comparison of Reconstruction

One possibility how to compare the original and the reconstructed signals is using the cepstral distance (2). First, both signals are divided in segments and each segment is separated in bands by a filter bank, which is the same as the one used for the analysis in strategy SPEAK, ACE or CIS. Second, the vectors of the cepstral distance between the input signal $c_i[k]$ and the reconstructed $c_r[k]$ signal is computed for each band. Finally, the formula (2) is used only for bands representing maxima in N-th segment.

$$d[N]) = 4.3429 \cdot \sqrt{\sum_{k=1}^{M} (c_i[k]-c_r[k])^2}, \qquad (2)$$

where M is the order of approximation. The cepstral coefficients were computed using the fast Fourier transform. The order of approximation was set to 30.

For the comparison we used 35 speech signals from Czech speech audiometry database. Segmental signal-to-noise ratio of the analyzed signals was 10-15 dB. We analyzed voiced and unvoiced parts of speech and pauses separately because of their different behavior.

Results

This section presents the results for both types of the speech reconstruction.

Synthesis Using Filter Bank

Figure 6 shows the average cepstral distance in different parts of speech in case of the SPEAK strategy. Cepstral distance for voiced parts is increasing with the number of selected maxima and the cepstral distance for the unvoiced parts is decreasing. Cepstral distance for pauses has the minimum for 7 maxima.

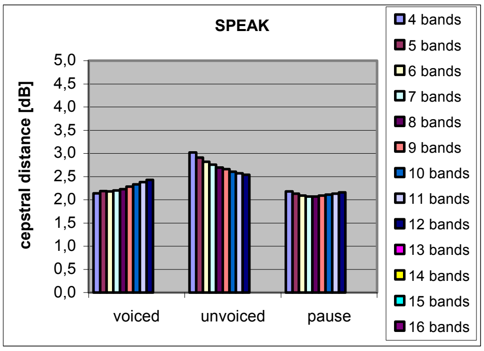
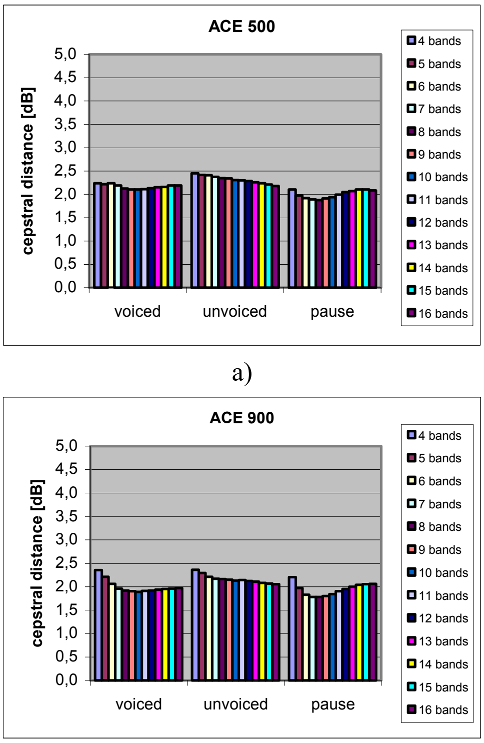


Figure 6: Cepstral distance for SPEAK strategy.

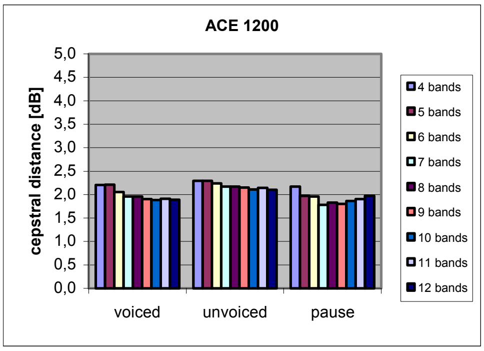
The dependency of the cepstral distance on a varying stimulating rate for ACE strategy is shown in Figure 7. The cepstral distance for the stimulating rate 500 and 900 Hz was computed for 4 to 16 bands. For 1200 Hz we were limited by total stimulating rate of the implant so we used only 12 bands. The cepstral distance decreases with the increasing stimulating rate.

In case of the CIS strategies (Figure 8) the cepstral distance reached bigger alteration in comparison with the ACE or SPEAK strategies.



