Improved perception of speech in noise and Mandarin tones with acoustic simulations of harmonic coding for cochlear implants^{a)}

Xing Li

Department of Electrical Engineering, University of Washington, Seattle, Washington 98195

Kaibao Nie,^{b)} Nikita S. Imennov, Jong Ho Won, Ward R. Drennan, and Jay T. Rubinstein

Virginia Merrill Bloedel Hearing Research Center, Department of Otolaryngology-Head and Neck Surgery, University of Washington, Box 357923, Seattle, Washington 98195

Les E. Atlas

Department of Electrical Engineering, University of Washington, Seattle, Washington 98195

(Received 21 November 2011; revised 13 August 2012; accepted 11 September 2012)

Harmonic and temporal fine structure (TFS) information are important cues for speech perception in noise and music perception. However, due to the inherently coarse spectral and temporal resolution in electric hearing, the question of how to deliver harmonic and TFS information to cochlear implant (CI) users remains unresolved. A harmonic-single-sideband-encoder [(HSSE); Nie *et al.* (2008). *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*; Lie *et al.*, (2010). *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*] strategy has been proposed that explicitly tracks the harmonics in speech and transforms them into modulators conveying both amplitude modulation and fundamental frequency information. For unvoiced speech, HSSE transforms the TFS into a slowly varying yet still noise-like signal. To investigate its potential, four- and eight-channel vocoder simulations of HSSE and the continuous-interleaved-sampling (CIS) strategy were implemented, respectively. Using these vocoders, five normal-hearing subjects' speech recognition performance was evaluated under different masking conditions; another five normal-hearing subjects' Mandarin tone identification performance was also evaluated. Additionally, the neural discharge patterns evoked by HSSE- and CIS-encoded Mandarin tone stimuli were simulated using an auditory nerve model. All subjects scored significantly higher with HSSE than with CIS vocoders. The modeling analysis demonstrated that HSSE can convey temporal pitch cues better than CIS. Overall, the results suggest that HSSE is a promising strategy to enhance speech perception with CIs.

© 2012 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4756827>]

PACS number(s): 43.66.Ts, 43.66.Hg, 43.64.Ri [PBN]

Pages: 3387–3398

I. INTRODUCTION

Despite the great advances in efficacy of cochlear implants (CIs) during the past two decades, most CI users still have trouble understanding speech in the presence of noise or multiple talkers (e.g., Stickney *et al.*, 2004). Additionally, the fundamental frequency (F_0) cues transmitted by contemporary CIs are unlikely to be sufficient for tonal pattern identification, an important aspect of speech in tonal languages (e.g., Xu *et al.*, 2011). A number of studies have suggested that temporal fine structure (TFS) information is important to the above-mentioned tasks (Moore, 2008). However, fine structure information is largely discarded in conventional CI signal processing strategies, e.g., the continuous-interleaved-sampling (CIS) strategy. This indicates a need for a new approach to deliver TFS information to CI users (Wilson and Dorman, 2008).

The importance of TFS has been demonstrated in a variety of listening tasks. Pitch perception (Oxenham, 2008),

lexical tone identification (Wang *et al.*, 2011), and speech recognition in noise (Lorenzi *et al.*, 2006; Hopkins *et al.*, 2008) all appear to be linked to TFS perception. The lack of TFS in CI encoding reduces temporal pitch cues and might partly account for CI users' speech perception difficulties in adverse situations. To deliver TFS by pulsatile stimulation, an important caveat to note is the reduced sensitivity to temporal modulation in electric hearing. Most CI users cannot discriminate changes in the repetition rate of the electric waveform above approximately 300 Hz (Shannon, 1992; Zeng, 2002), whereas TFS typically oscillates at a much higher rate.

Several approaches have been attempted to encode fine structure information for CI users. The HiRes strategy uses a relatively high envelope cutoff frequency and pulse rate to improve TFS representation. The HiRes Fidelity 120 uses a current steering paradigm to represent *spectral* fine structure and shows an improvement in frequency resolution over the HiRes strategy, although it is unclear if this provides significant benefits to speech or music perception (Han *et al.*, 2009; Drennan *et al.*, 2010). The fine structure processing (FSP) strategy bases the pulse triggering pattern in a particular channel on the zero crossings of the respective band waveform to encode TFS; its effectiveness for enhancing

^{a)}Portions of this work were presented at the 161st Meeting of the Acoustical Society of America, Seattle, WA, May 2011.

^{b)}Author to whom correspondence should be addressed. Electronic mail: niek@u.washington.edu

speech or music perception is yet to be demonstrated (Schatzer *et al.*, 2010; Riss *et al.*, 2011). Nie *et al.* (2005) proposed to convert TFS into a frequency modulation signal and use it to frequency modulate the pulse rate; they showed by vocoder simulations that the frequency modulation information was beneficial to speech recognition in noise. Laneau *et al.* (2006) suggested modulating the channel envelope at the input signal's F_0 with 100% modulation depth; this "F0mod" method showed improvement over the advanced-combination-encoder (ACE) strategy on music perception and Mandarin tone perception but no advantage for sentence recognition (Milczynski *et al.*, 2012). Finally, analog strategies can be used to transmit TFS, but these strategies suffer from increased electrode interaction, which ultimately reduces listeners' perceptual abilities.

In addition to the lack of TFS, contemporary CIs provide little or no representation of harmonics of complex sounds, such as those produced by human voice and musical instruments; this might also contribute to CI users' difficulties in perceiving music and recognizing speech in noise (Darwin, 2008; Oxenham, 2008). The lower harmonics of complex sounds can be resolved or separated on the basilar membrane and elicit a more salient and accurate pitch percept than the unresolved harmonics (e.g., Shackleton and Carlyon, 1994). In normal-hearing listeners, pitch cues have been found to be important for separating mixed speech signals (Summers and Leek, 1998). In short-electrode CI users (hybrid listeners), pitch cues obtained from low-frequency acoustic stimulation have been shown to benefit speech recognition in a competing-talker background (Turner *et al.*, 2004). In aggregate, the above-mentioned studies suggest that providing low-frequency harmonics might be beneficial to CI users.

To encode both TFS and harmonic information for CI users, a harmonic-single-sideband-encoder strategy [(HSSE); Nie *et al.*, 2008; Li *et al.*, 2010] has been proposed, which explicitly tracks the harmonics of complex sounds and linearly transforms harmonics into modulators conveying both amplitude modulation (AM) and TFS cues to electrodes. During unvoiced segments of speech, the fast-oscillating TFS is converted into a slowly varying yet still noise-like signal and then preserved in HSSE modulators. The main distinction of HSSE from CIS-like strategies is that the former uses frequency downshift operations to extract the modulators for electrodes, whereas the latter use incoherent approaches to extract temporal cues, e.g., the Hilbert envelope or half/full wave rectification followed by a low-pass filter. Frequency downshift is a linear operation and introduces no distortion to the resulting modulators (Schimmel and Atlas, 2005; Clark and Atlas, 2009), whereas incoherent approaches incur nonlinear distortions to the extracted temporal cues.

