A New Acoustic Model Incorporating Temporal Fine Structure Cue for Cochlear Implant

Fei Chen, *Member, IEEE,* and Yuan-ting Zhang, *Senior Member, IEEE*

***Abstract*— This paper introduces a new acoustic model incorporating temporal fine structure cue for cochlear implant (CI) in order to enhance the pitch perception of CI users. After bandpass filtering the speech signal into multiple frequency bands, a carrier signal is constructed for each band by using a train of high-rate sinusoidal pulses located at the peak positions of the fine structure. The carrier signal is then amplitude-modulated by the temporal envelope in the band so as to produce the band-specific decomposition output. Acoustic simulation experiment was conducted to investigate the contribution of the model-based speech processor for Mandarin tone identification. 12 Mandarin-speaking subjects participated in the experiment, and each subject listened to 40 sounds synthesized by the continuous-interleaved-sampling (CIS) processor and the proposed processor in their 6-band versions, respectively. The mean correct rates were 87.9% and 96.3% by using the CIS processor and the proposed processor, respectively, which indicated that the model-based speech processor might noticeably enhance the Mandarin tone identification. Therefore, it is believed that the proposed acoustic model would facilitate our design of novel CI speech processors to further improve the speech perception of CI users, particularly those speaking tonal languages.**

***Index Terms*— Temporal fine structure, acoustic simulation, cochlear implant.**

I. Introduction

Cochlear implant (CI) has been long accepted as the only medical intervention that could restore partial hearing to a profoundly deafened person [1]. Up to date, the CI technology has progressed to the stage that allows the majority of CI users to understand the conversational speech in quite surroundings, and talk on the phone. However, several studies have found that current CI subjects had difficulties in the recognition of music melodies and speakers, or other tasks requiring pitch perception [2]-[4]. Recently, it has attracted considerable attentions to improve the pitch perception for a large amount of potential CI users speaking tonal languages, particularly Mandarin [5]-[6]. Tonal languages are different from mono-tonal languages, e.g. English and German, as such that they use different tones to express the lexical meaning of the pronounced words. For instance, there are four tonal patterns in Mandarin, i.e. flat tone, rising tone, falling-rising tone, and falling tone. Studies have reported that cochlear implantees speaking Chinese showed poor results in identifying vowels and consonants compared with those speaking English [5]. Wei *et al.* investigated the Mandarin tone recognition of Chinese CI users. Their results suggested that the current CI speech processors did not provide sufficient acoustic cues to support adequate tone recognition [7].

According to Hilbert transform, a signal can be decomposed into a slowly varying envelope modulating a high-frequency carrier, or fine structure, of the original signal. Temporal envelope and fine structure have been recognized as two important acoustic cues for speech intelligibility [8]. Most of the current CI speech processors, such as the well-known continuous-interleaved-sampling (CIS) processor [9], basically emphasize the envelope cue, since it has been evidenced that using the slowly-varying temporal envelope extracted from only 3 to 4 frequency bands might produce a nearly perfect speech recognition in the quite environment [10]. However, the fine structure has been recently found to contain more acoustic cue for pitch recognition than the temporal envelope [11]. Xu and Pfingst studied the relative importance of temporal envelope and fine structure in the lexical-tone perception. They found that the lexical-tone recognition of Mandarin depended on the fine structure rather than the envelope when the number of frequency bands was between 4 and 16 [12].

Based on the contribution of fine structure for pitch perception in the acoustic simulation from normal-hearing subjects, it has been widely believed that modifying CI processor to deliver fine structure might improve the pitch perception of the implant patients. Studies are actively ongoing to seek effective strategies to extract the fine structure cue and, more importantly, encode it into the electrical signal used in CI, which might lead a breakthrough in CI design [13]. Proposing a novel acoustic model may facilitate the development of the novel speech processor to enhance the tone identification and melody appreciation of CI users. Therefore, the objective of this paper is to introduce a new acoustic model incorporating temporal fine structure cue for

Manuscript received September 6, 2006. This work was supported by the Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong.