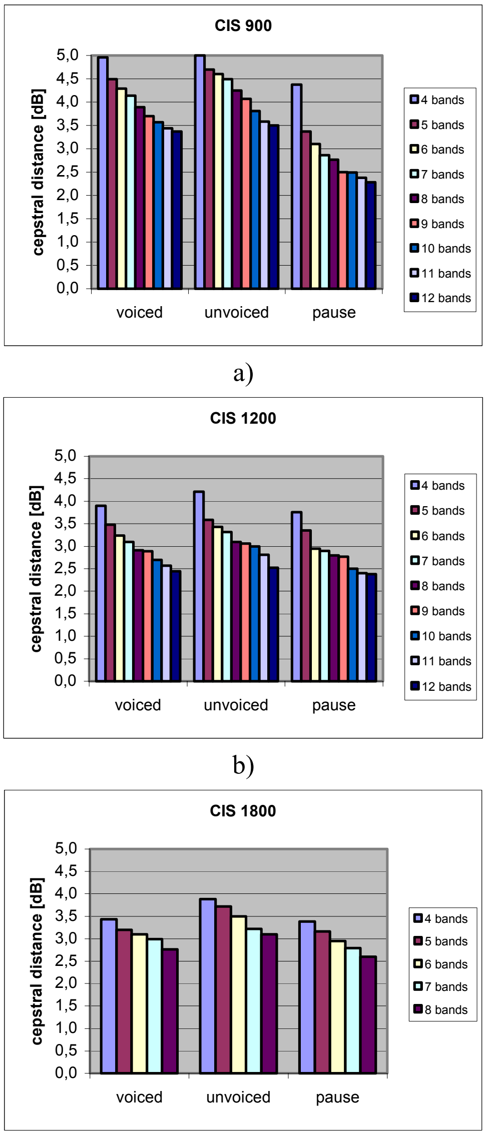
a)

b)



c)

Figure 7: Cepstral distance for ACE strategy for stimulating rate: a) 500Hz, b) 900Hz and c) 1200Hz.



a)

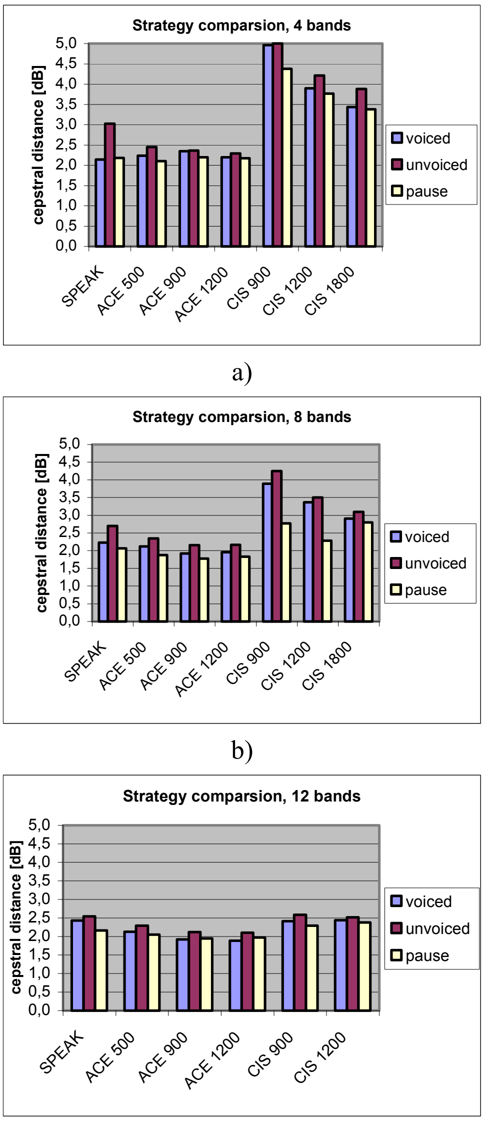
b)

c)

Figure 8: Cepstral distance for CIS strategy for stimulating rate: a) 900Hz, b) 1200Hz and c) 1800Hz.

Firstly, with an increasing stimulating rate, the computed cepstral distance decreased by 1 dB between 900 and 1200 Hz and 0.5 dB 1 dB between 1200 and 1800 Hz. Secondly, the cepstral distance decreased by 2 dB with increasing number of maxima. If the small number of selected maxima is used, the bandwidth of filters (Figure 2b) is very wide and the reconstructed signal is lees similar to the original one. For higher number of selected maxima the bandwidth decreased and the cepstral distance is lower. This decrease is also demonstrated in Figure 9.