A mathematical analysis of the nonlinear distortions caused by incoherent processing can be found in Flanagan (1980). As a simple illustration, Fig. 1 shows two tones at 800 and 1000 Hz, respectively. Combining the two tones together results in a bandpass signal. By swapping the amplitudes of the two tones, a different bandpass signal can be constructed. The spectrum and the waveform of each signal are displayed in the first and second column, respectively.

FIG. 1. (Color online) Illustration of the distortions introduced by incoherent operation. Note that the Hilbert envelope of a signal does not represent any specific tone component in the original signal.

Despite the evident difference between the two signals, they are mapped to identical temporal cues by taking the Hilbert envelope. The spectrum of each signal's envelope is displayed in the third column. Comparing a signal's spectrum with its envelope spectrum, one can hardly tell how a particular frequency component in the envelope is related to any specific tone in the signal, i.e., the envelope contains distortion components relative to the original signal. In contrast, if both bandpass signals are transformed by a frequency downshift operation like in HSSE, their uniqueness would be retained because the transformed signal contains the original tone components but no distortion component. More details on HSSE are provided in Sec. II.

The aim of this study was to investigate the potential benefits of HSSE for speech perception. In experiment 1, vocoder simulations of HSSE and CIS were implemented, respectively; their effects on sentence recognition in noise were compared using normal-hearing listeners. CIS was chosen as a comparison baseline because CIS is the basis of the encoding approaches used in virtually every modern CI speech processor. Nearly all recent strategies resemble CIS in terms of temporal envelope extraction. Although these strategies might be able to deliver more temporal or spectral fine structure cues, they seem to produce similar speech and music recognition performance as CIS (Drennan *et al.*, 2010; Schatzer *et al.*, 2010; Riss *et al.*, 2011). Typically CI users' perception performance can be gauged by a four to eight band CIS simulation (Friesen *et al.*, 2001). To prevent vocoders from presenting information that might not be accessible to CI users, both CIS envelopes and HSSE modulators were low-pass filtered at 300 Hz to correspond to the pitch saturation limit in electric hearing. The hypothesis was that under the same spectral and temporal constraints, listeners would recognize speech better with HSSE than with CIS vocoders, due to the advantage of HSSE in extracting non-distorted harmonic and TFS information.

Using the same vocoder simulations as in experiment 1, the effect of HSSE processing on Mandarin tone identification was compared with that of CIS encoding in experiment 2. Due to the important differences between acoustic and electric hearing, the neural discharge patterns evoked by HSSE- and CIS-encoded Mandarin tone stimuli were also

examined, using a population model of electrically stimulated auditory nerve fibers (Imennov and Rubinstein, 2009). This population model was chosen because the model's single-fiber response properties—e.g., spike latency, jitter, and relative refractory period—have been shown to closely correspond to the response quantities measured *in vivo* (e.g., Miller *et al.*, 1999). Moreover, the normalized response thresholds of a population of diameter-distributed model fibers have been shown to match that of the same number of *in vivo* fibers (Imennov and Rubinstein, 2009), suggesting that the model may be used to approximate the aggregate responses of the auditory nerve. Our reasoning was that if the advantage of HSSE can be observed in both vocoder and neural response simulations, then it is likely to be beneficial to CI users.

II. EXPERIMENT 1: SPEECH RECOGNITION IN NOISE WITH SIMULATED HSSE AND CIS STRATEGIES

A. Methods

1. HSSE processing

To encode harmonics for CI users, the F_0 of a given speech signal was first estimated such that F_0 's harmonics can be analyzed.

a. F_0 estimation. To track F_0 , an incoming signal was first segmented into short frames such that constant F_0 can be assumed within each frame. The frame size was empirically chosen as 20 ms to handle the possible F_0 range of human voices (>50 Hz). There was a 10 ms overlap between

contiguous frames. To further refine F_0 estimates, each frame was combined with its preceding and succeeding frames to produce a smooth F_0 trajectory during each 40 ms window. For simplicity, the following description is focused on the processing of a single frame.

A least-squares-harmonic model was used to track F_0 in this study (Li and Atlas, 2003). Given a signal frame, this method first detected its underlying harmonic structure in the frequency domain and then derived the F_0 accordingly. Pilot studies showed that this technique can reliably track F_0 even in adverse situations, e.g., at a signal to noise ratio of -10 dB. During unvoiced segment, an F_0 value was assigned by linearly interpolating the F_0 estimates from surrounding voiced frames. For example, in the top row of Fig. 2, the gaps in the estimated F_0 contour (related to unvoiced speech) would be linearly interpolated to generate a continuous F_0 trajectory. In this way, both voiced and unvoiced speech can be processed in the same harmonic-centered fashion, yet the concept of harmonic is only meaningful for voiced speech. It will be shown later that although an interpolated F_0 was assumed for unvoiced speech, the encoded TFS would be noise-like.

b. Harmonic analysis. Given the detected F_0 , a particular harmonic $h_k(t)$, with t being the time index and k the harmonic index, can be modeled as the following sinusoid:

$$h_k(t) = a_k(t) \cos(2\pi k F_0 t + \phi_k(t)), \quad (1)$$

where $a_k(t)$ represents the harmonic amplitude, kF_0 is the harmonic frequency, and $\phi_k(t)$ represents the phase information.

FIG. 2. (Color online) The extracted CIS envelopes and HSSE modulators for the “eas” segment of “easy” in quiet and in noise under four-channel condition. The spectrogram of the “eas” segment is displayed on the left. The waveform in the second row represents an IEEE sentence saying “it’s easy to tell the depth of a well,” for which the estimated F_0 contour is displayed in the first row. The speaker’s pitch is about 230 Hz.

The amplitude information is related to the envelope, while the harmonic frequency and phase information are related to the TFS. If $h_k(t)$ is an actual harmonic extracted from voiced speech and fluctuates nearly periodically, then $\varphi_k(t)$ stays approximately constant to let the overall TFS oscillate regularly at a rate of kF_0 . In contrast, if $h_k(t)$ is from unvoiced speech and fluctuates irregularly, then $\varphi_k(t)$ varies randomly over time to cause the overall TFS to oscillate irregularly (McAulay and Quatieri, 1986).

c. Harmonic selection. To extract the HSSE modulator for a particular channel, the first step is to identify which harmonics are contained in the channel. For example, the spectrogram in Fig. 2 shows how speech intensity (color scale) varies as a function of time and frequency: The evenly spaced frequency components represent harmonics, with the bottom one representing F_0 . The four-channel corner frequencies are overlaid on the spectrogram as dashed horizontal lines. One can see that because the channel numbers used in this study are comparatively small (four and eight), most bands are broad and often contain multiple harmonics.

