Revised submission -3

Effects of noise suppression and envelope dynamic range compression on the intelligibility of vocoded sentences for a tonal language

Fei Chen [a)]

Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Xueyuan Road 1088#, Xili, Nanshan District, Shenzhen 518055, China

Dingchang Zheng

Health and Wellbeing Academy, Faculty of Medical Science, Anglia Ruskin University, Chelmsford, UK

Yu Tsao

Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

Wednesday, August 2, 2017

[a)] Author to whom correspondence should be addressed:

Fei Chen, Ph. D.

Department of Electrical and Electronic Engineering

Southern University of Science and Technology

Xueyuan Road 1088#, Xili, Nanshan District, Shenzhen 518055, China

Phone: 0086-755-88018554 Email: fchen@sustc.edu.cn

PACS number: 43.71.Gv, 43.71.Es

Keywords: Noise suppression, envelope dynamic range compression.

Running title: Intelligibility of vocoded sentences

Abstract: Vocoder simulation studies have suggested that the carrier signal type employed affects the intelligibility of vocoded speech. The present work further assessed how carrier signal type interacts with additional signal processing, namely, single-channel noise suppression and envelope dynamic range compression, in determining the intelligibility of vocoder simulations. In Experiment 1, Mandarin sentences that had been corrupted by speech spectrum-shaped noise (SSN) or two-talker babble (2TB) were processed by one of four single-channel noise-suppression algorithms before undergoing tone- (TV) or noise-vocoded (NV) processing. In Experiment 2, dynamic ranges of multiband envelope waveforms were compressed by scaling of the mean-removed envelope waveforms with a compression factor before undergoing TV or NV processing. TV Mandarin sentences yielded higher intelligibility scores with normal-hearing (NH) listeners than did noise-vocoded sentences. The intelligibility advantage of noise-suppressed vocoded speech depended on the masker type (SSN vs. 2TB). NV speech was more negatively influenced by envelope dynamic range compression than was TV speech. These findings suggest that an interactional effect exists between the carrier signal type employed in the vocoding process and envelope distortion caused by signal processing.

I. INTRODUCTION

The perceptual contribution of the temporal envelope has attracted enduring research interest. Many studies have assessed the importance of the temporal envelope for speech intelligibility under various conditions (e.g., Shannon *et al*., 1995; Dorman *et al*. 1997; Chen and Loizou, 2011a). Vocoder simulations have long been used to extract the multiband temporal envelope waveforms while removing the underlying fine-structure information to synthesize envelope-based vocoded speech (e.g., Shannon *et al*., 1995; Dorman *et al*. 1997; Chen and Loizou, 2011a). With envelope information from up to four bands, normal-hearing (NH) listeners can have near-perfect speech understanding in quiet condition (Shannon *et al*., 1995).

In a cochlear implant (CI) device, incoming sound signals are received via a microphone and fed into a speech processor. Most of the existing CI speech processors capture multi-channel temporal envelopes of sound signal inputs, and then generate electric stimulations that excite patients' residual auditory nerves directly. Vocoders aim to transfer only those acoustic cues that are present for CI users, so they simulate the signal processing of a CI. Vocoder simulations have been applied to examine numerous factors that influence the intelligibility of envelope-based vocoded speech, including the number of channels (Shannon *et al*., 1995; Dorman *et al*. 1997), carrier signal type (Dorman *et al*. 1997; Fu *et al*., 2004; Gonzalez and Oliver, 2005; Whitmal, *et al*., 2007; Chen and Lau, 2014), envelope cutoff frequency (Shannon *et al*. 1995; Xu *et al*. 2005; Souza and Rosen 2009), and frequency spacing (Kasturi and Loizou, 2007), among other factors. For this reason, vocoder simulations have been used widely to assess the potential of new speech-processing and coding strategies for CIs before large-scale clinical evaluations with users are conducted. Vocoder simulation remains a valuable tool in the field of CI research because it can be used to assess the effects of acoustic factors in the absence of patient-specific confounds.

When performing vocoder simulations, the envelope waveform is extracted by steps of bandpass filtering (BPF), waveform rectification and low-pass filtering (LPF) (see Fig. 1). The envelope waveform is used to modulate a carrier signal. There are two common types of carrier signals used in synthesizing vocoded speech; pure-tone and white-noise signals yield tone- (TV) and noise-vocoded (NV) speech stimuli, respectively. A limited number of studies have compared the relative performance of these two vocoder types on speech intelligibility in English (e.g., Dorman *et al*. 1997; Whitmal, *et al*., 2007; Souza and Rosen, 2009; Rosen *et al*., 2015) and on the

listener's ability to distinguish gender and speaker identity (in English, Fu *et al*., 2004; in Spanish, Gonzalez and Oliver, 2005). Dorman *et al*. (1997) compared English speech intelligibility using a tone or noise vocoder with varying numbers of channels and found only small differences that did not reach statistical significance under most test conditions with vowels, consonants, and sentences. Their findings suggested that neither of the two vocoder types was superior to the other. However, in a more recent study, Whitmal *et al*. (2007) examined the intelligibility of English sentences and vowel-consonant-vowel syllables using a six-band vocoder and found that a tone vocoder produced more intelligible speech than a noise vocoder under both quiet and noisy conditions across different signal-to-noise ratios (SNRs) and two different masker types, namely, speech spectrum-shaped noise (SSN) and two-talker babble (2TB). In their study on the interaction between carrier type and cutoff frequency in the vocoding process, Souza and Rosen (2009) found that TV speech was less intelligible than NV speech for a low envelope cutoff frequency of 30 Hz, but more intelligible for a high envelope cutoff of 300 Hz. And Rosen *et al*. (2015) reported recently that using tone carriers with a denser spectrum improved the intelligibility of TV speech considerably over typical tone vocoders, equating and even surpassing the performance observed with noise vocoders.