Figure 9 presents the comparison of speech reconstructed using a filter bank in dependency on the number of selected maxima. For a small number of selected maxima (Figure 9a) CIS strategy gives two times greater cepstral distance.



a)

b)

c)

Figure 9: Comparison of SPEAK, ACE and CIS strategy for: a) 4 bands, b) 8 bands and c) 12 bands.

If more bands are selected (Figure 9b, 9c) cepstral distance quickly falled down. In case of the SPEAK or ACE strategies, the cepstral distance is decreased too, but not so evidently. The minimal cepstral distance for all cases of selected maxima was reached in the ACE strategy for stimulating rate 1200 Hz.

Synthesis with Superimposed Sinusoidal Signals

The synthesis with superimposed sinusoidal signals is independent to the stimulating rate, because the amplitudes of the reconstructing sinusoidal signals are given only by the amplitude of the stimulating pulses. Only one figure from each method is presented.

Figure 10 illustrates the average of the cepstral distance in different parts of speech in case of the SPEAK strategy. In comparison with the synthesis using a filter bank, the average cepstral distance for unvoiced parts of speech tends to increase with the increasing number of selected bands with maxima.

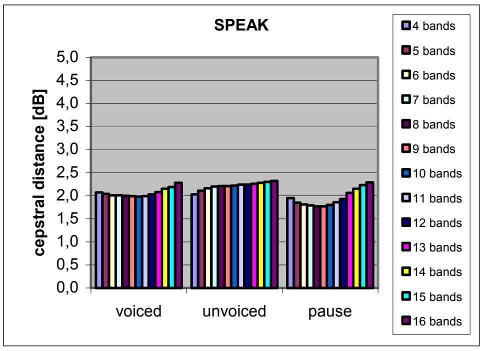


Figure 10: Cepstral distance for SPEAK strategy.

The speech reconstructed with this method in case of the ACE strategy (Figure 11) is more similar to original one for 4 to 8 bands. If more then 8 maxima are selected, the reconstruction using a filter bank gives better results.

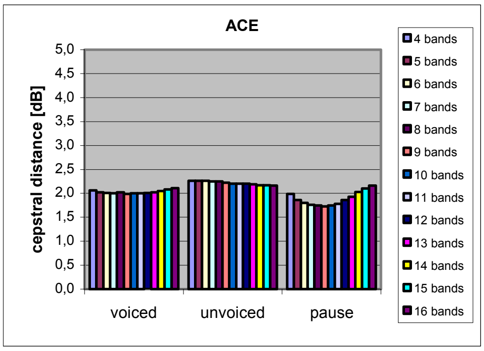


Figure 11: Cepstral distance for ACE strategy.

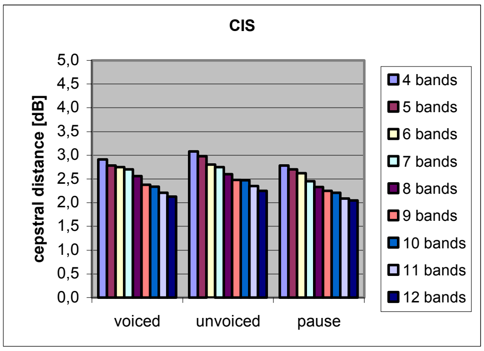
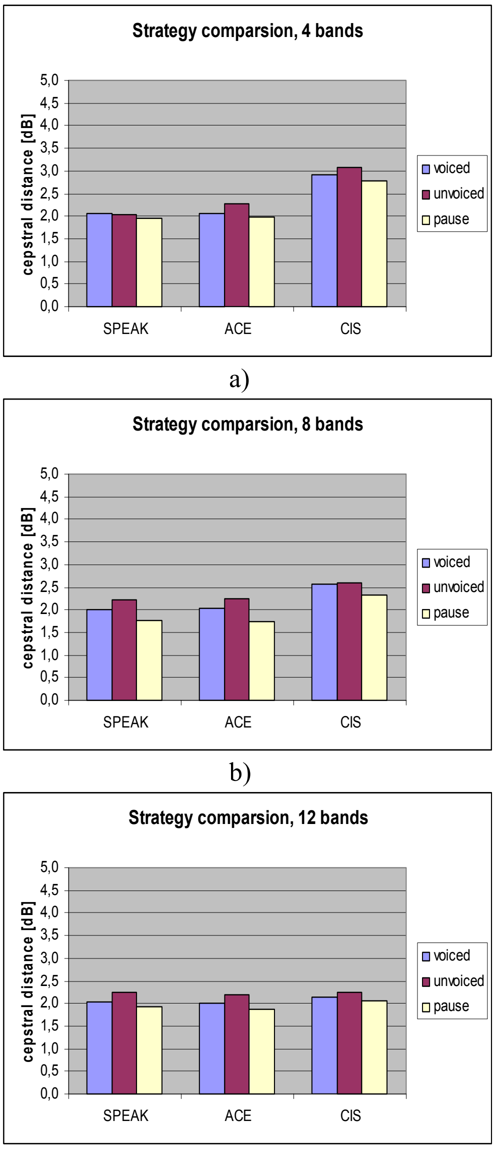