Fei Chen is with the Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong, Hong Kong.

Yuan-ting Zhang is with the Shun Hing Institute of Advanced Engineering, and the Joint Research Centre for Biomedical Engineering, Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong (corresponding author with e-mail: ytzhang@ee.cuhk.edu.hk; phone: +852 2609 8459; fax: +852 2603 5558).

CI, and also investigate its contribution for Mandarin tone identification in the acoustic simulation experiment.

II. Methodology

A. Hilbert Transform

For a real signal s_r(t), an analytic signal s(t) can be generated as

$$s(t) = s_r(t) + is_i(t), \tag{1}$$

where i is the imaginary number (i.e. $\sqrt{-1}$), and s_i(t) is the Hilbert transform of s_r(t). The Hilbert envelope is the magnitude of the analytic signal, as

$$a(t) = \sqrt{s_r^2(t) + s_i^2(t)}. \tag{2}$$

The Hilbert fine structure is $cos(\Phi(t))$, where Φ(t) is the phase of the analytic signal, as

$$\Phi(t) = tg^{-1}(\frac{s_i(t)}{s_r(t)}). \tag{3}$$

B. New Acoustic Model Incorporating Temporal Fine Structure Cue

The diagram of the proposed acoustic model is shown in Fig. 1. The model includes both speech decomposition and speech synthesis. The sampled speech is pre-amplified and decomposed into several frequency bands by using a bank of bandpass filters. Hilbert transform is then applied to get the temporal envelope a_j(t) and the fine structure Φ_j(t) of the decomposed signal in band j. A carrier signal C_j(t) is constructed by using a train of high-rate sinusoidal pulses, c(t), whose peaks are located at the peak positions of the temporal fine structure signal Φ_j(t), as

$$C_j(t) = \sum_k c(t - t_k), \tag{4}$$

where t_k is the occurrence time of the kth peak in Φ_j(t). Fig. 2 illustrates the construction of the carrier signal.

The carrier signal is then amplitude-modulated by the temporal envelope in the band to produce the band-specific decomposition output s'_j(t), as

$$s'_j(t) = a_j(t) \cdot C_j(t). \tag{5}$$

The speech synthesis is implemented by summing the above decomposition outputs from all the bands.

III. Experimental protocol

A CI speech processor was built up based on the proposed acoustic model. Acoustic simulation experiment was conducted to investigate the contribution of the model- based speech processor for Mandarin tone identification. The performance of the proposed processor was also compared with that of the CIS processor. It is noted that, for the acoustic simulation of the CIS processor, the carrier signal in each band used the sinusoid at the center frequency of the bandpass filter in that band [14]. 100 different Mandarin monosyllables, including a variety of combination of consonants and vowels, were recorded in a quiet environment pronounced by a male Mandarin speaker. The test voices were sampled at 22050 Hz, and decomposed into 6 frequency bands with cutoff frequencies of 25, 260, 600, 1240, 2420, 4650, and 8820 Hz.

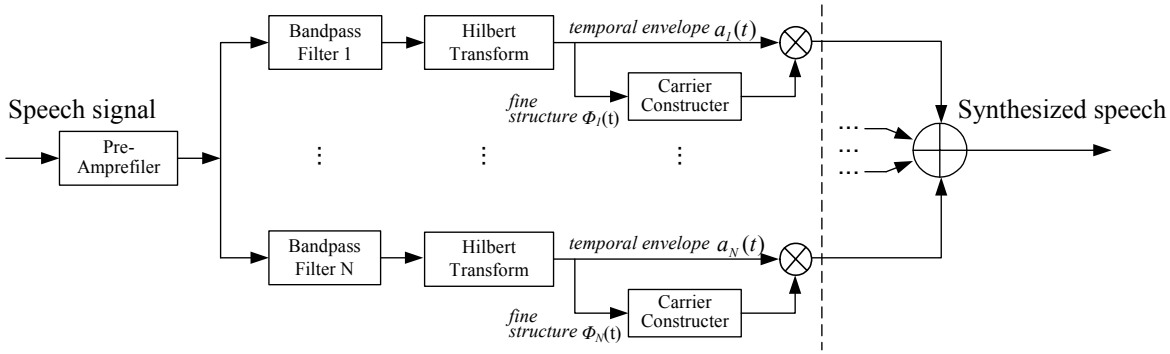

Fig. 1. The block diagram of the proposed acoustic model. The speech decomposition and synthesis parts are separated by the dashed line.