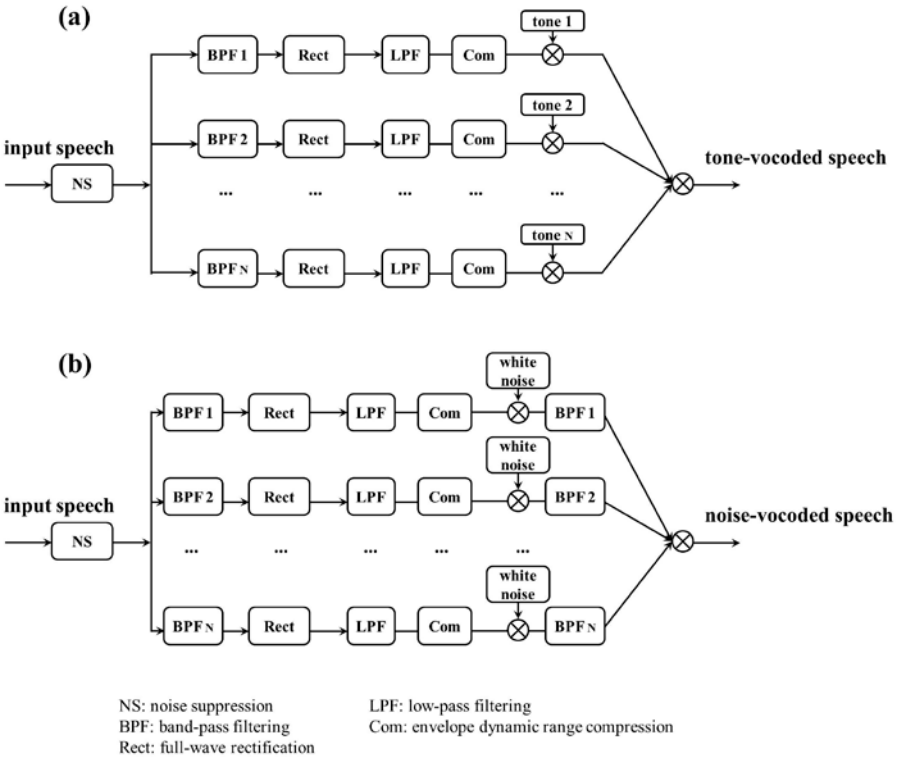

FIG 1. Block diagrams of (a) tone-vocoder and (b) noise-vocoder processes.

Studies on gender and speaker identification, in which good performance depends heavily on cues such as fundamental frequency (F0) and formant structure, have shown better performance for TV speech than for NV speech. Using one- and four-band noise vocoders, Fu *et al*. (2004) observed poor voice gender discrimination (approximately chance level). However, with a tone vocoder, they obtained better results that were more consistent with those of real CI users. In another study of gender and speaker identification in Spanish, Gonzalez and Oliver (2005) found that the tone vocoder performed substantially better than did the noise vocoder across conditions with different numbers of channels. Recently, Chen and Lau (2014) evaluated the effect of vocoder carrier signal type on the intelligibility of Mandarin Chinese, a tonal language, and found an advantage of tone over noise carriers on the intelligibility of vocoded Chinese speech.

Noise is prevalent in our daily lives and poses a great challenge to human speech perception. Alleviation of background noise interference is the goal of many single-channel noise-suppression algorithms, such as the spectral-subtraction (Kamath and Loizou 2002), statistical-model-based (Ephraim and Malah, 1985), and subspace (Hu and Loizou, 2003) algorithms. However, noise suppression may cause undesirable distortion (e.g., “musical noise”) of speech, which is detrimental to speech perception (Loizou, 2007). Certain noise-suppression algorithms (e.g., statistically based Wiener filtering) have been shown to improve speech quality *per se* without improving speech intelligibility for NH listeners (Hu and Loizou, 2007; Li *et al*., 2011).

Vocoded speech is synthesized with multiband envelope information from the original speech signal. When the original speech signal contains distortions due to noise-reduction processing, it is unclear how this distortion affects the intelligibility of envelope-based vocoded speech. In Experiment 1, our aim was to investigate whether the intelligibility advantage of tone over noise vocoders persists when noise-suppression processing is used. Comparisons between tone and noise vocoders have shown no inherent fluctuation in a tone carrier compared to a noise carrier. The combination of envelope distortion (caused by noise-suppression processing) and inherent fluctuation in a noise carrier may have a negative influence on the intelligibility of NV speech. Hence, we hypothesized that carrier signal type may affect the intelligibility of vocoded speech in the context of noise-suppression processing. In other words, we are supposing that the intelligibility advantage of tone over noise vocoders may occur when the vocoding process involves noise-suppression processing. In addition, the effect of noise masking is commonly

represented by two mechanisms: energetic masking by steady-state noise, and informational masking by other speech characteristics, such as fluctuations, in the noise (Carhart *et al*., 1967; Watson, 2005). Hence, in Experiment 1, we also examined whether the interactional effect of the vocoder type and noise-suppression processing depends on the masker type.

Dynamic range plays an important role in speech perception (e.g., Zeng *et al*., 2002). This fact provides a partial explanation for why CI users have poor speech perception (i.e., reduced hearing dynamic range of 5–10 dB), especially in adverse listening environments. Fitting the wide dynamic range of speech signals into the narrow range of the residual hearing of CI users requires dynamic range compression. Several vocoder simulation studies have assessed the effect of envelope dynamic range on speech intelligibility (Fu and Shannon, 1999; Loizou *et al*., 2000; Chen *et al*., 2013; Lai *et al*., 2015). Similarly, we are interested in clarifying whether reducing the dynamic range has a negative effect on envelope-based vocoded speech, and to what extent carrier signal type affects the intelligibility of vocoded speech with a dynamic range-compressed envelope. In Experiment 2, our aim was to investigate whether the intelligibility advantage of tone over noise vocoders persists in the context of envelope dynamic range compression. Earlier work has shown that the spectral sidebands contained in TV speech (due to the multiplication of pure-tone carrier and envelope waveform) carries additional cue which is beneficial for speech intelligibility (e.g., Whitmal et al., 2007; Stone et al., 2008). In addition, white-noise carrier has intrinsic envelope fluctuations that are absent in pure-tone carrier. Multiplying the white-noise carrier by the envelope waveform may have an additional temporal influence on the envelope waveform, which is detrimental to speech understanding (Stone et al., 2011). Given the potential negative effect of dynamic range compression and the intelligibility disadvantage of NV relative to TV speech, this work hypothesized that when envelope dynamic range compression is included in the vocoding process, the intelligibility of NV speech would drop at a higher rate than that of TV speech. NV speech would be far less intelligible than TV speech.

II. EXPERIMENT 1: EFFECT OF NOISE SUPPRESSION ON THE INTELLIGIBITY OF VOCODED SENTENCES

The purpose of Experiment 1 was to examine the effect of noise suppression on the intelligibility of TV and NV Mandarin sentences.

A. Methods

1. Subjects

Eight (five males and three females) native-Mandarin-Chinese listeners (18–23 years old) participated in the experiment. All participants were undergraduate students at Southern University of Science and Technology, and were paid for their participation. All subjects had NH, as determined by having measured pure-tone thresholds (250–8000 Hz) better or equal to a 20-dB hearing level. The study protocol was approved by the Human Research Ethics Committee for Non-Clinical Faculties of Southern University of Science and Technology.

2. Materials

The speech material consisted of sentences taken from the Mandarin Hearing in Noise Test (MHINT) database (Wong *et al.*, 2007), which includes 24 lists of 10 sentences, with each sentence containing 10 key words. All of the sentences were produced by a male speaker with an F0 range of 75–180 Hz.

Two types of masking were used to corrupt the sentences: steady-state SSN and 2TB. For SSN masking, a finite impulse response filter was designed based on the average spectrum of the MHINT sentences, and a white noise was filtered and scaled to the same long-term average spectrum and level as the sentences. The 2TB masker contained two equal-level interfering male talkers. A random noise segment of the same length as the clean speech signal was cut out of the noise recordings, appropriately scaled to reach the desired input SNR level, and finally added to the speech signals at -2-dB and 6-dB input SNR levels for the SSN and 2TB maskers, respectively. The input SNR levels were chosen based on known performance from a pilot study.

