32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

# Comparison of Speech Processing Strategies for the Design of an Ultra Low-Power Analog Bionic Ear

Chutham Sawigun, Wannaya Ngamkham, and Wouter A. Serdijn

Biomedical Electronics Group, Electronics Research Laboratory,
Delft University of Technology, the Netherlands
[c.sawigun; w.ngamkham; w.a.serdijn]@tudelft.nl

*Abstract*—**Miniaturizing area and power consumptions of cochlear prosthetic devices is strongly required for full implantation. In this paper, several speech encoding strategies are studied and compared in order to find a compact speech processor that allows for full implantation and is able to convey both time and frequency components of the incoming speech to a set of electrical pulse stimuli. The study covers the widely recognized continuous time interleaved sampling (CIS) and strategies that convey the temporal fine structure (TFS), including race-to-spike asynchronous interleaved sampling (AIS), phase-locking (PL) using zero-crossing detection (ZCD), and PL using a peak-picking (PP) technique. To estimate the performances of the four systems, a spike-based reconstruction algorithm is employed to retrieve the original sounds after being processed by different strategies. The correlation factors between the reconstructed and original signals imply that strategies that convey TFS outperform CIS. Among them, the peak picking technique combines good performance with great compactness since envelope detectors are not required.**

## I. Introduction

Recovering human hearing perception via electrical stimulation is the ultimate goal of contemporary Bionic Ear (BE) or Cochlear Implant (CI) devices. To some extent this goal has been achieved by the invention of a speech processing strategy called 'Continuous Interleaved Sampling, (CIS)' [1] which roughly emulates the behavior of the basilar membrane and inner and outer hair cells, and successfully prevents simultaneous interactions between electrodes using fixed-rate interleaved amplitude-modulated stimuli. CIS has been employed as a default strategy in several commercially available CI devices produced by different manufacturers, i.e., MED-EL GmbH, Cochlear Ltd., Advanced Bionic Corp., and the results obtained from clinical experiments have shown to offer reliable understanding of sentences in quiet environments but poor results obtained for simple melodies. In typical noisy environments, the patients (CI recipients) are still having difficulties to understand both sentences and melodies [2]. These indications imply that the temporal fine structure (fast varying components of the sound) is not being conveyed to the brain.

In the normal mammalian auditory nerve fibers, the spike trains synchronize with the stimulus waveform periodicity up to 5 kHz [3]. Beyond that frequency range, the spike trains are generated randomly [4]. The aforementioned mechanisms are missing in the CIS processor since the pulse stimulation rate is fixed. To gain the perception of tonal languages and music, an effort of realizing the BE processor that imitates the inner hair cells and the auditory nerve behavior more precisely is being considered. For this reason, the 'Hilbert Transform, (HT)' has been introduced to the BE processor to extract temporal envelope, instantaneous frequency and phase, and thereby the TFS [5]. Although extraction can be achieved, conveying all of the information to the brain via electrical pulse trains is still a challenge that remains. Besides, performing the HT requires a large computational cost for both digital [6] and analog [7] processors.

In this paper, we thus explore some enveloped-based processing strategies that fundamentally require the computationally intensive and power hungry HT, including CIS, AIS, PL-ZCD and PL-PP. We try to optimally balance the quality of the sound that can be conveyed via a set of pulse trains to the stimulation electrodes and hardware complexity of the processors. Then we found that to realize a CI processor based on PL-PP strategy, an envelope detector (ED) is no longer required. This leads to a great reduction of hardware complexity and power consumption. In Sec. II, the general concept behind all existing CI processors is described. According to the general concept, in Sec. III, all the different processing strategies are examined and compared in terms of complexity and quality of the coded waveforms. To estimate the sound coding performance, the spike-based reconstruction technique [8] is applied to find the correlation factors of the input signal and the coded output pulse sequences. The results are discussed in Sec. IV. Finally, the conclusions are given in Sec. V.