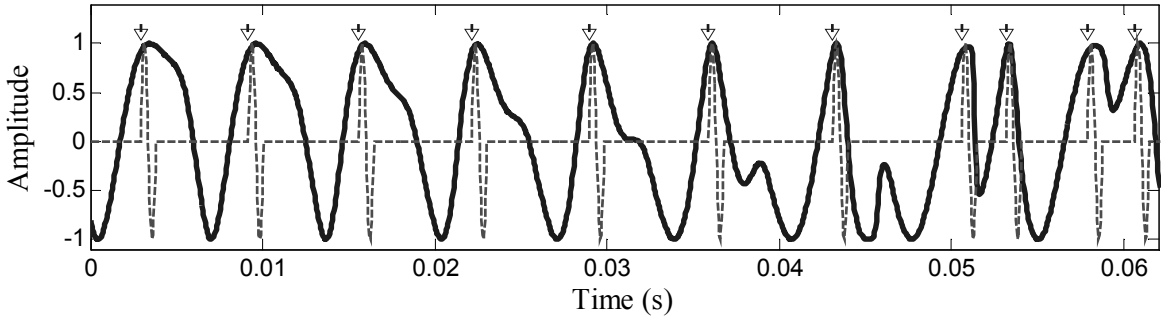

Fig. 2. The construction of the carrier signal C_j(t). The top arrows indicate the peak positions of the temporal fine structure signal Φ_j(t), and the high-rate sinusoidal pulses are located at the positions indicated by arrows. The period of the sinusoidal pulse is 0.9 ms.

12 Mandarin-speaking subjects, 4 male and 8 female adults from 25 to 35 years old, participated in the experiment. Before the test, each subject was administrated to a training session to be familiar with a couple of synthesized sounds. During the test, each subject listened to two sets of synthesized sounds generated by the CIS processor and the proposed processor, respectively, and completed a four-alternative forced-choice tone identification task. Each sound set contained 40 synthesized monosyllabic voices at a comfortable loudness.

IV. Experimental Results

Fig. 3 shows the performances of Mandarin tone identification by the CIS processor and the proposed processor. It is observed that the correct rate of the CIS processor is $87.9 \pm 13.3\%$, while that of the proposed processor is $96.3 \pm 4.9\%$. The higher correct rate and smaller standard deviation from the proposed processor suggested that the proposed speech processor performed better than the CIS processor in improving Mandarin tone identification. Furthermore, the paired *t*-test was calculated to determine the performance difference between the two processors. A scale of 100% was used to quantify the performance in identifying Mandarin tone. The null hypothesis that there was no performance difference between the two processors was accepted or rejected at a significance level of 95% confidence ($\alpha = 0.05$). The paired *t*-test result p=0.021 (<0.05) confirmed that the difference between the two processors was significant for Mandarin tone identification.

Fig. 4 gives the correct rates of identifying each Mandarin tonal pattern by the two processors. For identifying the synthesized sounds with the first Mandarin tonal pattern, results of the CIS processor and the proposed processor showed comparable correct rates. However, to recognize the synthesized voices with other three Mandarin tonal patterns, it is seen that the proposed processor consistently produced higher correct rates than the CIS processor. Particularly, a correct rate of $100 \pm 0\%$ was achieved in identifying the fourth Mandarin tonal pattern by the proposed processor.