3. Signal processing

The noise-suppressed vocoded speech generation processes are summarized in block diagrams in Figure 1. Input noise-corrupted speech signals were first processed by existing single-channel noise-suppression algorithms, followed by the tone- or noise-vocoding process. To process noise-corrupted sentences, we used four representative noise-suppression algorithms: the generalized Karhunen-Loeve transform (KLT) approach (Hu and Loizou, 2003), the Log

Minimum Mean Square Error (logMMSE) algorithm (Ephraim and Malah, 1985), the multiband spectral subtraction (MB) algorithm (Kamath and Loizou, 2002), and the Wiener algorithm based on *a priori* SNR estimation (Scalart and Filho, 1996). These four algorithms encompass the four most commonly used types of single-channel noise-suppression methods, namely the subspace, statistical-modeling, spectral-subtraction, and Wiener-filtering approaches, respectively (see review in Loizou, 2007).

For the KLT method, the noise-corrupted speech signal is projected into orthogonal subspaces; KLT parts representing the signal subspace are modified by a gain function, determined by the estimator; remaining KLT parts representing the noise subspace are nulled; and the enhanced signal is obtained from the inverse KLT of the modified parts (Hu and Loizou, 2003). The statistical-modeling approach employs statistical models with optimization criteria (e.g., minimum mean square error) to estimate the magnitude spectrum of the speech signal (Ephraim and Malah, 1985). The spectral-subtractive algorithm is implemented with an estimate of the clean signal spectrum, generated by subtracting an estimate of the noise spectrum from a noise-corrupted speech spectrum (Kamath and Loizou 2002). The Wiener filter uses *a priori* SNR statistics to design a gain function that suppresses low-SNR segments, while preserving high-SNR ones. Detailed descriptions of the algorithms including the exact parameters used in the current study can be found in Hu and Loizou (2007) and Loizou (2007). The Matlab code used to implement the four noise-suppression algorithms was obtained from Loizou (2007).

All noise-suppressed materials were further processed by a tone or noise vocoder (Fig. 1). To implement the tone vocoder, speech signals were first processed through a pre-emphasis filter (first-order high-pass filter with 1200-Hz cutoff frequency). Then, signals were bandpass-filtered into eight frequency bands between 80 Hz and 6000 Hz with sixth-order Butterworth filters. The cutoff frequencies for the channel allocation of bandpass filters were (in Hz): 80, 221, 426, 724, 1158, 1790, 2710, 4050, and 6000. From each band, the envelope was extracted by full-wave rectification and low-pass filtering with a 200-Hz cutoff frequency by way of a fourth-order Butterworth filter. Sine waves at the center frequencies of the bandpass filters were generated with amplitudes modulated by the extracted envelopes. All amplitude-modulated sine waves from the resultant set of bands were summed to generate a TV stimulus, whose amplitude was adjusted to have the same root-mean-square (RMS) energy as the original speech signal. RMS energy

scaling was performed with respect to the noisy and noise-suppressed input speech signals under the noisy and noise-suppressed conditions, respectively. Noise-suppression processing causes an RMS energy difference between noisy and noise-suppressed speech signals. The RMS energy scaling was done with respect to the energy of each original speech signal. Experimental results may vary when RMS energy is scaled with respect to the same energy (of either the noisy or noise-suppressed speech signal); this possibility warrants further investigation.

Implementation of the noise vocoder was similar to that of the tone vocoder, except that a white noise instead of a sine wave was used as the carrier signal, and amplitude-modulated by the extracted envelope. Output from each band was further band-limited with the same bandpass filter at that band. All amplitude-modulated noises (with band-limiting processing) were summed to generate the NV stimulus, with its amplitude adjusted to have the same RMS power as the original signal. Again, RMS energy scaling was performed with respect to the noisy and noise-suppressed input speech signals under the noisy and noise-suppression conditions, respectively. The envelope dynamic compression block (labeled 'Com' in Fig. 1) was deactivated (compression factor $\alpha = 1$; see Experiment 2) in the vocoding process.

4. Procedure

The experiment was performed in a sound booth, and stimuli were played to listeners diotically through an HD 650 circumaural headphone (Sennheiser, Germany) set at a comfortable listening level. Before the actual testing session, each subject participated in a 10-min training session and was given four lists of 10 MHINT sentences. The training session familiarized the subjects with the testing procedure and conditions. During the training session, the subjects were allowed to read transcriptions of the training sentences while they were listening to the sentences. Four testing conditions [= 2 masker types (i.e., SSN at -2 dB SNR and 2TB at 6 dB SNR) × 2 vocoder types (i.e., TV and NV) × 1 signal processing condition (i.e., noisy)] were used during training. In the testing session, the order of the conditions was randomized across subjects, and the subjects were asked to repeat orally all of the words they heard. In addition, the lists were randomized across listeners. The sentences used during testing were not the same as any of the training sentences. Each subject participated in a total of 20 conditions [= 2 masker types (i.e., SSN at -2 dB SNR and 2TB at 6 dB SNR) × 2 vocoder types (i.e., TV and NV) × 5 signal

processing conditions (i.e., KLT, logMMSE, Wiener, MB, and noisy)]. One list of 10 Mandarin sentences was used per tested condition, and none of the sentences was repeated across conditions. Subjects were allowed to listen to each stimulus a maximum of three times, and were asked to repeat as many words as they could recognize. A simple custom software interface was designed for the listening experiment, which each participant used to control the auditory delivery of the processed stimuli. During the testing session, a tester accompanied the participant and scored his/her response in the computer. A 5-minute break was given every 30 minutes to avoid listening fatigue. The intelligibility score for each condition was computed as the ratio between the number of correctly recognized words and the total number of words contained in each MHINT list. The total testing time was one hour and ten minutes (10-minute training and 60-minute testing).

5. Data analysis

The data were subjected to two-way repeated measures analyses of variance (rmANOVAs) with recognition score as the dependent variable and vocoder type and signal processing condition as within-subject factors. Recognition scores were first converted to rational arcsine units using the rationalized arcsine transform (Studebaker, 1985). A one-way rmANOVA was conducted for each type of vocoder to further analyze the effect of signal processing condition; the ANOVA alpha level was Bonferroni corrected, and only those tests with p values lower than 0.0125 (= 0.05/4) were considered significant. Paired t-tests were conducted in each signal processing condition to further analyze vocoder-type effects.

