Predicting the Intelligibility of Vocoded Speech

Fei Chen and Philipos C. Loizou

Objectives: The purpose of this study is to evaluate the performance of a number of speech intelligibility indices in terms of predicting the intelligibility of vocoded speech.

Design: Noise-corrupted sentences were vocoded in a total of 80 conditions, involving three different signal-to-noise ratio levels (−5, 0, and 5 dB) and two types of maskers (steady state noise and two-talker). Tone-vocoder simulations and combined electric-acoustic stimulation (EAS) simulations were used. The vocoded sentences were presented to normal-hearing listeners for identification, and the resulting intelligibility scores were used to assess the correlation of various speech intelligibility measures. These included measures designed to assess speech intelligibility, including the speech transmission index (STI) and articulation index based measures, as well as distortions in hearing aids (e.g., coherence-based measures). These measures employed primarily either the temporal-envelope or the spectral-envelope information in the prediction model. The underlying hypothesis in the present study is that measures that assess temporal-envelope distortions, such as those based on the STI, should correlate highly with the intelligibility of vocoded speech. This is based on the fact that vocoder simulations preserve primarily envelope information, similar to the processing implemented in current cochlear implant speech processors. Similarly, it is hypothesized that measures such as the coherence-based index that assess the distortions present in the spectral envelope could also be used to model the intelligibility of vocoded speech.

Results: Of all the intelligibility measures considered, the coherence-based and the STI-based measures performed the best. High correlations (r = 0.9 to 0.96) were maintained with the coherence-based measures in all noisy conditions. The highest correlation obtained with the STI-based measure was 0.92, and that was obtained when high modulation rates (100 Hz) were used. The performance of these measures remained high in both steady-noise and fluctuating masker conditions. The correlations with conditions involving tone-vocoded speech were found to be a bit higher than the correlations with conditions involving EAS-vocoded speech.

Conclusions: The present study demonstrated that some of the speech intelligibility indices that have been found previously to correlate highly with wideband speech can also be used to predict the intelligibility of vocoded speech. Both the coherence-based and STI-based measures have been found to be good measures for modeling the intelligibility of vocoded speech. The highest correlation (r = 0.96) was obtained with a derived coherence measure that placed more emphasis on information contained in vowel/consonant spectral transitions and less emphasis on information contained in steady sonorant segments. High (100 Hz) modulation rates were found to be necessary in the implementation of the STI-based measures for better modeling of the intelligibility of vocoded speech. We believe that the difference in modulation rates needed for modeling the intelligibility of wideband versus vocoded speech can be attributed to the increased importance of higher modulation rates in situations where the amount of spectral information available to the listeners is limited (eight channels in our study). Unlike the traditional STI method that has been found to perform poorly in terms of predicting the intelligibility of processed speech wherein nonlinear operations are involved, the STI-based measure used in the present study has been found to perform quite well. In summary, the present study took the first step in modeling the intelligibility of vocoded speech. Access to such intelligibility measures is of high significance as they can be used to guide the development of new speech coding algorithms for cochlear implants.

(Ear & Hearing 2011;32;1–●)

Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas.

INTRODUCTION

Numerous factors (e.g., electrode insertion depth and placement, duration of deafness, and surviving neural pattern) may affect the performance of cochlear implant (CI) users in quiet and noisy conditions; hence, it is not surprising that there exists a large variability in performance among implant users. Unfortunately, it is not easy to assess or delineate the impact of each of those factors on speech perception due to interaction among these factors. Vocoder simulations (Shannon et al. 1995) have been used widely as an effective tool for assessing the effect of some of these factors in the absence of patient-specific confounds. In these simulations, speech is processed in a manner similar to the CI speech processor and presented to normal-hearing (NH) listeners for identification (see review in Loizou 2006). Vocoder simulations have been used to evaluate, among others, the effect of number of channels (Shannon et al. 1995; Dorman et al. 1997a), envelope cutoff frequency (Shannon et al. 1995; Xu et al. 2005; Souza & Rosen 2009), F_0 discrimination (Qin & Oxenham 2005), electrode insertion depth (Dorman et al. 1997b), filter spacing (Kasturi & Loizou 2007), background noise (Dorman et al. 1998), and spectral "holes" (Shannon et al. 2001; Kasturi et al. 2002) on speech intelligibility. Vocoder simulations have also been used to assess factors influencing the performance of hybrid and electric-acoustic stimulation (EAS) users (Dorman et al. 2005; Qin & Oxenham 2006) and to provide valuable insights regarding the intelligibility benefit associated with EAS (Kong & Carlyon 2007; Li & Loizou 2008). For the most part, vocoder simulations have been shown to predict well the pattern or trend in performance observed in CI users. The 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 cutoff frequency) or property of the acoustic signal (e.g., F_0) is varied. This is so because no patient-specific factors are taken into account in the simulations. Nevertheless, vocoder simulations have proven, and continue, to be an extremely valuable tool in the CI field.

