Original Paper

ORL

ORL 2009;71:221–227
DOI: 10.1159/000229302

Published online: August 26, 2009

Simulated Phase-Locking Stimulation: An Improved Speech Processing Strategy for Cochlear Implants

Jing Chen[a] Xihong Wu[a, b] Liang Li[a–c] Huisheng Chi[a, b]

[a]National Key Laboratory on Machine Perception, Speech and Hearing Research Center, and [b]Department of Psychology, Peking University, Beijing, PR China; [c]Department of Psychology, Centre for Research on Biological Communication Systems, University of Toronto at Mississauga, Mississauga, Ont., Canada

Key Words
Speech processing strategy · Cochlear implant · Phase information · Chinese speech

Abstract
The continuous interleaved sampling (CIS) speech-processing strategy has been widely used for cochlear implants to extract speech envelope information without preserving phase information. In this study, a novel simulated phase-locking stimulation (SPLS) strategy, which detects zero-crossing times of the narrow-band signal of each band, was developed to extract both phase and amplitude-envelope information from a bank of frequency bands of speech sounds. The advantage of the SPLS strategy over the CIS strategy was confirmed by the results of our psychophysical experiments, showing that normal-hearing Chinese listeners' performance in recognizing SPLS-processed Chinese speech was significantly better than their performance of recognizing CIS-processed Chinese speech under quiet, noise-masking, or speech-masking conditions. Thus, the results suggest that if the SPLS strategy is used to modulate the interval of electrical stimulation pulses in cochlear-implant devices according to extracted phase information, the speech-processing functions of cochlear implant devices would be improved for Chinese cochlear implant users.

Introduction

Cochlear implant (CI) devices have been applied successfully to help profoundly deaf patients achieve hearing through electrical stimulation of the auditory nerve with fine electrodes inserted into the scala tympani of the cochlea [1]. The performance of listeners using CI devices depends largely on the signal processor transforming speech signals to electrical stimuli. Several signal-processing techniques have been developed over the past 30 years, and have been classified into 2 major types: waveform representation and feature extraction. As a typical waveform representation approach, the continuous interleaved sampling (CIS) strategy developed by researchers at the Research Triangle Institute shows a high level of speech recognition for the CI users speaking monotonal languages, such as English and German [2–4].

However, it has been reported that CI users who speak Chinese have poor identification of vowels and consonants [5, 6]. Chinese is a tonal language, which has 4 tonal patterns as defined by the fundamental frequency (F0) of voiced speech. For example, changing the tone in the syllable 'ma' from flat to rising, or to falling and rising, or to falling, changes the meaning of the word. Using the CIS strategy, Xu et al. [7] studied how signal-processing parameters, such as the low-pass cutoff frequency for

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2009 S. Karger AG, Basel
0301–1569/09/0714–0221$26.00/0

Accessible online at:
www.karger.com/orl

Xihong Wu, PhD
National Key Laboratory on Machine Perception
Speech and Hearing Research Center, Peking University
Beijing 100871 (PR China)
Tel./Fax +86 10 6275 9989, E-Mail wxh@cis.pku.edu.cn

extracting amplitude envelopes and the number of channels of the band-pass filter bank, affect tonal recognition. The results of their studies show that recognition of the 4 Mandarin tonal patterns depends on both the number of channels and the low-pass cutoff frequency, and temporal cues can compensate for diminished spectral cues in tone recognition and vice versa. In addition, the importance of pitch and periodicity information in Chinese speech recognition have also been confirmed in the study by Fu et al. [8], in which 3 carrier band conditions were tested, including noise-band carrier for all speech segments, pulse train carriers for the voiced speech segment whose rate followed the F0 of the speech signals, and fixed-rate pulse train carriers for voiced speech segments. The results show that the F0-controlled pulse train carriers produce the best performance, indicating the need to provide adequate amounts of both pitch and periodicity information to Chinese-speaking CI patients.

Although some CI users perform well in speech recognition as normal listeners in a quiet environment, they have considerable difficulties in performance when maskers, especially fluctuating maskers, are presented [9]. F0 information has long been thought to play an important role in perceptually segregating sound sources [10]. A reduction in F0 cues produced by cochlear-implant processing leads to difficulty in segregating different sources. Moreover, fine structure information is also important for sound localization and pitch perception [11]. So, it is important to study how to convey more fine structure information of the speech signal to CI users.

