Taylor & Francis
Taylor & Francis Group

Cochlear Implants International
An Interdisciplinary Journal

ISSN: 1467-0100 (Print) 1754-7628 (Online) Journal homepage: http://www.tandfonline.com/loi/ycii20

Implementation and preliminary evaluation of 'C-tone': A novel algorithm to improve lexical tone recognition in Mandarin-speaking cochlear implant users

Lichuan Ping, Ningyuan Wang, Guofang Tang, Thomas Lu, Li Yin, Wenhe Tu & Qian-Jie Fu

To cite this article: Lichuan Ping, Ningyuan Wang, Guofang Tang, Thomas Lu, Li Yin, Wenhe Tu & Qian-Jie Fu (2017) Implementation and preliminary evaluation of 'C-tone': A novel algorithm to improve lexical tone recognition in Mandarin-speaking cochlear implant users, Cochlear Implants International, 18:5, 240-249, DOI: 10.1080/14670100.2017.1339492

To link to this article: https://doi.org/10.1080/14670100.2017.1339492

Implementation and preliminary evaluation of 'C-tone': A novel algorithm to improve lexical tone recognition in Mandarin-speaking cochlear implant users

Lichuan Ping[1], Ningyuan Wang[2], Guofang Tang[2], Thomas Lu[1], Li Yin[2], Wenhe Tu[2], Qian-Jie Fu[3]

[1]Nurotron Biotechnology, Inc., Irvine, CA, USA, [2]Zhejiang Nurotron Biotechnology Co., Ltd, Zhejiang, PR China, [3]Department of Head and Neck Surgery, David Geffen School of Medicine, UCLA, Los Angeles, 2100 West Third Street, Suite 100, Los Angeles, CA 90057, USA

Objectives: Because of limited spectral resolution, Mandarin-speaking cochlear implant (CI) users have difficulty perceiving fundamental frequency (F0) cues that are important to lexical tone recognition. To improve Mandarin tone recognition in CI users, we implemented and evaluated a novel real-time algorithm (C-tone) to enhance the amplitude contour, which is strongly correlated with the F0 contour.

Methods: The C-tone algorithm was implemented in clinical processors and evaluated in eight users of the Nurotron NSP-60 CI system. Subjects were given 2 weeks of experience with C-tone. Recognition of Chinese tones, monosyllables, and disyllables in quiet was measured with and without the C-tone algorithm. Subjective quality ratings were also obtained for C-tone.

Results: After 2 weeks of experience with C-tone, there were small but significant improvements in recognition of lexical tones, monosyllables, and disyllables ($P < 0.05$ in all cases). Among lexical tones, the largest improvements were observed for Tone 3 (falling–rising) and the smallest for Tone 4 (falling). Improvements with C-tone were greater for disyllables than for monosyllables. Subjective quality ratings showed no strong preference for or against C-tone, except for perception of own voice, where C-tone was preferred.

Discussion: The real-time C-tone algorithm provided small but significant improvements for speech performance in quiet with no change in sound quality. Pre-processing algorithms to reduce noise and better real-time F0 extraction would improve the benefits of C-tone in complex listening environments.

Conclusions: Chinese CI users' speech recognition in quiet can be significantly improved by modifying the amplitude contour to better resemble the F0 contour.

Keywords: Cochlear implant, Lexical tone, Mandarin speech, C-tone, Nurotron

Introduction

While 4–8 spectral channels may be sufficient for cochlear implant (CI) listeners to understand speech in quiet (Shannon *et al.*, 1995; Wilson *et al.*, 1991), many more channels are required for complex listening tasks and conditions (Shannon *et al.*, 2004; Smith *et al.*, 2002). Owing to channel interaction, CI users' functional spectral resolution is much poorer than the number of implanted electrodes (Friesen *et al.*, 2001). The coarse spectral resolution restricts CI listeners' access to fine-structure cues that are important to complex pitch perception, which in turn limits perception of speech in noise (Fu and Nogaki, 2005; Nelson and Jin, 2004; Nelson *et al.*, 2003), music (Galvin *et al.*, 2007, 2009), speech prosody (Chatterjee and Peng, 2008), vocal emotion (Luo *et al.*, 2007), and lexical tones (Fu and Zeng, 2000; Fu *et al.*, 1998; Wang *et al.*, 2011, 2012; Xu *et al.*, 2011).

Mandarin Chinese is one of the most popular languages in the world, and there are a growing number of Mandarin-speaking CI users. Unlike English, Mandarin Chinese is a tonal language, in which syllables with the same vowel–consonant combination but produced with different tones can convey different meanings (Liang, 1963; Lin, 1988).

Correspondence to: Qian-Jie Fu, Department of Head and Neck Surgery, David Geffen School of Medicine, UCLA, 2100 West Third Street, Suite 100, Los Angeles, CA 90057, USA.
Email: qfu@mednet.ucla.edu

DOI 10.1080/14670100.2017.1339492

There are four tonal patterns in Mandarin Chinese, which are characterized by the variation in fundamental frequency (F0) or F0 contours during voiced speech: Tone 1 (high–flat), Tone 2 (mid-rising), Tone 3 (low-falling–rising), and Tone 4 (high-falling). F0 is the most dominant acoustic cue for tone recognition (Abramson, 1972). Previous studies have shown that Mandarin-speaking CI users are capable of performing moderately good tone recognition (Fu *et al.*, 2004; Luo *et al.*, 2008, 2009; Tao *et al.*, 2015; Wang *et al.*, 2011; Xu *et al.*, 2011). CI performance has been shown to be similar to that of normal hearing (NH) subjects listening to acoustic CI simulations with 1–6 channels (Fu and Zeng 2000; Fu *et al.*, 1998; Xu *et al.*, 2002).

Several commercial CI signal processing strategies have attempted to improve the spectro-temporal resolution, but have shown no significant benefit for Chinese CI users' tone recognition (e.g. Advanced Bionics' HiRes 120 versus HiRes strategies in Han *et al.*, 2009; MED-EL's FSP versus CIS strategies in Schatzer *et al.*, 2010). Various research groups have also evaluated experimental CI signal processing strategies aimed at improving perception of temporal periodicity cues. Milczynski *et al.* (2012) measured Chinese tone recognition with the clinical ACE strategy and the experimental F0 modulation (F0mod) strategy, which aims to enhance amplitude modulation information across channels (Geurts and Wouters, 2004; Laneau *et al.*, 2006; Milczynski *et al.*, 2009). Results showed significantly better tone recognition with a male talker for the F0mod strategy, but no significant improvement for sentence recognition. Similar improvements for pitch perception in CI users were found by Vandali and van Hoesel (2012) for the experimental enhanced-envelope-encoder (eTone) strategy over the clinical ACE strategy. For 8-channel noise-band CI simulations, Green et al. (2005) found that sharpening amplitude modulation improved pitch perception, but worsened vowel recognition.