Figure 12: Cepstral distance for CIS strategy.

Figure 12 represents results for the CIS strategy. The average cepstral distance between the original and reconstructed speech is, in comparison with the reconstruction using a filter bank, very low (see Figure 7). Thus for the CIS strategy the reconstruction with superimposed sinusoidal signals is preferable.



a)

b)

c)

Figure 13: Comparison of SPEAK, ACE and CIS strategy for: a) 4 bands, b) 8 bands and c) 12 bands.

Figure 13 presents the comparison of speech reconstructed using a superimposed sinusoidal signals in dependency on the number of selected maxima. As in case of reconstruction using filter bank, the cepstral distance in case of CIS strategy is decreased if more bands are selected.

Discussion

In this section we summarize the results of speech reconstruction by the synthesis using a filter bank and the synthesis using superimposed sinusoidal signals.

The ACE, SPEAK and CIS strategies and the speech reconstruction methods were implemented in Matlab programming environment with Nucleus Neural Toolbox [3].

The lowest cepstral distance in both reconstruction methods can be reached with the ACE strategy (Figure 7, 9, 11). On the other hand the CIS strategy seems to be worst of all strategies (Figure 8, 9, 12). For a low number of selected maxima or for a low stimulating rate the cepstral distance is two times greater than for the other strategies. If a high stimulating rate is used, the CIS strategy works better. But the usage of a high stimulating rate could be limited due to so called refractory period (the short time immediately after the generation of the action potential, in which a neuron cannot respond to another stimulus).

The filter bank method (SFD) reconstructs explosive consonants like “p”, “t”, “d” with difficulties. The superimposed sinusoidal signals (SSS) works better. Both SFD and SSS methods reconstruct sibilants worst with CIS strategy. The SSS method reconstructs pauses in speech worst. For the SPEAK strategy, SSS method is better than SFD method due to a low stimulation frequency. The reconstruction by the SSS method has a metal sound, but it is comprehensible. A signal reconstructed by the SFD method sounds more naturally, but is less comprehensible. The quality of the signal reconstructed by the filter bank depends on the stimulation frequency. A higher frequency makes the output signal more natural.

Conclusions

The speech reconstruction from current samples could be very useful in the future research of the speech preprocessing for cochlear implants. A new coding strategy could be preliminary verified before demanding tests with patients.

The SSS reconstruction method is independent on the stimulating rate, but is easy to implement. The SFD method takes into account the stimulating rate but it is very demanding. In the future work both methods could be used.

Acknowledgements

This work has been supported by the research program “Transdisciplinary Research in Biomedical Engineering 2” No. MSM 6840770012 and by grant „Biological and Speech Signal Modeling”, No 102/03/H085 of the Czech University in Prague.

References

[1] CLARK G. (2003): ‘Cochlear implants, fundamentals and applications’. Springer NY, pp 405-415

[2] ‘Nucleus Reference manual’ (2001): Cochlear Ltd.

[3] ‘Nucleus Matlab Neural Toolbox’ (2004): Cochlear Ltd.

[4] MOCEK V. (2002): ‘Evaluation of Performance of Speech Coding Strategies Used in Cochlear Implant Systems in Noisy Environment’, FMBE Proceedings. Vienna, Austria, p. 51-55

[5] LOIZOU P. C. (1998): ‘Mimicking the Human Ear’, IEEE Signal Processing, pp. 101-129

[6] LAN N., NIE K. B. (2004): ‘A Novel Speech-Processing Strategy Incorporatng Tonal Information for Cochlear Implants’. IEEE transactions on biomedical engeneering vol. 51