Vocoded speech is typically presented to NH listeners for identification or discrimination. Given the large algorithmic parametric space, and the large number of signal-to-noise ratio (SNR) levels (each possibly with a different type of masker) needed to construct psychometric functions in noisy conditions, a large number of listening tests with vocoded speech are often needed to reach reliable conclusions. In the study by Xu et al. (2005), for instance, a total of 80 conditions were examined, requiring about 32 hrs of testing per listener. Alternatively, a speech intelligibility index could be used to

0196/0202/11/3202-0001/0 • Ear & Hearing • Copyright © 2011 by Lippincott Williams & Wilkins • Printed in the U.S.A.

predict the intelligibility of vocoded speech. While a number of such indices (e.g., the articulation index [AI] [French & Steinberg 1947] and speech transmission index [STI] [Steeneken & Houtgast 1980; Houtgast & Steeneken 1985]) are available for predicting (wideband) speech intelligibility by NH and hearing-impaired listeners, only a limited number of studies (Goldsworthy & Greenberg 2004; Chen & Loizou 2010) proposed or considered new indices for vocoded speech.

Vocoded speech is degraded at many levels (spectrally and temporally) and it is not clear whether conventional measures would be good at predicting its intelligibility. Most intelligibility models, for instance, were developed assuming equivalent rectangular bandwidth (Glasberg & Moore 1990) or critical-band (Fletcher 1940) spectral representations intended for modeling normal auditory frequency selectivity. In contrast, in most vocoder studies (except the ones investigating the effect of number of channels), speech is spectrally degraded into a small number (4 to 8) of channels of stimulation. This is done based on the outcomes from several studies (Friesen et al. 2001), indicating that most CI users receive a limited number of channels of frequency information despite the relatively larger number (16 to 22) of electrodes available. To mimic to some extent the amount of information presented to CI users with current CI devices, vocoded speech was intended to preserve envelope cues while eliminating temporal fine-structure cues (Shannon et al. 1995). Fine-structure information in speech often refers to amplitude fluctuations with rates $>$500 Hz (Rosen 1992). Eliminating fine-structure information, however, might affect the correlation between existing speech intelligibility indices and speech perception scores, since most indices integrate the fine-spectral information contained in each critical band to derive a compact, auditory-motivated spectral representation of the signal. The present study will examine the extent to which the above concern on the importance of fine-structure information is true.

The development of a speech intelligibility index that would predict reliably the intelligibility of vocoded speech could be used to guide and accelerate the development of new speech coding algorithms for CIs. In the present study, we evaluate the correlation of several measures with intelligibility scores obtained by NH listeners presented with vocoded speech in a number of noisy conditions involving steady noise or fluctuating maskers. More precisely, we examine the correlations between intelligibility scores of vocoded speech and several coherence-based, AI-based, and STI-based measures. 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 measures such as the coherence-based index that assess the distortions present in the spectral envelope could also be used to model the intelligibility of vocoded speech.

Given the influence of the range of envelope modulation frequencies on speech perception (Drullman et al. 1994; Xu et al. 2005; Stone et al. 2008), the present study will also assess the performance of the STI-based measures for different ranges of modulation frequencies. The experiments are designed to answer the question as to whether higher modulation rates should be used for modeling the intelligibility of vocoded speech. For wideband speech, it is known that access to low envelope modulation rates ($<$12.5 Hz) is sufficient to obtain high correlations with speech intelligibility in noise (Steeneken & Houtgast 1980; Ma et al. 2009). Vocoded speech, however, contains limited spectral information; hence, it is reasonable to expect that higher modulation rates would be required in the prediction model to somehow compensate for the degraded spectral information. This is supported by several studies showing that there is an interaction between the importance of high modulation rates and the number of frequency channels available to the listeners (Healy & Steinbach 2007; Xu & Zheng 2007; Stone et al. 2008; Souza & Rosen 2009). Stone et al. (2008), for instance, assessed the effect of envelope low-pass frequency on the perception of tone-vocoded speech (in a competing-talker task) for a different number of channels (6 to 18) and found that higher envelope frequencies (180 Hz) were particularly beneficial when the number of channels was small (6 to 11). This was attributed to the better transmission of F_0-related envelope information, which is particularly important in competing-talker listening scenarios. Similar outcomes were observed by Healy and Steinbach (2007) and Souza and Rosen (2009) in quiet conditions. The importance of high modulation rates seems to also depend on the listening task, with low rates (16 Hz) being sufficient for steady-noise conditions (Xu & Zheng 2007) and higher rates (180 Hz) for competing-talker tasks (Stone et al. 2008). The influence of different modulation rates on the correlation of STI-based measures with vocoded speech was investigated in the present study in both steady-noise and competing-talker tasks. This was done using vocoder intelligibility data in which listeners had access to high modulation rates (up to 400 Hz), at least in the higher-frequency ($>$700 Hz) channels. The present modeling experiments thus probe the question as to whether a portion of this range of envelope frequencies is necessary or sufficient to predict performance. To answer this question, we varied systematically the modulation rate (from 12.5 to 180 Hz) in the STI-based models and examined the resulting correlation with the intelligibility of vocoded speech.

The EAS vocoder performs no processing on the low-frequency (typically $<$600 Hz) portion of the signal, which is in turn augmented with tone-vocoded speech in the high-frequency portions of the signal. The performance of all measures was evaluated with both tone-vocoded (Dorman et al. 1997a) and EAS-vocoded speech (Dorman et al. 2005) corrupted by steady noise or two competing talkers at several SNR levels. Given the differences in information contained in EAS-vocoded and tone-vocoded speech, it is important to assess and compare the performance of the proposed measures with both tone-vocoded and EAS-vocoded speech.