Although the principal acoustic feature for Chinese tones is the F0 contour (Abramson, 1972), other acoustic characteristics may co-vary with changes in F0. Vowel duration differs across tones (Fu and Shannon, 1998; Fu and Zeng, 2000), with the longest duration typically for Tone 3 and the shortest duration for Tone 4. The overall peak amplitude also varies across tones, with the lowest peak amplitude typically for Tone 3 and the highest for Tone 4 (Lin, 1988; Massaro *et al.*, 1985). Significant correlations have been observed between the amplitude contour and F0 contour (Fu and Zeng, 2000; Garding *et al.*, 1986; Sagart, 1986; Whalen and Xu, 1992). When F0 cues are available, as in the NH case, the contribution of duration and amplitude cues to Mandarin tone perception is negligible (Lin, 1988). However, some tone recognition is possible using duration, periodicity, and amplitude contour cues when F0 cues are partially or totally removed (Fu and Zeng, 2000; Fu *et al.*, 1998; Liang, 1963; Whalen and Xu, 1992). Luo and Fu (2004) found significantly better tone recognition in NH subjects listening to CI simulations when amplitude contour was modified to better resemble the F0 contour; different from Green *et al.* (2005), vowel recognition was largely unaffected by the modification.

Given the limited access to F0 cues, enhancement of the amplitude contour that co-varies with F0 may improve Chinese CI users' lexical tone recognition. In this study, a novel real-time tone enhancement algorithm (C-tone) was implemented and evaluated in users of the Nurotron NSP-60 CI system (Gao *et al.*, 2016; Zeng *et al.*, 2015). In the C-tone algorithm, the F0 contour is used to enhance the co-varying amplitude contour in real time. Recognition of Chinese tones, monosyllables, and disyllables in quiet were measured with the default clinical strategy (Advanced Peak Selection, or APS), with and without the C-tone algorithm. Subjective quality ratings were also obtained for C-tone.

Methods

Participants

Eight post-lingually deafened adult (2 males, 6 females; mean age = 44.4 years. range = 37–49 years) CI users participated in the study; all were native speakers of Mandarin Chinese. All CI subjects had at least 1 year of experience with their device; all used the Nurotron NSP-60B/C device, all used the APS strategy in their clinical processors, and all but subject S2 were implanted on the right side. Detailed demographic information for CI subjects is shown in Table 1. All subjects provided informed consent and all were paid for their participation.

Table 1 CI subject demographic information

Subject	Gender	Etiology	Age at testing (yrs)	CI experience (yrs)	Nurotron device (ear)
S1	M	Head trauma	47	3.1	NSP-60B (R)
S2	F	Unknown	45	2.1	NSP-60B (L)
S3	F	Unknown	49	3.1	NSP-60C (R)
S4	M	Drug induced	43	2.3	NSP-60B (R)
S5	F	Unknown	37	2.2	NSP-60B (R)
S6	F	Unknown	42	1.7	NSP-60B (R)
S7	F	Drug induced	46	3.5	NSP-60B (R)
S8	F	Unknown	46	1.4	NSP-60B (R)

APS and C-tone

Figure 1 shows block diagrams of the APS strategy and C-tone algorithm. APS is the default clinical sound processing strategy for the Nurotron CI, and is an '*n-of-m*' strategy. In APS, the spectral envelope of the acoustic input is first extracted by the Fast Fourier Transform (FFT) and grouped into 'm' analysis bands based on the corner frequencies of these spectral channels; the value of 'm' is typically equal to the number of active electrodes ($m = 24$ in the Nurotron device). In each stimulation cycle, the temporal envelope is extracted from each band and the 'n' number of bands with the largest amplitude are selected ($n < m$) for stimulation (typically, $n = 8$ in the Nurotron device). The amplitude contour from the selected bands is logarithmically compressed to scale the large acoustic dynamic range (DR) to fit within the small electric DR. The stimulation rate per channel in APS ranges from 510 to 1024 pulses per second (pps), with a default setting of 890 pps/channel. Trains of biphasic pulses are interleaved among the n channels in each stimulation frame.

The C-tone algorithm modifies the amplitude contour based on the F0 contour, and was implemented as a pre-processing scheme within the APS strategy. As shown in Fig. 1, there are three functional blocks of C-tone: (1) F0 contour tracking module, (2) noise floor estimator, and (3) envelope modification decision module.

(1) Fo contour tracking module

The F0 contour tracking module estimates the rate of change in F0 (i.e. a difference function algorithm) rather than the exact value of F0. This module consists of three steps: (1) pre-processing, (2) decimated difference (*dd*) function, and (3) post-processing.

In the pre-processing step, the input signal is first down-sampled from 16 to 8 kHz. A 4th-order Butterworth low-pass filter (LPF; cut-off frequency: 900 Hz; filter slope: −25 dB/octave) is then used to remove high-order harmonics. Next, the first derivative of the signal, derived by a 2-point central difference (Bahill *et al.*, 1982), is used to minimize the DR of the input signal while not affecting its periodicity; the first derivative could also be considered to be a high-pass filter (HPF) that reduces DC noise and very-low-frequency components of signal.

In the *dd* function step, a low-complexity modification of difference function (Molero *et al.*, 2011) was implemented by introducing a decimating factor S:

$$dd(i, \tau) = \sum_{i=1}^{W/S} |x(i*S) - x(i*S + \tau)| \qquad (1)$$

where $dd(i,\tau)$ is the difference function of lag τ calculated at time index (i), x is the input signal after the pre-processing step, and W is the integration window size. In the present C-tone implementation, the calculation window contains 256 samples, $W = 128$ samples, the maximum searching value $\tau_{\max} = 128$ samples (with the F0 searching lower limit = 62.5 Hz), and decimating factor $S = 4$. Theoretically, the time lag corresponding to the global near-zero minima in *dd* is the period of signal.