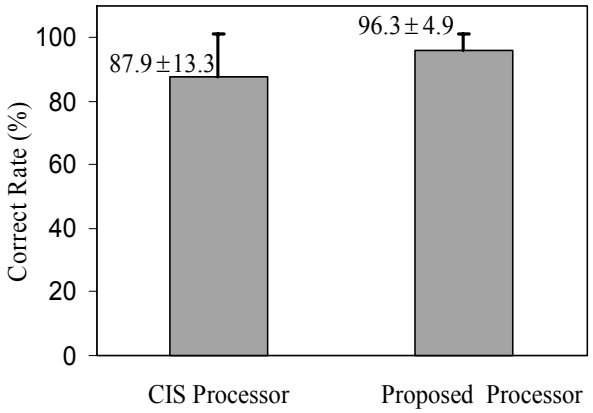

Fig. 3. The correct rates of Mandarin tone identification by the CIS processor and the proposed processor, respectively. The error bars represent the standard deviations in the test.

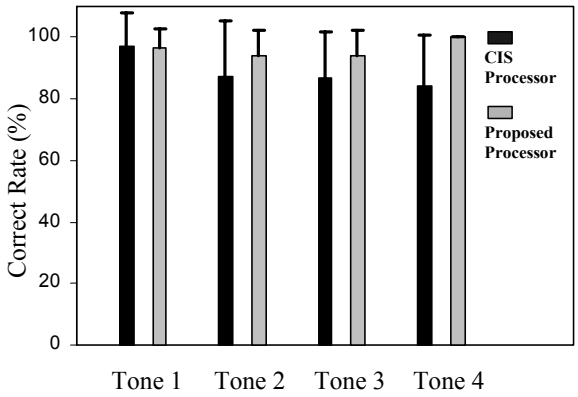

Fig. 4. The correct rates of identifying each Mandarin tonal pattern by the CIS processor and the proposed processor, respectively.

V. Discussion

It has been suggested that neither temporal nor spectral cues have been adequately and appropriately extracted and encoded in current CI devices [7]. Current CI devices can deliver up to 22 frequency channels. However, it was found that CI users could not functionally use more than 8 frequency channels [15]-[16]. Therefore, it was believed unlikely that electrode modifications in the near future would deliver the number of channels necessary to produce high levels of pitch perception using current signal processing techniques [17]. On the other hand, improving the fine structure encoding seems feasible and possible to soon produce improvements in pitch perception, and has been suggested as a new direction in designing CI speech processor [13].

In the proposed acoustic model, the fine structure was characterized by its peak position, and represented by a train of high-rate sinusoidal pulses. The peak interval carries the information on the instantaneous frequency, so that the proposed peak-position encoding method is compatible to the suggested frequency modulation (FM) strategy to deliver pitch information [5]-[6]. Experimental results in this paper also confirmed that the fine structure cue from peak-position could effectively convey tonal information and enhance Mandarin tone identification.

In order to extract the fine structure cue, several other acoustic models have been recently proposed [5], [6], [17]. Since the fundamental frequency (F0) of Mandarin speech was found to be associated with the perceived pitch, the trajectory of the F0 has been extracted, and used to modulate the frequency of the carrier signal in each band [5]. Nie *et al.* proposed a frequency-amplitude-modulation- encoding strategy. Besides preserving the amplitude modulation by the temporal envelope, they transformed the fast-varying fine structure into a slowly-varying FM signal to modulate the

center frequency in each band [6]. Both of the above two acoustic models used the carrier signals of the continuous triangular functions with frequency modulation. Otherwise, the proposed acoustic model produced the carrier signal by applying a train of high-rate sinusoidal pulses, as shown in Fig. 2. As we know, present CI devices use electrical pulse train to stimulate the auditory nerves so as to elicit speech perception [14]. Therefore, the proposed model features a more effective simulation of the process of electrical pulse stimulation in the actual CI devices than those models using continuous carrier signal. It may also be potential to put forward a novel acoustic simulation platform to study the effect of pulse-width on speech perception. The significance of pulse-width on speech recognition has been recently reported by Loizou *et al.* They found that, among several speech processing parameters, pulse-width, together with pulse-rate, had the largest effect on speech recognition [18]. Therefore, besides being able to deliver the temporal fine structure cue, the proposed model is potential to facilitate our study on the effect of pulse-width for speech perception in the future.