Due to auditory masking, the strongest harmonic within a channel usually dominates the perception for the related spectral region, whereas the weak components are masked (Darwin, 2008). Similarly, in HSSE processing, the harmonic with the largest magnitude in a channel was selected as the representative of the channel. The magnitude of each harmonic was estimated as the spectrum magnitude at the associated harmonic frequency. For instance, among the three harmonics contained in channel 2, the second harmonic is the strongest and will be selected accordingly.

d. Frequency downshift. Provided that the k th harmonic was selected for a given channel, it would be transformed into the modulator for the respective channel by a frequency downshift operation. Specifically, the spectrum of $h_k(t)$ would be transposed from its original location kF_0 to the F_0 , which is equivalent to multiplying $h_k(t)$ by an F_0 -dependent complex exponential function in the time domain (Li *et al.*, 2010).

To facilitate the frequency-shifting analysis, the harmonic model in Eq. (1) was converted into the following analytic form:

$$h_k(t) = \text{Re}\{a_k(t)e^{j(2\pi kF_0 t + \varphi_k(t))}\}, \quad (2)$$

where “Re” means taking the real part of a complex signal and symbol $j = \sqrt{-1}$. The analytic representation of $h_k(t)$ was first multiplied by a complex exponential function $e^{-j2\pi(k-1)F_0 t}$. The real part of the complex result was then taken to yield the transposed harmonic $\tilde{h}_k(t)$:

$$\begin{aligned} \tilde{h}_k(t) &= \text{Re}\{a_k(t)e^{j(2\pi kF_0 t + \varphi_k(t))}e^{-j2\pi(k-1)F_0 t}\} \\ &= a_k(t) \cos(2\pi F_0 t + \varphi_k(t)). \end{aligned} \quad (3)$$

To differentiate it from the traditional non-negative envelope, $\tilde{h}_k(t)$ was called the HSSE modulator in this study. Comparing Eqs. (1) and (3), one can see that $\tilde{h}_k(t)$ conveys the same AM cues as the original harmonic but oscillates at

a much slower rate. For voiced speech, $\tilde{h}_k(t)$ would oscillate regularly at the rate of F_0 instead of kF_0 . During the unvoiced segments, $\tilde{h}_k(t)$ would appear noise-like because $\varphi_k(t)$ varied randomly over time and caused the overall fluctuation to be irregular, although an interpolated F_0 was used in the frequency downshift operation.

e. HSSE modulator extraction. Figure 3(A) shows the functional blocks of HSSE processing. The incoming sound was first filtered into N channels. Within each channel, the strongest harmonic was identified [as described previously, yet not included in Fig. 3(A)] and then frequency downshifted, represented as multiplications between band signals and complex exponential functions. As a result, the strongest harmonic within each channel was transposed to the F_0 . Next, each transposed band signal was passed through a filter to keep only the strongest harmonic in the modulator. Because the F_0 of human voice is typically above 50 Hz, each transposed signal was first high-pass filtered at 50 Hz (third-order Butterworth), and then low-pass filtered at 300 Hz (third-order Butterworth) to limit the temporal information in the modulator. The combined effect is equivalent to a bandpass filter (50–300 Hz, third-order Butterworth), as shown in Fig. 3(A). Finally, the real part of the filter output was taken to yield the HSSE modulator for each channel, as implied in Eq. (3).

As a visual example, Fig. 2 shows the extracted HSSE modulators of a speech signal whose spectrogram is displayed on the left. For the voiced segment, the extracted HSSE modulators oscillate at a common rate of F_0 and exhibit coherent AM cues across channels, suggesting that they are from the same source. For the unvoiced segment, noise-

FIG. 3. (Color online) Acoustic simulation schemes of HSSE (A) and CIS (B), with the total number of channels being N . In (A), the multiplications between band signals and complex exponential functions (e.g., $e^{-j2\pi K F_0 t}$) represent frequency downshift operations; the block $\text{Re}\{\cdot\}$ means taking the real part of a complex signal.

like TFS cues are conveyed in the extracted HSSE modulators. One can see that the spectral profile of the speech signal is represented in the relative amplitudes of HSSE modulators between different channels.

In electric stimulation, where a non-negative stimulation signal is required, HSSE modulators can be further half-wave rectified. Alternatively, it is also possible to deliver HSSE modulators without rectification, e.g., by analog stimulation. Given that the purpose of vocoder simulation is to investigate the maximum potential of a strategy, it is not sensible to include rectification in the HSSE simulation, because rectification is an incoherent operation (introduces distortions) and the goal is to instead investigate the potential benefits of nondistorted TFS information to speech perception. In addition, there is a difference between acoustic simulation and electric stimulation regarding how rectification can be used. In acoustic simulation, if a 300-Hz-wide modulator is rectified, it must be low-pass filtered again to constrain the temporal information within 300 Hz; whereas in electric stimulation, it is unnecessary to low-pass filter a rectified modulator, given that the modulator starts as a 300 Hz signal. Because rectification does not necessarily make the simulation more closely resemble CI percepts and distorts the TFS cues, it was not included in the HSSE simulation.

2. Acoustic simulations of HSSE and CIS

The simulation diagrams of HSSE and CIS are shown in Figs. 3(A) and 3(B), respectively. First, the incoming sound was passed through a four- or eight-channel analysis filterbank (third-order Butterworth) spaced from 80 to 6000 Hz according to the Greenwood map (Greenwood, 1990). The corner frequencies used in the four-channel simulation were set to 80, 384, 1065, 2588, and 6000 Hz. In eight-channel simulation, the filter cutoff frequencies were 80, 202, 384, 657, 1065, 1675, 2588, 3955, and 6000 Hz.

Next, the temporal encoding of each band was executed. In HSSE, a 300 Hz HSSE modulator was extracted, as described previously, whereas in CIS, a Hilbert envelope was extracted and then low-pass filtered at 300 Hz (third-order Butterworth). Because the pitch saturation limit in electric hearing is typically restricted to 300 Hz, to prevent vocoders from presenting information that may not be accessible to CI users, both HSSE modulators and CIS envelopes were low-pass filtered at 300 Hz, such that normal-hearing listeners receive similar information as CI users, or at least represent the upper bound of CI performance.

For a visual comparison, the CIS and HSSE encodings of an identical sound, the “*eas*” segment from “*easy*,” are displayed side by side in Fig. 2. In the quiet condition, the F_0 cues are conveyed with both strategies yet by different mechanisms. With CIS, the F_0 cues result from the beating of unresolved harmonics, whereas in HSSE the F_0 cues lie in the transformed TFS of selected harmonics. Despite a possible difference in F_0 salience between CIS and HSSE, they seem to convey similar AM cues within each channel. In the noise condition, both AM and F_0 cues are evidently distorted in CIS encoding, with the first two channels being disturbed to a greater extent than the last two channels, because

speech-shaped noise produces more interference in low frequencies than in high frequencies. In contrast, the AM and F_0 cues in HSSE encoding appear to be less distorted, because HSSE modulators are extracted from predominant harmonics that are stronger and thus more resilient to interference (Darwin, 2008).