To avoid sub-harmonic errors, in the post-processing step, candidate periods, identified from the minima in *dd*, from the current and previous calculation windows are used to detect any changes in F0. Pairs of candidate periods correspond to the time lag of the first and the second minima in the difference function. The difference in time between the period candidates of the two consecutive windows is denoted as ΔT. The time shift is 8 ms between calculation windows. Based on the assumption that the values of period candidates should change slowly, the absolute value of ΔT is defined as follows:

$$|\Delta T| = \min(|C1-P1|, \quad |C1-P2|, \quad |C2-P1|, \quad |C2-P2|) \qquad (2)$$

The sign and magnitude of ΔT indicate the direction and amount of F0 change. ΔT is used to roughly estimate F0 change (Equation (3)).

$$\Delta F(i) = -\frac{\Delta T(i)}{T1*T2} \qquad (3)$$

where $\Delta F(i)$ is the F0 change at the ith analysis window, and $T1$ and $T2$ denote the selected adjacent

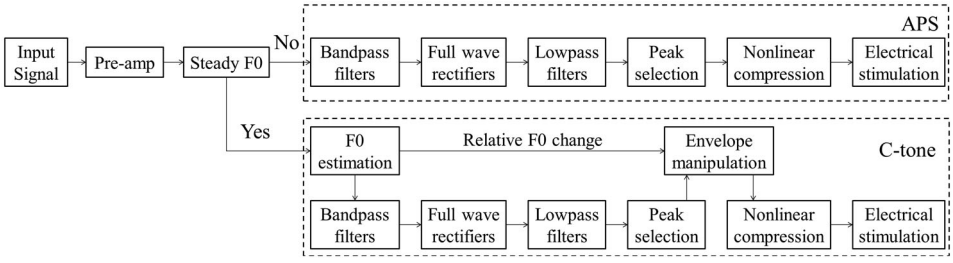

Figure 1 Block diagram for the clinical APS strategy and the C-tone algorithm.

periods. If $|\Delta T|$ is less than a criterion threshold (e.g. 3 ms) in three consecutive analysis windows, the input signal is assumed to be a steady pitch. This implementation of C-tone did not have a specific module to detect whether the speech is voiced or unvoiced. If F0 is undetectable or unsteady, no amplitude manipulation is performed.

(2) Noise floor estimator module

To implement a real-time F0 estimator within a clinical CI processor, the algorithm must have the lowest possible computational cost and memory demand. While the C-tone algorithm may able to track the F0 contour of Mandarin tones in quiet, accuracy may be negatively affected by noisy backgrounds, which might result in erroneous amplitude contour manipulations. For the present implementation of C-tone, a simple noise floor estimator was used to limit the activity of C-tone in noise. The noise floor estimator module constantly monitors the input signal energy. The input signal energy is updated by averaging the energy values for several consecutive frames (the averaging window is 64 ms). The minimum value of the input energy or the background noise level is updated over a period of several seconds (8 s in the present implementation).

(3) Envelope modification decision module

In this implementation of C-tone, the amplitude contour was modified only when: (1) the F0 contour tracking has detected a steady pitch, and (2) the background noise is below a criterion threshold (65 dB SPL). In each frame, the amplitude of the original envelope is modified as

$$Mod\_Env(i) = Ori\_Env(i) * \frac{\Delta F(i)}{SF} + Ori\_Env(i) \quad (4)$$

where $Ori\_Env(i)$ is the original envelope in the ith analysis window and $Mod\_Env(i)$ refers to the modified envelope in this analysis window. The original envelope in the ith analysis window is estimated by the FFT. $\Delta F(i)$ is the change of F0 (which has already been calculated by the F0 contour tracking module, see Equation (3), and SF is the shaping factor used to control the magnitude of the envelope modification. A relatively small shaping factor ($SF = 30$) was used to evaluate C-tone in this study. The envelope modification was done before the peak selection in the APS strategy, similar to the broadband signal manipulation in Luo and Fu (2004), which showed slightly better vowel and tone recognition.

For the real-time signal processing in C-tone, it was important to avoid excessive manipulations of the amplitude contour that might result in unwanted distortions to the speech signal. The degree of modification to the amplitude contour differed across tones according to change in F0. In this study, the mean change in F0 across the female talkers and vowels was 5.3, 40.0, 31.5, and 56.6% for Tones 1, 2, 3, and 4, respectively. For Tone 4, if the ratio of F0 change was applied to the modified amplitude contour, the reduction in amplitude might result in 'broken' speech due to the large drop in amplitude at the end of the vowel. Accordingly, the degree of amplitude contour manipulations was constrained to avoid such undesired artifacts in this module. Note also that the envelope manipulations were implemented before the amplitude mapping function (see Fig. 1). As such, the increase in peak amplitude for Tone 2 may have been compressed, limiting the effect of the contour manipulation, but also ensuring that sounds were not too loud.

Verification of C-tone implementation

Custom software was created to access the memory of the CI processor, and the output F0 contour and the modified amplitude contour were extracted in real time. A clinical speech processor was placed on the ear of a crystal mannequin head. Speech stimuli were presented at 65 dB SL from a single loudspeaker located directly in front of the mannequin head 1 m away. F0 contour and amplitude contour values were extracted from each stimulation frame. Figure 2 illustrates the amplitude contour extracted with APS (solid lines), the amplitude contour extracted with C-tone (dashed lines), and the extracted F0 change with C-tone for the vowel /a/ produced for all four lexical tones by a female talker; $SF = 30$ in this example. For Tone 1, the APS and C-tone amplitude contours were very similar, as there was little change in F0 across stimulation frames. For Tones 2, 3, and 4, the amplitude contour better follows the F0 contour with C-tone than with APS. For the vowel /a/ produced by this female talker, correlation coefficient r between the amplitude contour and F0 contour increased from 0.22 (APS) to 0.86 (C-tone) for Tone 2, from 0.94 (APS) to 0.97 (C-tone) for Tone 3, and from 0.62 (APS) to 0.80 (C-tone) for Tone 4.

Test procedures and materials

Speech performance was first measured with the subjects' clinical processors fitted with APS. After testing with APS was completed, C-tone was enabled in the clinical processors. All other clinical settings with APS were preserved in C-tone. Subjects were given 2 weeks of experience with C-tone before testing.