VI. Conclusion

This paper introduced a new acoustic model incorporating temporal fine structure cue for CI in order to enhance the pitch perception of CI users. The results from the acoustic simulation experiment showed that the model- based speech processor was able to noticeably improve the correct rate of Mandarin tone identification, which would facilitate our design of novel CI speech processors to further enhance the speech perception of CI users, particularly those speaking tonal languages.

References

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

[2] Y.Y. Kong, R. Cruz, J.A. Jones, and F.G. Zeng., "Music perception with temporal cues in acoustic and electric hearing," *Ear. Hearing*, vol. 25, pp. 173-185, April 2004.

[3] M. Vongphoe and F.G. Zeng, "Speaker recognition with temporal cues in acoustic and electric hearing," *J. Acoust. Soc. Amer.*, vol. 118, pp. 1055-1061, August 2005.

[4] B. Townshend, N. Cotter, and D.V. Compernoll, "Pitch perception by cochlear implant subjects," *J. Acoust. Soc. Amer.*, vol. 82, pp. 106-115, July 1987.

[5] N. Lan, K.B. Nie, S.K. Gao, and F.G. Zeng, "A novel speech- processing strategy incorporating tonal information for cochlear implants," *IEEE Trans. Biomed. Eng.*, vol. 5, pp. 752-760, May 2004.

[6] K.B. Nie, G. Stickney, and F.G. Zeng, "Encoding frequency modulation to improve cochlear implant performance in noise," *IEEE Trans. Biomed. Eng.*, vol. 52, pp. 64-73, January 2005.

[7] C.G. Wei, K.L Cao, and F.G. Zeng, "Mandarin tone recognition in cochlear-implant subjects," *Hear. Res.*, vol. 197, pp. 87-95, November 2004.

[8] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 97, pp. 585-592, January 1995.

[9] P.C. Loizou, "Signal-processing techniques for cochlear implants," *IEEE Eng. Med. Biol. Mag.*, vol. 18 (3), pp. 34-46, May-June 1999.

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

[11] Z.M. Smith, B. Delgutte, and A.J. Oxenham. "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, pp. 87-90, March 2002.

[12] L. Xu and B.E. Pfingst, "Relative importance of temporal envelope and fine structure in lexical-tone perception," *J. Acoust. Soc. Amer.*, vol. 114, pp. 3024-3027, December 2003.

[13] B.S. Wilson, R. Schatzer, E.A. Lopez-Poveda, X.A. Sun, D.T. Lawson, and R.D. Wolford, "Two new directions in speech processor design for cochlear implants," *Ear. Hearing*, vol. 26, pp. 73-81, August 2005.

[14] P.C. Loizou, "Introduction to cochlear implants," *IEEE Eng. Med. Biol. Mag.*, vol. 18 (1), pp. 32-42, January-February 1999.

[15] K. Fishman, R.V. Shannon, and W.H. Slattery, "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor," *J. Speech Hear. Res.*, vol. 40, pp. 1201-1215, October 1997.

[16] L. Friesen, R.V. Shannon, D. Baskent, and X.S. Wang, "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *J. Acoust. Soc. Amer.*, vol. 110, pp. 1150-1163, August 2001.

[17] J.T. Rubinstein and C. Turner, "A novel acoustic simulation of cochlear implant hearing: Effects of temporal fine structure," in *Proc. of the 1st IEEE EMBS Conference on Neural Eng.*, pp. 142-145, 2003.

[18] P.C. Loizou, O. Poroy, and M. Dorman, "The effect of parametric variations of cochlear implant processors on speech understanding," *J. Acoust. Soc. Amer.*, vol. 108, pp. 790-802, August 2000.