## II. Speech Processing in Cochlear Implants

Fig. 1 shows a general block diagram that can be used to describe all strategies that are considered in this paper. The processor comprises three layers of operation. On Layer-1, indicated by the white background boxes, the incoming sound is pre-emphasized by non-linearly amplifying before entering the bank of band pass filters (BPFs). This mechanism is adapted from the role of the outer hair cells that map the wide range of the incoming sound pressure onto the limited dynamic range of the ear. The BPF bank roughly mimics the basilar membrane behavior by decomposing the signal into a limited number (N) of frequency bands (channels). The signal strength of each channel will be extracted in the form of the temporal envelope (roughly emulates the role of inner hair cell) and then modulated with the generated pulse trains to further stimulate the nerve fibers. This is common for all envelope-based processors. As mentioned in Sec. I, the

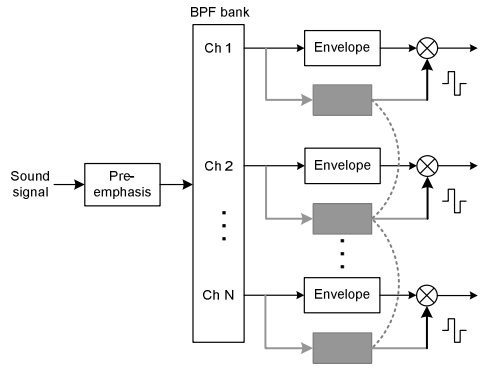


Figure 1. General block diagram for envelope-based speech processing

spiking pattern depends somewhat on the input frequency so that Layer-2 (indicated by the gray boxes) is introduced. Particular features (frequency, phase, TFS) of the output waveform of each channel will be detected and combined with the envelope to define the suitable stimulation pulse features. On this layer, the pulses generated from each channel are independent from each other and the stochastic spiking behavior of the auditory nerve is ignored. Layer-3 is therefore added to include this phenomenon by somehow conditioning the features detected from different channels to create a stimulation pattern that avoids electrode interaction and preserves the relevant extracted features. Note that the attempt to convey all the features of the incoming signal to the stimulation electrode is based on the assumption that the brain can interpret this information but in practice there are factors that deviate from what really happens along the auditory pathway. So the number of layers (system complexity) does not guarantee the quality of perception in real patients [9] but serves only as a first order estimation.

## III. REVIEW AND COMPARISON OF THE EXISTING PROCESSING STRATEGIES

### A. Continuous Interleaved Sampling

From the default setting of several CI models [2], it can be said that CIS is the most widely used strategy to date. CIS acquires the 1st layer of operation. There is some evidence that quality of speech perception obtained from a CIS processor strongly depends on the precision of the extracted envelopes [5, 10]. Accordingly, an attempt to replace the simple ED comprising a rectifier and a low-pass filter (LPF) by a HT based ED is of interest. This issue needs to be carefully considered for an analog processor since in order to perform the HT, a high complexity of its constituting electronic circuitry is unavoidable [7]. Fig. 2a shows a fraction of the speech signal from the word 'die' after 4th-order butter-worth BP filtering with a center frequency of 150 Hz. The envelopes are extracted by a simple ED with 200 Hz LPF cutoff frequency (the pink line of Figs. 2b-2c and 2e) and by the HT-ED (green line of Figs. 2b-2e). The positive pulse train generated within the CIS processor is represented by the blue line in Fig. 2b. In this case, we can clearly see that the accuracy of the amplitude of the pulses highly depends on