The stimuli and procedures were consistent throughout all tests. All stimuli were presented at 70 dB SPL via single loudspeaker located in front of the subjects

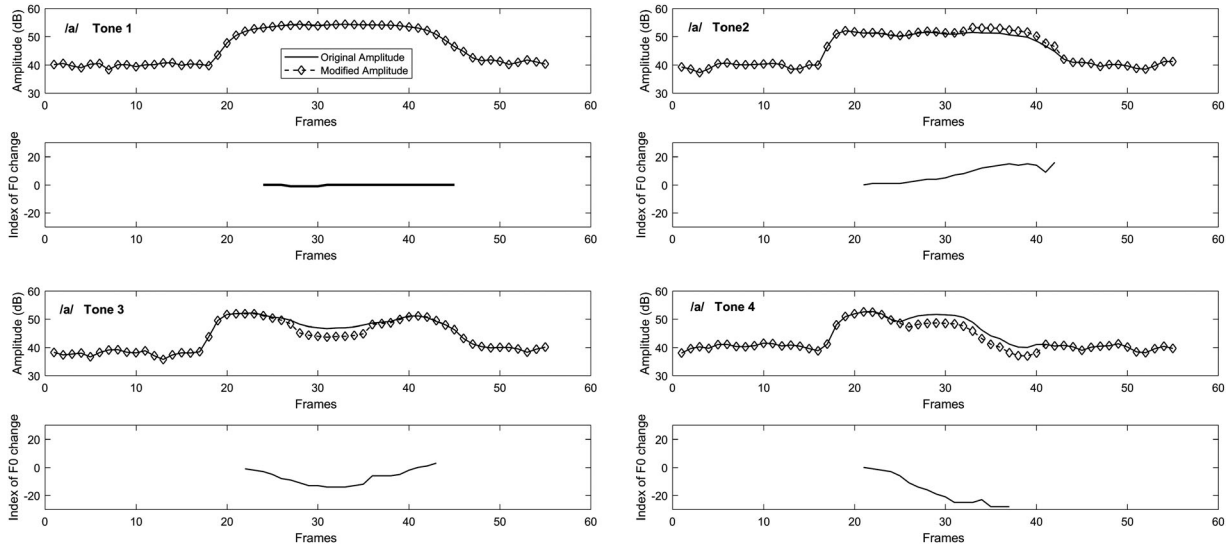

Figure 2 Examples of overall amplitude contours with APS (original amplitude, solid lines) and C-tone (modified amplitude, dashed lines) and the corresponding F0 change for 4 lexical tones for the vowel [a] produced by a female talker.

1 m away. All tests were conducted in a soundproof booth. Tone recognition stimuli consisted of four lexical tones for six Mandarin vowels ([a], [ɔ], [ɤ], [i], [u], [y]) produced by 10 talkers (5 males and 5 females), resulting in a total of 240 tokens for each test block; one test block each was tested with APS and C-tone. During each test, a stimulus was randomly chosen (without replacement) from among the 240 tokens. Subjects were allowed to repeat the stimulus presentation up to three times. Subjects responded by clicking on the response box (labeled Tone 1, 2, 3, or 4) that best matched the stimulus. No feedback was provided during testing. Open-set recognition of Mandarin monosyllables and disyllables was also tested with APS and C-tone using words from the Mandarin Speech Perception (MSP) materials produced by a single male talker (Li *et al.*, 2016). The MSP monosyllable and disyllable materials consist of 10 lists with 35 disyllables or 50 monosyllables in each list; test lists were explicitly balanced in terms of difficulty for CI users. During each monosyllable or disyllable test, a test list was randomly selected (without replacement) and a stimulus from the list was randomly selected (without replacement) and presented to the subject, who responded by repeating the stimulus as accurately as possible. The experimenter scored whether each stimulus was correctly identified. One monosyllable and disyllable list was tested for APS and C-tone.

After testing with C-tone was completed, subjects were asked to give quality ratings for C-tone using a questionnaire. Subjects were asked to mark the degree to which they agreed with different statements (e.g. 'Quality is better with C-tone', 'Speech is more natural', 'Speech is easier to understand in noise.) according to a five-point scale: 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree).

Results

Figure 3 shows scatterplots of tone recognition scores with C-tone as a function of scores with APS; scores above the diagonal indicate better performance with C-tone. Mean recognition with APS was 62.5 (SE = 5.3), 47.3 (SE = 7.5), 63.5 (SE = 5.9), and 79.4 (SE = 5.0) percent correct for Tones 1, 2, 3, and 4, respectively. Mean recognition with C-tone was 68.3 (SE = 4.7), 50.0 (SE = 6.0), 71.2 (SE = 6.1), and 78.1 (SE = 5.8) percent correct for Tones 1, 2, 3, and 4, respectively. A two-way repeated-measures analysis of variance (RM ANOVA) was performed on the tone recognition data, with processing (APS, C-tone) and tone (1, 2, 3, 4) as factors. Results showed significant effects for both processor [$F_{(1,21)} = 5.9$, $P = 0.046$] and tone [$F_{(3,21)} = 10.0$, $P < 0.001$]; there were no significant interactions [$F_{(3,21)} = 1.8$, $P = 0.186$]. Note that power was somewhat low for processor (0.47). Post hoc Bonferroni pairwise comparisons showed that performance was significantly better with C-tone than with APS only for Tone 3 ($P < 0.05$). With APS, performance with Tone 4 was significantly better than with Tone 2 ($P < 0.05$); there were no significant differences among the remaining tones ($P > 0.05$ in all cases). With C-tone, performance was significantly poorer with Tone 2 than with the other tones ($P < 0.05$ in all cases); there were no significant differences among the remaining tones ($P > 0.05$ in all cases). Table 2 shows the error patterns with the APS and the C-tone strategies. In both strategies, Tone 1 was most confused with Tone 4, Tone 2 was most confused with Tone 1, Tone 3 was most confused with Tone 2, and Tone 4 was most confused with Tone