Although in some CI strategies, such as MPEAK (multi-peak), F0, the first formant, and the second formant are extracted and used to modulate the electrical pulse's firing, errors are induced in formant extractions, especially in the situations where the speech signals are embedded in noise [1]. According to the CIS strategy, the envelope information of band-pass filtered speech sounds are extracted and used to modulate the amplitude of electrical stimulation pulses of implanted electrodes without preserving the phase information in speech sounds. Since the phase information is potentially useful for improving CI listeners' speech perception [12], the present study proposes a new CI speech-processing strategy, the simulated phase-locking stimulation (SPLS) strategy, which preserves part of phase information in original speech and would be useful for upgrading the function of a CI device by introducing phase-related modulation of stimulation-pulse intervals. To experimentally evaluate the efficacy of the SPLS strategy in processing Mandarin Chinese speech, we presented the acoustic stimulation of the SPLS strategy to normal-hearing Chinese listeners under either noise-masking or competing-speech-masking conditions.

Methods

Simulated Phase-Locked Stimulation Strategy

Figure 1 illustrates how the SPLS strategy extracts envelopes of band-pass filtered signals and uses phase information to modulate pulse rates [1, 6]. A signal is pre-emphasized first and then decomposed into multiple frequency bands by a bank of band-pass filters. Because in the present study the filter-bank should not distort phases of input signal components, the zero-phase transfer function is used in the stage of band-pass filtering [13]. After that, the signal in each band goes through 2 signal pathways: envelope extraction and phase extraction. To extract envelope information, the filtered signal is processed by the Hilbert transform and the extracted envelope is then logarithmically compressed to an acceptable dynamic range for CI. The compressed envelope will be used to modulate the amplitude of pulse trains that are interleaved among electrodes. To extract phase information, the 'zero-crossing detection' process was used to record every zero-crossing time of the narrow-band signal in each band. The phase information will be used to decide the firing time of pulse trains.

The pulse-firing strategy of SPLS simulates the neural mechanism of human hearing. In the human auditory system, the nerve firings occur at roughly the same phase of the waveform each time. However, there is also a difference between low and high frequencies. In detail, a single auditory nerve fiber fires on every cycle of tone stimulus in the low-frequency range and does not necessarily fire on every cycle of tone stimulus in the high-frequency range. In SPLS, the electrical stimulation pulses of each channel occur at the zero-phase of the signal in the corresponding channel. For a given channel whose center frequency is below 1,200 Hz, pulses fire at every zero-crossing time detected from the band-pass filtering signal. Otherwise, pulses fire once every $\lceil f/1{,}200 \rceil$ zero-crossing times, where *f* is the center frequency, and $\lceil . \rceil$ means the smallest integer bigger than f/1,200. The amplitude of the pulse is modulated by the extracted envelope.

For the CIS strategy, the periods between pulses in each channel are fixed and simultaneous firing across channels can be avoided. However, for the SPLS strategy, the pulse rate in each channel is changed according to phase information, and simultaneous firing between 2 adjacent channels will happen. So, we measured the possibility of simultaneous firing between 2 adjacent channels on a 49-second piece of sound (including male or female English speech, Chinese speech, and a piece of music), which was processed by the SPLS strategy with 8 channels. When 2 pulses of 2 adjacent channels, respectively, fired at the same time, this firing was counted as a simultaneous firing. The final percent of simultaneous firing was 1.9%, which was too small to use additional inhibitory procedures.