B. Results

Mean recognition scores for all conditions in Experiment 1 are shown in Figure 2, with data for the SSN and 2TB maskers shown in panels a and b, respectively. For the results of the SSN masker at -2 dB SNR condition (Fig. 2a), a two-way rmANOVA indicated significant effects of vocoder type ($F_{1, 7} = 34.13$, $p < 0.005$) and signal processing condition ($F_{4, 28} = 19.83$, $p < 0.001$), but no significant interaction between these two variables ($F_{4, 28} = 0.397$, $p = 0.81$). One-way rmANOVAs showed significant differences in performance between Wiener-processed and noisy (i.e., no noise suppression) vocoded speech for both vocoder types ($p < 0.01$), and paired t-tests revealed performance differences ($p < 0.05$) between paired TV and NV speech under all signal

processing conditions.

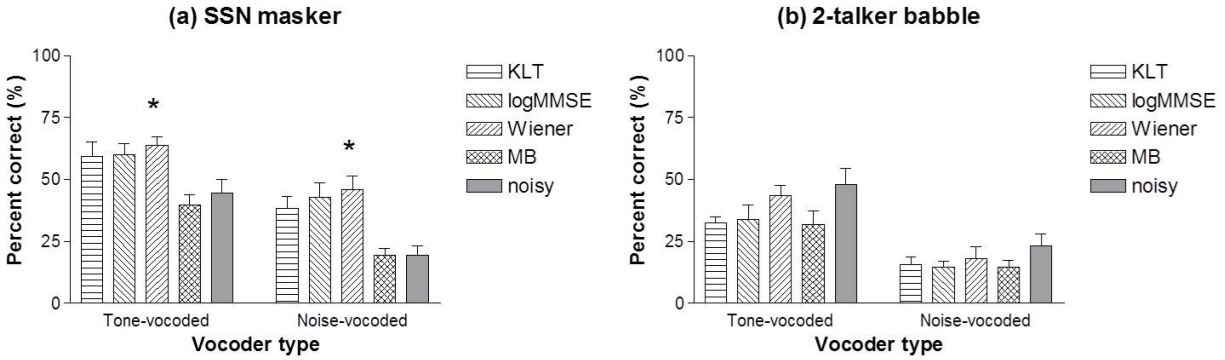

FIG 2. Sentence recognition scores for all conditions with and without noise reduction algorithms with (a) a -2-dB SNR SSN masker and (b) a 6-dB SNR 2TB masker. The error bars denote ±1 standard error of the mean. The asterisk denotes that the intelligibility score is significantly (p<0.01) larger than that in the noisy condition.

For the results of the 2TB masker at 6 dB SNR condition (Fig. 2b), a two-way rmANOVA indicated significant effects of vocoder type ($F_{1, 7}$ = 69.14, p < 0.001) and signal processing condition ($F_{4, 28}$ = 3.65, p < 0.05), but not a significant interaction ($F_{4, 28}$ = 0.40, p = 0.81) between vocoder type and signal processing condition. Again, a one-way rmANOVA revealed no significant performance difference (p > 0.02) between noise-suppressed and noisy vocoded speech. Paired t-tests revealed significant performance differences (p < 0.05) between paired TV and NV speech under all signal processing conditions.

III. EXPERIMENT 2: EFFECT OF ENVELOPE DYNAMIC RANGE COMPRESSION ON THE INTELLIGIBITY OF VOCODED SENTENCES

The purpose of Experiment 2 was to examine the effect of envelope dynamic range compression on the intelligibility of TV and NV Mandarin sentences.

A. Methods

1. Subjects and Materials

Seven (four males and three females, 19–20 years old) new (i.e., did not participate in

Experiment 1) NH native-Mandarin listeners participated in this experiment. All participants were undergraduate students at Southern University of Science and Technology, and were paid for their participation.

The speech materials were the same as in Experiment 1, and the SSN masker was used to corrupt the MHINT sentences at 3 dB and -3 dB input SNR levels.

2. Signal processing

We implemented a simple compression method (Chen *et al*., 2013). Letting x and y denote input and output amplitude envelopes, respectively, the output compressed amplitude envelope y was computed as:

$$y = \alpha \times (x - \overline{x}) + \overline{x}, \quad (1)$$

where $\overline{x}$ is the mean of the input amplitude envelope x, and α is the compression factor constant chosen for compressing the output amplitude envelope dynamic range. Mean values of the output and input amplitude envelopes were equal (i.e., $\overline{y} = \overline{x}$), regardless of the value of α. A small compression factor α denotes a large compression ratio and *vice versa*. When α = 0 in Eq. (1), the compressed amplitude envelope becomes a DC signal with a constant value of $\overline{x}$ (i.e., $\overline{y} = \overline{x}$), and the dynamic range is 0 dB. When α = 1.0, the output amplitude envelope maintains the original dynamic range of the input (i.e., no envelope compression). Figure 3 shows the three compressed amplitude envelope waveforms, with compression factor α = 1.0, 0.5, and 0.2, respectively. Note that the three compressed amplitude envelope waveforms have the same mean values (dashed lines in the three panels in Fig. 3). We employed α values of 1, 0.5, and 0.2, which reduced the input envelope dynamic range by 0 dB, 6 dB, and 14 dB, respectively.

The compressed envelope was multiplied by the carrier signal (i.e., tone or noise) to generate vocoded stimuli, as in Experiment 1. The noise suppression (NS) block in Fig. 1 was deactivated in the vocoding process. The compression strategy in Eq. (1) was motivated by preserving the loudness of processed speech signals, while reducing the dynamic range of envelope variation selectively (Chen *et al*., 2013). The compression strategy in Eq. (1) is different from those used in actual CIs, wherein a nonlinear function is used to limit the speech envelope into the range restricted by the threshold and most comfortable levels of a CI listener.

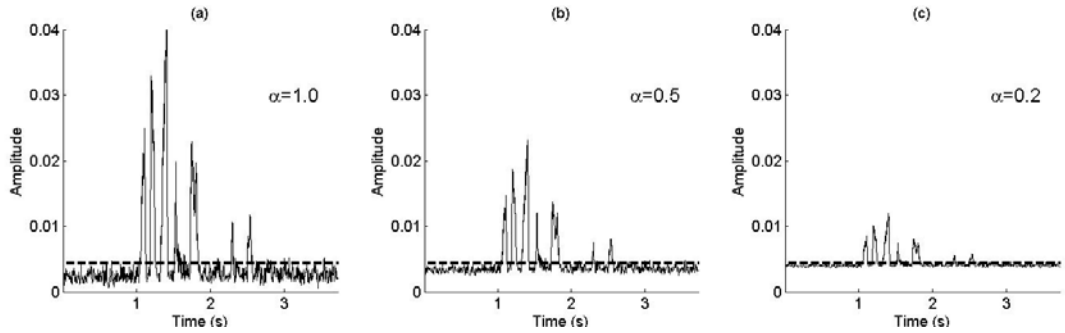

FIG 3. Example waveforms of compressed amplitude envelope with compression factor α of values (a) 1.0, (b) 0.5, and (c) 0.2. The dashed line in each panel denotes the mean value of the amplitude envelope waveform.