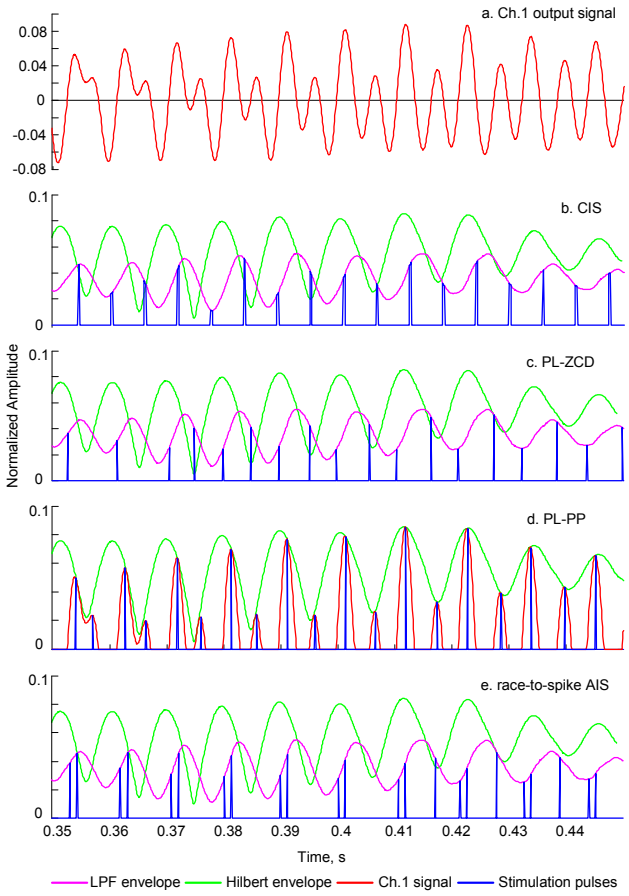


Figure 2. Waveforms obtained from different strategies

the accuracy of the ED. Also, it is hardly possible that the brain can recognize frequency, phase and TFS from the fixed timing interval of the pulse train.

### B. Zero-Crossing Detection

In this strategy, the 2nd layer is put on top of the 1st layer to introduce a phase locking amplitude modulated pulse train [11]. At the moment that the input signal crosses zero from negative to positive values, a pulse is generated and will be modulated with the momentary value of the envelope at that moment to create the stimulation pulse train. As we can see from the blue line of Fig. 2c, the magnitudes of the pulse are also defined by the quality of the ED but the real-time period of the fundamental frequency ($F_0$) can be roughly encoded. This processor thus requires high precision zero crossing instant and high accuracy envelope detectors.

### C. Peak-Picking Technique

This strategy also contains two layers of operation (1 and 2). But instead of detecting the zero-crossing moments to create the phase-locked pulse train, the occurrences of peaks in the input signal are detected [12]. There are two main features different from the PL-ZCD. First, the number of peaks detected is higher than the number of zero crossing moments which can be seen from the blue lines in Figs. 2c

and 2d during 0.35s < $t$ < 0.37s. This implies that more instantaneous frequency information other than $F_0$ can be conveyed to the stimulation electrodes. Second, as we can see from the peaks that always touch the Hilbert envelope, the BPF output signal and the detected peaking moments can be used to generate the stimulation pulses directly without the need for an ED. For this reason, the precision of the stimulation pulse amplitudes is relayed onto the precision of a peak-instant detector.

### *D. Race-to-Spike Asynchronous Interleaved Sampling*

In this case the 3rd layer is introduced. It has been proposed in [8] that to achieve the stochastic stimulation behavior, the gray boxes of Fig. 1 are replaced by half-wave rectifier circuits. Then, at particular repetitive time instants, the amplitudes from all channels will be sent to a winner takes all (WTA) network letting only the strongest amplitude pass to enable the pulse generator. To avoid successive stimulation within one channel that violates the bio-realism spiking behavior [13], additional circuit blocks are inserted to create an inhibition. At the moment that a stimulation pulse is being generated, there will be a signal created and applied to inhibit the signal from the half-wave rectifier (within the channel that is being stimulated), so that it will not be processed by the WTA network at the next time step. Even if the signal strength of that channel is highest, it will be ignored. Within this processor, the amplitude of each pulse is still specified by the ED of each channel but the location of the stimulated electrode is defined by the strength of the signal at that moment. The pulse waveform obtained from this processor is shown by the blue line in Fig. 2e. To some extent, encoding sound using this strategy can emulate the random spiking behavior of the normal auditory nerve fiber and the perception of music is expected. It is unfortunate that the system is very complicated requiring two more additional circuits blocks.

## IV. SYSTEM STIMULATIONS

MATLAB was used to simulate all encoding strategies. Three kinds of sound samples were used for the simulation with a sampling frequency of 11,025 Hz including the word '*die*', the sentence '*the discrete Fourier transform of a real value signal is conjugate symmetric'* and the sung phrase '*Hallelujah*'. An 8 channel 4th-order Butterworth BPF bank is used for all strategies with center frequencies ranging from 150 Hz to 4,000 Hz arranged according to the ERB scheme [14]. Each envelope detector is formed by a full-wave rectifier followed by a 4th-order Chebyshev LPF with 200 Hz cutoff frequency. The envelope detector is applied for all processors except PL-PP since this strategy does not need one. The stimulation pulses obtained from channels 1 to 8 of all strategies (shown in Fig. 2 only from the 1st channel) are collected for reconstruction using the spike-based technique [8] (in this context, spike refers to pulse signal). This technique has its foundation in prior neurophysiology work showing that the original analog waveform can be accurately reconstructed from a spiking waveform [15]. We therefore use this technique for the signal reconstruction. Fig. 3 shows

