Predicting the intelligibility of vocoded and wideband Mandarin Chinese

Fei Chen and Philipos C. Loizou[a)]
Department of Electrical Engineering, EC 33, University of Texas at Dallas, 800 West Campbell Road, Richardson, Texas 75083-0688

(Received 11 May 2010; revised 16 February 2011; accepted 6 March 2011)

Due to the limited number of cochlear implantees speaking Mandarin Chinese, it is extremely difficult to evaluate new speech coding algorithms designed for tonal languages. Access to an intelligibility index that could reliably predict the intelligibility of vocoded (and non-vocoded) Mandarin Chinese is a viable solution to address this challenge. The speech-transmission index (STI) and coherence-based intelligibility measures, among others, have been examined extensively for predicting the intelligibility of English speech but have not been evaluated for vocoded or wideband (non-vocoded) Mandarin speech despite the perceptual differences between the two languages. The results indicated that the coherence-based measures seem to be influenced by the characteristics of the spoken language. The highest correlation ($r = 0.91$–0.97) was obtained in Mandarin Chinese with a weighted coherence measure that included primarily information from high-intensity voiced segments (e.g., vowels) containing F0 information, known to be important for lexical tone recognition. In contrast, in English, highest correlation was obtained with a coherence measure that included information from weak consonants and vowel/consonant transitions. A band-importance function was proposed that captured information about the amplitude envelope contour. A higher modulation rate (100 Hz) was found necessary for the STI-based measures for maximum correlation ($r = 0.94$–0.96) with vocoded Mandarin and English recognition.
© *2011 Acoustical Society of America*. [DOI: 10.1121/1.3570957]

PACS number(s): 43.71.Gv, 43.71.Ky [RSN] Pages: 3281–3290

I. INTRODUCTION

As of 2008, over 5000 patients have benefited from cochlear implants (CIs) in China, and the number of Chinese implantees is foreseen to experience an exponential growth over the next decade (Li, 2009). Tong and Lee (2009) estimated that approximately 36 000–192 000 new implant surgeries will be performed each year, should China become fully developed. Despite the huge need for CIs in China, there are still several anticipated difficulties and challenges that need to be overcome in order to develop a successful CI program in China.

The first challenge lies in the need for the design of new speech coding strategies tailored for Chinese, as current coding strategies (originally designed for English) do not perform satisfactorily. Such a need originates from the difference in the linguistic cues used by native speakers of Western languages (e.g., English) vs Chinese for word recognition. Mandarin is a tonal language and differs from the non-tonal languages (e.g., English) in that different tones are used to express the lexical meaning of words. There are four lexical tones in Mandarin Chinese, namely, the flat, the rising, the falling-rising, and the falling tone, each characterized by their pattern in fundamental frequency (F0) variation (or F0 contour) during voiced segments of speech (e.g., Howie, 1976). Although the F0 contour is the dominant cue for lexical tone recognition (e.g., Lin, 1988), other temporal cues that covary with the F0 contour also contribute to tone recognition. These cues include the vowel duration and amplitude envelope/contour (Liang, 1963; Whalen and Xu, 1992; Fu *et al.*, 1998a; Fu and Zeng, 2000). The duration of tone 3, for instance, is significantly longer than that of the other tones (Fu and Zeng, 2000). The amplitude contour has been found to correlate with the F0 contour and contributes primarily to tone-3 and tone-4 discrimination (Whalen and Xu, 1992; Fu and Zeng, 2000). In the English language, the F0 contour conveys information about the gender of the speaker, information about intonation, and also serves as a segregating cue in competing-talker listening scenarios. It conveys, however, no lexical meaning, suggesting that speech can be understood even if the CI does not transmit accurate or reliable F0 information (Qin and Oxenham, 2005). In brief, substantial efforts are warranted to develop and perhaps redesign existing speech coding strategies to deliver important and language-specific perceptual cues. This is necessary in order to improve speech understanding, and particularly tone recognition, by Chinese CI patients (e.g., Luo and Fu, 2004).

The second related challenge regards the limited number of Chinese CI patients available for testing new speech coding algorithms specially designed for Mandarin Chinese. The population of Chinese CI listeners represents only a very small fraction of the total implanted population. Most importantly, the number of Mandarin-speaking CI users is scarce in countries (e.g., US and Europe) other than China. In fact, some studies conducted in the US recruited CI patients from Chinese hospitals (e.g., Zhou and Xu, 2008).

[a)] Author to whom correspondence should be addressed. Electronic mail: loizou@utdallas.edu

Downloaded 20 Jul 2011 to 129.110.5.92. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

In some respects, it is also challenging to recruit Mandarin-speaking normal-hearing (NH) listeners for intelligibility studies from certain geographical regions.

In summary, without having access to a large pool of Mandarin-speaking CI patients, it is extremely difficult to design and evaluate novel speech coding strategies tailored for Mandarin Chinese. In place of CI testing, vocoded speech, presented to NH listeners, is often used to evaluate the influence of various parameters involved in speech coding algorithms for CIs. Vocoder simulations have been shown to predict well the pattern or trend in performance observed in CI users, including the effects of number of electrodes (Dorman *et al.*, 1997a,b; Friesen *et al.*, 2001), background noise (Fu *et al.*, 1998b), types of speech maskers (Stickney *et al.*, 2004), spectral "holes" (Shannon *et al.*, 2001) on speech intelligibility, etc. Vocoder simulations have proven, and continue, to be an extremely valuable tool in the CI field. It should be stressed that vocoder simulations are not expected to predict the absolute levels of performance of individual users, but rather the trend in performance when a particular speech-coding parameter (e.g., envelope cut-off frequency) or property of the acoustic signal (e.g., F0) is varied. This is so because no patient-specific factors (e.g., neuron survival patterns, etc.) are taken into account in the vocoder simulations.

An intelligibility index that could reliably predict the intelligibility of vocoded Mandarin Chinese might be a viable solution to address the above problem of not having access to sufficient Mandarin-speaking CI patients in countries other than China. Such an index would accelerate the development of new speech coding strategies as it can be used to optimize the set of parameters (e.g., envelope cut-off frequency, F0 detection, etc.) involved in the implementation of such strategies. While a number of speech intelligibility indices have been proposed to predict speech intelligibility, these measures have not been evaluated with vocoded speech produced by either native English or Chinese speakers. These measures, which were originally developed for English, have not been evaluated with wideband (non-vocoded) Mandarin Chinese. Some of the commonly used measures for predicting speech intelligibility include the articulation index (AI) (Kryter, 1962; ANSI, 1997), speech-transmission index (STI) (Houtgast and Steeneken, 1985), and coherence-based index (Kates and Arehart, 2005). The development of an intelligibility index for predicting vocoded Mandarin Chinese poses some challenges. For one, most indices were developed for the Western languages (and for non-vocoded speech), hence it is not clear whether existing indices would be appropriate or fare well for tonal languages as well. Only a few studies have evaluated the STI measure with Chinese speech (Kang, 1998; Peng, 2007). Houtgast and Steeneken (1984) conducted a multi-language evaluation of a simplified STI measure [rapid speech transmission index (RASTI)] and found out that while the STI measure provided a good correlation with the intelligibility of speech from the majority of languages examined, some variability was noted in 4 of the 11 Western languages examined, which did not include a carrier phrase in their articulation tests. The study by Houtgast and Steeneken (1984) revealed some variability in the performance of the STI measure even between the 11 Western languages examined and did not assess the performance of the STI measure when applied in tonal languages.