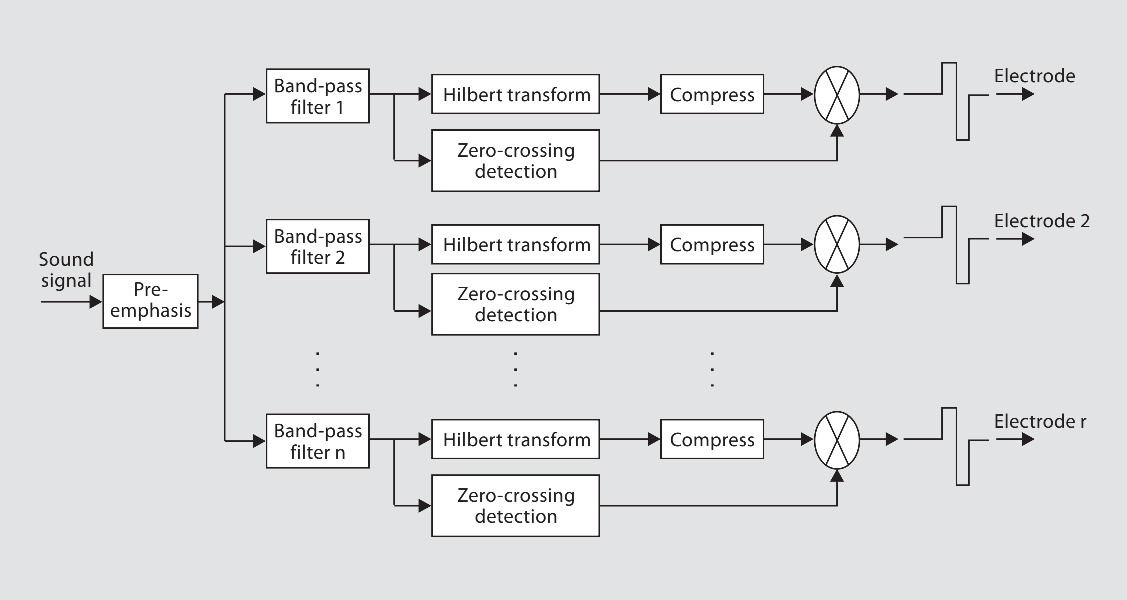

Fig. 1. Diagram showing how the SPLS method extracts envelopes of band-pass filtered signals and modulates pulse rates according to phase information.

Acoustic Simulation

Previous studies have confirmed that examination of normal-hearing listeners' responses to acoustic simulation of a CI processing strategy is useful for evaluating these strategies [14]. Thus, the effectiveness of the SPLS strategy was investigated in the same way.

Chinese speech was processed first through a high-pass filter (6 dB/octave slope at 1,200 Hz) implemented by a 1-order FIR filter to simulate the compensation, and then by a bank of 8 adjacent band-pass filters. All the band-pass filters were 1/3 octave, fourth-order Butterworth filters, and the bandwidths were determined following the equivalent rectangular bandwidth model [15]. The center frequencies of the 8 bands were equally spaced on a log scale from 215 to 4,891 Hz with the successive ratio factor of 1.25. The envelope of each band was extracted by the Hilbert transform, and was used to modulate the amplitude of the sine wave.

For the CIS strategy, the simulation could be described by the following formula:

$$y(t)=\sum_{i=1}^{n}A_i(t)\cdot\sin\left(2\pi f_i t+\phi_i\right), \tag{1}$$

where *y(t)* is the synthesized signal, subscript *i* means frequency band *i* and there are total *n* bands, A_i means the envelope of the band *i*, f_i means the center frequency of the band *i*, ϕ_i means the initial phase of the sine wave of the band *i*. Here, n = 8 and $\phi_i = 0$.

For the SPLS strategy, the formula (1) is modified into the following:

$$y(t)=\sum_{i=1}^{n}A_i(t)\cdot\sin\left(2\pi f_i(t)\cdot t+\phi_i\right), \tag{2}$$

where $f_i(t)$ reflects the firing rate of electronic pulse as a function of time:

$$f_i(t)=\frac{1}{T_i(t)}, \tag{3}$$

where $T_i(t)$ is the interval between 2 adjacent zero-crossing times.

Test Materials

The present psychophysical study aimed to compare the effectiveness of the SPLS and CIS strategies in 12 Chinese subjects (4 males, 8 females) with normal hearing confirmed by audiometry. Listeners were seated in a chair at the center of an anechoic chamber (Beijing CA Acoustics), which was 560 cm in length, 400 cm in width, and 193 cm in height. All acoustic signals were digitized at the sampling rate of 22.05 kHz using the 24-bit Creative sound blaster PCI128 (which had a built-in antialiasing filter) and audio editing software (Cooledit Pro 2.0), under the control of a computer with a Pentium IV processor. The analog outputs were delivered from a loudspeaker (Dynaudio Acoustics; BM6 A), which was in the frontal azimuthal plane at 0° position with respect to the median plane. The loudspeaker height was 106 cm, which was approximately ear level for a seated listener with average body height. The distance between the loudspeaker and the center of the participants' head was 150 cm.