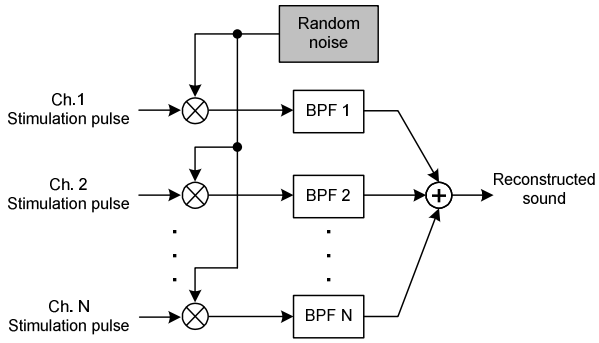


Figure 3. Spike-based reconstruction

a block diagram of this reconstruction technique. The stimulation pulses of each channel are multiplied by uniformly distributed random noise before injecting into the BPF with the same center frequency as in the processor. The resulting signals from all channels were added to produce the output sound. To exemplify the reconstructed waveforms, Figs. 4 and 5 show the reconstructed sounds of the word "*die*" from the CIS and PL-PP strategies, respectively. The original and reconstructed signals are represented by the green and blue lines, respectively. Roughly, it is visible that the reconstructed signal from PL-PP is closer to its origin than that of CIS. The correlation factor ($r$) between the original signal and the reconstructed signal was used to estimate the quality of the signals encoded from different strategies. The correlation coefficient is computed from the following equation

$$r = \frac{\sum_{i=o}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=o}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=0}^{n}(Y_i - \bar{Y})^2}}, \qquad (1)$$

where $X_i$, $Y_i$, $\bar{X}$ and $\bar{Y}$ are the original, the reconstructed signals, the mean value of $X_i$ and the mean value of $Y_i$, respectively. The correlation coefficient varies in the range of -1 to 1, where 1 indicates a perfect correlation, -1 shows the inversely perfect correlation and 0 represents no correlation. The resulting correlation coefficients obtained from different strategies are shown in Table I. It is clear that CIS performs worst of all. Besides, within the results from CIS, the values of $r$ depend on the complexity of the original sounds. The highest value of $r$ = 0.11 is from the simple word (single tone) and the lowest $r$ = 0.02 is from the song which contains several tones that CIS could not capture. Among the PL processors, as expected from the coding mechanism that conveys more instantaneous information without loss from the non-ideality of the ED (see Fig. 2d,) the PL-PP provides a better value of $r$ than PL-ZCD for all cases. The race-to-spike AIS processor gives the best values of $r$ for less complicated sounds (word and sentence) but for multi-tone sounds (song), the highest value of $r$ is given by the PL-PP processor.

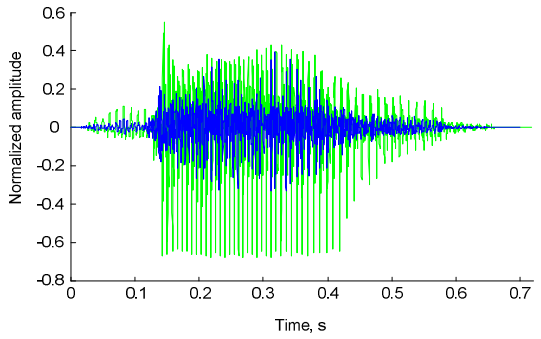


Figure 4. Reconstructed waveform of the word "*die*" from the CIS.

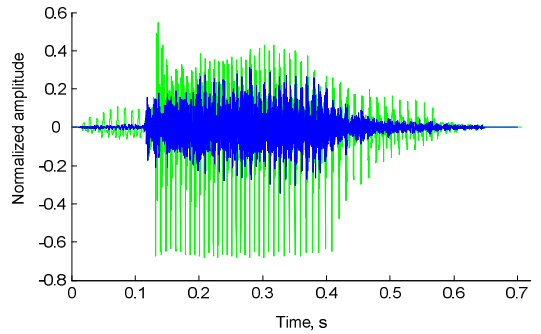


Figure 5. Reconstructed waveform of the word "*die*" from the PL-PP.