Second, most STI-based measures assume a low modulation rate (0–12.5 Hz), since this has been found to be sufficient for speech intelligibility (e.g., Drullman *et al.*, 1994), at least in Western languages. Such a narrow modulation rate might not be sufficient, however, for tone recognition, where access to higher-frequency modulations is necessary for capturing information regarding the shape (e.g., rising, falling, etc.) of the F0 contour or amplitude contour. Fu *et al.* (1998a), for instance, have shown that tone-recognition scores improved from 67 to 81% correct when the envelope cut-off frequency (affecting the representation of the temporal envelope) increased from 50 to 500 Hz. In view of this finding, we hypothesize that STI-based measures need to accommodate a higher (>12.5 Hz) modulation rate for Mandarin speech. Furthermore, it is not clear whether different band-importance functions need to be applied in the implementation of existing speech intelligibility measures. Most of the existing band-importance functions were developed for the English language and are tabulated in the ANSI (1997) for different speech materials. Given the difference in linguistic cues used by Mandarin-speaking listeners, it is reasonable to examine alternative band-importance functions that would be more appropriate for Mandarin Chinese. Palva (1965), for instance, observed differences in the band-importance functions estimated for different Western languages. Finally, it is of interest to examine whether existing intelligibility measures perform differently for Western (e.g., English) vs tonal languages. The parameters of several intelligibility measures were varied in the present study to answer the above questions.

II. INTELLIGIBILITY DATA

A. Subjects and materials

Nine (five males and four females) NH native-Mandarin-Chinese listeners participated in the experiment. The subjects' age ranged from 23 to 42 yr, with the majority being graduate students at The University of Texas at Dallas. All subjects had normal hearing, as determined by having pure tone thresholds (at 250–8000 Hz) lower than 25 dB hearing level (HL).

The speech material consisted of Mandarin sentences taken from the Sound Express database (TIGERSPEECH TECHNOLOGY, 2008), a self-paced software program designed for CI and hearing aid users to practice and develop their listening skills. All the sentences were produced by a female speaker. Two types of maskers were used to corrupt the sentences. The first masker was continuous steady-state noise (SSN), and the second was two equal-level interfering female talkers (two-talker). The fundamental frequencies of the target and two interfering talkers were 230 ± 65, 232 ± 58, and 235 ± 46 Hz, respectively.

B. Signal processing

The corrupted Mandarin sentences were presented in two vocoder conditions. The first was designed to simulate eight-

Downloaded 20 Jul 2011 to 129.110.5.92. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

channel electrical stimulation and used an eight-channel sinewave-excited vocoder. The reasons for choosing the eight-channel sinewave-excited vocoder in this study are as follows: first, in most vocoder studies, speech is spectrally degraded into a small number (4–8) of channels of stimulation. This is done based on the outcomes from several studies (e.g., Friesen *et al.*, 2001) indicating that most CI users receive a limited number of channels of frequency information, despite the relatively larger number (16–22) of electrodes available. Second, Dorman *et al.* (1997a) found that the nature of the output signal, noise bands or sine waves, made only a small difference in performance in vowel, consonant, and sentence recognition. The sinewave-vocoder has also been used by others for studies on tonal language recognition (e.g., Lan, *et al.*, 2004). Our implementation of the sinewave-excited vocoder is identical to that used by Dorman *et al.* (1997a).

In implementing the eight-channel tone-vocoder, signals were first processed through a pre-emphasis filter (2000 Hz cutoff) with a 3 dB/octave roll-off and then band-passed into eight frequency bands between 80 and 6000 Hz (see Table I) using sixth-order Butterworth filters. The envelope of the signal was extracted by full-wave rectification and low-pass (LP) filtering using a second-order Butterworth filter (400 Hz cutoff). Sinusoids were generated with amplitudes equal to the root-mean-square (rms) energy of the envelopes (computed every 4 ms) and frequencies equal to the center frequencies of the bandpass filters. The sinusoids of each band were finally summed up and the level of the synthesized speech segment was adjusted to have the same rms value as the original speech segment.

The second processing condition simulated combined electric and acoustic stimulation (EAS). In EAS patients, an electrode array is implanted only partially into the base region of cochlea, so as to preserve the residual acoustic hearing at low frequencies (typically 20–60 dB HL up to 750 Hz and severe to profound hearing loss at 1000 Hz and above), which many patients still have (e.g., Gantz *et al.*, 2006). The low-frequency and high-frequency (>1000 Hz) speech information is provided to these patients via a hearing aid and a CI, respectively. Thus, these patients perceive speech via a combined EAS mode. The signal was first LP filtered to 600 Hz using a sixth-order Butterworth filter. To simulate the effects of EAS for patients with residual hearing below 600 Hz, we combined the LP stimulus with the upper five channels of the eight-channel vocoder from condition 1. It is noted that individual patients may have different upper limits of low-frequency residual hearing. This study chose 600 Hz as upper limit based on existing clinical data (Dorman *et al.*, 2008) and based on configurations used in prior EAS vocoder simulations (e.g., Qin and Oxenham, 2005; Chen and Loizou, 2010).

TABLE I. Filter cut-off (−3 dB) frequencies used for the V and LP + V conditions.

	Condition V		Condition LP + V	
Band	Low (Hz)	High (Hz)	Low (Hz)	High (Hz)
1	80	221	Unprocessed (80–600)	
2	221	426		
3	426	724		
4	724	1158	724	1158
5	1158	1790	1158	1790
6	1790	2710	1790	2710
7	2710	4050	2710	4050
8	4050	6000	4050	6000

We will refer to the above two conditions as: (1) vocoded speech alone (V) and (2) combined LP and vocoded speech (LP + V). As shown in Table I, there was no frequency overlap in the LP + V condition between LP and V (Chen and Loizou, 2010). The sentences in V condition were corrupted by the SSN and two-talker maskers at −4, 0, 4, 8, and 12 dB signal-to-noise-ratio (SNR) levels, while those in LP + V condition were corrupted by the SSN and two-talker maskers at −4, −2, 0, 2, and 4 dB SNR levels (lower SNR levels were used for the LP + V conditions to avoid ceiling effects).

In addition to vocoded Mandarin Chinese speech, we also used wideband (non-vocoded) Mandarin speech. Sentences, taken from the same database, were corrupted by the SSN and two-talker maskers at −14, −12, −10, −8, −6, −4, −2, and 0 dB SNRs. These SNR levels were chosen to avoid ceiling/floor effects. The corrupted Mandarin sentences were first processed through a pre-emphasis filter (2000 Hz cutoff) with a 3 dB/octave roll-off and then band-limited to the frequency range between 80 and 6000 Hz.