The fact that vocoded speech will be used raises the question as to whether the clean (wideband) waveform or the vocoded clean waveform should be used for reference when computing the intelligibility measures. Both reference signals will be investigated in the present study to answer this question. If the clean waveform is used as reference, then what is the influence of the number of bands used in the implementation of the intelligibility measures and should the number of analysis bands match those used in the vocoder? To address the above questions, intelligibility scores of vocoded speech were first collected from listening experiments and subsequently correlated with a number of intelligibility measures.

Table 1. Summary of subject and test conditions involved in the correlation analysis

Experiment	No. Subjects	Maskers	SNR Levels (dB)	No. Conditions: Tone Vocoder	No. Conditions: EAS Vocoder
1	7	SSN, two-talker	5, 0, −5	6	24
2	6	SSN, two-talker	5, 0	4	20
3	9	SSN	5, 0	6	20

SPEECH INTELLIGIBILITY DATA

Speech intelligibility data were collected from three listening experiments using NH listeners as subjects. The data in experiments 1 and 2 were taken from the study by Chen and Loizou (2010) that assessed the contribution of weak consonants to speech intelligibility in noise. These experiments involved tone-vocoded speech and EAS-vocoded speech corrupted in several noisy conditions. In addition, experiment 3 was introduced in this study to examine the effects of noise suppression on the speech intelligibility measures. It extended the study by Chen and Loizou (2010) by investigating the performance of two noise-suppression algorithms, which were used in a preprocessing stage to enhance tone-vocoded and EAS-processed speech in noisy environments. The same conditions examined in Chen and Loizou (2010) were used in experiment 3. The main difference is that the noisy sentences were first preprocessed by two different noise-suppression algorithms before tone or EAS vocoding. The Wiener filtering algorithm (Scalart & Filho 1996) and the algorithm proposed by Ephraim and Malah (1985) were used for noise suppression.

There are a total of 80 tone-vocoding and EAS-vocoding conditions tested in the three experiments, and each experiment included different numbers of tone-vocoding and EAS-vocoding conditions, as listed in Table 1. For the tone-vocoder simulation, signals were first processed through a pre-emphasis (high-pass) filter (2000 Hz cutoff) with a 3 dB/octave rolloff and then bandpassed into eight frequency bands between 80 and 6000 Hz using sixth-order Butterworth filters. The equivalent rectangular bandwidth scale (Glasberg & Moore 1990) was used to allocate the eight channels within the specified bandwidth. This filter spacing has also been used by Qin and Oxenham (2006) and is shown in Table 2. 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 energy of the envelopes (computed every 4 msecs) 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 root-mean-square value as the original speech segment. For the EAS-vocoder simulation, the signal was LP filtered to 600 Hz using a sixth-order Butterworth filter to simulate the LP acoustic stimulation alone. We then combined the LP (<600 Hz) stimulus with the upper five channels of the eight-channel tone vocoder, as shown in Table 2. The original speech signal was sampled at a rate of 16 kHz (i.e., 8-kHz bandwidth), and to discriminate it from the vocoded signal, it will be henceforth noted as wideband signal throughout the article.

Table 2. Filter cutoff (−3 dB) frequencies used for the tone- and EAS-vocoding processing

Channel	Tone Vocoding: Low (Hz)	Tone Vocoding: High (Hz)	EAS Vocoding: Low (Hz)	EAS Vocoding: High (Hz)
1	80	221		
2	221	426	Unprocessed (80–600)	
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

In summary, the three experiments were designed to cover a wide range of conditions, involving tone and EAS vocoding, corrupted speech, and preprocessed speech by noise-suppression algorithms. Two types of maskers, i.e., continuous steady state noise (SSN) and two competing talkers (two-talker), were used to corrupt the test sentences by additively mixing the target speech and masker signals. The SNR levels were chosen to avoid ceiling/floor effects. Table 1 summarizes the test conditions. More details regarding the test stimuli and experimental conditions can be found in Chen and Loizou (2010).

SPEECH INTELLIGIBILITY MEASURES

Figure 1 shows the proposed framework for computing the intelligibility measures of vocoded speech. As mentioned earlier, given the nature of the vocoded signals, it is not clear whether to use the clean wideband waveform or the vocoded clean waveform as a reference signal when computing the intelligibility measures. Both possibilities were thus examined in this study to answer this question. Furthermore, as shown in Figure 1, the input to all measures is the synthesized vocoded waveform. This facilitated the use of conventional speech intelligibility measures without the need to custom design each measure for different types of vocoding (tone versus noise bands). Use of synthesized vocoded waveform also allowed us to use and examine the same measures for EAS-vocoded speech.

The present intelligibility measures employ primarily either the temporal-envelope or the 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 (Goldsworthy & Greenberg 2004). For the spectral-envelope based measure, we investigated a short-term AI-based measure (AI-ST) (Ma et al. 2009) and a number of coherence-based measures including the coherence-based speech intelligibility index (CSII) measure (Kates & Arehart 2005), the three-level CSII measures ($CSII_{high}$, $CSII_{mid}$, and $CSII_{low}$), and measures derived from the three-level CSII measures, such as the I3 measure (Kates & Arehart 2005) and the Q3 measure (Arehart et al. 2007). Figure 2 shows the implementation of the coherence-based measures. As shown in Figures 1 and 2, the inputs to these measures are the clean signal waveform (or vocoded clean waveform) and the synthesized (processed or corrupted) vocoded waveform.