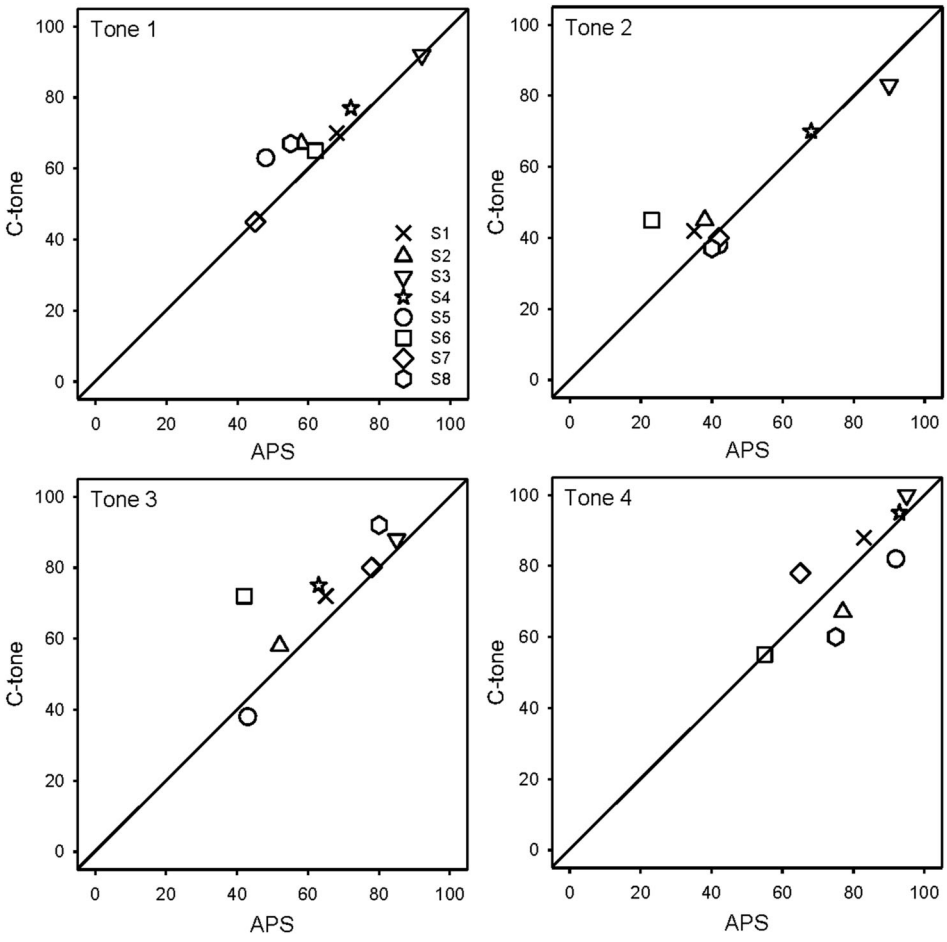

Figure 3 Performance for individual subjects (different symbols) with C-tone as a function of performance with APS for individual tones (shown in different panels). The diagonal line represents equal performance with APS and C-tone.

1. With C-tone, confusions were reduced between Tone 1 and Tone 4, between Tone 2 and Tone 1, between Tone 2 and Tone 4, between Tone 3 and Tone 1, and between Tone 3 and Tone 2.

Figure 4 shows scatterplots of monosyllable and disyllable recognition scores with C-tone as a function of scores with APS; scores above the diagonal indicate better performance with C-tone. Mean recognition with APS was 45.3 (SE = 4.3) and 46.8 (SE = 5.0) percent correct for monosyllables and disyllables, respectively. Mean recognition with C-tone was 51.5 (SE = 4.6) and 55.5 (SE = 6.6) percent correct for monosyllables and disyllables, respectively. A two-way RM ANOVA, with test (monosyllable, disyllable) and processing (APS, C-tone) as factors, showed that performance was significantly better with C-tone than with APS [$F_{(1,7)} = 9.0$, $P = 0.020$], but no significant difference between monosyllables and disyllables [$F_{(1,7)} = 1.0$, $P = 0.346$] and no significant interactions [$F_{(1,7)} = 0.7$, $P = 0.489$].

Table 2 Response error matrix for tone recognition with APS and C-tone

	Response tone							
	APS				C-tone			
Target tone	1	2	3	4	1	2	3	4
1	63	9	0	28	68	10	0	22
2	31	47	2	20	30	48	4	18
3	9	25	65	1	7	18	74	1
4	13	8	0	79	16	8	0	76

All values show the percent of responses.

Lexical tone, monosyllable, and disyllable scores were compared within APS and C-tone. With APS, tone recognition was significantly correlated with monosyllable recognition ($r = 0.76$, $r^2 = 0.58$, $P = 0.029$), and monosyllable recognition was significantly correlated with disyllable recognition ($r = 0.83$, $r^2 = 0.69$, $P = 0.011$); there was no significant correlation between tone and disyllable recognition ($r = 0.50$, $r^2 = 0.25$, $P = 0.210$). With C-tone, tone recognition was significantly correlated with monosyllable recognition ($r = 0.84$, $r^2 = 0.71$, $P = 0.009$), and monosyllable recognition was significantly correlated with disyllable recognition ($r = 0.88$, $r^2 = 0.77$, $P = 0.004$); there was no significant correlation between tone and disyllable recognition ($r = 0.57$, $r^2 = 0.32$, $P = 0.137$). Lexical tone, monosyllable, and disyllable scores were also compared to demographic factors age at testing and CI experience. With APS, age at testing

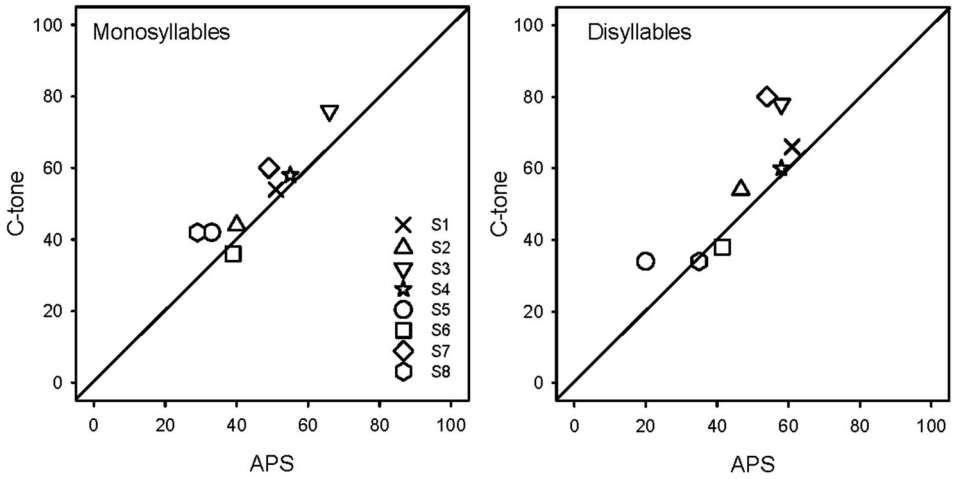

Figure 4 Performance for individual subjects (different symbols) with C-tone as a function of performance with APS for monosyllables (left panel) and disyllables (right panel). The diagonal line represents equal performance with APS and C-tone.

was significantly correlated with disyllable recognition ($r = 0.78$, $r^2 = 0.61$, $P = 0.030$); there were no significant correlations between age at testing and tone or monosyllable recognition ($P > 0.05$ in both cases). With C-tone, there were no significant correlations between age at testing and tone, monosyllable, or disyllable recognition ($P > 0.05$ in all cases). With APS, CI experience was significantly correlated with monosyllable recognition ($r = 0.73$, $r^2 = 0.53$, $P = 0.039$); there were no significant correlations between CI experience and tone or disyllable recognition ($P > 0.05$ in both cases). With C-tone, CI experience was significantly correlated with monosyllable ($r = 0.77$, $r^2 = 0.59$, $P = 0.025$) and disyllable recognition ($r = 0.91$, $r^2 = 0.83$, $P = 0.002$); there were no significant correlations between CI experience and tone recognition ($P > 0.05$).