Finally, the temporal encoding of each band is multiplied by a sine carrier at the respective channel center frequency. Tone vocoders were chosen because they are thought to more closely resemble CI percepts, whereas noise carriers introduce random fluctuations in the envelope and distort TFS cues (Whitmal *et al.*, 2007). In the last processing step, the modulated carriers are combined and then band-pass filtered again (80–6000 Hz, third-order Butterworth) to constrain the total information within the analysis spectral range.

3. Subjects

Ten native English speakers with normal hearing participated in the sentence recognition test. As in all of the tests in this study, subjects were seated in a double-walled, sound-insulated booth (IAC, Bronx, NY). A program written in MATLAB (Mathworks, Natick, MA) played simulation sounds to subjects and recorded their responses via a graphical user interface. Sounds were presented at 68 dB SPL through an Apple PowerMacG5 sound card connected to a Crown D45 amplifier (Crown International, Elkhart, IN) and a free-standing studio monitor (B&W DM303 speaker from Bowers & Wilkins, North Reading, MA). During the test, the speaker was placed at head level, about 1 m in front of subjects. The University of Washington Institutional Review Board approval and informed consent were obtained from all subjects.

4. Stimuli and test procedure

Two sets of materials, the Hearing in Noise Test [(HINT); Nilsson *et al.*, 1994] sentences and the Institute of Electrical and Electronic Engineers [(IEEE); Rothauser *et al.*, 1969] sentences, were used to sample a wide range of listening abilities and to avoid a potential ceiling effect during HINT testing (e.g., Dorman *et al.*, 1998) and a floor effect with IEEE sentences (e.g., Stickney *et al.*, 2004). The target HINT and IEEE sentences were produced by different male talkers with a mean F_0 of 110 and 108 Hz, respectively.

To investigate the effect of HSSE on speech perception, a pilot study was done using IEEE sentence recognition in quiet: A clear advantage of HSSE over CIS was observed in the four-channel condition (76% versus 52%), but the performance was too high with both strategies in the eight-channel condition (>90%). HINT sentence recognition was then tested in noise at +10 dB signal-to-noise-ratio (SNR): A ceiling effect was observed again in the eight-channel condition. Thus, a fixed SNR of +5 dB was chosen.

Two types of maskers were used in the HINT test: A speech-spectrum shaped noise (SSN) and a competing female talker (mean F_0 = 219 Hz). These two types of maskers were also used in the IEEE test; additionally, a male-talker masker (mean F_0 = 136 Hz) was included to explore the effect of masker type on speech perception. The

SSN masker was generated by filtering white noise with the target sentence's long-term spectral profile. The competing sentences were selected from the IEEE corpus and varied for each presentation. The masker always lasted longer than the target sentence. The target and the masker were first mixed at +5 dB SNR and then vocoder processed to generate a simulation sound.

To compare the effect of processing strategies, a within-subject test design was used. Each participant listened to both CIS- and HSSE-vocoded speech under the same masker conditions. During the HINT test, eight conditions were used, covering two strategies (CIS and HSSE), two sets of channel numbers (four and eight), and two types of maskers (SSN and female masker). The eight-channel with SSN masker condition was eventually dropped because subjects' performance was too high (>90%) to yield meaningful comparison. Previous studies (e.g., Culling and Zeng, 2008) showed similar performance levels. During the IEEE test, results were obtained under six conditions, two strategies (CIS and HSSE), and three types of maskers (SSN, male, and female maskers); the channel number was fixed at eight.

Five normal-hearing subjects participated in the HINT test and another five, different, normal-hearing subjects participated in the IEEE test. For each subject, the order of test conditions was randomized. Under a given condition, subjects began by listening to two practice sentences to familiarize themselves with the test stimuli. Afterwards, they began the actual test, which consisted of 20 new sentences. Each subject was presented with a different set of sentences for the same condition and was instructed to type in their responses using a computer keyboard. Listeners' performance was evaluated offline by the experimenter as the percentage of keywords correctly recognized.

B. Results

The average intra-subject performance on the HINT test is shown in Fig. 4(a). Across all testing conditions, subjects performed significantly better with HSSE than with CIS vocoders. The largest improvement was observed under the four-channel condition: Subjects scored 42% higher with HSSE in steady-state noise [paired $t(4)=9.7$, $p=0.001$], and 45% higher against a female competing talker [paired $t(4)=6.6$, $p=0.002$]. Under the eight-channel condition with a female-talker masker, subjects also improved significantly with HSSE over CIS vocoders [paired $t(4)=3.0$, $p=0.038$].

Figure 4(b) shows subjects' average performance on the IEEE test, during which the channel number was fixed at eight. In all of the three masker conditions, subjects scored higher with HSSE than with CIS vocoders. Of the three maskers, subjects performed the worst in the male-masker condition, yet for the same condition, HSSE showed the largest advantage over CIS, approximately 30% [paired $t(4)=6.8$, $p=0.002$]. A significant benefit of HSSE was also observed for speech recognition in SSN [paired $t(4)=4.8$, $p=0.009$] and in the female-masker condition [paired $t(4)=3.1$, $p=0.036$].

FIG. 4. The mean intra-subject performance on HINT (a) and IEEE (b) sentence recognition tests with CIS (hatched bars) and HSSE (filled bars) vocoders. The SNR was fixed at 5 dB across all masking conditions. Error bars represent the standard error of the mean.

III. EXPERIMENT 2: MANDARIN TONE IDENTIFICATION WITH SIMULATED HSSE AND CIS STRATEGIES

Understanding of a tonal language depends not only on phoneme recognition, but also on identification of changing fundamental frequencies. For example, in Mandarin Chinese, the monosyllabic word “ji” can be pronounced with either a flat (jī 鸡), rising (jí 急), falling then rising (jǐ 几), or a falling (jì 记) fundamental frequency. Each word has an entirely different meaning. In experiment 2, the potential effect of HSSE on Mandarin tone identification was investigated, using both acoustic simulation and computational modeling approaches.

A. Methods

1. Behavioral test

a. Subjects. Five native Mandarin Chinese speakers with normal hearing participated in experiment 2.

b. Stimuli and test procedure. The test procedure and stimuli were adopted from Xu *et al.* (2002). All of the stimuli were vocoder processed in the same way as in experiment 1. The stimuli were drawn from ten lists. Each list consists of four words, all of which have the same syllable but different tonal patterns. The stimuli were pronounced by both female and male speakers. Their F_0 profiles in producing the four words of “ji” list are displayed in Fig. 5, to show the dependence of the F_0 profile on the gender. To prevent subjects from using duration as one of the identification cues, all syllables in a particular list were made the same in duration through careful selection of multiple recordings.