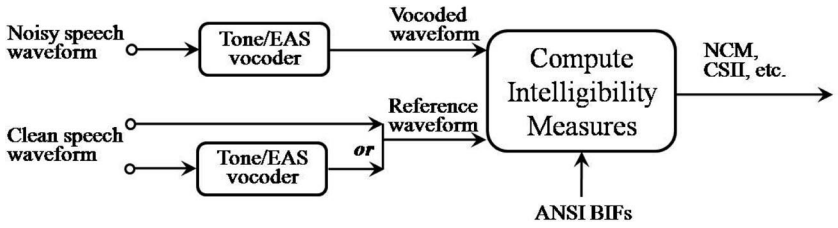

Fig. 1. Framework used in the present study for the computation of intelligibility measures of vocoded speech.

The NCM measure is similar to the STI (Steeneken & Houtgast 1980) in that it computes the STI as a weighted sum of transmission index values determined from the envelopes of the probe and response signals in each frequency band (Goldsworthy & 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 (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 & Greenberg 2004). In computing the NCM measure, the stimuli were first bandpass filtered into K bands spanning the signal bandwidth. The envelope of each band was computed using the Hilbert transform, antialiased using low-pass filtering, and then downsampled to f_d Hz, thereby limiting the envelope modulation frequencies to $0 - f_d/2$ Hz.

The AI-ST measure (Ma et al. 2009) is a simplified version of the AI measure and operates on a short-time (30 msecs) frame basis. This measure differs from the traditional SII measure (ANSI 1997) in many ways: (a) it does not require as input the listener's threshold of hearing, (b) it does not account for upward spread of masking, and (c) it does not require as input the long-term average spectrum (sound pressure) levels of the speech and masker signals. The AI-ST measure divides the signal into short (30 msecs) data segments, computes the AI value for each segment, and averages the segmental AI values over all frames. The AI-ST measure (Ma et al. 2009) also differs from the extended (short-term) AI measure proposed by Rhebergen and Versfeld (2005) in that (a) it uses the same duration window for all critical bands and (b) it does not require as input the listener's threshold of hearing.

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 & Arehart 2005) and have been shown in the study by Ma et al. (2009) to yield high correlations with speech intelligibility.

All the above measures have been implemented assuming an SNR dynamic range of −15 to 15 dB for mapping the SNR computed in each band to the range of 0 to 1. In this way, the resultant intelligibility measure averaged from all bands was limited to the range of 0 to 1. The ANSI AI weights (ANSI 1997) were used as the band-importance function (BIF) in computing the intelligibility measures. The influence of modulation rate, the number of bands, and the choice of reference signal (wideband clean waveform versus vocoded clean speech waveform) used in the implementation of the above measures are investigated in the present study.

RESULTS

Two statistical measures were used to assess the performance of the above speech intelligibility measures, the Pearson's correlation coefficient (r) and the estimate of the standard deviation (SD) of the prediction error (σ_e). The average intelligibility scores obtained by NH listeners for each condition (Table 1) were subjected to correlation analysis with the corresponding average values obtained from the intelligibility measures described in the Speech Intelligibility Measures section. That is, the intelligibility scores were first averaged across the whole listening group for each condition and then subjected to correlation analysis with the corresponding intelligibility measures. As summarized in Table 1, these conditions involved vocoded speech corrupted by two maskers (SSN and a two-talker masker) at three SNR levels and corrupted speech processed by two different noise-suppression algorithms. Intelligibility scores obtained from a total of 80 vocoded conditions, involving EAS-vocoded and tone-vocoded speech, were included in the correlation analysis.

The resulting correlation coefficients and prediction errors are tabulated in Table 3. For all the data given in Table 3, the clean wideband waveform was used as the reference waveform (discussed later in this article). Of all the intelligibility measures considered, the coherence-based and the NCM measures performed the best, accounting for >80% of the variance in speech intelligibility scores. Among the three-level CSII measures, the mid-level CSII ($CSII_{mid}$) measure yielded the highest correlation ($r = 0.91$). This is consistent with the outcomes reported by Arehart et al. (2007) and Ma et al. (2009). Similar to the approach taken in the study by Kates and Arehart (2005), a multiple regression analysis was run on the three CSII

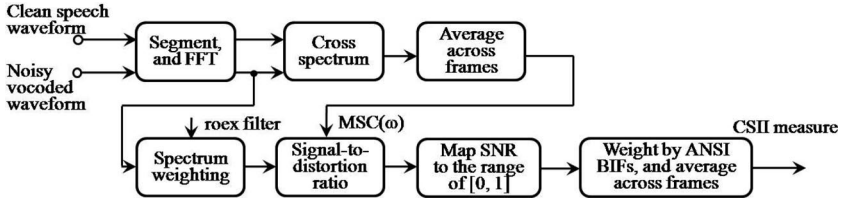

Fig. 2. Signal processing steps involved the implementation of the coherence-based measures.