Speech stimuli were Chinese 'nonsense' sentences that were syntactically correct but not meaningful. These sentences are similar but not identical to the English sentences that were developed by Helfer [16] and were also used in the study by Li et al. [17]. Each of the Chinese nonsense sentences has 3 key components: subject, predicate, and object, which are also the 3 key words with 2 syllables for each. For example, for the nonsense sentence 'An

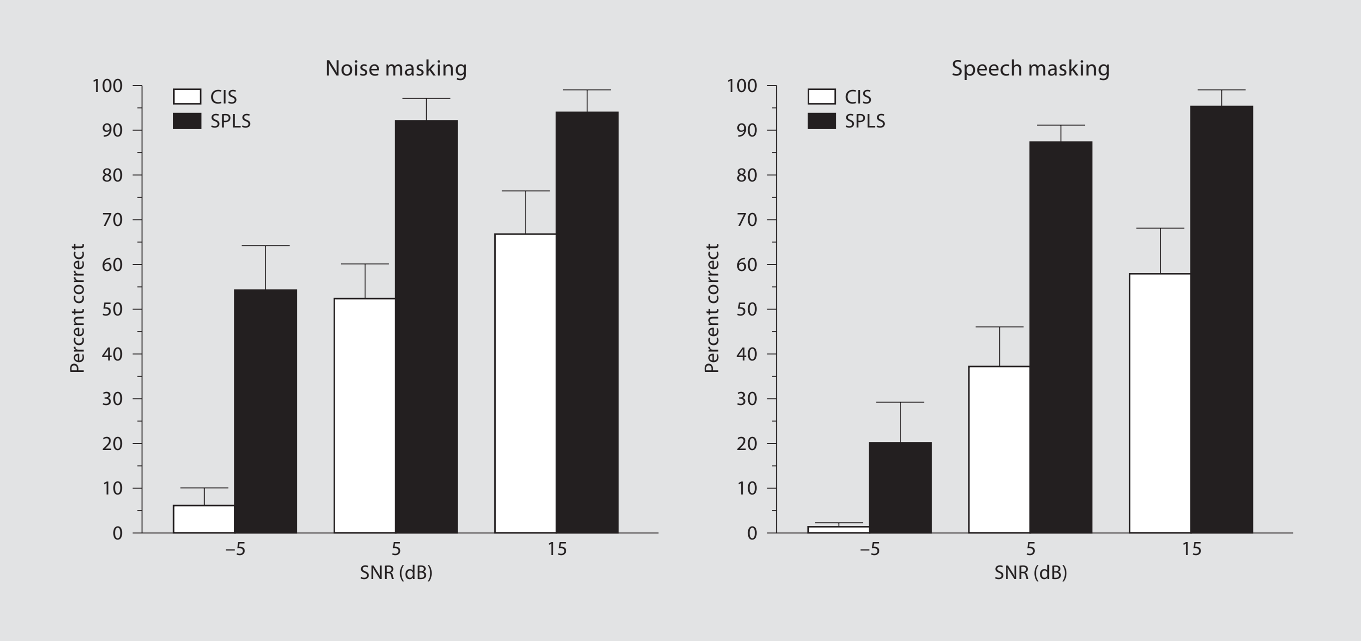

Fig. 2. Mean percent-correct identification of key words across 12 subjects as a function of SNR for each of the 2 processing strategies under 2 masking conditions: steady-spectrum-noise masking and speech masking. The error bars indicate the SD of the mean.

ant is quarrelling with a bag', whose direct Chinese translation sounds like: 'Yi1 zhi1 ma3 yi3 zheng4 zai4 xuan1 nao4 yi1 ge1 shu1 bao1', all the 3 underlined words are the key words [18]. Target speech stimuli were spoken by a young female speaker, and tested in a quiet environment or 1 of the 2 masking conditions, including steady-speech-spectrum noise masking and speech masking. Masking speech was a recording of nonsense sentences spoken by 2 other young females, with contents different from the targets.