Tone recognition was evaluated under four conditions, consisting of two strategies (CIS and HSSE) and two sets of channel numbers (four and eight). The test order of four conditions was randomized within each subject. Under each condition, a four-interval, four-alternative forced-choice paradigm was used. In each particular trial, a word list was first

FIG. 5. (Color online) The estimated F_0 profile of the female (dashed lines) and the male (solid lines) speaker in pronouncing jī, jí, jǐ, and jì, respectively. The F_0 variation amount for each tone is displayed along with the F_0 profile. ST stands for semitone.

randomly chosen from either the female or the male recordings. One of its four words was then selected, again randomly, as the trial stimulus. Overall, both the 10 female lists and the 10 male lists were presented twice in random order, resulting in a total of 40 lists and 160 trials per condition. Subjects' performance was calculated as the percent of correctly identified tones.

2. Computational modeling

a. Description of the model. Implementation details of the model as well as its parameter set have been presented in Mino *et al.* (2004) and Imennov and Rubinstein (2009), respectively. The input to the model is the electric pulse train generated by a CI processor for a particular electrode. The output is the simulated neural discharge patterns evoked by the input pulse train.

The model is based on the morphology and electrophysiology of spiral ganglion cells (SGCs). In CI users, SGCs serve as a locus where electric stimuli are first converted into neural responses in CI users. Given that subsequent neurological processing can only extract, but not add, information from the incoming stimuli, neural responses at SGCs provide an upper bound on the auditory information available to CI users. Because the normalized response properties of a population of diameter-distributed model fibers have been shown to match that of the same number of *in vivo* fibers (Imennov and Rubinstein, 2009), the same distribution of 250 fibers was used to generate all of the neural outputs in this study.

b. Generation of electric pulse train. Analogous to the vocoder processing, eight-channel CIS and HSSE implementations were used to generate the electric encoding of a particular stimulus. Because the model was inherently single-channel, neural responses in each spectral channel would be simulated independently. To gauge the best potential of a strategy, the third band ([384 657] Hz) was selected for simulation, because visual observation suggested that clear F_0

cues were present in this band. With CIS, the F_0 fluctuations were due to beats of unresolved harmonics, while in HSSE the F_0 cues were represented in the TFS of a resolved harmonic.

As described earlier, a 300-Hz-wide CIS envelope and HSSE modulator were first extracted from the third band; they were then logarithmically compressed and converted into an electric pulse train, respectively, using the Nucleus MATLAB toolbox (Swanson and Mauch, 2006). The negative samples in HSSE modulators were set to zero during log-compression, i.e., HSSE modulators were half-wave rectified. To generate a faithful representation of TFS, the per channel stimulation rate was set to 1900 Hz. Each pulse was biphasic and 25 μ s wide. The cathodic phase was applied first, followed by an 8 μ s gap and an equal-amplitude anodic phase of the pulse.

c. Stimuli and analysis of results. Due to the high computational cost of simulating neural responses for the whole set of tone stimuli, only the "jī" list was selected, for which both the CIS and the HSSE pulse trains exhibit F_0 cues to allow a further comparison of F_0 encoding with the evoked spike trains. Each stimulus of the list is a 600 ms recording of a female speaker pronouncing the word "jī" with one of the four tonal patterns. The estimated F_0 profile for each stimulus is shown in Fig. 5.

To examine the amount of F_0 information captured by the auditory nerve, raster-plots of the simulated neural spikes along 250 fibers were generated and evaluated qualitatively. Furthermore, the intervals between successive spikes were analyzed. Given that the F_0 cues were encoded temporally by both strategies, an increase in the speaker's F_0 would produce a decrease in the interspike intervals (ISIs). The converse also holds true: As F_0 decreases, the spikes should occur more sparsely, producing larger ISIs. Therefore, by measuring how the ISI changes as the speaker's F_0 evolves, a strategy's ability to convey F_0 cues was evaluated.

B. Results

1. Behavioral results

Overall, subjects could identify Mandarin tone better with HSSE than with CIS vocoders. Their mean score, averaged across all four tonal patterns, is shown in Fig. 6(a). Subjects scored 24% higher with HSSE than with CIS vocoders in the four-channel condition [paired $t(4) = 8.8$, $p = 0.001$] and 14% higher in the eight-channel condition [paired $t(4) = 8.3$, $p = 0.001$]. Their mean score on recognizing each individual tonal pattern is displayed in Figs. 6(b)–6(e), respectively. With CIS, the recognition performance on rising tone was the worst; HSSE demonstrated the greatest benefit in identifying rising tones. A three-way repeated measures analysis of variance (strategy, channel number, tonal pattern) revealed that processing strategy [$F(1, 4) = 200.3$, $p = 0.001$], channel number [$F(1, 4) = 30.2$, $p = 0.005$], and tonal pattern [$F(3, 12) = 19.5$, $p = 0.001$] all had a significant effect on tone identification. There was a significant interaction between strategy and channel number ($p = 0.030$), resulting from the fact that subjects' score with CIS varied substantially as a

FIG. 6. (a) The average intra-subject performance on Mandarin tone identification with CIS (hatched bars) and HSSE (filled bars) vocoders. (b)–(e) Subjects' mean score on recognizing flat, rising, falling then rising, and falling tones, respectively. Error bars represent the standard error of the mean.

function of channel number [see Fig. 6(c)]. A significant interaction was also observed between strategy and tonal pattern, reflecting that subjects' score with CIS was largely affected by tonal pattern. In contrast, subjects' performance with HSSE was consistently good across all of the four tone patterns under both channel conditions (ranged 95%–100%).

2. Modeling results

a. Raster-plots. Due to space constraints, only the raster-plot for the $\tilde{j}$ stimulus is provided, which shows both directions of F_0 glide. Figure 7 is divided into two subpanels. The CIS encoding and the evoked neural responses are displayed in Fig. 7(A), while Fig. 7(B) shows the results of HSSE. Within Figs. 7(A) and 7(B), the top row shows the input electric pulse train; the bottom shows the simulated neural spike train by placing a dot for every occurrence of a spike.

Comparing Figs. 7(A) and 7(B), one can see that the CIS and the HSSE evoked spike trains exhibit similar envelope cues, e.g., the synchronized onset and offset patterns, but they convey different timing cues. The HSSE-evoked spike train displays clear troughs and peaks following the F_0 cues in the HSSE pulse train, whereas such a timing pattern is missing in the CIS-evoked spike train. Although the F_0 cues in the CIS pulse train are visible, the modulation depth is comparatively shallow. Consequently, the CIS stimulation causes saturation in large-diameter fibers ($>3.6\ \mu\text{m}$), forcing most of the F_0 cues to reside only in small-diameter fibers ($<3.6\ \mu\text{m}$). In contrast, HSSE might encode F_0 in the duration of a pulse burst and in the interval between successive bursts. Both properties are captured in the simulated HSSE spike trains—e.g., as the speaker's F_0 decreases between 100 and 290 ms, both the width of the spike bursts and the interval between successive bursts increase noticeably.