Table 3. Correlation coefficients (*r*) and SDs of the prediction error (σ_e) between sentence recognition scores and various intelligibility measures

Intelligibility Measure	*r*	σ_e (%)
Spectral-envelope based		
CSII	0.90	9.8
$CSII_{high}$	0.81	13.2
$CSII_{mid}$	0.91	9.3
$CSII_{low}$	0.66	16.9
I3	0.90	9.7
Q3	0.93	8.4
mI3	0.96	6.3
AI-ST	−0.06	22.5
Temporal-envelope based		
NCM	0.92	8.8

The waveform of clean speech was used as the reference waveform in computing the intelligibility measures. The modulation rate used in implementing the NCM measure was 100 Hz.

measures, yielding the following predictive model for intelligibility of vocoded speech:

$$c = -0.28 + 0.86 \cdot \mathrm{CSII}_{low} + 1.96 \cdot \mathrm{CSII}_{mid} + 0.37 \cdot \mathrm{CSII}_{high},$$
$$\mathrm{mI3} = 1/(1 + e^{-c}). \quad (1)$$

The new composite measure, called mI3, improved the I3 correlation from 0.90 to 0.96, making it the highest correlation obtained in the present study. Consistent with the outcomes from the studies by Kates and Arehart (2005) and Ma et al. (2009), the regression analysis yielded the highest coefficient (1.96), i.e., largest weight, for the $CSII_{mid}$ measure in Eq. (1). The $CSII_{mid}$ measure captures information about envelope transients and spectral transitions, critical for the transmission of information regarding place of articulation. Hence, in this regard, the mI3 measure in Eq. (1) places more emphasis on information contained in vowel/consonant spectral transitions and less emphasis on information contained in steady sonorant segments, as captured by the $CSII_{high}$ measure. Figure 3A shows the scatter plot of the predicted mI3 scores against the listeners' recognition scores. Figures 3B, C show the individual scatter plots for the SSN and two-talker masker conditions. As can be seen, high correlations ($r = 0.97$) were maintained consistently for both masker conditions.

Influence of Modulation Rate

To further assess whether including higher (>12.5 Hz) modulation frequencies (i.e., $f_d/2$) would improve the correlation of the NCM measure, we examined additional implementations that included modulation frequencies up to 180 Hz. The correlations obtained with different modulation rates are tabulated in Table 4. As can be seen, there was a notable improvement in the correlation when the modulation rate increased. The correlation improved to $r = 0.92$ when extending the modulation frequency range to 100 Hz, compared with $r = 0.85$ obtained with a modulation rate of 12.5 Hz. The outcomes reported in Table 4 are partly consistent with those reported in the study by van Wijngaarden and Houtgast (2004), which showed that increasing the modulation rate (to 31.5 Hz) improved the correlation of the STI

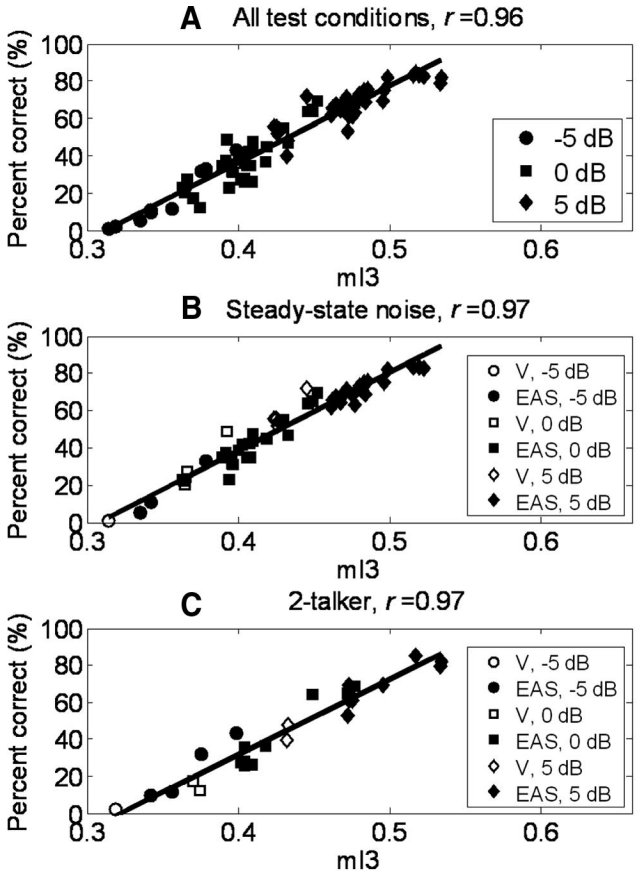

Fig. 3. Scatter plots of intelligibility scores against the predicted values of the proposed mI3 measure for (A) all test conditions, and conditions involving vocoded speech corrupted by (B) steady-state noise and (C) two-talker masker. "V" and "EAS" denote tone-vocoder and EAS-vocoder conditions, respectively.

index with the intelligibility of conversational-style speech. Furthermore, the data are consistent with the conclusions from the study by Stone et al. (2008), who showed benefits of higher modulation rates (180 Hz) particularly when speech was vocoded into a small number of channels. Souza and Rosen (2009) also reported benefit with modulation rates up to 300 Hz. In our study, however, the use of modulation rates >100 Hz did not improve further the correlation of the NCM measure. It should also be noted that, when we consider the influence of the range of envelope modulation frequencies on intelligibility prediction, fewer channels carry envelope information with high modulation

Table 4. Correlation coefficients obtained with the NCM measure for different modulation rates ranging from 12.5 to 180 Hz

Modulation Rate (Hz)	*r*
12.5	0.85
20	0.90
31.5	0.90
40	0.90
60	0.91
80	0.91
100	0.92
140	0.92
180	0.92