All the sound stimuli were presented by the loudspeaker. The sound level was calibrated using a B&K sound level meter (type 2230) whose microphone was placed at the central location of the listener's head when the listener was absent, using a 'slow'/'RMS' meter response. The pressure level of target speech remained constant at 60 dB(A) throughout the experiment. The sound pressure levels of the masker were adjusted to produce 3 signal-to-noise ratios (SNR): –5, 5 and 15 dB.

The signal of each acoustic stimulus was produced by adding the signal of a target speech to that of the corresponding masker, and then the mixed signal was processed by both the CIS and SPLS strategies.

Psychophysical Experiment

There were 3 within-subject variables in the present experiment: (1) masker type (steady noise, speech); (2) processing strategy (SPLS, CIS); (3) SNR (–5 dB, 5 dB, 15 dB). In total, there were 12 (2 × 2 × 3) conditions for each listener, and a list of 18 target sentences was randomly assigned to a condition. The 4 masker/strategy combinations were counterbalanced across 12 listeners using a complete Latin square order, and the 3 SNR were arranged randomly at each masker/strategy combination.

Before the formal test, 2 lists of clear speech processed by SPLS and CIS, respectively without masking, and 4 lists of masker/strategy combinations at 15 dB SNR were delivered to train the listeners. For every strategy, listeners were also tested in the quiet condition without any maskers.

In each trial, the listener pressed a button of the response box to start the masking sound. About 1 s later, a single target sentence was presented. The masker was gated off with the target. Listeners were instructed to repeat the target sentence as best as they could immediately after the sentence was completed. The experimenters indicated the key words which had been identified correctly on a marking sheet. The number of correctly identified words was tallied later.

Results

Figure 2 shows mean percent-correct identification of key words across 12 subjects as a function of SNR for each of the 2 processing strategies (sparse columns, CIS; dense columns, SPLS) under 2 masking conditions.

In the quiet condition, the mean percent-correct recognition (85.88%) of the speech processed by SPLS was much larger than that (70.92%) processed by CIS. An

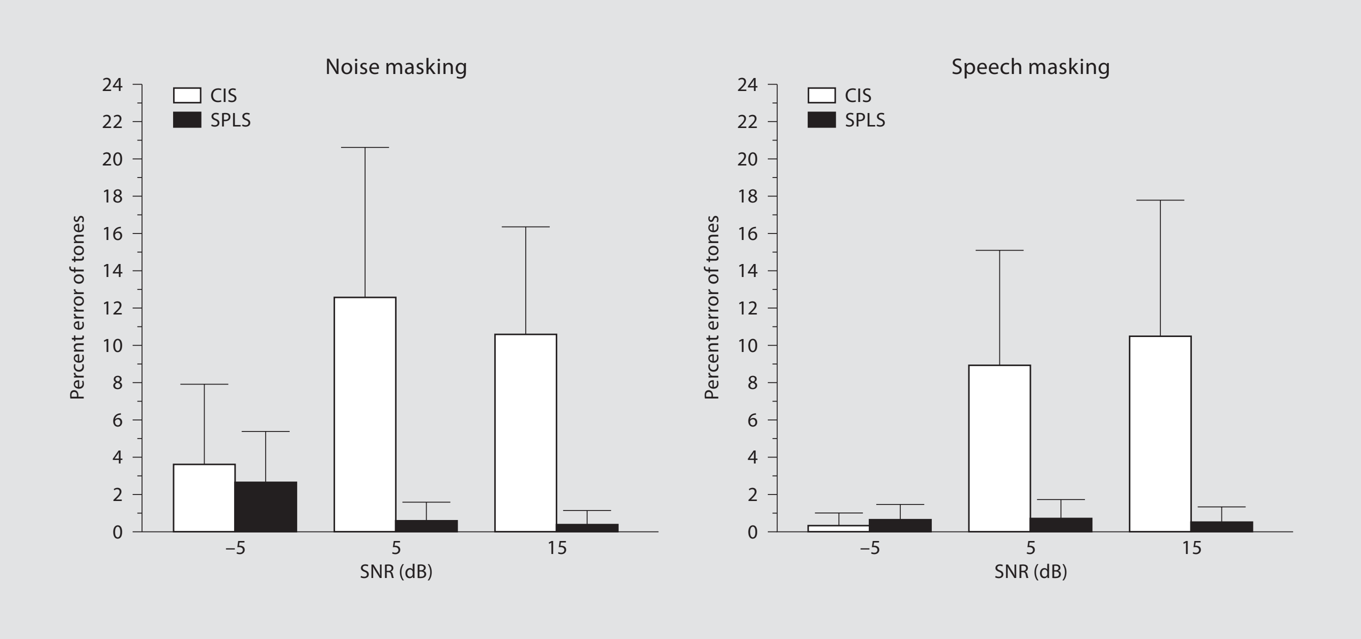

Fig. 3. Mean percent-error in recognition of tones across 12 subjects as a function of SNR for each of the 2 processing strategies under 2 masking conditions: steady-spectrum-noise masking and speech masking. The error bars indicate the SD of the mean.