Table 3 shows the distribution of responses for various aspects of sound quality collected after 2 weeks of experience with C-tone. For most sound quality ratings, most subjects reported a 'Neutral' response, indicating no improvement but also no decrement in quality. For some aspects ('Quality is better,' 'Speech is more natural,' and 'Understanding own voice is better'), a sizeable proportion of subjects rated C-tone favorably. For 'Understanding male voice,' a sizeable proportion of subjects rated C-tone unfavorably.

Discussion

The present real-time manipulation of the amplitude contour according to the F0 contour resulted in small but significant improvements in CI users' recognition of Mandarin tones, monosyllables, and disyllables in quiet. Mean tone recognition with APS or C-tone was comparable to that observed in previous studies with CI subjects (Fu *et al.*, 2004; Luo *et al.*, 2009; Wang *et al.*, 2011) or with CI simulations (Fu and Zeng, 2000; Fu *et al.*, 1998; Xu *et al.*, 2002).

Real-time Fo contour tracking

Real-time F0 estimation is essential for lexical tone enhancement in a CI strategy (e.g. F0mod, eTone, C-tone). There are two major difficulties in implementing a real-time F0 estimation algorithm: (1) limited computational power in the processor (more complex algorithms require greater processing power), (2) optimal window length (with longer windows providing better accuracy but also introducing a delay).

The eTone strategy from Vandali and van Hoesel (2012) was implemented in real time, but in a research processor (SPEAR 3). The F0 estimator of eTone strategy (Vandali and van Hoesel, 2011) was based on the 'harmonic sieve' (Duifhuis *et al.*, 1982; Zakis *et al.*, 2007); a 32 ms analysis window was used. The total processing delay was 18 ms for the

Table 3 Distribution of subjective ratings for various sound quality categories after weeks of experience with C-tone

With C-tone…	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Quality is better			62.5	37.5	
Speech is more natural		12.5	50.0	37.5	
Understanding male voices is better		37.5	50.0	12.5	
Understanding female voices is better		12.5	62.5	25.0	
Understanding own voice is better			50.0	50.0	
Environment sounds are easier to distinguish			87.5	12.5	
Speech is easier to understand in noise		12.5	75.0	12.5	

quiet processing mode and 40 ms for the noise processing mode, and required approximately 5000 words of memory space. Vandali and van Hoesel (2011) estimated that the computational complexity could be double for the F0 estimation based on autocorrelation. Francart *et al.* (2015) implemented the F0mod strategy in a real-time system (Simulink/xPC system), but not within a clinical processor. An autocorrelation-based F0-estimator with a frame length of 22.5 ms and a hop size of 9.2 ms was used in F0mod. The computational capacity of the Nurotron-60 processor is 16 million instructions per second. For C-tone F0 contour tracking, with a 16-ms analysis window and decimating factor *dd* in the difference equation, the processing delay was approximately 4 ms and required 800 words of memory space. While the lower complexity of the proposed algorithm could reduce accuracy in estimating extract values of F0, the main purpose of the F0 estimator is to track only changes in the F0 contour between analysis frames.

Speech performance with APS and C-tone

The largest mean improvement with Pfingst C-tone was observed for Tone 3 and the smallest for Tone 4. In this study and in Fu and Zeng (2000), the amplitude contour and F0 contour were most strongly correlated for Tone 3. As such, the amplitude contour with C-tone may have more faithfully represented the dynamics in the F0 contour. The limited improvement for Tone 4 may have been due to ceiling performance effects, with half the subjects scoring between 80 and 100% correct with either strategy. Performance gains with C-tone were somewhat variable across tones and across subjects. It is unclear why mean performance improved with C-tone for Tone 1, as the amplitude contour manipulation had little effect on Tone 1 (see Fig. 2); reduced confusion between Tones 1 and 4 may have contributed to the C-tone advantage for Tone 1. Similar results were observed in Luo and Fu (2004), where recognition of Tone 1 improved by as much as 10 percentage points. Given the relative poor performance for Tone 2 with APS and the greatly improved correlation between the amplitude contour and F0 contour with C-tone, we hoped that C-tone would most benefit recognition of Tone 2, but this was not the case. As shown in Fig. 2, while the F0 contour rises almost continuously, the amplitude contours with APS and C-tone rise then fall near the end. This may have to produce a conflicting cue. Indeed, there was substantial confusion between Tone 2 (rising) and Tones 1 (flat) and 4 (falling).

While performance for all speech measures was significantly better for C-tone, mean performance gains were modest: 3.9 percentage points across all tones, 6.3 percentage points for monosyllables, and 8.7 percentage points for disyllables. However, performance gains were quite variable across tests and across subjects, ranging from −1 to 14 percentage points for tone recognition, −3 to 13 percentage points for monosyllable recognition, and −4 to 26 percentage points for disyllable recognition. Performance gains with C-tone were not always consistent across tests (e.g. a 14-point gain for tone recognition but a 3-point decrement for disyllables for subject S6). It is possible that differences in test stimuli and paradigms (closed-set for tones, open-set for monosyllables and disyllables) may have contributed to the overall variability in performance across tests and C-tone benefit. CI experience was significantly correlated with monosyllable recognition (with C-tone) and disyllable recognition (with APS and C-tone), suggesting that greater gains in performance might be observed as patients gain more experience with their device and/or processing strategy. It should be noted that this study did not use a true cross-over design; baseline performance was measured with APS, then with C-tone, with no re-testing of APS to check for procedural learning. While subjects were very familiar with the APS processing, it is possible that some procedural learning may have occurred.

As shown in Table 2, confusion between Tone 3 and Tone 2 (and to a lesser extent, Tone 1) was reduced with C-tone. However, great confusion between Tone 2 and Tone 1 persisted with C-tone. As shown in Fig. 2, the envelope modification provided only about 3 dB of gain near the end of the vowel, which may not have been sufficient to convey the change in amplitude with F0. It is possible that larger enhancements are needed to convey changes in amplitude at the end of the vowels in Tones 2 and 4.

Quality ratings

While there was no strong preference for C-tone, the majority of subjects reported no deficit in quality with C-tone. Indeed, some subjects reported better quality with C-tone for some categories. This suggests that under everyday listening conditions, the quality of C-tone was as good as (and sometimes better than) that of the default clinical APS processor.