C. Procedure

The experiment was performed in a sound-proof room (Acoustic Systems, Inc.) using a PC connected to a Tucker-Davis system 3 (Tucker Davis Technologies, Inc. Alachua, FL). Stimuli were played to listeners monaurally through a Sennheiser HD 250 Linear II circumaural head-phone (Sennheiser electronic GmbH & Co. KG, Wedemark, Germany) at a comfortable listening level. Prior to the test, each subject participated in a 10-min training session and listened to a set of V and LP + V stimuli. The training session familiarized the subjects with the testing procedure. Feedback was provided to listeners only during the training session, so that they can get accustomed to listening to vocoded speech.

In the testing session involving vocoded Mandarin Chinese speech, the subjects were asked to write down the words they heard. Each subject participated in a total of 20 conditions (=two maskers × five SNR levels for the V condition + two maskers × five SNR levels for the LP + V condition). A total of 20 Mandarin sentences were used per condition, and none of the sentences were repeated across the conditions. The order of the test conditions was randomized across subjects. Subjects were given a 5-min break every 30 min during the testing session.

The same procedure was used, after 5 months, to test wideband Mandarin Chinese speech. The test sentences (taken from the same database) were corrupted by the SSN and two-talker maskers at eight SNR levels (ranging from −14 to 0 dB), and each listener participated in a total of 16 conditions (= two maskers × eight SNR levels). Twenty sentences were used per condition. The intelligibility score for each condition was computed as the ratio between the

Downloaded 20 Jul 2011 to 129.110.5.92. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

number of the correctly recognized words and the total number of words contained in 20 sentences.

III. SPEECH INTELLIGIBILITY MEASURES

Present intelligibility measures employ primarily either temporal-envelope or spectral-envelope information to compute the intelligibility index. For the temporal-envelope based measure, we examined the performance of the normalized covariance metric (NCM) measure, which is an STI-based measure (see review in Goldsworthy and Greenberg, 2004). For the spectral-envelope based measure, we investigated the coherence-based speech intelligibility index (CSII) measure and the three-level CSII measures ($\mathrm{CSII}_{\mathrm{high}}$, $\mathrm{CSII}_{\mathrm{mid}}$, and $\mathrm{CSII}_{\mathrm{low}}$) (Kates and Arehart, 2005). The three-level measures were computed by first dividing speech into short-term (20 ms) segments and classifying each segment into one of three regions according to its relative rms power. The CSII measure is computed separately for each region. The high-level region consisted of segments at or above the overall rms level of the whole utterance. The mid-level region consisted of segments ranging from the rms level to 10 dB below and the low-level region consisted of segments ranging from rms-10 dB to rms-30 dB. We adopted the same threshold levels (i.e., 0, -10, and -30 dB) as proposed by Kates and Arehart (2005). The three-level CSII measures obtained for the low-, mid-, and high-level segments were denoted as $\mathrm{CSII}_{\mathrm{low}}$, $\mathrm{CSII}_{\mathrm{mid}}$, and $\mathrm{CSII}_{\mathrm{high}}$, respectively. Figure 1 shows an example segmentation of a Mandarin-spoken sentence based on the above rms threshold values.

The underlying hypothesis in the present study is that measures that assess temporal envelope distortions (e.g., STI-based) should correlate highly with the intelligibility of vocoded speech. This is based on the fact that vocoder simulations preserve and convey primarily envelope information (Shannon *et al.*, 1995) to the listeners. Similarly, it is hypothesized that the spectral envelope representation, as implemented in the coherence-based indices, could also be used to model the intelligibility of vocoded speech. This is so because the spectral representation, as used in the implementation of the coherence-based indices, does not capture reliable F0 information (particularly in the high frequencies) or accurate formant frequency information. The derived spectral envelope provides only a gross representation of the spectrum, as that available in vocoded speech. Assuming that the STI-based measure (NCM) utilizes the same number of bands as the CSII measure, the resulting spectral representation would be similar and consequently would expect the resulting correlations of these two measures with vocoded speech to be comparable. Unlike the CSII measure, the NCM measure captures segmental level information signifying word/syllable boundaries and manner/voicing. Similar information is extracted explicitly by the CSII measure via the three-level segmentation of the utterances.

The NCM measure is similar to the STI (Steeneken and Houtgast, 1980) in that it computes the STI as a weighted sum of transmission index (TI) values determined from the envelopes of the probe and response signals in each frequency band (Goldsworthy and Greenberg, 2004). Unlike the traditional STI measure, however, which quantifies the change in modulation depth between the probe and response envelopes using the modulation transfer function, the NCM measure is based on the covariance between the probe

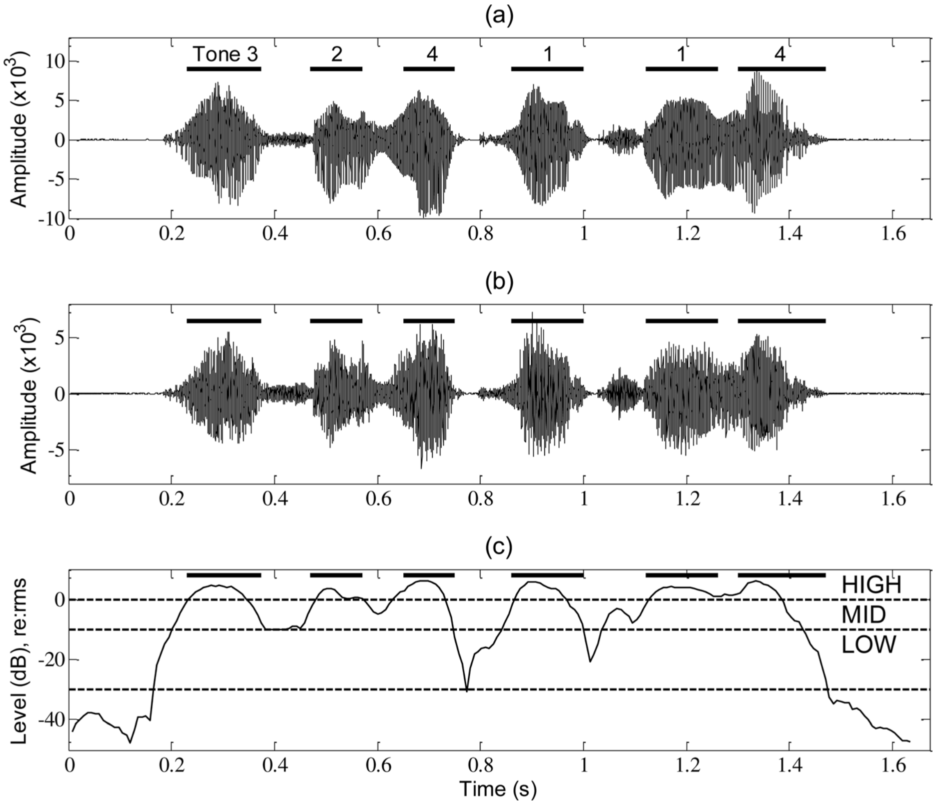

FIG. 1. Example waveforms of: (a) the wideband (non-vocoded) Chinese sentence and (b) vocoded sentence of: "wo (in tone 3) xue (tone 2) hui (tone 4) kai (tone 1) che (tone 1) le (tone 4)," and (c) relative rms energy of the vocoded sentence expressed in decibel relative to the rms level of the whole utterance. The solid lines shown in (a) and (b) delineate the voiced segments of the six Chinese words wherein the tones reside, while the numbers above the solid lines denote the tones of those words. Dashed lines in (c) show the boundaries of the high-, mid-, and low-level regions used for computing the three-level CSII measures.