ANOVA analysis shows that the difference is significant, F(1, 11) = 55.288, MSE = 372.09, p = 0.000.

As shown in figure 2, under masking conditions speech recognition increased with the increase of the SNR in all conditions, and the recognition of the target speech processed by SPLS was much larger than that processed by CIS in both noise and speech-masking conditions. The main effect of SNR was significant, F(1, 11) = 448.33, MSE = 4.642, p = 0.000, the main effect of processing strategy was significant, F(1, 11) = 656.473, MSE = 4.821, p = 0.000, and the main effect of masking type was significant, F(1, 11) = 102.406, MSE = 0.471, p = 0.000.

To examine whether the SPLS strategy was also beneficial to recognition of tones, we analyzed the 'tone error' in sentence repeating across 12 subjects. The percent error in recognizing tones was defined as the percentage of the number of Chinese characters whose syllable was correctly recognized but whose tone was not correctly pronounced out of the number of 108, which was the total number of keyword characters in each list. Under the quiet condition, the mean percent-error in recognition of tones was 0.29% for the SPLS strategy and 8.17% for the CIS strategy. Under masking conditions, the percent error in recognition of tones was much less for the SPLS strategy than for the CIS strategy. Under the low SNR condition (SNR = –5 dB), the difference between the SPLS strategy and the CIS strategy was not significant. However, when the SNR was increased to 5 or 15 dB, the percent error in recognition of tones was decreased more for the SPLS strategy than for the CIS strategy (fig. 3).

Discussion

As pointed out by Fu et al. [19], there are additional needs for developing speech-processing strategies to specifically improve functions of cochlear implant devices for recognizing tonal languages, such as Chinese. Phase information is presented in speech for normal listeners, and is important not only for sound localization, but also for signal recognition in noise [12]. In the present study, adding phase information with the SPLS method into target speech remarkably improved listeners' recognition performance in quiet. More importantly, additional phase information presented in target speech released the speech from noise and speech maskers.

It is well known that firings of the auditory nerve to pure tones are phase locked in the low-frequency range. CI devices create auditory sensation of sounds by directly stimulating the auditory nerve. If the interval of stimulation pulses at a stimulated site is modulated by phase information provided by the SPLS strategy developed in this study, the function of CI devices for processing tonal speech and even music would be improved. In addition, it would be interesting to study whether the SPLS is also beneficial for processing western languages, such as English and German, particularly under noisy listening conditions.

Lan et al. [20] also proposed a strategy for improving the performance of CI users speaking tonal languages, such as Chinese. In their speech-processing strategy, the envelopes of the narrow-band signals were extracted to modulate the amplitude of stimulation pulses, while the F0 was extracted to modulate the rate of stimulation pulses. The results of their psychophysical studies using normal-hearing listeners show that their processing strategy produced significantly better speech perception performance than the CIS strategy. In particular, the largest performance improvement occurred when the 4-channel processor was used in the test of sentence recognition. It should be noted that although both the SPLS strategy and the strategy of Lan et al. [20] can markedly improve the sentence recognition performance, there are several substantial differences. First, in the SPLS strategy, the modulation of stimulation pulses potentially used in CI devices temporally alters the stimulation pulses according to dynamic phase changes in speech but not F0. The better sentence recognition performance in quiet suggests that phase information is also important for recognizing Chinese speech. Second, in the study of Lan et al. [20], behavioral tests were not conducted under noisy conditions. It is not clear whether the fast Fourier transform-based F0 extraction algorithm still works appropriately when the target speech signal is embedded in noise or in competing speech, especially when the SNR is relatively low. The results of the present study clearly show that the SPLS strategy has a marked advantage in recognizing Chinese speech in both quiet and noisy environments.