Table 5. Individual analysis of correlations (*r*) for the two types of vocoding used and for the two types of maskers used

		r		*r*	
Intelligibility Measure	*r*	Tone-Vocoder	EAS-Vocoder	SSN Masker	Two-Talker Masker
Spectral-envelope based					
CSII	0.90	0.98	0.87	0.91	0.88
$CSII_{high}$	0.81	0.87	0.77	0.81	0.80
$CSII_{mid}$	0.91	0.93	0.89	0.93	0.89
I3	0.90	0.93	0.88	0.92	0.88
Q3	0.93	0.93	0.94	0.93	0.94
mI3	0.96	0.93	0.96	0.97	0.97
Temporal-envelope based					
NCM	0.92	0.95	0.90	0.92	0.91

The modulation rate used in the implementation of the NCM measure was 100 Hz.

rate. For instance, as shown in Table 2, only channels 2 and above carry envelope information with modulation rate up to 100 Hz.

Influence of Vocoding Type

The correlations shown in Table 3 included conditions involving both EAS-vocoded and tone-vocoded speech. We also examined separately the correlations obtained with EAS vocoding (64 conditions) and tone vocoding (16 conditions), and Table 5 shows the resulting correlation coefficients. For the most part, the correlations obtained with tone-vocoding were higher than the corresponding correlations obtained with EAS-vocoding. The number of tone-vocoding conditions, however, was significantly lower than the number of EAS-vocoding conditions. Nevertheless, the EAS correlations were modestly high, ranging from a low of $r = 0.77$ to a high of $r = 0.96$. We also examined separately the correlations obtained with the two types of maskers, and the results are shown in Table 5. As can be seen, consistently high correlation coefficients (ranging from $r = 0.80$ to 0.97) were obtained for both types of maskers tested.

Influence of Number of Bands

Given the computational framework used in the present study (Fig. 1), one can vary the number of bands used in the analysis of the clean reference signal without paying attention to the number of channels used to synthesize the vocoded signals. This raises the question whether the number of bands used in the analysis of the reference clean signal should match the number used to produce the vocoded speech. We thus assessed the impact of the number of critical bands used in the computation of the proposed measures. In Table 3, a total of $K = 20$ critical bands spanning the bandwidth of 100 to 8000 Hz were used in the implementation of all the measures tested. Vocoded speech, on the other hand, was processed using a relatively smaller number (eight) of bands. We thus implemented the above measures using the same number of bands (eight) and the same frequency spacing (Table 2) utilized in the tone vocoder. The resulting correlation coefficients are shown in Table 6. This table also shows the absolute difference in correlations, indicated as $|\Delta r|$, between the 20-band and 8-band implementations of the proposed measures. As can be seen from this table, the number of bands used in the implementation of the NCM, AI-ST, and coherence-based measures had relatively little impact on the correlations. The difference in correlations $|\Delta r|$ was rather small, ranging from 0 to 0.1.

Table 6. Correlation coefficients (*r*) between sentence recognition scores and various intelligibility measures implemented using eight bands spaced the same way as the analysis filters used in the tone-vocoder

Intelligibility Measure	*r*	$\lvert\Delta r\rvert$
Spectral-envelope based		
CSII	0.90	0.00
$CSII_{high}$	0.84	0.03
$CSII_{mid}$	0.90	0.01
$CSII_{low}$	0.61	0.05
I3	0.90	0.00
Q3	0.89	0.04
mI3	0.92	0.04
AI-ST	−0.05	0.01
Temporal-envelope based		
NCM	0.82	0.10

The fourth column shows the absolute difference in correlations $|\Delta r|$ between 8-band and 20-band implementations (Table 3) of the intelligibility measures. The modulation rate used in implementing the NCM measure was 100 Hz.

Influence of Reference Waveform

So far, we have reported correlations when using the clean wideband signal as the reference waveform (see Fig. 1) for the computation of the intelligibility measures. As shown in Figure 1, alternatively, the vocoded clean waveform can be used as the

Table 7. Correlation coefficients (*r*) and SDs of the prediction error (σ_e) between sentence recognition scores and various intelligibility measures

Intelligibility Measure	*r*	σ_e (%)
Spectral-envelope based		
CSII	0.86	11.5
$CSII_{high}$	0.77	14.4
$CSII_{mid}$	0.85	11.9
$CSII_{low}$	0.62	17.7
I3	0.84	12.1
Q3	0.85	11.7
mI3	0.91	9.5
Temporal-envelope based		
NCM	0.88	10.5

The waveform of the vocoded clean speech was used as reference when computing the intelligibility measures (see Fig. 1). The modulation rate in implementing the NCM measure was 100 Hz.

reference waveform. Table 7 shows the resulting correlations obtained when the vocoded clean speech was used as the reference waveform. All parameters were the same as those used in Table 3. Comparing the correlations given in Tables 3 and 7, it is observed that the two sets of results shared the same pattern. The STI-based and coherence-based measures consistently performed the best, with correlations of 0.88 and 0.86, respectively.

DISCUSSION AND CONCLUSIONS

The results in this study share several common findings with those obtained with wideband (nonvocoded) speech by Ma et al. (2009). The majority of the measures that were found by Ma et al. (2009) to predict reliably the intelligibility of (wideband) nonvocoded speech in noise were also the ones found in the present study to predict well the intelligibility of vocoded speech. These included the STI-based (NCM) and coherence-based measures.