TABLE I. CORRELATION FOR DIFFERENT STRATEGIES

| Strategy | "*die*" | Sentence | '*hallelujah*' |
|---|---|---|---|
| CIS | 0.11 | 0.05 | 0.02 |
| PL-ZCD | 0.36 | 0.14 | 0.50 |
| PL-PP | 0.47 | 0.25 | 0.59 |
| race-to-spike AIS | 0.49 | 0.38 | 0.52 |

## V. DISCUSSION AND CONCLUSIONS

The system complexities and quality of the reconstructed signals from different signal processing strategies have been investigated and compared. Targeting the design of a fully implantable analog BE with an ability of tone recognition, the PL-PP provides us the best solution, both in terms of compactness and correlation factors. Since the information of frequency, phase and TFS cannot be conveyed to the stimulation electrodes by CIS, it is really hard to believe that the brain can recognize any tone without proper input information. CIS is therefore removed from our optimistic consideration. One may object that the values of the correlation factor cannot 100% guarantee the quality of hearing perception in real BE recipients. Still, we are that the brain can interpret multi-tone sounds from the fast varying information conveyed to the stimulation

## ACKNOWLEDGMENT

The authors would like to acknowledge the financial support for parts of this work by STW, the Dutch Technology Foundation, under project grant 10056.

## REFERENCES

[1] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz, "Better speech recognition with cochlear implants," *Nature*, vol. 352, pp. 236–238, 1991.

[2] Y. Y. Kong, R. Cruz, J. A. Jones, and F. G. Zeng, "Music perception with temporal cues in acoustic and electric hearing," *Ear and Hearing,* vol. 25, no. 2, pp. 173–185, 2004.

[3] I. Tasaki, "Nerve impulses in individual auditory nerve fibers of guinea pig," *Journal of Neurophysiology*, no. 17, pp. 97-122, 1954.

[4] N. Y. S. Kiang, T. Watanabe, E. C. Thomas, and L. F. Andclark, "Discharge patterns of single fibers in the cat's auditory nerve," *M.I.T Press*, Research Monograph no. 35, Cambridge, Mass., 1965.

[5] F. G. Zeng, "Trends in cochlear implants," *Trends In Amplif.*, vol. 8, pp. 1–34, 2004.

[6] S. K. Padala and K. M. M. Prabhu, "Systolic arrays for the discrete Hilbert transform," *IEE Proceedings-Circuits Devices Systems*., vol. 144, no. 5, pp. 259-264, 1997.

[7] W. Ngamkham, C. Sawigun, S. Hiseni and W. A. Serdijn, "Analog complex gammatone filter for cochlear implant channels," *Proc. IEEE ISCAS*, France, 2010.

[8] J. J. Sit, A. M. Simonson, A. J. Oxenham, M. A. Faltys, and R. Sarpeshkar, "A low-power asynchronous interleaved sampling algorithm for cochlear implants that encodes envelope and phase information," *IEEE Transactions on Biomedical Engineering*, vol. 54, pp. 138–149, Jan. 2007.

[9] P. Schleich, "Pulsatile Cochlear Implant Stimulation Strategy," *Patent Application Publication*, US20090264961A1, Oct. 22, 2009.

[10] K. Nie, A. Barco, and F. G. Zeng, "Spectral and temporal cues in cochlear implant speech perception," *Ear and Hearing*, vol. 27, pp. 208–217, 2006.

[11] J. Chen, X. Wu, L. Li, and H. Chi, "Simulated phase-locking stimulation: An improved speech processing strategy for cochlear implants," *Journal for Oto-Rhino-Laryngology*, vol. 71, pp. 221-227, 2009.

[12] B. S. Wilson, D. T. Lawson, C. C. Finley, and M. Zerbi, "Speech processors for auditory prostheses," *10th Quarterly Progress Report*, NIH project N01-DC-9-2401, National Institutes of Health, 1991.

[13] M. White, M. Merzenich, and J. Gardi, "Multichannel cochlear implants: Channel interactions and processor design," *Archives of Otolaryngology*, vol. 110, pp. 493-501, 1984.

[14] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103-108, 1990.

[15] R. Wessel, C. Koch, and F. Gabbiani, "Coding of time-varying electric field amplitude modulations in a wave-type electric fish," *Journal of Neurophysiology*, vol. 75, pp. 2280–2293, 1996.