Signal Processing 88 (2008) 2693–2699

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

A novel temporal fine structure-based speech synthesis model for cochlear implant ☆

Fei Chen [a,b,*], Yuan-Ting Zhang [a,b,c,d]

[a] Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong
[b] Joint Research Centre for Biomedical Engineering, Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong
[c] Institute of Biomedical and Health Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
[d] Key Laboratory for Biomedical Informatics and Health Engineering, Chinese Academy of Sciences, China

ARTICLE INFO

Article history:
Received 20 August 2007
Received in revised form
17 May 2008
Accepted 20 May 2008
Available online 28 May 2008

Keywords:
Temporal fine structure
Speech synthesis
Mandarin tone identification
Speech intelligibility
Speaker recognition
Cochlear implant

ABSTRACT

It has widely been recognized that delivering the temporal fine structure information to the cochlear implant (CI) users might significantly improve their speech perception performance and lead a breakthrough for the CI design. This paper introduces a new speech synthesis model incorporating the temporal fine structure cue for CI. After bandpass filtering the speech signal, band-specific carriers are constructed by placing the high-rate sinusoidal pulses at the peak positions of the fine structures. The carriers are then amplitude-modulated by their envelopes, and summed to generate the synthesized speech. Mandarin-speaking subjects participated in the acoustic simulation experiment by listening to the voices synthesized by the continuous-interleaved-sampling (CIS) processor and the model-based algorithm in their six-band versions. The experimental results indicated that the model-based algorithm produced significant improvements in Mandarin tone identification, subjective assessment of speech intelligibility and speaker recognition. The proposed model should be helpful for the development of novel CI speech processing strategies to improve the speech perception of cochlear implantees, particularly those speaking tonal languages.

1. Introduction

Cochlear implant (CI) has long been accepted as the only medical treatment to restore partial hearing to a severely-to-profoundly deafened person [1,2]. CI device inserts the electrodes into the scala tympani of the cochlea, bypasses the hair cells and directly stimulates the residual auditory nerves (ANs). Although the CI technology has advanced from its early single-channel, analog to current multi-channel, pulsatile stimulation state, its speech processor still plays an important role for the effective extraction and delivery of the information from the speech signal.

Temporal envelope and fine structure (FS) have been recognized as two important acoustic cues for speech intelligibility [3,4]. Most of the current CI speech processors, exemplified by the well-known continuous-interleaved-sampling (CIS) processor [5], basically emphasize the envelope cue while discarding the temporal FS, which is supported by the fact that, in the quiet environment, only four channels of envelope information were adequate to produce high levels of speech intelligibility [6]. However, recent studies have discovered that CI subjects faced difficulties in speech identification in the presence of background noise, the recognition of music melodies and speakers, or other tasks requiring pitch perception [7–10], which motivated the work to improve the pitch

☆ Part of this work was presented at the 5th International Special Topic Conference on ITAB in Ioannina, Greece, 2006.

* Corresponding author at: Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong.
E-mail addresses: fchen@ee.cuhk.edu.hk (F. Chen), ytzhang@ee.cuhk.edu.hk (Y.-T. Zhang).

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.05.011

perception of CI users, particularly the tone identification for a large amount of potential cochlear implantees speaking tonal languages, such as Mandarin [11–16]. Tonal languages are different from mono-tonal languages, e.g. English, as such that they use different tones to express the lexical meanings of the pronounced words. Recently, it was found that the FS was more important for pitch recognition than the temporal envelope [3]. Xu et al. reported that the lexical-tone recognition of Mandarin depended on the FS rather than the envelope when the number of frequency bands was between 4 and 16 [17].

Nowadays, studies have been actively ongoing to seek effective strategies to extract the FS cue and, more importantly, represent it into a way utilizable for CI devices [4,11,14–15,18]. The acoustic simulation of CI has been long used to evaluate the performance of the novel speech processing strategy. It is designed to synthesize a speech signal according to the CI speech processing strategy, and present the synthesized degraded speech to the normal-hearing ear, which simulates the hearing process of CI users. Rubinstein et al. introduced a method to generate an "FS noise" by taking a bandlimited noise signal and extracting its envelope to modulate the randomized FS of the input signal [18]. However, to our knowledge, its performance, such as in pitch perception, has not been widely demonstrated. Nie et al. proposed a frequency–amplitude-modulation-encoding (FAME) strategy decomposing speech into the amplitude modulation (AM) and frequency modulation (FM) components [11]. Similarly, based on the association between the fundamental frequency (F0) of the Mandarin speech signal and the perceived pitch, the trajectory of the F0 was extracted to convey the tonal information by Lan et al. [12]. The continuous functions, such as the triangular functions, were used to synthesize the voice during the acoustic simulation, whose behavior differed from the pulsatile electrical stimulation strategy in present CI devices to convey the acoustic information [1].