FIG. 7. Simulated neural discharge patterns under CIS (A) and HSSE stimulation (B), respectively. Within (A) and (B), the top row shows the electric encoding of stimulus $\tilde{j}$ by a CIS/HSSE processor for one electrode; the bottom displays the raster-plot of the evoked neural spike train. The regions labeled ROIs 1–3 will be sampled in the analysis of interspike intervals.

b. ISI histogram. Because the speaker's F_0 is time-varying (Fig. 5), each neural spike train was accordingly time-divided into three segments, designated as regions of interest (ROIs) 1, 2, and 3. The location of each ROI is indicated in Fig. 7 as black bars placed on top of the raster-plots. For each ROI, the intervals between successive spikes were calculated and the overall ISI distribution is shown in Fig. 8. Figures 8(A)–8(D) correspond to stimuli $\bar{j}\bar{i}$, $j\acute{i}$, $j\check{i}$, and $j\grave{i}$, respectively. Within each panel, the histograms on the left were sampled from the CIS-evoked spike trains, and those on the right were from the HSSE-evoked responses. The peak of each ISI histogram, corresponding to the speaker's F_0 at the respective ROI, is indicated by a dashed vertical line. Because

the speaker's F_0 was lower than 500 Hz, only ISIs longer than 2 ms were considered in locating a histogram's peak.

Comparing the histograms within each panel, for an identical stimulus, the peak location is the same between CIS and HSSE histograms, but the peak height is notably lower with CIS, sometimes even invisible. This difference in peak height arises because fewer F_0 cues are available in the CIS- than in the HSSE-evoked spike trains. With HSSE, the ISI histograms exhibit clear peaks, correctly capturing the evolution of a particular tonal pattern: The F_0 -related peak remained in the same location in response to $\bar{j}\bar{i}$, shifted to the right and left when stimulated with $j\acute{i}$ and $j\check{i}$, respectively, and exhibited a clear displacement in response to $j\grave{i}$. In fact,

FIG. 8. Histograms of interspike intervals. Panels (A)–(D) correspond to stimulus $\bar{j}\bar{i}$, $j\acute{i}$, $j\check{i}$, and $j\grave{i}$, respectively. Within each panel, the histograms on the left were sampled from a CIS-evoked spike train and those on the right were from an HSSE-evoked spike train. From top to bottom, the sample regions are sequentially ROIs 1–3. Vertical dashed lines indicate histogram peaks; only ISIs longer than 2 ms were considered in locating the peaks, because the stimuli's F_0 is below 500 Hz.

one can identify the tone supplied to the model based solely on the HSSE yielded ISI distributions. A comparable interpretation is considerably harder with CIS: While it might be possible to deduce the F_0 profile of jí, the absence of histogram peaks makes the remaining identifications difficult.

IV. DISCUSSION

A. The effects of HSSE on simulated speech recognition in noise

This study showed that by encoding nondistorted harmonic and TFS information, HSSE can potentially improve perception of speech in a variety of masking conditions as well as Mandarin tones. During both four- and eight-channel simulations, HSSE demonstrated a clear advantage over CIS for all listening tasks tested. Specifically, a larger benefit of HSSE encoding was observed in the four-channel than in the eight-channel condition, suggesting an interaction between spectral and temporal cues for speech perception. Xu *et al.* (2002) reported that under a given channel condition, Mandarin tone identification with CIS can be improved by increasing the envelope cutoff frequency; the improvement was relatively larger when the channel number was smaller. Stone *et al.* (2008) observed a similar interaction between spectral and temporal cues for speech recognition against a competing talker. Presumably, when the channel number was decreased from eight to four, the temporal cues in HSSE can better compensate for the diminished spectral cues for speech perception; although the same temporal cutoff frequency was applied in both HSSE and CIS simulations.

Among the three types of maskers tested, the female and the male talker were found to generate more masking than the SSN masker, which is consistent with previous studies on the effect of masker type on speech recognition with CIs (*e.g.*, Stickney *et al.*, 2004). Between the female (mean $F_0 = 219$ Hz) and the male (mean $F_0 = 136$ Hz) talker, the male-masker condition was more difficult, because the target speaker (mean $F_0 = 108$ Hz) was also a male and the F_0 difference between he and the male masker was much smaller than that between he and the female masker. This effect of the F_0 difference on speech recognition in a competing-talker background was also reported by Cullington and Zeng (2008). Presumably, TFS information plays a larger role in speech recognition with a competing talker of the same gender, thus a greater benefit of HSSE encoding was observed in the male-masker condition than in the other two conditions.

There are several possible reasons why HSSE is advantageous over CIS for speech perception. First, HSSE uses linear operations to transform speech harmonics into modulators, whereas CIS uses nonlinear operations to extract channel envelopes. As illustrated in Fig. 1, nonlinear operations distort the frequency components in a signal, thus speech harmonics cannot be properly represented in CIS as in HSSE. Consequently, the potential benefit of harmonic information to speech perception is more likely to be restored by HSSE than by CIS. Second, HSSE appears to encode F_0 cues better than CIS. In CIS, the F_0 fluctuation is due to the beating of harmonics, which would be obviously distorted in the presence of noise (see Fig. 2), whereas the F_0 cues in HSSE modulators

are extracted from predominant harmonics, which are relatively resilient to noise. Previous studies have shown that the F_0 information is an important segregation cue for speech recognition against a competing talker (Summers and Leek, 1998; Turner *et al.*, 2004). The advantage of HSSE relies on the F_0 tracking accuracy, which is generally good at +5 dB SNR if the target and the competing-talkers' F_0 are apart to some degree. Therefore, the F_0 cues are more likely to be available to HSSE than to CIS listeners as a segregation cue for speech recognition in noise. Third, HSSE might encode TFS information better than CIS under the 300 Hz cutoff frequency. In the noise condition, because the speech AM cues are obscured by noise in the extracted envelope (see Fig. 2), additional TFS information is needed to assist the separation of intelligibility-related AM cues (Lorenzi *et al.*, 2006; Hopkins *et al.*, 2008). The temporal details in CIS envelopes contain distortions caused by incoherent operations, whereas the TFS cues in HSSE modulators are linearly extracted and thus free of distortions. Presumably, the TFS information in HSSE modulators is more beneficial to speech perception than the information in CIS envelopes.

B. The effects of HSSE on simulated Mandarin tone identification

Using HSSE vocoders, subjects could reliably identify each tonal pattern (>95%) under both four- and eight-channel conditions, whereas with CIS, their performance was comparatively poor under both channel conditions. The fact that subjects scored higher with a four-channel HSSE vocoder than with an eight-channel CIS vocoder indicates that under a spectral constraint of four to eight channels, tone identification relies primarily on temporal cues, while place cues can only assist F_0 discrimination to a lesser degree (Xu *et al.*, 2002).