3. Procedure

The experimental procedure used in Experiment 2 was essentially the same as that used in Experiment 1. Again, in the training session in which subjects were familiarized with the testing procedure and conditions, each subject was given four lists of 10 sentences (different from those used in the testing session) and allowed to read transcriptions while listening to the sentences. However, in Experiment 2, each subject was exposed to a total of 12 conditions [= 2 input SNR levels (i.e., 3 dB and -3 dB) × 2 vocoder types (i.e., TV and NV) × 3 values of compression factor (i.e., α=1.0, 0.5 and 0.2)], which were randomized across the subjects. As in Experiment 1, one list of 10 sentences was presented per condition, and none of the sentences was repeated across the conditions. The total testing time was 50 minutes (10-minute training and 40-minute testing).

6. Data analysis

The data were analyzed as in Experiment 1. The three within-subject factors for the three-way rmANOVAs were vocoder type, SNR level and compression factor; and paired *t*-tests were conducted in each compression condition to further analyze vocoder-type effects.

B. Results

The mean recognition scores of Mandarin sentences for all conditions are shown in Figure 4. A three-way rmANOVA indicated significant effects of vocoder type ($F_{1, 6}$ = 167.47, p < 0.005), SNR level ($F_{1, 6}$ = 230.66, p < 0.005) and compression factor ($F_{2, 12}$ = 107.75, p < 0.005), as well

as a non-significant interaction between vocoder type and SNR level ($F_{1, 6}$ = 5.9, p =0.06), a non-significant interaction between SNR level and compression factor ($F_{2, 12}$ = 153.80, p =0.3), a significant interaction between vocoder type and compression factor ($F_{1, 6}$ = 6.69, p < 0.05), and a non-significant interaction among vocoder type, SNR level and compression factor ($F_{2, 12}$ = 3.08, p =0.09). Paired t-tests showed that performance differed significantly (p < 0.001) between TV and NV speech under all test conditions with the same SNR level and compression factor value.

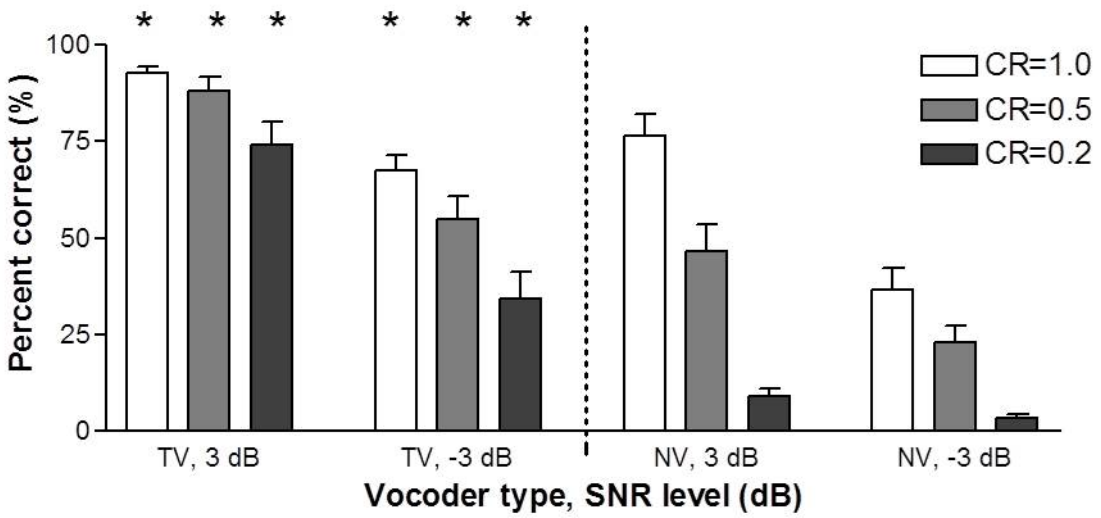

FIG 4. Sentence recognition scores for all test conditions. The error bars denote ±1 standard error of the mean. The asterisk denotes that the intelligibility score of tone-vocoded speech is significantly (p<0.005) larger than that of noise-vocoded speech.

The significant interaction between vocoder type and compression factor appears to be due to the ceiling/flooring effect on the intelligibility scores of TV/NV speech in Fig. 4. To further analyze the interactional effect between vocoder type and compression factor, Figure 5 displays the scores of TV and NV speech in near-linear range (as a function of compression factor), and excludes the effect of ceiling/flooring on data analysis. The SNR levels in Fig. 5 are -3 dB and 3 dB for TV and NV speech, respectively. It is seen in Fig. 5 that at uncompressed condition (i.e., CR=1.0), the intelligibility scores of TV and NV speech are similar. Envelope dynamic range compression causes decreased intelligibility to both TV and NV speech. However, it is noted that the intelligibility score of NV speech drops at a higher rate than that of TV speech does, indicating that the effect of compression is different between TV and NV speech.

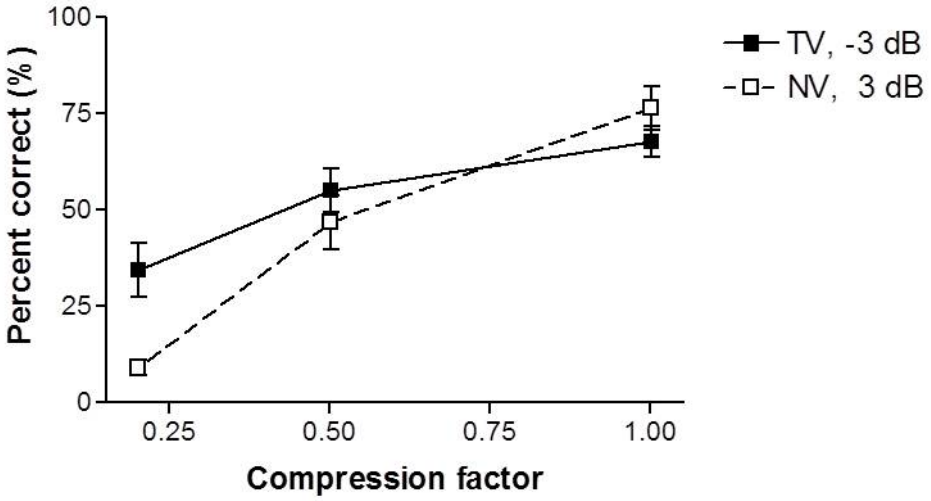

FIG 5. Sentence recognition scores for selected test conditions in Fig. 4. The solid and dashed lines show the scores for TV speech at -3 dB and NV speech at 3 dB, respectively. The error bars denote ±1 standard error of the mean.

IV. DISCUSSION AND CONCLUSIONS

Prior vocoder simulation studies have demonstrated a perceptual contribution of the temporal envelope to speech intelligibility (e.g., Shannon *et al*. 1995; Dorman *et al*. 1997; Whitmal *et al.*, 2007; Stone *et al.*, 2008; Stone *et al.*, 2011; Chen and Loizou, 2011a). Several factors can be manipulated to control the amount of information that is included in the multiband envelope. In the present work, we assessed how noise suppression and envelope dynamic range compression affect the intelligibility of vocoded speech. We also investigated the effect of carrier signal type on the intelligibility of noise-suppressed and envelope dynamic range-compressed vocoded speech.