The traditional STI method has been found to perform poorly in terms of predicting the intelligibility of processed speech wherein nonlinear operations are involved (Ludvigsen et al. 1993; Van Buuren et al. 1998; Goldsworthy & Greenberg 2004). The NCM measure, albeit an STI-based measure, is computed differently than the STI measure and has been shown by Goldsworthy and Greenberg (2004) to perform better than the conventional STI method in predicting the effects of nonlinear operations such as envelope thresholding or distortions introduced by spectral-subtractive algorithms. This was also confirmed by Ma et al. (2009) who evaluated the performance of the NCM measure with noise-suppressed speech, which generally contains various forms of nonlinear distortions including the distortions introduced by spectral-subtractive algorithms. The correlation of the NCM measure with noise-suppressed speech was found to be quite high ($r = 0.89$) (Ma et al. 2009). The present study extends the utility of the NCM measure to vocoded speech and further confirms that it can be used for nonlinearly processed speech.

One of the main conclusions that can be drawn from the present study is that higher (up to 100 Hz) modulation rates are needed in the STI-based measure (NCM) for better prediction of the intelligibility of vocoded speech. Alternatively, and perhaps equivalently, we can say that for the vocoded data considered in this study (based on a 400-Hz envelope cutoff frequency), it is not necessary to include modulation rates as high as 400 Hz to predict reliably speech intelligibility. Including modulation rates up to 100 Hz seems to be sufficient. In contrast, this was not found to be true for wideband speech, as modulation rates of 12.5 Hz were found to be sufficient to achieve high correlations (Houtgast & Steeneken 1985; Ma et al. 2009). We believe that the difference in modulation rates needed for modeling intelligibility of wideband versus vocoded speech can be attributed to the increased importance of higher modulation rates when the spectral information is limited (eight channels in our study), as is the case in most vocoder studies and demonstrated in several studies (Healy & Steinbach 2007; Stone et al. 2008). In addition, as demonstrated by the difference in outcomes between the studies by Xu and Zheng (2007) and Stone et al. (2008), higher modulation rates are needed for better modeling of the intelligibility of vocoded speech in competing-talker listening tasks.

Slightly higher correlations were obtained when 20 critical bands were used in the implementation of the intelligibility measures (Table 3). We believe that this was because the 20-band spectral representation of vocoded speech captures additional spectral information associated with the use of tone vocoders (Whitmal et al. 2007; Stone et al. 2008). Tone vocoders inherently contain spectral sideband information, particularly when the number of vocoded channels is small and the channel bandwidths are large. These sidebands can be easily resolved by NH listeners and subsequently be used in noisy conditions (Whitmal et al. 2007; Stone et al. 2008). Hence, use of 20 bands better captures additional information contained in the tone-vocoded signal that might otherwise not be present in the eight-channel implementations of the intelligibility measures (as reported in Table 6). It is thus recommended to use the 20-band implementations of the examined intelligibility measures for better modeling of the intelligibility of vocoded speech. Given that noise-band vocoders (not examined in the present study) do not contain spectral sideband information, we do not expect to see a difference in correlations between 20- and 8-band implementations of the measures examined. Further work is needed to confirm this.

The majority of the measures examined in the present study are based on either a temporal-envelope representation (NCM measure) or a spectral-envelope representation (e.g., AI-ST and coherence-based measures) of vocoded speech. The former measure discards temporal fine-structure as done by current CI processing strategies and in fact employs only envelope information (e.g., up to 100 Hz modulations). High correlation ($r = 0.92$) was obtained in the present study when modulation rates up to 100 Hz were included. The spectral-envelope based measures (AI-ST and coherence-based measures) integrate fine-spectral information contained within critical bands to derive an auditory-motivated representation of the signal. In our study, fine-spectral information was available in the low-frequency (<600 Hz) portion of the acoustic signal in the EAS conditions. Yet, when comparing the correlations obtained with EAS-vocoded signals versus CI-vocoded signals (see Table 5), we observe that the correlations obtained in the CI-vocoded conditions were in fact higher than those obtained in the EAS conditions. Hence, integrating fine-spectral information in the low frequencies, as required for the implementation of the spectral-envelope measures (e.g., CSII), did not seem to be beneficial in terms of improving the correlation. Further, and perhaps clearer, evidence was provided with the experiments reported in Table 6. No fine-spectral information was present in the eight-channel vocoded speech, yet the coherence-based measures yielded high ($r > 0.84$) correlation (see Table 6) despite the fact that both the clean and noisy signals were vocoded into eight channels. A high correlation ($r = 0.92$) was maintained with the derived mI3 measure [Eq. (1)]. These findings taken together with the outcomes reported in Ma et al. (2009) suggest that it is not necessary, in terms of predicting the intelligibility of vocoded or wideband speech, to develop measures that incorporate fine-structure information. Measures that incorporate primarily (temporal or spectral) envelope information can account for $>70\%$ of the variance in speech intelligibility scores. This is not to say that fine-structure information is not important or should not be included (if available) in the implementation of intelligibility measures or implementation of speech coding strategies, but rather that it is not necessary, at least

in the context of predicting the intelligibility of vocoded speech with eight channels. The question of how measures incorporating fine structure information will further improve the intelligibility prediction of vocoded speech remains to be answered.