With CIS, subjects' performance appeared to be dependent on tonal patterns: The recognition score on rising tones was considerably worse than that on falling tones, which was in line with the performance pattern observed in CI users (Han *et al.*, 2009). Subjects' performance with HSSE, however, was consistently high across all of the four tonal patterns. Presumably, there might be a smaller F_0 change during the evolution of a rising tone than a falling tone (Fig. 5). Because the F_0 cues conveyed in CIS envelopes were not salient, it was hard for CIS listeners to identify small F_0 changes during a rising tone. For example, Han *et al.* (2009) found that CI users often misperceived rising tones as flat, suggesting that the F_0 cues in CIS envelopes were too weak to elicit an accurate pitch percept.

The difference between CIS and HSSE in temporal F_0 encoding was also demonstrated in neural response simulation. For an identical tone stimulus, the CIS- and the HSSE-evoked spike trains exhibited a similar AM pattern, but different timing pattern. Although the F_0 cues in the CIS pulse train were visible, the modulation depth was comparatively shallow. Under a constant pulse rate, the CIS stimulation caused saturation in large-diameter fibers (>3.6 μm), forcing most of the F_0 cues to reside only in small-diameter fibers (<3.6 μm). In contrast, HSSE may encode F_0 in the

duration of a pulse burst and in the interval between successive bursts. Both properties were captured in the simulated HSSE spike trains. As found in the ISI histogram analysis, the dominant interval in an HSSE-evoked spike train corresponded to the stimulus' fundamental frequency period, whereas such a temporal code was not evident in the CIS-evoked responses. Thus, the modeling results supported the behavioral test results, indicating that HSSE is potentially advantageous over CIS for enhancing tonal pattern identification with CIs.

C. Implications for cochlear implants

Due to the inherently coarse spectral and temporal resolution in electric hearing, how to encode harmonic and TFS information for CI uses is still an open question. Various strategies have been proposed or implemented for improving TFS representation in CIs, most of which follow a scheme of extracting the AM and TFS cues of a band separately both by nonlinear operations, e.g., the strategy proposed by Nie *et al.* (2005) and the FSP strategy (Schatzer *et al.*, 2010). During electric stimulation, the AM cues are encoded in the amplitude of pulses, while the TFS cues are represented in the timing pattern of pulses. The F0mod strategy follows a similar scheme, except that it directly encodes F_0 instead of TFS cues in addition to the AM information (Laneau *et al.*, 2006). These strategies are like CIS in terms of AM extraction: The operation involved introduces nonlinear distortions and causes misrepresentation of speech harmonics. In contrast, HSSE modulators can closely resemble the original harmonics. Because CI users typically have a small number of effective channels, HSSE delivers only predominant harmonics that are resilient to interference and thus important to speech recognition in noise (Darwin, 2008). The fact that a subset of predominant harmonics—instead of a complete set—can be sufficient to benefit speech perception suggests the feasibility of HSSE with CIs.

Another distinction of HSSE is that the TFS cues in HSSE modulators are never separated from the AM cues except being transformed from high to low frequency as a whole, whereas CIS-like strategies extract AM and TFS cues separately. Schatzer *et al.* (2010) showed that the FSP strategy produced similar performance as CIS on Cantonese tone recognition, regardless of the TFS representation from 100 to 800 Hz, suggesting that the TFS cues encoded in FSP might not be accessible or beneficial to CI uses. Milczynski *et al.* (2012) demonstrated that the F0mod strategy outperformed the ACE strategy on Mandarin tone perception but not on sentence recognition. Typically, CI users can perceive temporal information up to about 300 Hz. Under the temporal constraint of 300 Hz, HSSE produced better sentence recognition and Mandarin tone identification performance than CIS in vocoder simulations. Moreover, the spike data showed that the temporal cues in HSSE modulators can be translated into the simulated neural responses (see Fig. 7). Given that the model's response properties have been shown to closely resemble the response quantities measured *in vivo* (Miller *et al.*, 1999; Imennov and Rubinstein, 2009), these results suggest that HSSE is a promising strategy to enhance

speech perception with CIs. The reduced accuracy in phase locking poses a great challenge for CI users to perceive TFS. Compared with the zero crossing information encoded by FSP, the low-frequency TFS cues in HSSE modulators are potentially more accessible and beneficial to CI users.

HSSE is aimed to improve temporal encoding in CIs and the present study has demonstrated its potential benefit to speech perception. However, it might be subject to several limitations in electric hearing, such as neural degeneration and poor temporal resolution. If there is neural degeneration in patients, their perception is likely to be adversely affected. The extent of neural degeneration would influence the extent to which the improved temporal information could be effectively used, which would result in individually variable benefits for HSSE. On the other hand, to implement HSSE in real time, an efficient F_0 tracker is required. Pitch tracking is technically solvable in high SNR conditions, e.g., >5 dB, although the tracking process increases the computational cost (Vandali and van Hoesel, 2011). Once the F_0 is known, the extraction of each individual harmonic can be executed in parallel. Given the computational resource of a modern processor, it is feasible to implement HSSE in real time.

More than a quarter of the world's population use tonal languages. The relatively poor tone identification performance with CIS-like strategies poses a significant challenge to CI users relying on tonal languages. The proposed HSSE strategy can potentially make a great impact on their life, given that the F_0 cues conveyed by a four-channel HSSE vocoder can support reliable tone identification and these F_0 cues can translate to the timing patterns of simulated neural responses. In addition to the F_0 cues for voiced speech, HSSE can also preserve noise-like TFS cues for unvoiced speech, possibly producing a more natural speech signal than CIS. By improving the encoding of harmonic and TFS information, HSSE appears to be a promising strategy to improve speech perception with CIs.

ACKNOWLEDGMENTS

The authors thank Professor Li Xu of the Ohio University for providing the Mandarin tone stimuli and all of the subjects for participating in the study. We also thank the anonymous reviewers for their helpful and constructive comments. This research was supported by National Institutes of Health Grants No. R01-DC007525, No. DC-010148, No. P30-DC004661, and No. T32-DC005361; Air Force Office of Scientific Research Grant No. FA95500910060; University of Washington Commercialization Gap Fund 657635; Institute of Translational Health Sciences Grant No. 620491; and National Science Foundation Grant No. TG-IBN090004.