Future work

In this study, C-tone was first evaluated in quiet listening conditions to ensure that the algorithm did not degrade speech performance relative to the clinical processor. However, most CI users encounter noisy complex listening environments in everyday listening. The noise floor estimator implemented in this version of C-tone disabled the algorithm when noise was above a certain level. A robust voice activity detector (VAD) may provide some advantage over the

current noise floor. Mao and Xu (2017) showed that CI users' tone recognition performance was more susceptible to noise than was NH performance. Pre-processing noise-reduction schemes may benefit F0 contour estimation, and thus allow for the beneficial amplitude contour enhancements with C-tone. It is also unclear how competing speech might affect the F0 contour tracking or the saliency of the amplitude contour manipulations. Luo and Fu (2009) showed that competing tones could result in distortion to the amplitude contour (e.g. Tone 2 concurrent with Tone 4 resulted in a flat amplitude contour associated with Tone 1). Further modifications are needed to improve F0 extraction and to optimize the envelope enhancements (e.g. the falling contour in Tone 4). While these preliminary data are promising, further evaluation is needed with a greater number of subjects and listening conditions, using a true cross-over design (APS – C-tone – APS; C-tone – APS – C-tone).

C-tone relies on accurate F0 estimation. However, due to the computation and memory limits in the current Nurotron speech processor, the current implementation using real-time F0 extraction is relatively simple and requires low levels of background noise to achieve accurate estimation. Previous studies used off-line processing or custom speech processors that had no realistic limits in terms of computation and/or memory, allowing for advanced F0 extraction methods. It is difficult to directly compare the accuracy of F0 estimation across studies when the resources (computation power and memory) are so different. This study represents a first step to show the possibility of implementing real-time amplitude contour modification to improve tone recognition. With the increased computation and memory in future generations of Nurotron speech processors, it will be possible to implement more advanced F0 extraction methods and improve the accuracy of F0 estimation in quiet and in noise.

Conclusions

A novel algorithm (C-tone) that manipulated the amplitude contour according to the F0 contour was implemented in real time and evaluated in 8 Mandarin-speaking CI subjects. Preliminary data suggest that real-time enhancement of the amplitude contour may benefit Chinese CI users' speech performance. Major findings include:

1. Lexical tone recognition was significantly better with C-tone. The largest performance gains with C-tone were observed for Tone 3 (falling–rising), most likely due to reduced confusion between Tone 3 and Tone 2 (rising).
2. Recognition of monosyllables and disyllables was significantly better with C-tone.
3. Although mean performance gains with C-tone were modest, there was substantial inter-subject variability, with substantial performance gains or decrements in some subjects.
4. For most subjects, quality ratings showed no deterioration with C-tone, relative to the default clinical strategy, with some subjects reporting better quality with C-tone for some listening categories.

Acknowledgements

We thank all the subjects for participating in this research. We thank John J. Galvin III for editorial assistance.

Disclaimer statements

Contributors None.

Funding This study is sponsored by Nurotron Biotechnology and/or Zhejiang Nurotron Biotechnology.

Conflicts of interest In accordance with Taylor & Francis policy, we report that authors Ping, Wang, Tang, Lu, Yin, and Tu are all current or former employees of Nurotron Biotechnology and/or Zhejiang Nurotron Biotechnology. Dr Fu is a scientific advisor for Zhejiang Nurotron Biotechnology. The authors have disclosed those interests fully to Taylor & Francis, and have in place an approved plan for managing any potential conflicts arising from their relationship to Nurotron.

Ethics approval None.

References

Abramson, A.S. 1972. Tonal experiments with whispered Thai. In: Valdman A, (ed.) *Papers in linguistics and phonetics to the memory of pierre delattre*. The Hague: Mouton, p. 31–44.

Bahill, A.T., Kallman, J.S., Lieberman, J.E. 1982. Frequency limitations of the two-point central difference differentiation algorithm. *Biological Cybernetics*, 45: 1–4.

Chatterjee, M., Peng, S.C. 2008. Processing F0 with cochlear implants: modulation frequency discrimination and speech intonation recognition. *Hearing Research*, 235: 143–156.

Duifhuis, H., Willems, L.F., Sluyter, R.J. 1982. Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. *Journal of the Acoustical Society of America*, 71(6): 1568–1580.

Francart, T., Osses, A., Wouters, J. 2015. Speech perception with F0mod, a cochlear implant pitch coding strategy. *International Journal of Audiology*, 54: 424–432.

Friesen, L.M., Shannon, R.V., Baskent, D., Wang, X. 2001. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *Journal of the Acoustical Society of America*, 110: 1150–1163.

Fu, Q.J., Nogaki, G. 2005. Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing. *Journal of the Association for Research in Otolaryngology*, 6: 19–27.

Fu, Q.-J., Shannon, R.V. 1998. Effects of amplitude nonlinearity on phoneme recognition by cochlear implant users and normal-hearing listeners. *Journal of the Acoustical Society of America*, 104(5): 2570–2577.

Fu, Q.J., Zeng, F.G. 2000. Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language, and Hearing*, 5: 13.

Fu, Q.J., Zeng, F.G., Shannon, R.V., Soli, S.D. 1998. Importance of tonal envelope cues in Chinese speech recognition. *Journal of the Acoustical Society of America*, 104: 505–510.

Fu, Q.J., Hsu, C.J., Horng, M.J. 2004. Effects of speech processing strategy on Chinese tone recognition by nucleus-24 cochlear implant users. *Ear and Hearing*, 25: 501–508.

Galvin, J.J. 3rd, Fu, Q.J., Nogaki, G. 2007. Melodic contour identification by cochlear implant listeners. *Ear and Hearing*, 28: 302–319.

Galvin, J.J. 3rd, Fu, Q.J., Shannon, R.V. 2009. Melodic contour identification and music perception by cochlear implant users. *Annals of the New York Academy of Sciences*, 1169: 518–533.

Gao, N., Xu, X.D., Chi, F.L., Zeng, F.G., Fu, Q.J., Jia, X.H., *et al.* 2016. Objective and subjective evaluations of the Nurotron Venus cochlear implant system via animal experiments and clinical trials. *Acta Otolaryngologica*, 136: 68–77.

Garding, E., Kratochvil, P., Svantesson, J.O., Zhang, J. 1986. Tone 4 and Tone 3 discrimination in modern Standard Chinese. *Language and Speech*, 29: 281–293.

Geurts, L., Wouters, J. 2004. Better place-coding of the fundamental frequency in cochlear implants. *Journal of the Acoustical Society of America*, 115: 844–852.

Green, T., Faulkner, A., Rosen, S., Macherey, O. 2005. Enhancement of temporal periodicity cues in cochlear implants: effects on prosodic perception and vowel identification. *Journal of the Acoustical Society of America*, 118: 375–385.