Downloaded 20 Jul 2011 to 129.110.5.92. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

(input) and response (output) envelope signals. The NCM measure is expected to correlate highly with the intelligibility of vocoded speech due to the similarities in the NCM calculation and CI processing strategies; both use information extracted from the envelopes in a number of frequency bands, while discarding fine-structure information (Goldsworthy and Greenberg, 2004). The coherence-based measures have been used extensively to assess subjective speech quality (Arehart *et al.*, 2007) and speech distortions introduced by hearing aids (Kates, 1992; Kates and Arehart, 2005) and have been shown in the study by Ma *et al.* (2009) to yield high correlations with intelligibility of noise-masked English speech processed by various noise reduction algorithms.

Ma *et al.* (2009) also proposed several signal-dependent band-weighting functions (BWFs) for predicting the intelligibility of speech corrupted by fluctuating maskers. Unlike the band-importance functions used in the ANSI (1997), the BWFs are dynamic in that they are computed on a short-term basis (i.e., every 20–30 ms) and vary from segment to segment. This was found to be necessary since the CSII measure is computed using short-time (20–30 ms) segments and averaged over all segments present in each utterance. The BWFs are flexible in that they can be designed to emphasize spectral peak (e.g., F1 and F2) and/or valley information. To distinguish between the American National Standards Insitute (ANSI) band-importance functions and our importance functions, we refer to the band-importance functions used in our study as BWFs. The BWFs tested in the present study are described in the Appendix. At issue is whether different BWFs should be used for Mandarin speech and particularly whether any of the proposed BWFs capture any useful information related to the F0 or amplitude envelope contours, which are known to contribute to tone recognition (Whalen and Xu, 1992; Fu and Zeng, 2000).

IV. RESULTS AND DISCUSSION

Two figures of merit were used to assess the performance of the above speech intelligibility measures, namely, the Pearson's correlation coefficient (r) and the standard deviation of the prediction error (σ_e). The average intelligibility scores obtained by NH listeners in each condition were subjected to correlation analysis with the corresponding average values obtained in each condition by the intelligibility measures described in Sec. III.

A. Predicting the intelligibility of vocoded Mandarin Chinese

To assess the extent to which the modulation frequency range (or modulation rate) influences the correlation of the NCM measure with speech intelligibility scores, we varied the modulation rate from 30 to 400 Hz. The results obtained with the NCM measure for different modulation rates are shown in Fig. 2. As can be seen, there is a notable improvement in the correlation when the modulation rate increased from 30 to 100 Hz. The correlation improved from $r = 0.88$, obtained when the modulation rate was 30 Hz, to $r = 0.94$ when the modulation rate was extended to 100 Hz. Increasing further the modulation rate beyond 100 Hz did not seem to improve significantly the correlation.

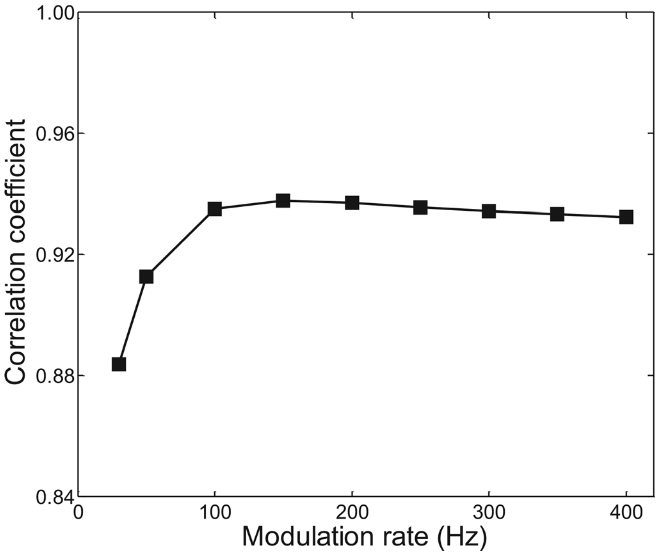

FIG. 2. Correlation coefficients obtained with the NCM measure for different modulation rates for predicting the intelligibility of vocoded Mandarin Chinese. The BWF $W_i^{(2)}$ given in Eq. (A2) with $p = 0.12$ was used in the implementation of the NCM measure.

The resulting correlation coefficients and prediction errors for the remaining intelligibility measures tested are tabulated in Table II. Both the coherence-based and the NCM measures performed relatively well (see also scatter plots in Fig. 3), with the NCM measure yielding the highest correlation ($r = 0.94$). This high correlation was obtained after using the signal-dependent BWF given in Eq. (A2) (Appendix). Similar improvements with the proposed BWFs were also noted for the coherence-based measures. The CSII$_{\text{high}}$ and CSII$_{\text{mid}}$ measures, in particular, yielded equally high correlation ($r = 0.91$) when the BWF W_2 [Eq. (A4) in Appendix] was applied (the BWF W_2 corresponds to the target excitation spectrum, computed for each segment, raised to a power.). This result differs from the outcomes reported by Arehart *et al.* (2007) and Ma *et al.* (2009), who found that, among the three-level CSII measures, the middle-level CSII measure (CSII$_{\text{mid}}$) yielded the highest correlation with the intelligibility of English stimuli. The CSII$_{\text{high}}$ measure is computed using information extracted primarily from the high-intensity (e.g., vowels and semivowels) speech segments (e.g., see Fig. 1). Hence, on this regard it is not surprising that the highest correlation was observed when information extracted from the voiced segments was included in its computation. Aside from formant frequency (F1/F2/F3) information, these segments carry F0 contour information, which is critically needed for lexical tone recognition (Whalen and Xu, 1992; Fu and Zeng, 2000). Note that the BWF [W_2 with $p = 4$ in Eq. (A4), Appendix] used in the weighted CSII$_{\text{high}}$ measure provides more emphasis on the dominant low-frequency spectral peaks and less emphasis on spectral valleys. As discussed later in Sec. IV D, the BWF [W_2 with $p = 4$ in Eq. (A4), Appendix] used in the CSII$_{\text{high}}$ measure captures information about the amplitude envelope contour of the four tones.