The present study took the first step in modeling the intelligibility of vocoded speech. High correlations were obtained with both STI-based and coherence-based measures. High correlations with the intelligibility of vocoded speech were also obtained in our previous study (Chen & Loizou 2010) using a measure that was originally designed to predict subjective speech quality. The speech intelligibility measures examined in the present study as well as in our previous study (Chen & Loizou 2010) could be used to guide the development of new speech processing strategies for CIs. These measures could be used, for instance, to tune the parameters of novel noise-suppression algorithms or assist in the selection of other signal processing parameters (e.g., shape of compression map) involved in the implementation of CI speech coding strategies.

ACKNOWLEDGMENTS

The authors thank the Associate Editor, Dr. Gail Donaldson, and the two reviewers who provided valuable feedback that significantly improved the presentation of the manuscript.

This work was supported by Grant No. R01 DC007527 from the National Institute of Deafness and other Communication Disorders, NIH.

Address for correspondence: Philipos C. Loizou, PhD, University of Texas at Dallas, Department of Electrical Engineering, EC 33, 800 West Campbell Road, Richardson, TX 75080-3021. E-mail: loizou@utdallas.edu.

Received October 21, 2009; accepted September 13, 2010.

REFERENCES

ANSI (1997). *Methods for Calculation of the Speech Intelligibility Index. Report No. ANSI S3.5-1997*. New York, NY: American National Standards Institute.

Arehart, K., Kates, J., Anderson, M., et al. (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., & Loizou, P. (2010). Contribution of consonant landmarks to speech recognition in simulated acoustic-electric hearing. *Ear Hear*, *31*, 259–267.

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

Dorman, M., Loizou, P., Fitzke, J., et al. (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. *J Acoust Soc Am*, *104*, 3583–3585.

Dorman, M., Spahr, A., Loizou, P., et al. (2005). Acoustic simulations of combined electric and acoustic hearing (EAS). *Ear Hear*, *26*, 371–380.

Drullman, R., Festen, J. M., Plomp, R. (1994). Effect of temporal envelope smearing on speech perception. *J Acoust Soc Am*, *95*, 1053–1064.

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

Fletcher, H. (1940). Auditory patterns. *Rev Mod Phys*, *12*, 47–55.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *J Acoust Soc Am*, *19*, 90–119.

Friesen, L. M., Shannon, R. V., Baskent, D., et al. (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.

Glasberg, B., & Moore, B. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear Res*, *47*, 103–138.

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

Healy, E., & Steinbach, H. (2007). The effect of smoothing filter slope and spectral frequency on temporal speech information. *J Acoust Soc Am*, *121*, 1177–1181.

Houtgast, T., & 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.

Kasturi, K., & Loizou, P. (2007). Effect of frequency spacing on melody recognition: Acoustic and electric hearing. *J Acoust Soc Am*, *122*, EL29–EL34.

Kasturi, K., Loizou, P., Dorman, M., et al. (2002). The intelligibility of speech with holes in the spectrum. *J Acoust Soc Am*, *112*, 1102–1111.

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

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

Kong, Y., & Carlyon, R. (2007). Improved speech recognition in noise in simulated binaurally combined acoustic and electric stimulation. *J Acoust Soc Am*, *121*, 3717–3727.

Li, N., & Loizou, P. (2008). A glimpsing account for the benefit of simulated combined acoustic and electric hearing. *J Acoust Soc Am*, *123*, 2287–2294.

Loizou, P. (2006). Speech Processing in Vocoder-Centric Cochlear Implants. In A. Moller (Ed). *Cochlear and Brainstem Implants* (pp. 109–143). Basel, New York: Karger.

Ludvigsen, C., Elberling, C., Keidser, G. (1993). Evaluation of a noise reduction method—Comparison of observed scores and scores predicted from STI. *Scand Audiol Suppl*, *38*, 50–55.

Ma, J., Hu, Y., 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.

Qin, M., & Oxenham, A. (2005). Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification. *Ear Hear*, *26*, 451–460.

Qin, M., & Oxenham, A. (2006). Effects of introducing unprocessed low-frequency information on the reception of the envelope-vocoder processed speech. *J Acoust Soc Am*, *119*, 2417–2426.

Rhebergen, K. S., & Versfeld, N. J. (2005). A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *J Acoust Soc Am*, *117*, 2181–2192.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philos Trans R Soc Lond B Biol Sci*, *336*, 367–373.

Scalart, P., & Filho, J. (1996). Speech enhancement based on a priori signal to noise estimation. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Atlanta, GA, *2*, 629–632.

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

Shannon, R. V., Zeng, F. G., Kamath, V., et al. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Souza, P., & Rosen, S. (2009). Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech. *J Acoust Soc Am*, *126*, 792–805.

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

Stone, M., Füllgrabe, C., Moore, B. (2008). Benefit of high-rate envelope cues in vocoder processing: Effect of number of channels and spectral region. *J Acoust Soc Am*, *124*, 2272–2282.

van Buuren, R. A., Festen, J. M., Houtgast, T. (1998). Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality. *J Acoust Soc Am*, *105*, 2903–2913.

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

Whitmal N., Poissant, S., Freyman, R., et al. (2007). Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. *J Acoust Soc Am*, *122*, 2376–2388.

Xu, L., Thompson, C. S., Pfingst, B. E. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *J Acoust Soc Am*, *117*, 3255–3267.

Xu, L., & Zheng, Y. (2007). Spectral and temporal cues for phoneme recognition in noise. *J Acoust Soc Am*, *122*, 1758–1764.