- Clark, P., and Atlas, L. E. (2009). "Time-frequency coherent modulation filtering of nonstationary signals," *IEEE Trans Signal Process* **57**, 4323–4332.
- Cullington, H. E., and Zeng, F. G. (2008). "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," *J. Acoust. Soc. Am.* **123**, 450–461.
- Darwin, C. J. (2008). "Listening to speech in the presence of other sounds," *Philos. Trans. R. Soc. London, Ser. B* **363**, 1011–1021.
- Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (1998). "The recognition of sentences in noise by normal-hearing listeners using simulations of

- cochlear-implant signal processors with 6–20 channels,” *J. Acoust. Soc. Am.* **104**, 3583–3585.
- Drennan, W. R., Won, J. H., Nie, K., Jameyson, E., and Rubinstein, J. T. (2010). “Sensitivity of psychophysical measures to signal processor modifications in cochlear implant users,” *Hear. Res.* **262**, 1–8.
- Flanagan, J. L. (1980). “Parametric coding of speech spectra,” *J. Acoust. Soc. Am.* **68**, 412–419.
- Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (2001). “Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants,” *J. Acoust. Soc. Am.* **110**, 1150–1163.
- Greenwood, D. D. (1990). “A cochlear frequency-position function for several species—29 years later,” *J. Acoust. Soc. Am.* **87**, 2592–2605.
- Han, D., Liu, B., Zhou, N., Chen, X., Kong, Y., Liu, H., Zheng, Y., and Xu, L. (2009). “Lexical tone perception with HiResolution and HiResolution 120 sound-processing strategies in pediatric Mandarin-speaking cochlear implant users,” *Ear Hear.* **30**, 169–177.
- Hopkins, K., Moore, B. C. J., and Stone, M. A. (2008). “Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech,” *J. Acoust. Soc. Am.* **123**, 1140–1153.
- Imennov, N. S., and Rubinstein, J. T. (2009). “Stochastic population model for electrical stimulation of the auditory nerve,” *IEEE Trans. Biomed. Eng.* **56**, 2493–2501.
- Laneau, J., Wouters, J., and Moonen, M. (2006). “Improved music perception with explicitly pitch coding in cochlear implants,” *Audiol. Neuro-Otol.* **11**, 38–52.
- Li, Q., and Atlas, L. E. (2003). “Time-variant least squares harmonic modeling,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, pp. II - 41–44.
- Li, X., Nie, K., Atlas, L., and Rubinstein, J. (2010). “Harmonic coherent demodulation for improving sound coding in cochlear implants,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, TX, pp. 5462–5465.
- Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (2006). “Speech perception problems of the hearing impaired reflect inability to use temporal fine structure,” *Proc. Natl. Acad. Sci. U.S.A.* **103**, 18866–18869.
- McAulay, R. J., and Quatieri, T. F. (1986). “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, Signal Process.* **ASSP-34**, 744–754.
- Milczynski, M., Chang, J. E., Wouters, J., and van Wieringen, A. (2012). “Perception of Mandarin Chinese with cochlear implants using enhanced temporal pitch cues,” *Hear. Res.* **285**, 1–12.
- Miller, C. A., Abbas, P. J., and Rubinstein, J. T. (1999). “An empirically based model of the electrically evoked compound action potential,” *Hear. Res.* **135**, 1–18.
- Mino, H., Rubinstein, J. T., Miller, C. A., and Abbas, P. J. (2004). “Effects of electrode-to-fiber distance on temporal neural response with electrical stimulation,” *IEEE Trans. Biomed. Eng.* **51**, 13–20.
- Moore, B. (2008). “The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people,” *J. Assoc. Res. Otolaryngol.* **9**, 399–406.
- Nie, K., Atlas, L., and Rubinstein, J. (2008). “Single sideband encoder for music coding in cochlear implants,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, pp. 4209–4212.
- Nie, K., Stickney, G., and Zeng, F.-G. (2005). “Encoding frequency modulation to improve cochlear implant performance in noise,” *IEEE Trans. Biomed. Eng.* **52**, 64–73.
- Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). “Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and noise,” *J. Acoust. Soc. Am.* **95**, 1085–1099.
- Oxenham, A. J. (2008). “Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants,” *Trends Amplif.* **12**, 316–331.
- Riss, D., Hamzavi, J. S., Selberherr, A., Kaider, A., Blineder, M., Starlinger, V., Gstottner, W., and Arnoldner, C. (2011). “Envelope versus fine structure speech coding strategy: A crossover study,” *Otol. Neurotol.* **32**, 1094–1101.
- Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (1969). “I.E.E.E. recommended practice for speech quality measurements,” *IEEE Trans. Audio Electroacoust.* **AU-17**, 227–246.
- Schatzer, R., Krenmayr, A., Au, D. K., Kals, M., and Zierhofer, C. (2010). “Temporal fine structure in cochlear implants: Preliminary speech perception results in Cantonese-speaking implant users,” *Acta Oto-Laryngol.* **130**, 1031–1039.
- Schimmel, S. M., and Atlas, L. (2005). “Coherent envelope detection for modulation filtering of speech,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 221–224.
- Shackleton, T. M., and Carlyon, R. P. (1994). “The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination,” *J. Acoust. Soc. Am.* **95**, 3529–3540.
- Shannon, R. (1992). “Temporal modulation transfer functions in patients with cochlear implants,” *J. Acoust. Soc. Am.* **91**, 2156–2164.
- Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (2004). “Cochlear implant speech recognition with speech maskers,” *J. Acoust. Soc. Am.* **116**, 1081–1091.
- Stone, M. A., Fullgrabe, C., and Moore, B. C. (2008). “Benefit of high-rate envelope cues in vocoder processing: Effect of number of channels and spectral region,” *J. Acoust. Soc. Am.* **124**, 2272–2282.
- Summers, V., and Leek, M. R. (1998). “F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss,” *J. Speech Lang. Hear. Res.* **41**, 1294–1306.
- Swanson, B., and Mauch, H. (2006). *Nucleus MATLAB Toolbox 4.20 Software User Manual* (Cochlear Ltd., Lane Cove, Australia).
- Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A., and Henry, B. A. (2004). “Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing,” *J. Acoust. Soc. Am.* **115**, 1729–1735.
- Vandali, A. E., and van Hoesel, R. J. (2011). “Development of a temporal fundamental frequency coding strategy for cochlear implants,” *J. Acoust. Soc. Am.* **129**, 4023–4036.
- Wang, S., Xu, L., and Mannell, R. (2011). “Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners,” *J. Assoc. Res. Otolaryngol.* **12**, 783–794.
- Whitmal, N. A., Poissant, S. F., Freyman, R. L., and Helfer, K. S. (2007). “Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience,” *J. Acoust. Soc. Am.* **122**, 2376–2388.
- Wilson, B. S., and Dorman, M. F. (2008). “Cochlear implants: A remarkable past and a brilliant future,” *Hear. Res.* **242**, 3–21.
- Xu, L., Chen, X., Lu, H., Zhou, N., Wang, S., Liu, Q., Li, Y., Zhao, X., and Han, D. (2011). “Tone perception and production in pediatric cochlear implants users,” *Acta Oto-Laryngol.* **131**, 395–398.
- Xu, L., Tsai, Y., and Pfungst, B. E. (2002). “Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses,” *J. Acoust. Soc. Am.* **112**, 247–258.
- Zeng, F.-G. (2002). “Temporal pitch in electric hearing,” *Hear. Res.* **174**, 101–106.