Downloaded 20 Jul 2011 to 129.110.5.92. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

TABLE II. Correlation coefficients (r) and standard deviations of the prediction error (σ_e) between sentence recognition scores and various intelligibility measures for vocoded Mandarin Chinese, vocoded English, and wideband (non-vocoded) Mandarin Chinese. The modulation rate (MR) used in the implementation of the NCM measure is indicated.

	Vocoded Mandarin Chinese			Non-vocoded Mandarin Chinese			Vocoded English		
Intelligibility measures	Band-weighting function	r	σ_e(%)	Band-weighting function	r	σ_e (%)	Band-weighting function	r	σ_e (%)
CSII	ANSI	0.88	10.2	ANSI	0.90	12.6	ANSI	0.94	8.7
CSII$_{\text{high}}$	ANSI	0.74	14.5	ANSI	0.97	6.5	ANSI	0.88	12.4
CSII$_{\text{mid}}$	ANSI	0.90	9.6	ANSI	0.95	8.8	ANSI	0.95	8.4
CSII$_{\text{low}}$	ANSI	0.59	17.5	ANSI	0.65	21.7	ANSI	0.78	16.5
CSII	$W_2, p=0.5$	0.89	9.7	$W_2, p=0.5$	0.93	10.6	$W_2, p=0.5$	0.95	8.6
CSII$_{\text{high}}$	$W_2, p=4$	0.91	9.2	$W_2, p=4$	0.97	7.2	$W_2, p=0.05$	0.88	12.4
CSII$_{\text{mid}}$	$W_2, p=0.5$	0.91	9.1	$W_2, p=0.5$	0.95	8.9	$W_2, p=0.5$	0.95	8.2
CSII$_{\text{low}}$	$W_2, p=4$	0.72	15.1	$W_2, p=1$	0.71	19.9	$W_1, p=4$	0.86	13.4
NCM	$W_i^{(2)}, p=0.12$ (MR = 100 Hz)	0.94	7.5	$W_i^{(1)}, p=1.5$ (MR = 100 Hz)	0.96	8.0	$W_i^{(1)}, p=1.5$ (MR = 100 Hz)	0.96	7.1
	ANSI (MR = 12.5 Hz)	0.85	11.5	ANSI (MR = 12.5 Hz)	0.82	16.3	ANSI (MR = 12.5Hz)	0.86	13.4

Given the high correlation obtained with the CSII$_{\text{high}}$ measure, we wanted to examine the influence (if any) of the rms threshold used in its implementation. As mentioned earlier, segments with relative rms level greater than 0 dB were used in the implementation of the CSII$_{\text{high}}$ measure (e.g., see Fig. 1). To assess the influence of the rms threshold, we varied the rms threshold from −5 to 0 dB and examined the correlation of the modified CSII$_{\text{high}}$ measure with the intelligibility scores. The results, obtained using ANSI weights and the proposed W_2 BWF, are plotted in Fig. 4. As can be seen, the correlation improved from $r=0.74$ (0 dB rms threshold) to $r=0.82$ (−5 dB rms threshold) when the ANSI weights are used. However, no significant improvements in correlation were obtained when the proposed W_2 BWF was used.

Speech processed via a sinewave-vocoder was used in the present study to evaluate the NCM and coherence-based intelligibility measures. Given that differences in outcomes might exist between noise-vocoded and sinewave-vocoded speech, at least for some tasks (e.g., gender discrimination) (Gonzalez and Oliver, 2005), it is not clear whether the measures evaluated in the present study can be used to predict the intelligibility of noise-vocoded speech. Further studies are needed to assess that.

B. Comparison with intelligibility prediction of vocoded English

For comparative purposes, Table II shows the correlation coefficients obtained when the same intelligibility measures were used for the prediction of the intelligibility of vocoded English speech. This comparison is important as it could unveil any potential language effects. That is, this comparison will assess whether the prediction power of the measures tested is influenced by the characteristics of the language (Western vs tonal). The intelligibility scores used in the correlation analysis were taken from the studies by Chen and Loizou (2010, 2011). The speech material consisted of IEEE sentences corrupted by SSN and two-talker maskers at −5, 0, and 5 dB SNR. The speech stimuli were processed in a total of 30 conditions using the same tone-vocoder and EAS-vocoder settings (e.g., filter spacing and filter cut-off frequencies) used in this study. When comparing the two sets of correlations for vocoded Mandarin Chinese and vocoded English, we note the following common findings: (1) The coherence-based measures predict equally well the intelligibility of vocoded tonal and non-tonal languages (i.e., Mandarin Chinese and English), i.e., 0.90 and 0.95 for vocoded Mandarin Chinese and vocoded English, respectively; and (2) the STI-based measure (e.g., NCM) also

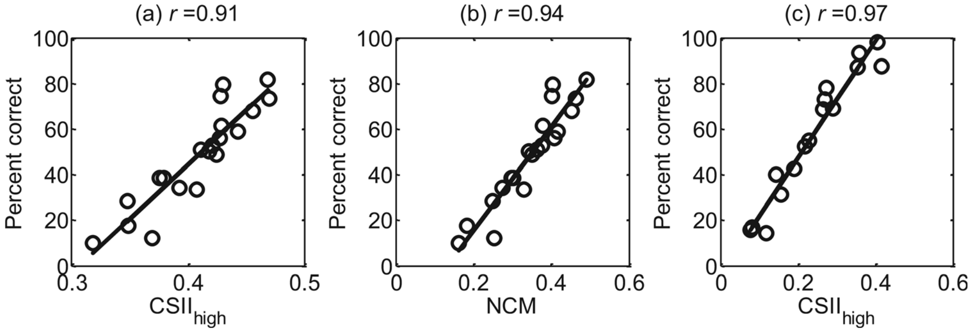

FIG. 3. Scatter plots of recognition scores against the predicted scores for vocoded Mandarin Chinese (panels a and b) and non-vocoded Mandarin Chinese (panel c). The W_2 BWF ($p=4$) was used in the implementation of the CSII$_{\text{high}}$ measure, and $W_i^{(2)}$ with $p=0.12$ was used in the implementation of the NCM measure. The modulation rate of the NCM measure was set to 100 Hz.