A. Intelligibility advantage of tone- over noise-vocoded speech

In Experiments 1 and 2, TV speech was found to be more intelligible than NV speech under the same signal-processing conditions, consistent with earlier studies reporting a perceptual advantage of a tone vocoder over a noise vocoder (e.g., Whitmal *et al*., 2007; Chen and Lau, 2014). We further showed that this advantage persisted even under conditions of envelope waveform distortion (i.e., noise suppression and narrowing of the dynamic range). Taken together,

these results provide evidence for the notion that there is a perceptional advantage of TV speech not only when processed by the raw vocoder simulation model, but also when there is additional signal processing, such as noise suppression and envelope dynamic range compression.

Two mechanisms may account for the perceptual advantage of TV speech. The first potential mechanism concerns the spectral sidebands that are contained in TV speech when a pure tone is multiplied by the envelope waveform (e.g., Whitmal *et al.*, 2007; Stone *et al.*, 2008). The amplitude-modulated tone carrier has two spectral sidebands, and these sidebands impose a periodic temporal structure in voiced speech segments on the tone-vocoder's output, with the talker's pitch being preserved over most voiced segments (Whitmal *et al.*, 2007). Hence, the spectral sidebands contain an additional cue that is beneficial for speech intelligibility, even when the noise-suppressed envelope contains nonlinear distortions due to noise-suppression processing. The second potential mechanism is related to the difference in intrinsic temporal fluctuations between sine-wave and white-noise carriers. White-noise carriers have intrinsic envelope fluctuations that are absent in sine-wave carriers. Hence, the white-noise carrier, when multiplied by the envelope waveform, may have an additional temporal influence on the envelope waveform and cause a detrimental effect on speech understanding (Stone *et al.*, 2011).

Another factor that might account for the intelligibility difference observed between TV and NV speech might be the tonal quality of Mandarin Chinese. Mandarin differs from English in that a syllable's tone (or F0 contour) is used to differentiate meaning between otherwise similar lexical items (Howie, 1976; Chen and Loizou, 2011b;). Although the F0 contour is the primary cue for lexical tone identification, the tonal envelope waveform also carries important information for tone identification (Luo and Fu, 2004). In this study, the F0 of the target MHINT sentence ranged from 75 Hz to 180 Hz, and the envelope cutoff frequency was set to 200 Hz. Hence, the envelope waveform and the spectral sidebands of tone-vocoded speech may carry important tonal information. However, noise-vocoded speech, due to its use of noise carriers, may influence or distort the envelope waveform.

B. Dependence of masker type on the intelligibility of noise-suppressed vocoded speech

For noise-suppressed wideband speech signals, no improvements in speech intelligibility have been observed for NH listeners (Hu and Loizou, 2007; Li *et al.*, 2011). When the

noise-suppressed wideband speech signal was processed by a vocoder, we observed masker-dependent intelligibility performance. With an SSN masker, a single-channel noise-suppression algorithm (i.e., Wiener filtering, see Loizou and Kim, 2011) may improve the intelligibility of vocoded speech, regardless of vocoder type. However, when the masker is competing speech (i.e., 2TB), no intelligibility improvement was observed with processing by existing noise-suppression algorithms.

The exact mechanism underlying the presently observed masker-dependent intelligibility performance in vocoded speech is unclear. We hypothesize that noise-suppression processing causes less envelope distortion of speech signals with an SSN masker than with a 2TB masker. When the noise-suppression algorithm was integrated into the vocoding process, this difference in envelope distortion could have accounted for the beneficial effect of noise suppression with steady-state noise corruption and the lack of intelligibility improvement by noise suppression with competing masker corruption. When Chen *et al.* (2015) evaluated the performance of noise-suppression (i.e., the same four single-channel noise-suppression algorithms used in this work) for improving speech recognition by Mandarin-speaking CI users, they tested three types of maskers: SSN, babble, and car noise. They found that although most noise-suppression algorithms could improve Mandarin speech recognition in the presence of noise (e.g., SSN), the algorithms performed differently across different environmental noise conditions. They used an envelope-distortion based objective intelligibility measure (i.e., the normalized covariance measure) to predict CI speech recognition scores and found that an envelope-distortion based intelligibility index could predict the intelligibility of noisy and noise-suppressed speech by CI listeners modestly well (i.e., correlation coefficient 0.81). Similarly, when Baumgärtel *et al*. (2015) evaluated the performance of single-channel noise reduction in the listening scenarios of stationary speech-shaped noise and competing speech, they also found better and worse performance in the stationary noise and competing speech scenarios, respectively. They attributed this masker-differentiated performance to the errors to estimate speech and noise power based on the speech presence probability in single-channel noise reduction processing. Baumgärtel *et al*. noted that in the stationary noise condition, speech and noise power estimates (or the separation of a noisy signal into speech and noise components) worked quite well, whereas in the nonstationary noise (e.g., competing speech) condition, estimation errors occurred and little performance

improvement was found. Future studies should investigate the degree of envelope waveform distortion generated by processing with existing single-channel noise-suppression algorithms. In addition, two different SNR levels were used for the SSN and 2TB conditions in Experiment 1, i.e., a negative SNR of -2 dB for SSN and a positive SNR of 6 dB for 2TB. It remains to be resolved how this SNR level difference interacts with masker type in determining the intelligibility of noise-suppressed vocoded speech.

C. Influence of envelope dynamic range compression on the intelligibility of vocoded speech

In addition to demonstrating a perceptual advantage of employing a tone carrier over employing a noise carrier in the vocoding process, the present work showed that these two types of vocoded speech were associated with differing responses to envelope dynamic range compression. Dynamic range narrowing has been shown repeatedly to impede speech intelligibility (Fu and Shannon, 1999; Loizou *et al*., 2000; Chen *et al*., 2013). Fu and Shannon (1999) measured phoneme recognition in CI users when the dynamic range of the input speech signals was reduced by either peak clipping or center clipping. The compression strategy in Eq. (1) in the present study follows that in Chen *et al*. (2013), and is similar to that developed in Loizou *et al.* (2000). That is, both compression strategies use a linear transformation to convert the range of the input amplitude envelope to a smaller range of the output amplitude envelope; however, the main differences lie in (1) how the minimum envelope amplitude of the input signal is determined and (2) how the linear transformation is designed. In addition, the compression strategy in Eq. (1) preserves the loudness of the processed speech signal.

The present work further showed that noise-vocoded speech was more negatively affected by reducing the envelope dynamic range. With the same compressed envelope waveform (e.g., compression factor of α = 0.5 or 0.2 in Fig. 4), noise-vocoded speech showed a much larger drop in intelligibility than did tone-vocoded speech relative to the uncompressed condition (i.e., compression factor of α = 1.0). For instance, at 3 dB SNR level, compared to the uncompressed condition, a 6-dB drop of envelope dynamic range reduced intelligibility by 4.6% and 29.8%, and a 14-dB drop reduced intelligibility by 18.8% and 67.5%, for TV and NV speech, respectively (Fig. 5). This result indicates that narrowing the envelope dynamic range has a more negative influence on noise- than on tone-vocoded speech. This finding may not be fully attributed to the

confounding factor of saturation or flooring effect when comparing the intelligibility of TV and NV speech (see Fig. 4). Analysis in Fig. 5 excluded the effect of saturation/flooring in intelligibility scores by choosing two different SNR levels for TV and NT speech (i.e., -3 dB and 3 dB, respectively). Again, it is observed in Fig. 5 that NV speech is more susceptible to the influence of reduced envelope dynamic range than TV speech, and its intelligibility score drops at a higher range than TV speech does.

D. Implications of vocoder-based acoustic simulation for studies with CIs

Vocoder simulations have been used for inferring systematically effects of noise suppression and dynamic range compression on speech intelligibility for the purpose of implications in CI listeners (e.g., Lai *et al*., 2015). Researchers have also developed speech-processing strategies for tonal languages (e.g., Mandarin Chinese) and have applied vocoder simulations for assessing their performance (e.g., Luo and Fu, 2006, Lan *et al.*, 2004). Establishing an optimal vocoder for acoustic simulation in CI studies remains an important issue. Although tone-vocoding yields an intelligibility advantage over noise-vocoding, both simulation types may reflect speech intelligibility performance trends with respect to manipulations of acoustic cues. The present study provides evidence of an intelligibility difference between NV and TV speech for NH listeners when an extra signal-processing block is involved in the vocoding process.

The better intelligibility of TV sentences relative to NV sentences may be due, at least in part, to the spectral sidebands contained in TV speech and the absence of intrinsic temporal fluctuations in sine wave carriers. Accordingly, when the vocoding process is combined with another signal-processing block, such as noise suppression or envelope dynamic range compression, it is necessary to consider potential interactions between the nature of the carrier signal and distortion produced during signal processing and how such interactions may impede the performance of the signal. The lower intelligibility of the NV speech might be attributable to envelope distortion caused by noise suppression and/or increased envelope distortion when the noise-suppressed envelope is multiplied by a noise carrier containing noise-like amplitude fluctuation. Conversely, the higher intelligibility of TV Mandarin speech may be due in part to a potential contribution of the spectral sidebands in the tone-vocoding process.

Although the tone and noise vocoders implemented in this work mimic speech processing in a CI device, many patient- and device-specific confounds were not addressed, including electrode array insertion, spread of the electrical field generated by the implant, etc. Williges *et al.* (2015) used a modified vocoder to sample the envelope waveform in each channel with either sequential or randomized pulse train. Spatial spread of the electrical field was simulated by multiplying each pulse with a two-sided exponential decaying function; additionally, an auralization step was implemented to mimic the transfer of signals in each channel to their respective positions along the cochlea. This vocoder implementation provides a realistic simulation of the technical and physiological steps of signal processing in CI listeners. Future work should investigate the effect of modelling such physiologically-inspired features on the results presented here.

E. Limitations of the present work

First, the present work was focused selectively on the effects of noise-suppression and envelope dynamic range compression on the intelligibility of vocoded sentences. Many other factors that may affect the performance of these two vocoder types were not considered, such as envelope cutoff frequency, the number of channels, and filter width. Notably, Rosen *et al*. (2015) showed that a noise vocoder yielded a higher intelligibility than a tone vocoder for a small number of channels (i.e., 2–5). Second, the contribution of the selected cutoff frequency (200 Hz in this study) for extracting the envelope waveform needs to be further investigated. With a 200-Hz cutoff frequency, the original signal (envelope, and a portion of full-wave rectified fine structure waveform) is preserved through channel one (with cutoff frequencies of 80 Hz and 221 Hz) with the tone vocoder, but not with the noise vocoder which adds noisy fluctuations. Third, the tone vocoder modulates the carrier sinusoids in each frequency channel, i.e., the narrow-band signals. The noise vocoder, however, modulates white noise in each channel and then the amplitude-modulated white noises are bandpass-filtered; or the noise vocoder modulates wideband signals. Hence, it is possible that the aforementioned relative intelligibility deficit of noise-vocoded speech simulations may be due, perhaps in part, to the additional narrowband filtering of the amplitude modulation that occurs at the end of the noise vocoding process. Fourth, the present work used the envelope dynamic range compression strategy developed by Chen *et al.*

(2013). It is possible that a different pattern of results would be obtained with the use of alternative compression strategies.

In conclusion, the present work assessed the effects of noise suppression and envelope dynamic range compression on the intelligibility of vocoded Mandarin sentences, and compared the intelligibility of tone- versus noise-vocoded speech. The following conclusions can be drawn:

1) Under all test conditions, tone-vocoded Mandarin sentences showed higher intelligibility scores than did noise-vocoded sentences. This perceptual advantage is consistent with earlier findings. The present study extends this result to vocoded speech that was processed through a noise-suppression algorithm and through envelope dynamic range compression. The perceptual advantage of tone-vocoded Mandarin speech might be attributable to the spectral sidebands contained in tone-vocoded speech and the influence of the amplitude fluctuation of a noise carrier.
2) The intelligibility benefit of noise suppression on both tone- and noise-vocoded speech was dependent upon the masker type employed. When corrupted by a steady-state noise, existing single-channel noise-reduction algorithms (e.g., Wiener filtering) might cause intelligibility improvement. However, when corrupted by a competing masker (e.g., two-talker babble), most existing noise-suppression algorithms did not yield intelligibility improvement.
3) While the envelope dynamic range was narrowed, both tone- and noise-vocoded speech showed reduced intelligibility performance. However, noise-vocoded speech was more negatively influenced by envelope dynamic range compression, yielding a substantial intelligibility gap between tone- and noise-vocoded speech.
4) When additional signal processing is involved in vocoder simulations, interpreting the functional contribution of this processing should be done cautiously. The nature of the carrier signal in the vocoding process and the envelope distortion caused during signal processing may jointly affect the intelligibility of vocoded speech.

ACKNOWLEDGEMENTS

This work was supported by the National Nature Science Foundation of China (Grant No. 61571213), and the Basic Research Foundation of Shenzhen (Grant No. JCYJ20160429191402782).

Baumgärtel, R. M., Krawczyk-Becker, M., Marquardt, D., Völker, C., Hu, H., Herzke, T., Coleman, G., Adiloğlu, K., Ernst, S. M, Gerkmann, T., Doclo, S., Kollmeier, B., Hohmann, V., Dietz, M. (**2015**). "Comparing Binaural Pre-processing Strategies I: Instrumental Evaluation," Trends Hear. **19**, 1–16.

Carhart, R., Tillman, T. W., and Johnson, K. R. (**1967**). "Release of masking for speech through interaural time delay," J. Acoust. Soc. Am. **42**, 124–138.

Chen, F., Hu, Y., and Yuan, M. (**2015**). "Evaluation of noise reduction methods for speech recognition by Mandarin-speaking cochlear implant listeners," Ear Hear. **36**, 61–71.

Chen, F., and Lau, A. H. Y. (**2014**). "Effect of vocoder type to Mandarin speech recognition in cochlear implant simulation," in *International Symposium on Chinese Spoken Language Processing*, pp. 551–554.

Chen, F., Wong, L. L., Qiu, J., Liu, Y., Azimi, B., and Hu, Y. (**2013**). "The contribution of matched envelope dynamic range to the binaural benefits in simulated bilateral electric hearing," J. Speech Language Hear. Research **56**, 1166–1174.

Chen, F., and Loizou, P. C. (**2011a**). "Predicting the intelligibility of vocoded speech," Ear Hear. **32**, 3281–3290.

Chen, F., and Loizou, P. C. (**2011b**). "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," J. Acoust. Soc. Am. **129**, 3281–3290.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.

Ephraim, Y., and Malah, D. (**1985**). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process. ASSP-**33**, 443–445.

Fu, Q. J., Chinchilla, S., and Galvin, J. J. (**2004**). "The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users," J. Asso.

Research Oto. **5**, 253–260.

Fu, Q. J., and Shannon, R. V. (**1999**). “Effect of acoustic dynamic range on phoneme recognition in quiet and noise by cochlear implant users,” J. Acoust. Soc. Am. **106**, EL65–EL70.

Gonzalez, J., and Oliver, J. C. (**2005**). “Gender and speaker identification as a function of the number of channels in spectrally reduced speech,” J. Acoust. Soc. Am. **118**, 461–470.

Howie, J.M. (**1976**). *Acoustical studies of Mandarin Vowels and Tones* (Cambridge University Press, Cambridge, England), pp. 1–308.

Hu, Y., and Loizou, P. (**2003**). “A generalized subspace approach for enhancing speech corrupted by colored noise,” IEEE Trans. Speech Audio Process. **11**, 334–341.

Hu, Y., and Loizou, P. C. (**2007**). “A comparative intelligibility study of single-microphone noise reduction algorithms,” J. Acoust. Soc. Am. **122**, 1777–1786.

Kamath, S., and Loizou, P. (**2002**). “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. IV–4164.

Kasturi, K., and Loizou, P. C. (**2007**). “Effect of filter spacing on melody recognition: acoustic and electric hearing,” J. Acoust. Soc. Am. **122**, 29–34.

Lai, Y. H., Tsao, Y., and Chen, F. (**2015**). “Effects of adaptation rate and noise suppression on the intelligibility of compressed-envelope based speech,” PLOS ONE **10**, e133519.

Lan, N., Nie, K., Gao, S., and Zeng, F. G. (**2004**). “A novel speech-processing strategy incorporating tonal information for cochlear implants,” IEEE Trans. Biomed. Eng. **52**, 752–760.

Li, J., Yang, L., Zhang, J., Yan, Y., Hu, Y., Akagi, M., and Loizou, P. C. (**2011**). “Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English,” J. Acoust. Soc. Am. **129**, 3291–3301.

Loizou, P. C. (**2007**). Speech Enhancement: Theory and Practice (CRC, Boca Raton, FL), pp. 1–689.

Loizou, P. C., and Kim, G. (**2011**). “Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions,” IEEE Trans. Audio Speech Lang Process. **19**, 47–56.

Loizou, P. C., Dorman, M., and Fitzke, J. (**2000**). “The effect of reduced dynamic range on speech

630 understanding: implications for patients with cochlear implants," Ear Hear. **21**, 25–31.
631 Luo, X., and Fu, Q. J (**2006**). "Contribution of low-frequency acoustic information to Chinese
632 speech recognition in cochlear implant simulations," J. Acoust. Soc. Am. **120**, 2260–2266.
633 Luo, X., and Fu, Q. J (**2004**). "Enhancing Chinese tone recognition by manipulating amplitude
634 envelope: Implications for cochlear implants," J. Acoust. Soc. Am. **116**, 3659–3667.
635 Rosen, S., Zhang, Y., and Speers, K. (**2015**). "Spectral density affects the intelligibility of
636 tone-vocoded speech: Implications for cochlear implant simulations," J. Acoust. Soc. Am. **138**,
637 EL318–EL323.
638 Scalart, P., and Filho, J. (**1996**). "Speech enhancement based on a priori signal to noise estimation,"
639 in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal*
640 *Processing*, pp. 629–632.
641 Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition
642 with primarily temporal cues," Science **270**, 303–304.
643 Souza, P., and Rosen, S. (**2009**). "Effects of envelope bandwidth on the intelligibility of sine- and
644 noise-vocoded speech," J. Acoust. Soc. Am. **126**, 792–805.
645 Stone, M. A., Füllgrabe, C., Mackinnon, R. C., and Moore, B. C. (**2011**). "The importance for
646 speech intelligibility of random fluctuations in 'steady' background noise," J. Acoust. Soc. Am.
647 **130**, 2874–2881.
648 Stone, M. A., Füllgrabe, C., Moore, B. C. (**2008**). "Benefit of high-rate envelope cues in vocoder
649 processing: Effect of number of channels and spectral region," J. Acoust. Soc. Am. **124**,
650 2272–2282.
651 Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear Research **28**,
652 455–462.
653 Watson, C. S. (**2005**). "Some comments on informational masking," Acta Acust **91**, 502–512.
654 Whitmal, N. A., Poissant, S. F., Freyman, R. L., and Helfer, K. S. (**2007**). "Speech intelligibility in
655 cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience,"
656 J. Acoust. Soc. Am. **122**, 2376–2388.
657 Williges, B., Dietz, M., Hohmann, V., and Jürgens, T. (**2015**). "Spatial release from masking in
658 simulated cochlear implant users with and without access to low-frequency acoustic hearing,"
659 Trends Hear. **19**, 1–14.

660 Wong, L. L., Soli, S. D., Liu, S., Han, N., and Huang, M. W. (**2007**). “Development of the Mandarin
661 Hearing in Noise Test (MHINT),” Ear Hear. **28**, 70S–74S.
662 Xu, L., Thompson, C. S., and Pfingst, B. E. (**2005**). “Relative contributions of spectral and temporal
663 cues for phoneme recognition,” J. Acoust. Soc. Am. **117**, 3255–3267.
664 Zeng, F. G., Grant, G., Niparko, J., Galvin, J., Shannon, R., Opie, J., and Segel, P. (**2002**). “Speech
665 dynamic range and its effect on cochlear implant performance,” J. Acoust. Soc. Am. **111**,
666 377–386.