In summary, the SPLS strategy not only improves the recognition of words, but also reduces errors in recognition of tones in Chinese speech. The limitation of the traditional CIS strategy is that it only encodes the speech envelope information. However, envelope information is not enough either for processing tonal language or for unmasking speech. The SPLS strategy developed in the present study can be used to encode both envelope information and phase information, and therefore has the potential to improve CI users' speech recognition, particularly in noisy environments.

Compared to previous phase-extracting strategies, such as MPEAK, the phase information of the SPLS strategy is extracted by zero-crossing detection from the waveform in each channel, not from the original full-band speech. So, in future we will conduct additional psychophysical experiments to study which phase-extracting strategy is more powerful in improving the performance of CI.

Acknowledgments

We are grateful to Ying Huang and Wen-Jie Wang for data collection, to Qiang Huang for technical support. This work was supported by China Natural Science Foundation for Young Scientists (60305004); the Key Project of China Natural Science Foundation (60535030, 60435010), the Project of New Century Excellent Talents in Universities, the Ministry of Science and Technology of China (No. 2002CCA01000), a '985' grant from Peking University, and the Canadian Institutes of Health Research.

References

1 Loizou PC: Mimicking the human ear. IEEE Signal Process Mag 1998;15:101–130.

2 Wilson B, Finley C, Lawson D, Wolford R, Eddington D, Rabinowitz W: Better speech recognition with cochlear implants. Nature 1991;352:236–238.

3 Boex C, Pelizzone M, Montandon A: Speech recognition with a CIS strategy for the Ineraid multichannel cochlear implant. Am J Otol 1996;17:61–68.

4 Shannon RV, Zeng FG, Kamath V: Speech recognition with primarily temporal cues. Science 1995;270:303–304.

5 Zeng FG: Cochlear implants in China. Audiology 1995;34:61–75.

6 Zeng FG, Cao KL, Wang ZZ: Progress in cochlear implants. Chin J Otolaryngol 1998;33: 123–125.

7 Xu L, Tsai YJ, Pfingst BE: Features of stimulation affecting tonal-speech perception: implication for cochlear prostheses. J Acoust Soc Am 2002;112:247–258.

8 Fu QJ, Zeng FG, Shannon RV: Importance of tonal envelope cues in Chinese speech recognition. J Acoust Soc Amer 1998;104:505–510.

9 Qin MK, Oxenham AJ: Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J Acoust Soc Am 2003;114:446–454.

10 Bregman AS: Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MIT, 1991.

11 Smith ZM, Delgutte B, Oxenham AJ: Chimaeric sounds reveal dichotomies in auditory perception. Nature 2002;416:87–90.

12 Clopton BM, Spelman FA: Technology and the future of cochlear implants. Ann Otol Rhinol Laryngol Suppl 2003;191:26–32.

13 Mitra SK: Digital Signal Processing: A Computer-Based Approach. New York, McGraw-Hill, 2002.

14 Roggero MA, Robles L, Rich NC, Costalupes JA: Basilar membrane motion and spike initiation in the cochlear nerve; in Moore BCJ, Patterson RD (eds): Auditory Frequency Selectivity. New York, Plenum, 1986.

15 Glasberg BR, Moore BCJ: Derivation of auditory filter shapes from notched-noise data. Hear Res 1990;47:103–138.

16 Helfer KS: Auditory and auditory-visual perception of clear and conversational speech. J Sp Lan Hear Res 1997;40:432–443.

17 Li L, Daneman M, Qi JG, Schneider BA: Does the information content of an irrelevant source differentially affect speech recognition in younger and older adults? J Exp Psychol Hum Percept Perform 2004;30:1077–1091.

18 Wu XH, Wang C, Chen J, Qu HW, Li WR, Wu YH, Schneider BA, Li L: The effect of perceived spatial separation on informational masking of Chinese speech. Hear Res 2005;199:1–10.

19 Fu QJ, Hsu CJ, Horng MJ: Effects of speech processing strategy on Chinese tone recognition by Nucleus-24 cochlear implant users. Ear Hear 2004;25:501–508.

20 Lan N, Nie KB, Gao SK, Zeng FG: A novel speech processing strategy incorporating tonal information for cochlear implants. IEEE Trans Biomed Eng 2004;51:752–760.