The purpose of this paper is to introduce a new temporal FS-based speech synthesis model for CI in order to develop novel speech processor for CI users speaking tonal languages, particularly Mandarin. The model extracts the FS cue by using the characteristic points straight from the temporal domain, and represents it in a way analogous to the pulsatile stimulation process in present CI devices. The acoustic simulation experiments were conducted to evaluate the contribution of the model-based algorithm for Mandarin speech perception, i.e. Mandarin tone identification, subjective assessment of speech intelligibility and speaker recognition.

2. A temporal FS-based speech synthesis model

2.1. Hilbert transform

The Hilbert transform (HT) implements the process to mathematically decompose a real signal into a slowly varying signal (envelope) modulating a high-frequency signal (carrier). Briefly, an analytic signal $s(t)$ can be generated from a real signal $s_r(t)$ as

$$s(t) = s_r(t) + \mathrm{i}s_i(t), \tag{1}$$

where i is the imaginary number, i.e. $\sqrt{-1}$, and $s_i(t)$ is the HT of $s_r(t)$ [2,3,19], as

$$s_i(t) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s_r(\tau)}{t-\tau}\,\mathrm{d}\tau. \tag{2}$$

The Hilbert envelope is defined as the magnitude of the analytic signal $s(t)$, as

$$a(t) = \sqrt{s_r^2(t) + s_i^2(t)}, \tag{3}$$

while the Hilbert FS or the carrier is $\cos(\Phi(t))$, where $\Phi(t)$ is the phase of the analytic signal $s(t)$, as

$$\Phi(t) = a\tan\left(\frac{s_i(t)}{s_r(t)}\right). \tag{4}$$

The FS cue is critical for speech recognition in quiet, and mostly in background noise [20]. A number of studies have also explored that the FS information carries more important cue for pitch perception and sound localization than the temporal envelope [3,17].

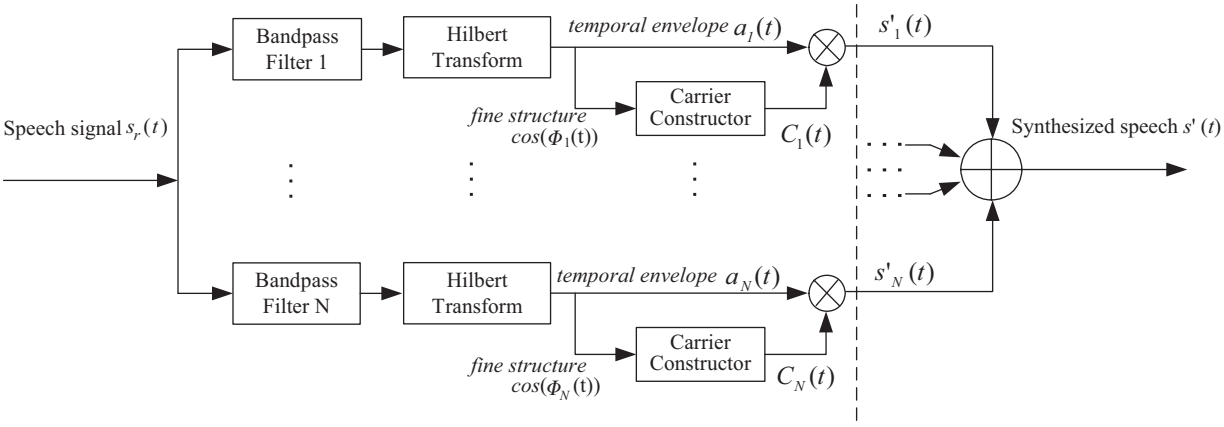

Fig. 1. The block diagram of the proposed speech synthesis model.