Downloaded 20 Jul 2011 to 129.110.5.92. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

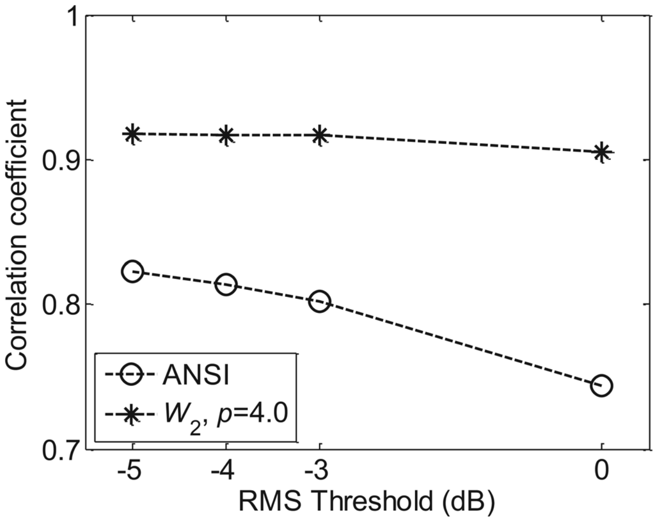

FIG. 4. Correlation coefficients obtained with the $\mathrm{CSII}_{\mathrm{high}}$ measure, implemented using different rms thresholds values, for the intelligibility of vocoded Mandarin Chinese. The $\mathrm{CSII}_{\mathrm{high}}$ measure was implemented using the ANSI weights and the W_2 [$p=4$, Eq. (A4)] BWF.

correlates highly with the intelligibility scores ($r>0.94$) in the two languages.

There is, however, one major difference in the outcomes. In vocoded English, the highest correlation is obtained with the $\mathrm{CSII}_{\mathrm{mid}}$ measure, whereas in vocoded Mandarin Chinese, the highest correlation was obtained with the $\mathrm{CSII}_{\mathrm{high}}$ measure (and the $\mathrm{CSII}_{\mathrm{mid}}$ measure). The two measures incorporate different phonetic segments in their computation. The $\mathrm{CSII}_{\mathrm{mid}}$ measure uses primarily weak consonants and vowel/consonant transitions, while the $\mathrm{CSII}_{\mathrm{high}}$ measure uses primarily high-energy voiced segments (e.g., see Fig. 1). This reflects to some extent the difference in the importance of linguistic cues used by the English and Chinese listeners for word recognition. The high-energy voiced segments, for instance, contain F0 information, which is highly needed for the perception of tonal languages, and as such they are more appropriate for predicting the intelligibility of Mandarin Chinese. These data suggest that the coherence-based measures seem to be influenced by the characteristics of the spoken language.

C. Predicting the intelligibility of wideband (non-vocoded) Mandarin Chinese

Table II shows the correlation coefficients obtained for predicting the intelligibility of wideband (non-vocoded) Mandarin Chinese. As shown in Table II, of all the intelligibility measures examined, the $\mathrm{CSII}_{\mathrm{high}}$ measure yielded the highest correlation coefficient ($r=0.97$), asserting once again the emphasis on information contained in high-energy voiced segments (e.g., vowels). This outcome was consistent with that obtained with vocoded Mandarin Chinese using the $\mathrm{CSII}_{\mathrm{high}}$ measure. That is, of the three CSII measures examined, highest correlation was obtained with the $\mathrm{CSII}_{\mathrm{high}}$ measure. Second, the STI-based measure (i.e., NCM) also correlated well with the intelligibility scores of non-vocoded Mandarin stimuli, i.e., $r=0.96$, when a high modulation rate (100 Hz) and the proposed BWF [Eq. (A1), Appendix] were used. The NCM measure uses primarily temporal envelope information and relatively no fine-structure information. Yet, the NCM measure yielded a high correlation in predicting the intelligibility of wideband Mandarin Chinese. This finding, when taken together with the outcomes predicting the intelligibility of wideband English (Ma *et al.*, 2009), vocoded English and Mandarin Chinese in Table II, suggests that it is not necessary, in terms of predicting the intelligibility of vocoded or wideband speech, to develop measures that incorporate fine-structure information. This is not to say that incorporating fine-structure information is not important, but rather that it is not necessary.

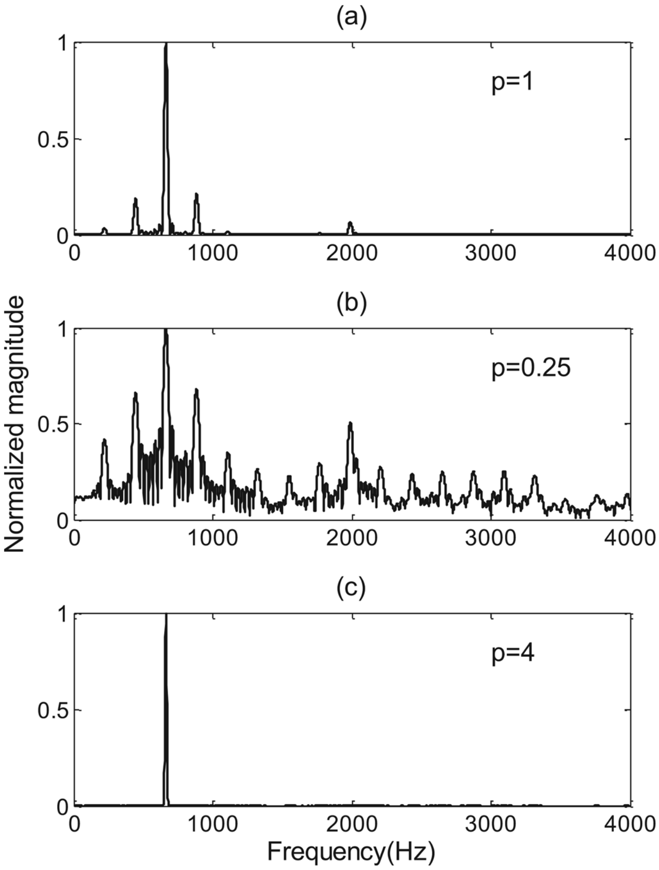

FIG. 5. (a) FFT magnitude spectrum of a segment taken from the vowel /ε/ (excised from the word "head") produced by a male talker, (b) the same spectrum raised to the power of 0.25 [see BWF W_2, Eq. (A4) in Appendix], and (c) the same spectrum raised to the power of 4. All spectra are shown in linear units and have been normalized by their maximum for better visual clarity.

D. Analysis of a Chinese-specific BWF

The data in Table II demonstrated that the selected BWF [W_2, $p=4$, Eq. (A4) in Appendix] has improved significantly the correlation of the $\mathrm{CSII}_{\mathrm{high}}$ measure. This BWF is computed by raising the short-term target excitation spectrum to the power 4, with its values changing from segment to segment owing to the dynamic nature of speech. The use of the proposed BWF improved the correlation from $r=0.74$ obtained when using the ANSI (1997) band-importance function to $r=0.91$ obtained with the proposed W_2 BWF [$p=4$, Eq. (A4)]. We conducted further analysis to

Downloaded 20 Jul 2011 to 129.110.5.92. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