Han, D., Liu, B., Zhou, N., Chen, X., Kong, Y., Liu, H., *et al.* 2009. Lexical tone perception with HiResolution and HiResolution 120 sound-processing strategies in pediatric Mandarin-speaking cochlear implant users. *Ear and Hearing*, 30: 169–177.

Laneau, J., Moonen, M., Wouters, J. 2006. Improved music perception with explicit pitch coding in cochlear implants. *Audiology and Neuro-Otology*, 11: 38–52.

Li, Y., Wang, S., Su, Q., Galvin, J.J. 3rd, Fu, Q.J. 2016. Validation of list equivalency for Mandarin speech materials to use with cochlear implant listeners. *International Journal of Audiology*, 14: 1–10.

Liang Z.A. 1963. The auditory perception of Mandarin Tones. *Acta Physiologica Sinica*, 26: 85–91.

Lin, M.C. 1988. The acoustic characteristics and perceptual cues of tones in Standard Chinese. *Chinese Linguistics*, 204: 182–193.

Luo, X., Fu, Q.J. 2004. Enhancing Chinese tone recognition by manipulating amplitude envelope: implications for cochlear implants. *Journal of the Acoustical Society of America*, 116: 3659–3667.

Luo, X., Fu, Q.J. 2009. Concurrent-vowel and tone recognitions in acoustic and simulated electric hearing. *Journal of the Acoustical Society of America*, 125: 3223–3233.

Luo, X., Fu, Q.J., Galvin, J.J. 2007. Vocal emotion recognition by normal-hearing listeners and cochlear implant users. *Trends in Amplification*, 11: 301–315.

Luo, X., Fu, Q.J., Wei, C.G., Cao, K.L. 2008. Speech recognition and temporal amplitude modulation processing by Mandarin-speaking cochlear implant users. *Ear and Hearing*, 29: 957–970.

Luo, X., Fu, Q.J., Wu, H.P., Hsu, C.J. 2009. Concurrent-vowel and tone recognition by Mandarin-speaking cochlear implant users. *Hearing Research*, 256: 75–84.

Mao, Y., Xu, L. 2017. Lexical tone recognition in noise in normal-hearing children and prelingually deafened children with cochlear implants. *International Journal of Audiology*, published online: 2016 Aug 26. doi:10.1080/14992027.2016.1219073

Massaro, D.W., Cohen, M.M., Tseng, C.Y. 1985. The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics*, 13: 267–289.

Milczynski, M., Wouters, J., van Wieringen, A. 2009. Improved fundamental frequency coding in cochlear implant signal processing. *Journal of the Acoustical Society of America*, 125: 2260–2271.

Milczynski, M., Chang, J.E., Wouters, J., van Wieringen, A. 2012. Perception of Mandarin Chinese with cochlear implants using enhanced temporal pitch cues. *Hearing Research*, 285: 1–12.

Molero, P.C., Reyes, N.R., Candeas, P.V., Bascon, S.M. 2011. Low-complexity F0-based speech/nonspeech discrimination approach for digital hearing aids. *Multimedia Tools and Applications*, 54: 291–319.

Nelson, P.B., Jin, S.H. 2004. Factors affecting speech understanding in gated interference: cochlear implant users and normal-hearing listeners. *Journal of the Acoustical Society of America*, 115: 2286–2294.

Nelson, P.B., Jin, S.H., Carney, A.E., Nelson, D.A. 2003. Understanding speech in modulated interference: cochlear implant users and normal-hearing listeners. *Journal of the Acoustical Society of America*, 113: 961–968.

Sagart, L. 1986. Tone production in modern standard Chinese: an electromyographic investigation. *Cahiers de Linguistique, Asie Orientale, Paris*, 15: 205–221.

Schatzer, R., Krenmayr, A., Au, D.K. 2010. Temporal fine structure in cochlear implants: preliminary speech perception results in Cantonese-speaking implant users. *Acta Otolaryngologica*, 130: 1031–1039.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M. 1995. Speech recognition with primarily temporal cues. *Science*, 270: 303–304.

Shannon, R.V., Fu, Q.J., Galvin, J.J. 2004. The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngologica*, 124 (Suppl): 50–54.

Smith, Z.M., Delgutte, B., Oxenham, A.J. 2002. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416: 87–90.

Tao, D., Deng, R., Jiang, Y., Galvin, J.J. III, Fu, Q.-J., Chen, B. 2015. Melodic pitch perception and lexical tone perception in mandarin-speaking cochlear implant users. *Ear and Hearing*, 36(1): 102–110.

Vandali, A.E., van Hoesel, R.J. 2011. Development of a temporal fundamental frequency coding strategy for cochlear implants. *Journal of the Acoustical Society of America*, 129: 4023–4036.

Vandali, A.E., van Hoesel, R.J. 2012. Enhancement of temporal cues to pitch in cochlear implants: effects on pitch ranking. *Journal of the Acoustical Society of America*, 132: 392–402.

Wang, W., Zhou, N., Xu, L. 2011. Musical pitch and lexical tone perception with cochlear implants. *International Journal of Audiology*, 50: 270–278.

Wang, S., Liu, B., Dong, R., Zhou, Y., Li, J., Qi, B., *et al.* 2012. Music and lexical tone perception in Chinese adult cochlear implant users. *Laryngoscope*, 122: 1353–1360.

Whalen, D. H., Xu, Y. 1992. Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, 49: 25–47.

Wilson, B.S., Finley, C.C., Lawson, D.T., Wolford, R.D., Eddington, D.K., Rabinowitz, W.M. 1991. Better speech recognition with cochlear implants. *Nature*, 352: 236–238.

Xu, L., Tsai, Y., Pfingst, B.E. 2002. Features of stimulation affecting tonal-speech perception: implications for cochlear prostheses. *Journal of the Acoustical Society of America*, 112: 247–258.

Xu, L., Chen, X., Lu, H., Zhou, N., Wang, S., Liu, Q., *et al.* 2011. Tone perception and production in pediatric cochlear implants users. *Acta Otolaryngologica*, 131: 395–398.

Zakis, J.A., Dillon, H., McDermott, H.J. 2007. The design and evaluation of a hearing aid with trainable amplification parameters. *Ear and Hearing*, 28(6): 812–830.

Zeng, F.G., Rebscher, S.J., Fu, Q.J., Chen, H., Sun, X., Yin L., *et al.* 2015. Development and evaluation of the Nurotron 26-electrode cochlear implant system. *Hearing Research*, 322: 188–199.