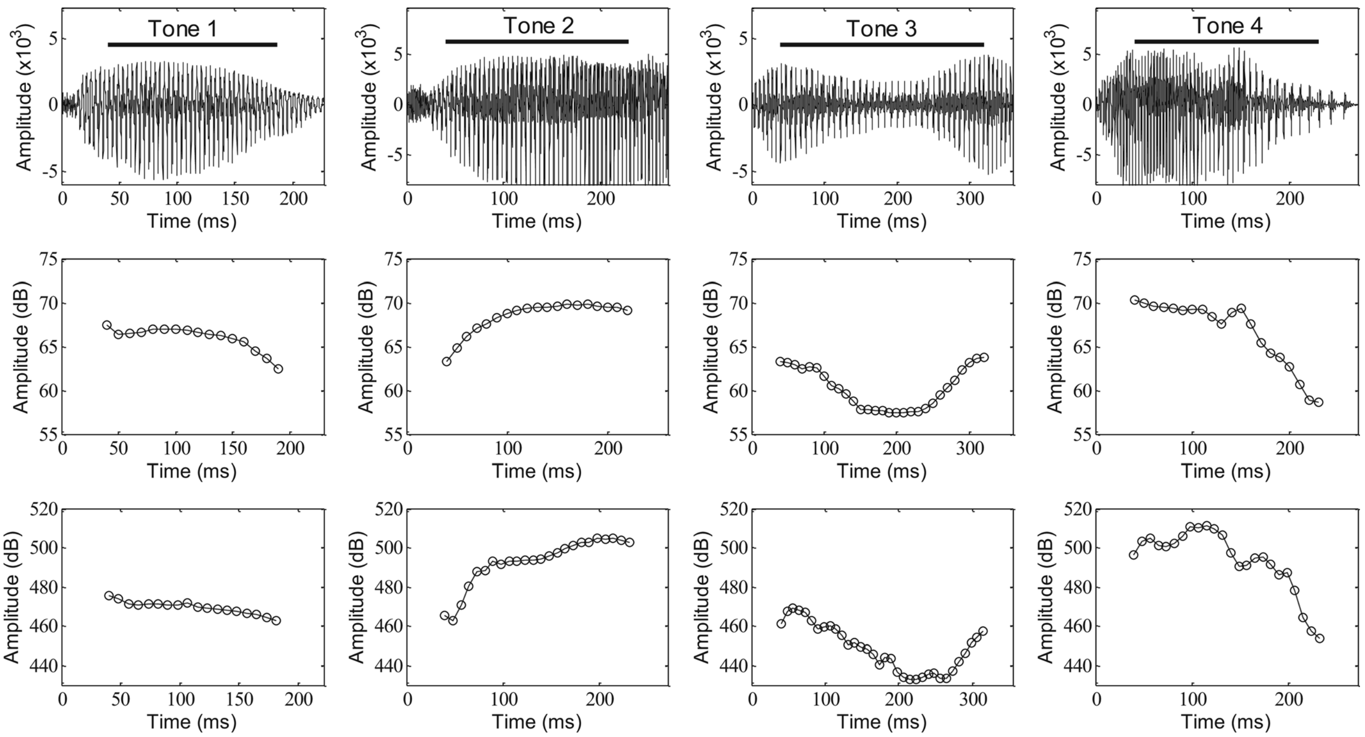

FIG. 6. The top row shows an example waveform of a Mandarin Chinese vowel /i/ produced in four tones. The middle row shows the amplitude envelope contour of the four tones and the bottom row shows the maximum of the W_2 BWF [$p=4$, Eq. (A4)] computed over time for the duration of the vowel /i/.

understand the reason behind the significant improvements in correlation obtained with this particular BWF. Use of $p>1$ in the BWF [Eq. (A4)] implements an expansive function and places more emphasis on the dominant spectral peaks while suppressing the valleys, whereas use of $p<1$ implements a log-compressive function and incorporates information from both spectral valleys and peaks. Figure 5 shows an example weighting function implemented with $p=0.25$ and $p=4$ for the vowel /ε/ excised from the word “head.” When $p=4$, only the first dominant peak is preserved, most likely reflecting information about F1. In other instances, the peaks corresponding to F2 and F3 might also be present in the BWF, but the valleys are always suppressed (see Fig. 5) when $p>1$. It should be noted that the proposed BWF used in the CSII$_{\text{high}}$ measure is applied primarily to the voiced segments of the utterance, wherein the tones reside (e.g., see Fig. 1).

We next computed (over time) the amplitude of the most dominant peak present in the BWF for each 20-ms frame and compared that with the shape of the amplitude envelope contour of the individual tones. Figure 6 shows an example of such analysis for each of the four tones. The amplitude envelope contour was computed by first re-sampling the wideband signal to 100 Hz (thus restricting the amplitude modulations to less than 50 Hz, Fu and Zeng, 2000), and extracting its envelope using the Hilbert transform. As can be seen clearly from Fig. 6, the amplitude of the most dominant peak present in the BWF follows closely the amplitude contour of the four tones. In fact, the correlations between the BWF-derived contour and the amplitude contour of the four tones were quite high and varied from $r=0.76$ (tone 3) to $r=0.94$ (tone 2). Note that the amplitude contours have been found previously (Whalen and Xu, 1992; Fu and Zeng, 2000) to correlate modestly high with the F0 contours and have been found to contribute significantly to the recognition of tones 3 and 4. Hence, we conclude that the chosen BWF [Eq. (A4), $p=4$] captures effectively information about the amplitude contour, which is known to be important for tone recognition (Fu and Zeng, 2000). We thus attribute the improvement in the correlation with the chosen BWF to the fact that this weighting function captures information about the shape of the amplitude contour of the four tones. In contrast, the BWF used for modeling the intelligibility of vocoded English uses values of p smaller than 1 (see Table II), thus implicating the importance of a different set of cues for English (note that the shape of the amplitude contour carries non-significant information about speech perception in English). On that regard, we consider the proposed BWF [Eq. (A4), $p=4$] as appropriate, and to some degree, tailored for Mandarin Chinese, at least when the CSII$_{\text{high}}$ measure is used.

V. CONCLUSIONS

The present study assessed the correlation of several speech intelligibility measures with vocoded and non-vocoded Mandarin Chinese. The following conclusions can be drawn:

(1) The coherence-based measures and STI-based measures (e.g., NCM) can be used to predict reliably vocoded tonal language (i.e., Mandarin) recognition as well as vocoded English recognition.

(2) The coherence-based measures seem to be influenced by the characteristics of the spoken language. In Mandarin Chinese, highest correlation was obtained with a weighted

Downloaded 20 Jul 2011 to 129.110.5.92. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

(by a segment-dependent BWF) coherence-based measure ($\mathrm{CSII}_{\mathrm{high}}$) that included primarily information from high-energy voiced segments (e.g., vowels). These segments contain F0 information, which is important for reliable lexical tone recognition. In contrast, in English highest correlation was obtained with the $\mathrm{CSII}_{\mathrm{mid}}$ measure that included information from weak consonants and vowel/consonant transitions.

(3) The proposed BWFs (see Appendix) significantly improved the correlations for the majority of the measures tested. In particular, the BWF [W_2, $p = 4$, Eq. (A4)], which was computed by raising the target excitation spectrum (computed on a short-term basis) to the power 4, was found to perform significantly better compared to the ANSI (1997) band-importance function. Analysis of this BWF [Eq. (A4), $p = 4$] indicated that it captures information about the amplitude envelope contour, which is known to be important for tone recognition (Fu and Zeng, 2000).

(4) A higher modulation rate (100 Hz) was found to be needed in the implementation of the STI-based measures for maximum correlation ($r = 0.94$) with vocoded tonal language (Mandarin) recognition and vocoded English recognition (Chen and Loizou, 2011). This differed from the modulation rate used (12.5–30 Hz) for modeling wideband English speech (Houtgast and Steeneken, 1985; Van Wijngaarden and Houtgast, 2004).

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC007527 from the National Institute of Deafness and other Communication Disorders, NIH. The authors are grateful to the Associate Editor, Dr. Rochelle Newman, and the three reviewers who provided valuable feedback that significantly improved the presentation of the manuscript.

APPENDIX

For the NCM measure, two signal-dependent BWFs were considered (Ma *et al.*, 2009):

$$W_i^{(1)} = \left(\sum_t x_i^2(t)\right)^p, \tag{A1}$$

$$W_i^{(2)} = \left(\sum_t (\max[x_i(t) - d_i(t), 0])^2\right)^p, \tag{A2}$$

where $x_i(t)$ and $d_i(t)$ denote the downsampled envelopes of the clean speech and scaled masker signals, respectively, in the ith band and p is the power exponent. The value of p was varied from 0.12 to 1.5 in this study.

For the implementation of the coherence-based measures, two types of BWFs were considered (Ma *et al.*, 2009):

$$W_1(j,m) = \begin{cases} X(j,m)^p, \text{if } X(j,m) > D(j,m) \\ 0, \text{else} \end{cases}, \tag{A3}$$

and

$$W_2(j,m) = X(j,m)^p, \tag{A4}$$

where $X(j,m)$ and $D(j,m)$ are the critical-band magnitude spectra (excitation spectra) of the clean speech and scaled masker signals, respectively, in the jth frequency band at the mth frame. The power exponent p in Eqs. (A3) and (A4) was varied from 0.05 to 4 in the present study.

ANSI (**1997**). S3.5, *Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).

Arehart, K., Kates, J., Anderson, M., and Harvey, L. (**2007**). "Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **122**, 1150–1164.

Chen, F., and Loizou, P. (**2010**). "Contribution of consonant landmarks to speech recognition in simulated acoustic-electric hearing," Ear Hear. **31**, 259–267.

Chen, F., and Loizou, P. (**2011**). "Predicting the intelligibility of vocoded speech," Ear Hear. **32**, (in press).

Dorman, M., Loizou, P., and Rainey, D. (**1997a**). "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.

Dorman, M., Loizou, P., and Rainey, D. (**1997b**). "Simulating the effect of cochlear implant electrode insertion-depth on speech understanding," J. Acoust. Soc. Am. **102**, 2993–2996.

Dorman, M., Gifford, R., Spahr, A., and McKarns, S. (**2008**). "The benefits of combining acoustic and electric stimulation for the recognition of speech, voice and melodies," Audiol. Neuro-Otol. **13**, 105–112.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.

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.

Fu, Q. J., and Zeng, F. G. (**2000**). "Identification of temporal envelope cues in Chinese tone recognition," Asia Pac. J. Speech, Lang. Hear. **5**, 45–57.

Fu, Q. J., Zeng, F. G., Shannon, R. V., and Soli, S. D. (**1998a**). "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am. **104**, 505–510.

Fu, Q. J., Shannon, R. V., and Wang, X. (**1998b**). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," J. Acoust. Soc. Am. **104**, 3586–3596.

Gantz, B. J., Turner, C., and Gfeller, K. E. (**2006**). "Acoustic plus electric speech processing: Preliminary results of a multicenter clinical trial of the Iowa/Nucleus Hybrid implant," Audiol. Neuro-Otol. **11**, 63–68.

Goldsworthy, R., and Greenberg, J. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679–3689.

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

Houtgast, T., and Steeneken, H. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069–1077.

Houtgast, T., and Steeneken, H. (**1984**). "A multi-language evaluation of the RASTI-method for estimating speech intelligibility in auditoria," Acustica **54**, 185–199.

Howie, J. M. (**1976**). *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge University Press, Cambridge, England).

Kang, J. (**1998**). "Comparison of speech intelligibility between English and Chinese," J. Acoust. Soc. Am. **103**, 1213–1216.

Kates, J. (**1992**). "On using coherence to measure distortion in hearing aids," J. Acoust. Soc. Am. **91**, 2236–2244.

Kates, J., and Arehart, K. (**2005**). "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. **117**, 2224–2237.

Kryter, K. D. (**1962**). "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698–1706.

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, D. (**2009**). "Cochlear implants in China," ORL **71**, 183.

Liang, Z. A. (**1963**). "The auditory perception of Mandarin tones," Acta Phys. Sin. **26**, 85–91.

Downloaded 20 Jul 2011 to 129.110.5.92. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

Lin, M. C. (**1988**). “The acoustic characteristics and perceptual cues of tones in Standard Chinese,” Chinese Yuwen **204**, 182–193.

Luo, X., and Fu, Q. J. (**2004**). “Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants,” J. Acoust. Soc. Am. **116**, 3659–3667.

Ma, J. F., Hu, Y., and Loizou, P. (**2009**). “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,” J. Acoust. Soc. Am. **125**, 3387–3405.

Palva, A. (**1965**). “Filtered speech audiometry: I. Basic studies with Finnish speech toward the creation of a method for the diagnosis of central hearing disorders,” Acta Oto-Laryngol., Suppl. **210**, 7–86.

Peng, J. X. (**2007**). “Relationship between Chinese speech intelligibility and speech transmission index using diotic listening,” Speech Commun. **49**, 933–936.

Qin, M., and Oxenham, A. (**2005**). “Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification,” Ear Hear. **26**, 451–160.

Shannon, R. V., Galvin, J. J. III, and Baskent, D. (**2001**). “Holes in hearing,” J. Assoc. Res. Otolaryngol. **3**, 185–199.

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

Steeneken, H., and Houtgast, T. (**1980**). “A physical method for measuring speech transmission quality,” J. Acoust. Soc. Am. **67**, 318–326.

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.

TIGERSPEECH TECHNOLOGY (**2008**). http://www.tigerspeech.com/ (Last viewed 15 February 2011).

Tong, C. F., and Lee, Y. S. (**2009**). “Do Chinese speakers need a specialized cochlear implant system?” ORL **71**, 184–186.

Van Wijngaarden, S., and Houtgast, T. (**2004**). “Effect of talker and speaking style on the speech transmission index,” J. Acoust. Soc. Am. **115**, 38L–41L.

Whalen, D. H., and Xu, Y. (**1992**). “Information for Mandarin tones in the amplitude contour and in brief segments,” Phonetica **49**, 25–17.

Zhou, N., and Xu, L. (**2008**). “Development and evaluation of methods for assessing tone production skills in Mandarin-speaking children with cochlear implants,” J. Acoust. Soc. Am. **123**, 1653–1664.

Downloaded 20 Jul 2011 to 129.110.5.92. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp