Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions

Jianfen Ma[a)]
College of Computer Engineering and Software, Taiyuan University of Technology, Shanxi 030024, China and Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083-0688

Yi Hu and Philipos C. Loizou[b)]
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083-0688

(Received 8 August 2008; revised 11 December 2008; accepted 14 February 2009)

The articulation index (AI), speech-transmission index (STI), and coherence-based intelligibility metrics have been evaluated primarily in steady-state noisy conditions and have not been tested extensively in fluctuating noise conditions. The aim of the present work is to evaluate the performance of new speech-based STI measures, modified coherence-based measures, and AI-based measures operating on short-term (30 ms) intervals in realistic noisy conditions. Much emphasis is placed on the design of new band-importance weighting functions which can be used in situations wherein speech is corrupted by fluctuating maskers. The proposed measures were evaluated with intelligibility scores obtained by normal-hearing listeners in 72 noisy conditions involving noise-suppressed speech (consonants and sentences) corrupted by four different maskers (car, babble, train, and street interferences). Of all the measures considered, the modified coherence-based measures and speech-based STI measures incorporating signal-specific band-importance functions yielded the highest correlations (r=0.89–0.94). The modified coherence measure, in particular, that only included vowel/consonant transitions and weak consonant information yielded the highest correlation (r=0.94) with sentence recognition scores. The results from this study clearly suggest that the traditional AI and STI indices could benefit from the use of the proposed signal- and segment-dependent band-importance functions.
 [DOI: 10.1121/1.3097493]

PACS number(s): 43.72.Ar, 43.72.Dv [DOS] Pages: 3387–3405

I. INTRODUCTION

A number of measures have been proposed to predict speech intelligibility in the presence of background noise. Among these measures, the articulation index (AI) (French and Steinberg, 1947; Fletcher and Galt, 1950; Kryter, 1962a, 1962b) and speech-transmission index (STI) (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985) are by far the most commonly used today for predicting speech intelligibility in noisy conditions. The AI measure was further refined to produce the speech intelligibility index (SII) (ANSI, 1997). The SII measure is based on the idea that the intelligibility of speech depends on the proportion of spectral information that is audible to the listener and is computed by dividing the spectrum into 20 bands (contributing equally to intelligibility) and estimating the weighted average of the signal-to-noise ratios (SNRs) in each band (Kryter, 1962a, 1962b; Pavlovic, 1987; Allen, 1994; ANSI, 1997). The SNRs in each band are weighted by band-importance functions (BIFs) which differ across speech materials (ANSI, 1997). The SII measure has been shown to predict successfully the effects of linear filtering and additive noise on speech intelligibility (e.g., Kryter, 1962a, 1962b). It has, however, a number of limitations. For one, the computation of the SII measure requires as input the levels of speech and masker signals at the eardrum of the listeners, something that might not be available in situations wherein we only have access to recorded (digitized) processed signals. Second, the SII measure has been validated for the most part only for steady (stationary) masking noise since it is based on the long-term average spectra (computed over 125-ms intervals) of the speech and masker signals. As such, it cannot be applied to situations in which speech is embedded in fluctuating maskers (e.g., competing talkers). Several attempts have been made to extend the SII measure to assess speech intelligibility in fluctuating maskers (Rhebergen *et al.*, 2005, 2006; Kates, 1987). Rhebergen *et al.* (2006), for instance, proposed to divide the speech and masker signals into short frames (9–20 ms), evaluate the instantaneous AI value in each frame, and average the computed AI values across all frames to produce a single AI metric. Their extended short-term AI (AI-ST) measure was found to predict speech intelligibility better than the traditional AI measure when evaluated with sentences embedded in artificial masking signals (e.g., periodically interrupted noise) and speech-like maskers, but the predictions with the latter maskers were found to be less accurate (Rhebergen and Versfeld, 2005).

[a)]Work done while Dr. Jianfen Ma visited Professor Loizou's laboratory as a research scholar.
[b)]Author to whom correspondence should be addressed. Electronic mail: loizou@utdallas.edu

Other extensions to the SII measure were proposed by Kates and Arehart (2005) for predicting the intelligibility of peak-clipping and center-clipping distortions in the speech signal, such as those found in hearing aids. The modified index, called the CSII index, used the base form of the SII procedure, but with the SNR estimate replaced by the signal-to-distortion ratio, which was computed using the coherence function between the input and processed signals. While a modest correlation was obtained with the CSII index, a different version was proposed that divided the speech segments into three level regions and computed the CSII index separately for each level region. The three-level CSII index yielded higher correlations for both intelligibility and subjective quality ratings (Arehart *et al.*, 2007) of hearing-aid type of distortions. Further testing of the CSII index is performed in the present study to examine whether it can be used (1) to predict the intelligibility of speech corrupted by fluctuating maskers and (2) to predict the intelligibility of noise-suppressed speech containing different types of non-linear distortions than those introduced by hearing aids.

The STI measure (Steeneken and Houtgast, 1980) is based on the idea that the reduction in intelligibility caused by additive noise or reverberation distortions can be modeled in terms of the reduction in temporal envelope modulations. The STI metric has been shown to predict successfully the effects of reverberation, room acoustics, and additive noise (e.g., Steeneken and Houtgast, 1982; Houtgast and Steeneken, 1985). It has also been validated in several languages (Anderson and Kalb, 1987; Brachmanski, 2004). In its original form (Houtgast and Steeneken, 1971), the STI measure used artificial signals (e.g., sinewave-modulated signals) as probe signals to assess the reduction in signal modulation in a number of frequency bands and for a range of modulation frequencies (0.6–12.5 Hz) known to be important for speech intelligibility. When speech is subjected, however, to non-linear processes such as those introduced by dynamic envelope compression (or expansion) in hearing aids, the STI measure fails to successfully predict speech intelligibility since the processing itself might introduce additional modulations which the STI measure interprets as increased SNR (Hohmann and Kollmeieir, 1995; Ludvigsen *et al.*, 1993; van Buuren *et al.*, 1999; Goldsworthy and Greenberg, 2004). For that reason, several modifications have been proposed to use speech or speech-like signals as probe signals in the computation of the STI measure (Steeneken and Houtgast, 1980; Ludvigsen *et al.*, 1990). Despite these modifications, several studies have reported that the speech-based STI methods fail to predict the intelligibility of nonlinearly-processed speech (van Buuren *et al.*, 1999; Goldsworthy and Greenberg, 2004). Several modifications were made by Goldsworthy and Greenberg (2004) to existing speech-based STI measures but none of these modifications were validated with intelligibility scores obtained with human listeners.

The SII and speech-based STI measures can account for linear distortions introduced by filtering and additive noise, but have not been tested extensively in conditions wherein non-linear distortions might be present, such as when speech is processed via hearing-aid algorithms or noise-suppression algorithms. Some of the noise-suppression algorithms (e.g., spectral subtractive), for instance, can introduce non-linear distortions in the signal and unduly increase the level of modulation in the temporal envelope (e.g., Goldsworthy and Greenberg, 2004). The increased modulation might be interpreted as increased SNR by the STI measure. Hence, it remains unclear whether the speech-based STI measures or the SII measure can account for the type of distortions introduced by noise-suppression algorithms and to what degree they can predict speech intelligibility. It is also not known whether any of the numerous objective measures that have been proposed to predict speech quality (Quackenbush *et al.*, 1988; Loizou, 2007, Chap. 10; Hu and Loizou, 2008) in voice communications applications can be used to predict speech intelligibility. An objective measure that would predict well both speech intelligibility and quality would be highly desirable in voice communication and hearing-aid applications. The objective quality measures are primarily based on the idea that speech quality can be modeled in terms of differences in loudness between the original and processed signals (e.g., Bladon and Lindblom, 1981) or simply in terms of differences in the spectral envelopes [e.g., as computed using a linear predictive coding (LPC) model] between the original and processed signals. The perceptual evaluation of speech quality (PESQ) objective measure (ITU-T, 2000; Rix *et al.*, 2001), for instance, assesses speech quality by estimating the overall loudness difference between the noise-free and processed signals. This measure has been found to predict very reliably ($r>0.9$) the quality of telephone networks and speech codecs (Rix *et al.*, 2001) as well as the quality of noise-suppressed speech (Hu and Loizou, 2008). Only a few studies (Beerends *et al.*, 2004, 2005) have tested the PESQ measure in the context of predicting speech intelligibility. High correlation ($r>0.9$) was reported, but it was for a relatively small number of noisy conditions which included speech processed via low-rate vocoders (Beerends *et al.*, 2005) and speech processed binaurally via beamforming algorithms (Beerends *et al.*, 2004). The speech distortions introduced by noise-suppression algorithms (based on single-microphone recordings) differ, however, from those introduced by low-rate vocoders. Hence, it is not known whether the PESQ measure can predict reliably the intelligibility of noise-suppressed speech containing various forms of non-linear distortions, such as musical noise.

The aim of the present work is two-fold: (1) to evaluate the performance of conventional objective measures originally designed to predict speech quality and (2) to evaluate the performance of new speech-based STI measures, modified coherence-based measures (CSII), as well as AI-based measures that were designed to operate on short-term (20–30 ms) intervals in realistic noisy conditions. A number of modifications to the speech-based STI, coherence-based, and AI measures are proposed and evaluated in this study. Much focus is placed on the development of band-importance weighting functions which can be used in situations wherein speech is corrupted by fluctuating maskers. This is pursued with the understanding that a single BIF,

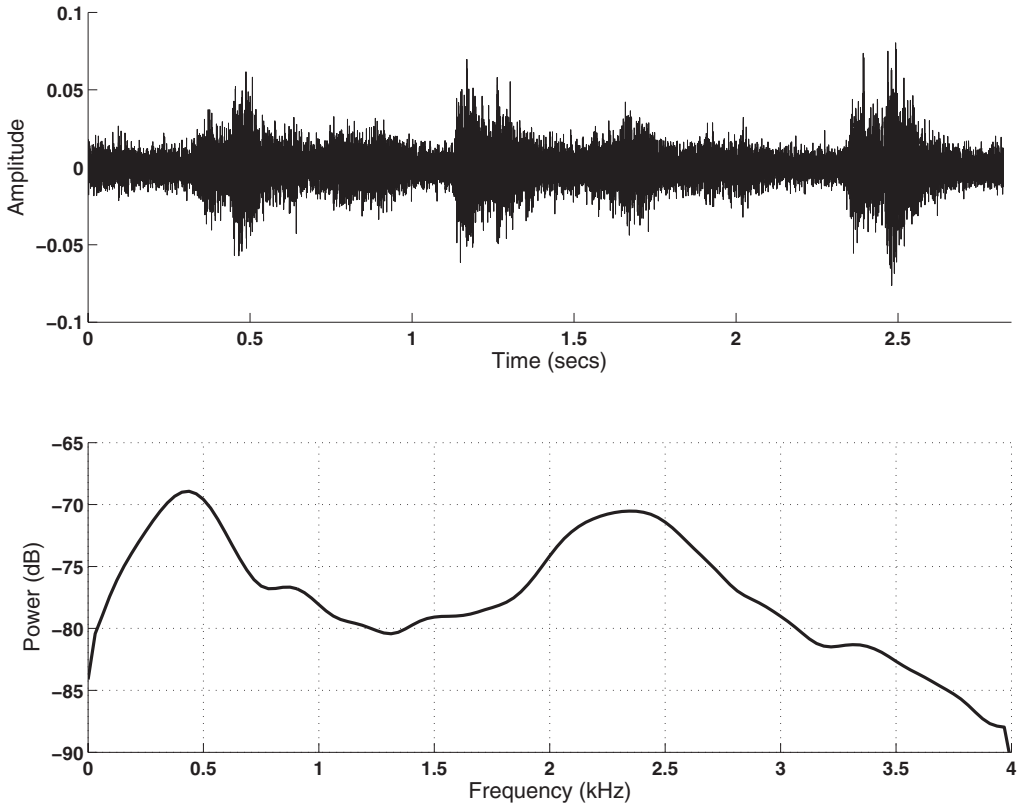

FIG. 1. Waveform (top panel) and long-term averaged spectrum (bottom panel) of the train noise used in the present study.

such as those used in STI and SII indices (ANSI, 1997), might not suitable for evaluating the intelligibility of speech embedded in fluctuating maskers.

II. METHODS

The intelligibility evaluation of noise-corrupted speech processed through eight different noise-suppression algorithms was reported in Hu and Loizou (2007) and is summarized briefly below.

A. Materials and subjects

IEEE sentences (IEEE, 1969) and consonants in/a C a/ format were used as test material. The consonant test included 16 consonants recorded in /a C a/ context, where C = /p,t,k,b,d,g,m,n,dh,l,f,v,s,z,sh,dj/. All consonants were produced by a female speaker, and all sentences were produced by a male talker. The sentences and consonants were originally sampled at 25 kHz and downsampled to 8 kHz. These recordings are available in Loizou (2007). The maskers were artificially added to the speech material. The masker signals were taken from the AURORA database (Hirsch and Pearce, 2000) and included the following real-world recordings from different places: babble, car, street, and train. Figure 1 shows the time-domain waveform and long-term average spectrum of the train noise, illustrating the modulating nature of this masker. The maskers were added to the speech signals at SNRs of 0 and 5 dB.

A total of 40 native speakers of American English were recruited for the sentence intelligibility tests, and 10 additional listeners were recruited for the consonant tests. All subjects were paid for their participation.

B. Noise reduction algorithms

The noise-corrupted sentences were processed by eight different noise-reduction algorithms which included the generalized subspace approach (Hu and Loizou, 2003), the perceptually-based subspace approach (Jabloun and Champagne, 2003), the log minimum mean square error (logMMSE) algorithm (Ephraim and Malah, 1985), the logMMSE algorithm with speech-presence uncertainty (Cohen and Berdugo, 2002), the spectral subtraction algorithm based on reduced-delay convolution (Gustafsson *et al.*, 2001), the multiband spectral-subtractive algorithm (Kamath and Loizou, 2002), the Wiener filtering algorithm based on wavelet-thresholded multitaper spectra (Hu and Loizou, 2004), and the traditional Wiener algorithm (Scalart and Filho, 1996). With the exception of the logMMSE-SPU algorithm which was provided by the authors (Cohen and Berdugo, 2002), all other algorithms were based on our own implementation. The parameters used in the implementation of these algorithms were the same as those published. MATLAB implementations of all noise reduction algorithms tested in the present study are available in Loizou (2007).

C. Procedure

A total of 40 native speakers of American English were recruited for the sentence intelligibility tests. The 40 listeners

were divided into four panels (one per type of noise), with each panel consisting of 10 listeners. Each subject participated in a total of 19 listening conditions (=2 SNR levels×8 algorithms+2 noisy references+1 quiet). Two IEEE sentence lists (ten sentences per list) were used for each condition, and none of the sentence lists were repeated. Additional ten listeners were recruited for the consonant recognition task. Subjects were presented with six repetitions of each consonant in random order. The processed speech files (sentences/consonants), along with the clean and noisy speech files, were presented monaurally to the listeners in a double-walled sound-proof booth (Acoustic Systems, Inc.) via Sennheiser's (HD 250 Linear II) circumaural headphones at a comfortable level.

The intelligibility study by Hu and Loizou (2007) produced a total of 72 noisy conditions including the noise-corrupted (unprocessed) conditions. The 72 conditions included distortions introduced by 8 different noise-suppression algorithms operating at two SNR levels (0 and 5 dB) in four types of real-world environments (babble, car, street, and train). The intelligibility scores obtained in the 72 conditions were used in the present study to evaluate the predictive power of a number of old and newly proposed objective measures.

III. OBJECTIVE MEASURES

A number of objective measures are examined in the present study for predicting the intelligibility of speech in noisy conditions. Some of the objective measures (e.g., PESQ) have been used successfully for the evaluation of speech quality (e.g., Quackenbush *et al.*, 1988; Rix et al, 2001), while others are more appropriate for intelligibility assessment. A description of these measures along with the proposed modifications to speech-based STI and AI-based measures is given next.

A. PESQ

Among all objective measures considered, the PESQ measure is the most complex to compute and is the one recommended by ITU-T (2000) for speech quality assessment of 3.2 kHz (narrow-band) handset telephony and narrow-band speech codecs (Rix *et al.*, 2001; ITU-T, 2000). The PESQ measure is computed as follows. The original (clean) and degraded signals are first level equalized to a standard listening level and filtered by a filter with response similar to that of a standard telephone handset. The signals are time aligned to correct for time delays, and then processed through an auditory transform to obtain the loudness spectra. The difference in loudness between the original and degraded signals is computed and averaged over time and frequency to produce the prediction of subjective quality rating. The PESQ produces a score between 1.0 and 4.5, with high values indicating better quality. High correlations ($r>0.92$) with subjective listening tests were reported by Rix *et al.* (2001) using the above PESQ measure for a large number of testing conditions taken from voice-over-internet protocol applications. High correlation ($r\approx 0.9$) was also reported in Hu and Loizou (2008) with the subjective quality judgments of noise-corrupted speech processed via noise-suppression algorithms.

B. LPC-based objective measures

The LPC-based measures assess, for the most part, the spectral envelope difference between the input (clean) signal and the processed (or corrupted) signal. Three different LPC-based objective measures were considered: the log likelihood ratio (LLR), the Itakura–Saito (IS), and the cepstrum (CEP) distance measures. All three measures assess the difference between the spectral envelopes, as computed by the LPC model, of the noise-free and processed signals. The LLR measure is defined as (Quackenbush *et al.*, 1988)

$$d_{\mathrm{LLR}}(\vec{a}_p,\vec{a}_c)=\log\left(\frac{\vec{a}_p\mathbf{R}_c\vec{a}_p^T}{\vec{a}_c\mathbf{R}_c\vec{a}_c^T}\right), \tag{1}$$

where $\vec{a}_c$ is the LPC vector of the clean speech signal, $\vec{a}_p$ is the LPC vector of the processed (enhanced) speech signal, and $\mathbf{R}_c$ is the autocorrelation matrix of the noise-free speech signal. Only the smallest 95% of the frame LLR values were used to compute the average LLR value (Hu and Loizou, 2008). The segmental LLR values were limited in the range of [0, 2] to further reduce the number of outliers (Hu and Loizou, 2008).

The IS measure is defined as (Quackenbush *et al.*, 1988)

$$d_{\mathrm{IS}}(\vec{a}_p,\vec{a}_c)=\frac{\sigma_c^2}{\sigma_p^2}\left(\frac{\vec{a}_p\mathbf{R}_c\vec{a}_p^T}{\vec{a}_c\mathbf{R}_c\vec{a}_c^T}\right)+\log\left(\frac{\sigma_c^2}{\sigma_p^2}\right)-1, \tag{2}$$

where σ_c^2 and σ_p^2 are the LPC gains of the clean and processed signals, respectively. The IS values were limited in the range of [0, 100] to minimize the number of outliers.

The CEP distance provides an estimate of the log spectral distance between two spectra and is computed as follows (Kitawaki *et al.*, 1988):

$$d_{\mathrm{CEP}}(\vec{c}_c,\vec{c}_p)=\frac{10}{\log 10}\sqrt{2\sum_{k=1}^{p}[c_c(k)-c_p(k)]^2}, \tag{3}$$

where $\vec{c}_c$ and $\vec{c}_p$ are the CEP coefficient vectors of the noise-free and processed signals, respectively. The CEP distance was limited in the range of [0, 10] to minimize the number of outliers (Hu and Loizou, 2008).

C. Time-domain and frequency-weighted SNR measures

The time-domain segmental SNR (SNRseg) measure was computed as per Hansen and Pellom (1998) as follows:

$$\mathrm{SNRseg}=\frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}x^2(n)}{\sum_{n=Nm}^{Nm+N-1}(x(n)-\hat{x}(n))^2}, \tag{4}$$

where $x(n)$ is the input (clean) signal, $\hat{x}(n)$ is the processed (enhanced) signal, N is the frame length (chosen to be 30 ms), and M is the number of frames in the signal. Only frames with SNRseg in the range of $[-10,35]$ dB were considered in the computation of the average (Hansen and Pellom, 1998).

The frequency-weighted segmental SNR (fwSNRseg) was computed using the following equation (Hu and Loizou, 2008):

$$\text{fwSNRseg} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)\log_{10}\dfrac{X(j,m)^2}{(X(j,m)-\hat{X}(j,m))^2}}{\sum_{j=1}^{K} W(j,m)}, \quad (5)$$

where $W(j,m)$ is the weight placed on the jth frequency band, K is the number of bands, M is the total number of frames in the signal, $X(j,m)$ is the critical-band magnitude (excitation spectrum) of the clean signal in the jth frequency band at the mth frame, and $\hat{X}(j,m)$ is the corresponding spectral magnitude of the enhanced signal in the same band. The critical-band spectra $X(j,m)$ in Eq. (5) were obtained by multiplying the FFT magnitude spectra by 25 overlapping Gaussian-shaped windows (Loizou, 2007, Chap. 11) spaced in proportion to the ear's critical bands and summing up the power within each band. Similar to the implementation in Hu and Loizou (2008), the excitation spectra were normalized to have an area of 1. The SNR term in the numerator of Eq. (5) was limited within the range of $[-15,15]$ dB. To assess the influence of the dynamic range on performance, we also considered limiting the SNR range to $[-15,20]$, $[-15,25]$, $[-15,30]$, $[-15,35]$, and $[-10,35]$ dB. The latter range $[-10,35]$ dB) was chosen for two reasons. First, to facilitate comparisons with the SNRseg measure [Eq. (4)], which was also limited to the same range. Second, it was chosen to be consistent with several studies (Boothroyd *et al.*, 1994; Studebaker and Sherbecoe, 2002) that showed that the speech dynamic range often exceeds 30 dB.

For the weighting function $W(j,m)$, we considered the AI weights (given in Table I) as well as the critical-band spectrum of the noise-free signal raised to a power, i.e.,

$$W(j,m) = X(j,m)^p, \quad (6)$$

where p is the power exponent, which can be varied for maximum correlation and can be optimized for different speech materials. In our experiments, we varied p from 0.5 to 4. The AI weights were taken from Table B.1 of the ANSI (1997) standard. For the consonant materials, we used the nonsense syllable weights and for the sentence materials we used the short-passage weights given in Table B.1 (ANSI, 1997). The weights were linearly interpolated to reflect the range of band center-frequencies adopted in the present study.

The value of p in Eq. (6) can control the emphasis or weight placed on spectral peaks and/or spectral valleys. Values of $p<1$, for instance, compress the spectrum, while values of $p>1$ expand the spectrum. Compressive values of $p(p<1)$ equalize the spectrum by boosting the low-intensity components (e.g., spectral valleys). Consequently, the effective dynamic range of the spectrum is reduced, and relatively uniform weights are applied to all spectral components. Figure 2 shows as an example the spectrum of a segment taken from the vowel /ɛ/ (as in "head"), along with the same spectrum raised to powers of 0.25 and 1.25. Note that prior to the compression, the F2 amplitude is very weak compared to the F1 amplitude (compare the top two panels). After the compression, the F2 peak gets stronger and closer in amplitude to F1's. Expansion ($p>1$), on the other hand, has the opposite effect in that it enhances the dominant spectral peak(s), while suppressing further the weak spectral components (see bottom panel in Fig. 2). In this example, the F2 amplitude was further weakened following the spectrum expansion. In brief, the value of p in Eq. (6) controls the steepness of the compression/expansion function, and in practice, it can be optimized for different speech materials.

TABLE I. AI weights (ANSI, 1997) used in the implementation of the fwSNRseg and AI-ST measures for consonants and sentence materials.

Band	Center frequencies (Hz)	Consonants	Sentences
1	50.0000	0.0000	0.0064
2	120.000	0.0000	0.0154
3	190.000	0.0092	0.0240
4	260.000	0.0245	0.0373
5	330.000	0.0354	0.0803
6	400.000	0.0398	0.0978
7	470.000	0.0414	0.0982
8	540.000	0.0427	0.0809
9	617.372	0.0447	0.0690
10	703.378	0.0472	0.0608
11	798.717	0.0473	0.0529
12	904.128	0.0472	0.0473
13	1020.38	0.0476	0.0440
14	1148.30	0.0511	0.0440
15	1288.72	0.0529	0.0470
16	1442.54	0.0551	0.0489
17	1610.70	0.0586	0.0486
18	1794.16	0.0657	0.0491
19	1993.93	0.0711	0.0492
20	2211.08	0.0746	0.0500
21	2446.71	0.0749	0.0538
22	2701.97	0.0717	0.0551
23	2978.04	0.0681	0.0545
24	3276.17	0.0668	0.0508
25	3597.63	0.0653	0.0449

The last conventional measure tested was the weighted spectral slope (WSS) measure (Klatt, 1982). The WSS distance measure computes the weighted difference between the spectral slopes in each frequency band. The spectral slope is obtained as the difference between adjacent spectral magnitudes in decibels. The WSS measure evaluated in this paper is defined as

$$d_{\text{WSS}} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W_{\text{WSS}}(j,m)(S_c(j,m)-S_p(j,m))^2}{\sum_{j=1}^{K} W_{\text{WSS}}(j,m)}, \quad (7)$$

where $W_{\text{WSS}}(j,m)$ are the weights computed as per Klatt (1982), $K=25$, M is the number of data segments, and $S_c(j,m)$ and $S_p(j,m)$ are the spectral slopes for the jth frequency band of the noise-free and processed speech signals, respectively.

Aside from the PESQ measure, all other measures were computed by segmenting the sentences using 30-ms duration

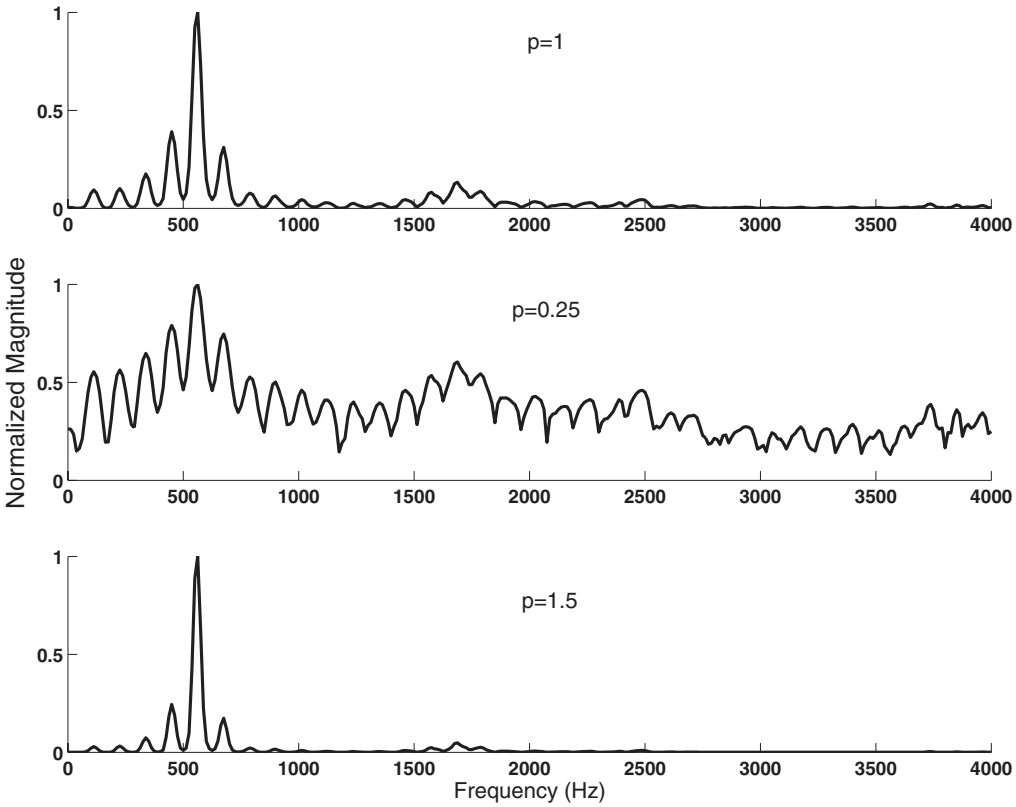

FIG. 2. (Top panel) FFT magnitude spectrum of a segment taken from the vowel /ɛ/ (excised from the word "head" and produced by a male talker). (Middle panel) Same spectrum raised to the power of 0.25. (Bottom panel) Same spectrum raised to the power of 1.5. All spectra are shown in linear units and have been normalized by their maximum for better visual clarity.

Hamming windows with 75% overlap between adjacent frames. This frame duration was chosen to be consistent with that used in our previous study (Hu and Loizou, 2008) which focused on evaluation of objective measures for predicting quality ratings. A tenth-order LPC analysis was used in the computation of the LPC-based objective measures (CEP, IS, and LLR).

D. Normalized covariance metric measures

From the various speech-based STI measures proposed (see review in Goldsworthy and Greenberg, 2004), we chose the normalized covariance metric (NCM) (Hollube and Kollmeier, 1996). This 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 (MTF), the NCM measure is based on the covariance between the probe (input) and response (output) envelope signals.

The NCM measure is computed as follows. The stimuli were first bandpass filtered into K bands spanning the signal bandwidth. The envelope of each band was computed using the Hilbert transform and then downsampled to 25 Hz, thereby limiting the envelope modulation frequencies to 0–12.5 Hz. Let $x_i(t)$ be the downsampled envelope in the ith band of the clean (probe) signal and let $y_i(t)$ be the downsampled envelope of the processed (response) signal. The normalized covariance in the ith frequency band is computed as

$$r_i = \frac{\Sigma_t (x_i(t) - \mu_i)(y_i(t) - \nu_i)}{\sqrt{\Sigma_t (x_i(t) - \mu_i)^2}\sqrt{\Sigma_t (y_i(t) - \nu_i)^2}}, \tag{8}$$

where μ_i and ν_i are the mean values of the $x_i(t)$ and $y_i(t)$ envelopes, respectively. Note that the r_i values are limited to $|r_i| \leq 1$. A value of r_i close to 1 would suggest that the input [i.e., $x_i(t)$] and processed [i.e., $y_i(t)$] signals are linearly related, while a value of r_i close to 0 would indicate that the input and processed signals are uncorrelated. The SNR in each band is computed as

$$\mathrm{SNR}_i = 10 \log_{10}\left(\frac{r_i^2}{1 - r_i^2}\right). \tag{9}$$

and subsequently limited to the range of $[-15, 15]$ dB (as done in the computation of the SII measure, ANSI, 1997). The TI in each band is computed by linearly mapping the SNR values between 0 and 1 using the following equation:

$$\mathrm{TI}_i = \frac{\mathrm{SNR}_i + 15}{30}. \tag{10}$$

Finally, the transmission indices are averaged across all frequency bands to produce the NCM index:

TABLE II. AI weights (ANSI, 1997) used in the implementation of the NCM measure for consonants and sentence materials.

Band	Center freq. (kHz)	Consonants	Sentences
1	0.3249	0.0346	0.0772
2	0.3775	0.0392	0.0955
3	0.4356	0.0406	0.1016
4	0.5000	0.0420	0.0908
5	0.5713	0.0433	0.0734
6	0.6502	0.0457	0.0659
7	0.7376	0.0472	0.0580
8	0.8344	0.0473	0.0500
9	0.9416	0.0471	0.0460
10	1.0602	0.0487	0.0440
11	1.1915	0.0519	0.0445
12	1.3370	0.0534	0.0482
13	1.4980	0.0562	0.0488
14	1.6763	0.0612	0.0488
15	1.8737	0.0684	0.0493
16	2.0922	0.0732	0.0491
17	2.3342	0.0748	0.0520
18	2.6022	0.0733	0.0549
19	2.8989	0.0685	0.0555
20	3.2274	0.0670	0.0514

$$\mathrm{NCM} = \frac{\Sigma_{i=1}^{K} W_i \times \mathrm{TI}_i}{\Sigma_{i=1}^{K} W_i}, \tag{11}$$

where W_i are the weights applied to each of the K bands. The denominator term is included for normalization purposes. The weights W_i are often called BIF in the computation of the SII measure (ANSI, 1997). Fixed weights (given in Table II), such as those used in AI studies, are often used in the computation of the STI measure (Steeneken and Houtgast, 1980). In our study, we consider making those weights signal and frequency (i.e., band) dependent. More precisely, we considered the following two weighting functions:

$$W_i^{(1)} = \left(\sum_t x_i^2(t) \right)^p, \tag{12}$$

$$W_i^{(2)} = \left(\sum_t (\max[x_i(t) - d_i(t), 0])^2 \right)^p, \tag{13}$$

where $d_i(t)$ denotes the (downsampled) scaled masker signal in the time domain. The power exponent p was varied from 0.12 to 1.5. The motivation behind the use of Eq. (12) is to place weight to each TI value in proportion to the signal energy in each band. The motivation behind the use of Eq. (13) is to place weight to each TI value in proportion to the excess masked signal.

To assess the influence of the SNR range used in the computation of the STI measure, we also considered limiting the SNR to the range of $[-15, 20]$, $[-15, 25]$, $[-15, 30]$, $[-15, 35]$, and $[-10, 35]$ dB. To accommodate for the new range in SNR values, the TI values in Eq. (10) were modified accordingly. So, for instance, to accommodate the $[-10, 35]$ dB range, the TI values in Eq. (10) were computed as follows:

$$\mathrm{TI}_i = \frac{\mathrm{SNR}_i + 10}{45}. \tag{14}$$

The above equation ensures that the SNR is linear mapped to values between 0 and 1.

The STI measure is typically evaluated for modulation frequencies spanning 0.63–12.5 Hz. To assess the influence of including higher modulation frequencies (>12.5 Hz), we also varied the modulation frequency range to 0–20 and 0–31 Hz. This was motivated by the study of Van Wijngaarden and Houtgast (2004) that showed that extending the modulation bandwidth to 31.5 Hz improved the correlation of the STI index for conversational-style speech.

The NCM computation in Eq. (11) takes into account a total of K bands spanning the signal bandwidth, which was 4 kHz in our study. To assess the contribution of low-frequency envelope information, spanning the range of 100–1000 Hz, we considered a variant of the above NCM measure in which we included only the low-frequency (<1000 Hz) bands in the computation. We refer to this measure as the low-frequency NCM measure and denote it as $\mathrm{NCM}_{\mathrm{LF}}$:

$$\mathrm{NCM}_{\mathrm{LF}} = \frac{\Sigma_{i=1}^{8} W_i \times \mathrm{TI}_i}{\Sigma_{i=1}^{8} W_i}. \tag{15}$$

Note that only the first eight low-frequency envelopes, spanning the frequency range of 100–1000 Hz, are used in the computation of the $\mathrm{NCM}_{\mathrm{LF}}$ measure. We considered using uniform weights for all frequency envelopes (i.e., $W_i = 1$ for all bands) as well as the weights given in Eq. (12). The $\mathrm{NCM}_{\mathrm{LF}}$ measure can be considered to be a simplified version of the NCM measure, much like the rapid STI (RASTI) measure is a simplified version of the STI measure. The RASTI measure is calculated using only the 500- and 2000-Hz octave bands (IEC 60268, 2003). In terms of prediction accuracy, the RASTI measure was found to produce comparable results to that obtained by the STI measure (Mapp, 2002; Larm and Hongisto, 2006).

E. AI-based measures

A simplified version of the SII measure is considered in this study that operates on a frame-by-frame basis. The proposed 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) does not account for spread of upward masking, and (c) does not require as input the long-term average spectrum (sound-pressure) levels of the speech and masker signals. The proposed AI-ST measure divides the signal into short (30 ms) data segments, computes the AI value for each segment, and averages the segmental AI values over all frames. More precisely, it is computed as follows:

$$\text{AI-ST} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\Sigma_{j=1}^{K} W(j,m) T(j,m)}{\Sigma_{j=1}^{K} W(j,m)}, \tag{16}$$

where M is the total number of data segments in the signal, $W(j,m)$ is the weight (i.e., band importance function, ANSI, 1997) placed on the jth frequency band, and

$$T(j,m) = \frac{\mathrm{SNR}(j,m) + 15}{30}, \tag{17}$$

$$\mathrm{SNR}(j,m) = 10 \log_{10} \frac{\hat{X}(j,m)^2}{D(j,m)^2}, \tag{18}$$

where $D(j,m)$ denotes the critical-band spectrum of the scaled masker signal (obtained before mixing) and $\hat{X}(j,m)$ denotes the enhanced signal's critical-band spectral magnitude in the jth band. Unlike the normalization used in the computation of the fwSNRseg measure [Eq. (5)], the excitation spectra were not normalized to have an area of unity. The SNR term in Eq. (18) was limited within the range of $[-15,15]$ dB and mapped linearly in each band to values between 0 and 1 using Eq. (17). For comparative purposes, we also considered limiting the SNR in Eq. (18) to $[-15,20]$, $[-15,25]$, $[-15,30]$, $[-15,35]$, and $[-10,35]$ dB.

Aside from using the AI weights for $W(j,m)$ (see Table I), the following four band-importance weighting functions were also considered for $W(j,m)$ in Eq. (16):

$$W_1(j,m) = \begin{cases} 1 & \text{if } X(j,m) > D(j,m) \\ 0 & \text{else,} \end{cases} \tag{19}$$

$$W_2(j,m) = \begin{cases} (X(j,m) - D(j,m))^p & \text{if } X(j,m) > D(j,m) \\ 0 & \text{else,} \end{cases} \tag{20}$$

$$W_3(j,m) = \begin{cases} X(j,m)^p & \text{if } X(j,m) > D(j,m), \\ 0 & \text{else} \end{cases} \tag{21}$$

$$W_4(j,m) = X(j,m)^p. \tag{22}$$

The motivation behind the use of the above BIFs [Eqs. (19)–(21)] was to include in the computation of the AI-ST measure only bands with positive SNR, i.e., only bands in which the target is stronger than the masker. The rather simplistic assumption made here is that bands with negative SNR contribute little, if anything, to intelligibility. As such, those bands should not be included in the computation of the AI-ST measure. The power exponent p in Eqs. (20)–(22) was varied from 0.5 to 4. As mentioned earlier, the value of p controls the emphasis or weight placed on spectral peaks and/or spectral valleys. Use of $p>1$, for instance, places more emphasis on the dominant spectral peaks (see example in Fig. 2).

Unlike the BIFs used in the traditional AI measure (ANSI, 1997) and in the extended (short-term) versions of the AI measure (Kates, 1987; Kates and Arehart, 2005; Rhebergen and Versfeld, 2005), the BIFs proposed in Eqs. (19)–(22) are signal and segment dependent. This was done to account for the fact that the AI-ST values are computed at a (short-duration) segmental level rather than on a global (long-term average spectrum) level. The speech-and masker-spectra vary markedly over time, and this variation is captured to some degree with the use of signal-dependent band-importance (weighting) functions.

F. Coherence-based measures

The magnitude-squared coherence (MSC) function is the normalized cross-spectral density of two signals and has been used to assess distortion in hearing aids (Kates, 1992). It is computed by dividing the input (clean) and output (processed) signals in a number (M) of overlapping windowed segments, computing the cross power spectrum for each segment using the FFT, and then averaging across all segments. For M data segments (frames), the MSC at frequency bin ω is given by

$$\mathrm{MSC}(\omega) = \frac{|\Sigma_{m=1}^{M} X_m(\omega) Y_m^*(\omega)|^2}{\Sigma_{m=1}^{M} |X_m(\omega)|^2 \Sigma_{m=1}^{M} |Y_m(\omega)|^2}, \tag{23}$$

where the asterisk denotes the complex conjugate and $X_m(\omega)$ and $Y_m(\omega)$ denote the FFT spectra of the $x(t)$ and $y(t)$ signals, respectively, computed in the mth data segment. In our case, $x(t)$ corresponds to the clean signal and $y(t)$ corresponds to the enhanced signal. The MSC measure takes values in the range of 0–1. The averaged, across all frequency bins, MSC was used in our study as the objective measure. The MSC was computed by segmenting the sentences using 30-ms duration Hamming windows with 75% overlap between adjacent frames. The use of a large frame overlap (>50%) was found by Carter *et al.* (1973) to reduce bias and variance in the estimate of the MSC.

It should be noted that the above MSC function can be expressed as a weighted MTF (see Appendix), which is used in the implementation of the STI measure (Houtgast and Steeneken, 1985). The main difference between the MTF (Houtgast and Steeneken, 1985) used in the computation of the STI measure and the MSC function is that the latter function is evaluated for all frequencies spanning the signal bandwidth, while the MTF is evaluated only for low modulation frequencies (0.5–16 Hz).

Extensions of the MSC measure were proposed by Kates and Arehart (2005) for assessing the effects of hearing-aid distortions (e.g., peak clipping) on speech intelligibility by normal-hearing and hearing-impaired subjects. More precisely, the new measure, called coherence SII (CSII), was proposed that used the SII index as the base measure and replaced the SNR term with the signal-to-distortion ratio term, which was computed using the coherence between the input and output signals. That is, the $\mathrm{SNR}(j,m)$ term in Eq. (18) was replaced with the following expression:

$$\mathrm{SNR}_{\mathrm{CSII}}(j,m) = 10 \log_{10} \frac{\Sigma_{k=1}^{N} G_j(\omega_k) \times \mathrm{MSC}(\omega_k) |Y_m(\omega_k)|^2}{\Sigma_{k=1}^{N} G_j(\omega_k) [1 - \mathrm{MSC}(\omega_k)] |Y_m(\omega_k)|^2}, \tag{24}$$

where $G_j(\omega)$ denotes the ro-ex filter (Moore and Glasberg, 1993) centered around the jth critical band, $\mathrm{MSC}(\omega)$ is given by Eq. (23), $Y(\omega_k)$ is the FFT spectrum of the enhanced signal, and N is the FFT size. The above SNR term is limited to $[-15,15]$ dB and mapped linearly between 0 and 1 using Eq. (17) to produce a new $T_{\mathrm{CSII}}(j,m)$ term. Finally, the latter term is substituted in Eq. (16) to compute the CSII value as follows:

$$\mathrm{CSII} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)T_{\mathrm{CSII}}(j,m)}{\sum_{j=1}^{K} W(j,m)}. \quad (25)$$

The above CSII measure is computed using all M speech segments of the utterance. Kates and Arehart (2005) found that a three-level version of the CSII measure yielded higher correlation with speech intelligibility than the above CSII measure. The three measures were computed by first dividing the M speech segments into three level regions and computing separately the CSII measure for each region. The high-level region consisted of segments at or above the overall root-mean-square (rms) level of the whole utterance. The mid-level region consisted of segments ranging from the overall rms level to 10 dB below, and the low-level region consisted of segments ranging from rms–10 dB to rms–30 dB. 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. A linear combination of the three CSII values followed by a logistic function transformation was subsequently used to model the intelligibility scores. The resulting intelligibility measure, termed I3 (Kates and Arehart, 2005), will be evaluated and compared against other measures in the present study. The I3 measure was later extended by Arehart *et al.* (2007) to model judgments of quality ratings of noise and hearing-aid type of distortions. A new measure, termed Q3, was developed based on a different linear combination of the three-level CSII measures (Arehart *et al.*, 2007).

The critical-band spacing was used in the implementation of the above CSII measures (Kates and Arehart, 2005). A total of 16 critical bands spanning the bandwidth of 100–3700 Hz were used in our implementation. The BIF given in Table B.1 (ANSI, 1997) were used in Eq. (25) for $W(j,m)$. In addition, the four band-importance weighting functions proposed in Eqs. (19)–(22) were tested.

IV. RESULTS

Two figures of merit were used to assess the performance of the above objective measures in terms of predicting speech intelligibility. The first figure of merit was Pearson's correlation coefficient, r, and the second figure of merit was an estimate of the standard deviation of the error computed as $\sigma_e = \sigma_d\sqrt{1-r^2}$, where σ_d is the standard deviation of the speech recognition scores in a given condition, and σ_e is the computed standard deviation of the error. A smaller value of σ_e indicates that the objective measure is better at predicting speech intelligibility.

The average intelligibility scores obtained by normal-hearing listeners in the 72 different noisy conditions (see Sec. II) were subjected to correlation analysis with the corresponding mean values obtained with the objective measures. As mentioned earlier, these conditions involved noise-suppressed speech (consonants and sentences) originally corrupted by four different maskers (car, babble, train, and street interferences) at two different SNR levels. The computed correlation coefficients (and prediction error) are tabulated separately for the consonants and sentence materials and are given in Tables III and IV, respectively.

TABLE III. Correlation coefficients, r, and standard deviations of the error, σ_e, between consonant recognition scores and the various objective measures examined. The BIFs used in some measures are indicated in the second column. In the implementation of the fwSNRseg, NCM, CSII, and AI-ST measures the SNR was restricted in the range of [−15, 15] dB.

Objective measure	Band-importance function	r	σ_e
PESQ		0.77	0.08
LLR		−0.51	0.10
SNRseg		0.40	0.12
WSS		−0.33	0.11
Itakura–Saito (IS)		−0.35	0.12
Cepstrum (CEP)		−0.48	0.11
Coherence (MSC)		0.76	0.08
CSII	ANSI (1997)	0.76	0.08
$\mathrm{CSII}_{\mathrm{high}}$	ANSI (1997)	0.80	0.07
$\mathrm{CSII}_{\mathrm{mid}}$	ANSI (1997)	0.80	0.07
$\mathrm{CSII}_{\mathrm{low}}$	ANSI (1997)	0.36	0.12
I3		0.80	0.07
Q3		0.79	0.07
mI3		0.82	0.07
CSII	W_4, p=0.5, Eq. (22)	0.77	0.08
$\mathrm{CSII}_{\mathrm{high}}$	W_4, p=0.5, Eq. (22)	0.80	0.07
$\mathrm{CSII}_{\mathrm{mid}}$	W_4, p=0.5, Eq. (22)	0.78	0.08
$\mathrm{CSII}_{\mathrm{low}}$	W_4, p=4, Eq. (22)	0.68	0.09
fwSNRseg	ANSI (Table I)	0.59	0.10
fwSNRseg	Eq. (6), p=4	0.68	0.09
$\mathrm{NCM}_{\mathrm{LF}}$	W_i=1	0.65	0.09
$\mathrm{NCM}_{\mathrm{LF}}$	$W_i^{(1)}$, p=1, Eq. (12)	0.72	0.09
NCM	ANSI (Table II)	0.66	0.09
NCM	$W_i^{(1)}$, p=0.5, Eq. (12)	0.77	0.08
NCM	$W_i^{(2)}$, p=1, Eq. (13)	0.72	0.09
AI-ST	ANSI (Table I)	0.39	0.11
AI-ST	W_1, Eq. (19)	0.56	0.10
AI-ST	W_2, p=4, Eq. (20)	0.68	0.09
AI-ST	W_3, p=4, Eq. (21)	0.67	0.09
AI-ST	W_4, p=4, Eq. (22)	0.52	0.11

A. Subjective quality measures

Of the seven measures designed for subjective quality assessment, the PESQ and fwSNRseg measures performed the best. When applied to the sentence materials, the fwSNRseg measure, based on the weighting function given in Eq. (6), performed better than the PESQ measure and yielded a correlation of r=0.81, compared to r=0.79 obtained with the PESQ measure. When applied to the consonant materials, the PESQ measure performed better than the fwSNRseg measure. The LLR measure, which was found in Hu and Loizou (2008) to yield a correlation coefficient that was nearly as good as that of the PESQ measure, performed comparatively worse than the PESQ measure. The MSC, which has been used to assess hearing-aid distortion, performed modestly well (r=0.71–0.77) for both sentence and consonant materials. We believe that the modest performance of the MSC measure can be attributed to the fact that the MSC function can be expressed as a weighted MTF (see Appendix), which is used in the implementation of the STI measure. Higher correlation (r=0.79–0.88) was obtained with the coherence-based Q3 measure, which was used by Arehart *et al.* (2007) for modeling subjective quality judgments of

TABLE IV. Correlation coefficients, r, and standard deviations of the error, σ_e, between sentence recognition scores and the various objective measures examined. The BIFs used in some measures are indicated in the second column. In the implementation of the fwSNRseg, NCM, CSII, and AI-ST measures the SNR was restricted in the range of $[-15, 15]$ dB.

Objective measure	Band-importance function	r	σ_e
PESQ		0.79	0.11
LLR		−0.56	0.15
SNRseg		−0.46	0.15
WSS		−0.27	0.17
Itakura–Saito (IS)		−0.22	0.17
Cepstrum (CEP)		−0.49	0.15
Coherence (MSC)		0.71	0.12
CSII		0.82	0.10
$CSII_{high}$		0.85	0.09
$CSII_{mid}$		0.91	0.07
$CSII_{low}$		0.86	0.09
I3		0.92	0.07
Q3		0.88	0.08
mI3		0.92	0.07
CSII	W_4, $p=4$, Eq. (22)	0.86	0.09
$CSII_{high}$	W_4, $p=2$, Eq. (22)	0.88	0.08
$CSII_{mid}$	W_4, $p=1$, Eq. (22)	0.94	0.06
$CSII_{low}$	W_4, $p=0.5$, Eq. (22)	0.86	0.09
fwSNRseg	ANSI (Table I)	0.78	0.11
fwSNRseg	Eq. (6), $p=1$	0.81	0.10
NCM_{LF}	$W_i=1$	0.81	0.10
NCM_{LF}	$W_i^{(1)}$, $p=2$, Eq. (12)	0.87	0.09
NCM	ANSI (Table II)	0.82	0.10
NCM	$W_i^{(1)}$, $p=1.5$, Eq. (12)	0.89	0.07
NCM	$W_i^{(2)}$, $p=1.5$, Eq. (13)	0.89	0.08
AI-ST	ANSI (Table I)	0.33	0.16
AI-ST	W_1, Eq. (19)	0.66	0.13
AI-ST	W_2, $p=3$, Eq. (20)	0.80	0.11
AI-ST	W_3, $p=3$, Eq. (21)	0.80	0.11
AI-ST	W_4, $p=4$, Eq. (22)	0.62	0.14

hearing-aid distortion. In summary, of all the measures tested previously (Hu and Loizou, 2008) for subjective quality predictions, the fwSNRseg and PESQ measures seem to predict modestly well both speech quality and speech intelligibility.

B. Intelligibility measures

Of all the intelligibility measures considered, the coherence-based (CSII) and NCM measures performed the best. The highest correlations were obtained with the CSII measures for both consonants and sentence materials. The I3 measure (Kates and Arehart, 2005), in particular, produced the highest correlation for consonants ($r=0.80$) and sentence ($r=0.92$) materials. Figure 3 shows the scatter plot of the predicted I3 scores against the listeners' recognition scores for consonants and sentences. Figures 4 and 5 show the individual scatter plots broken down by noise type for consonant and sentence recognition, respectively. As can be seen, a high correlation was maintained for all noise types, including modulated (e.g., train) and non-modulated (e.g., car) maskers. The correlations with consonant recognition scores ranged from $r=0.82$ with street noise to $r=0.85$ with car

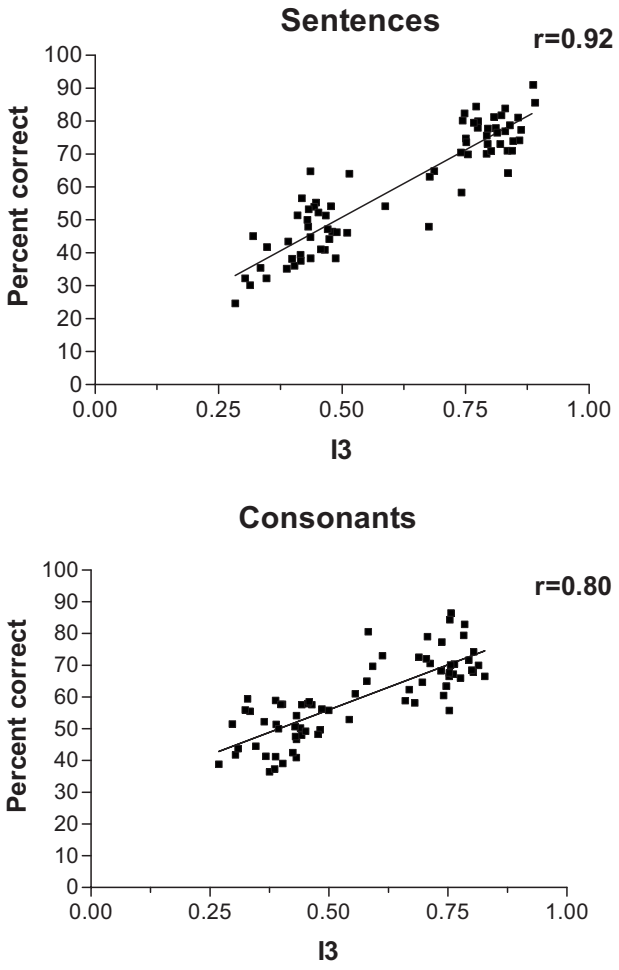

FIG. 3. Scatter plot of sentence recognition scores (top panel) and consonant recognition scores (bottom panel) against the predicted I3 values.

noise. The correlations with sentence recognition scores ranged from $r=0.88$ with train noise to $r=0.98$ with babble.

Among the three-level CSII measures, the mid-level CSII ($CSII_{mid}$) measure yielded the highest correlation for both consonant and sentence materials, consistent with the outcome reported by Kates and Arehart (2005). The $CSII_{mid}$ measure captures information about envelope transients and spectral transitions, critical for the transmission of information regarding place of articulation. Similar to the approach taken in Kates and Arehart (2005), a multiple-regression analysis was run on the three CSII measures, yielding the following predictive models for consonant and sentence intelligibility. For consonants, the modified I3 measure, indicated as mI3, is given by

$$m\text{I3} = 0.026 - 1.033 \times \text{CSII}_{low} + 0.822 \times \text{CSII}_{mid} + 0.506 \times \text{CSII}_{high}, \quad (26)$$

and for sentences, it is given by

$$m\text{I3} = -0.029 - 0.055 \times \text{CSII}_{low} + 2.206 \times \text{CSII}_{mid} - 0.349 \times \text{CSII}_{high}. \quad (27)$$

Subsequent logistic transformations of the mI3 measure did not improve the correlations. The correlations of the above mI3 measures with consonant and sentence recognition

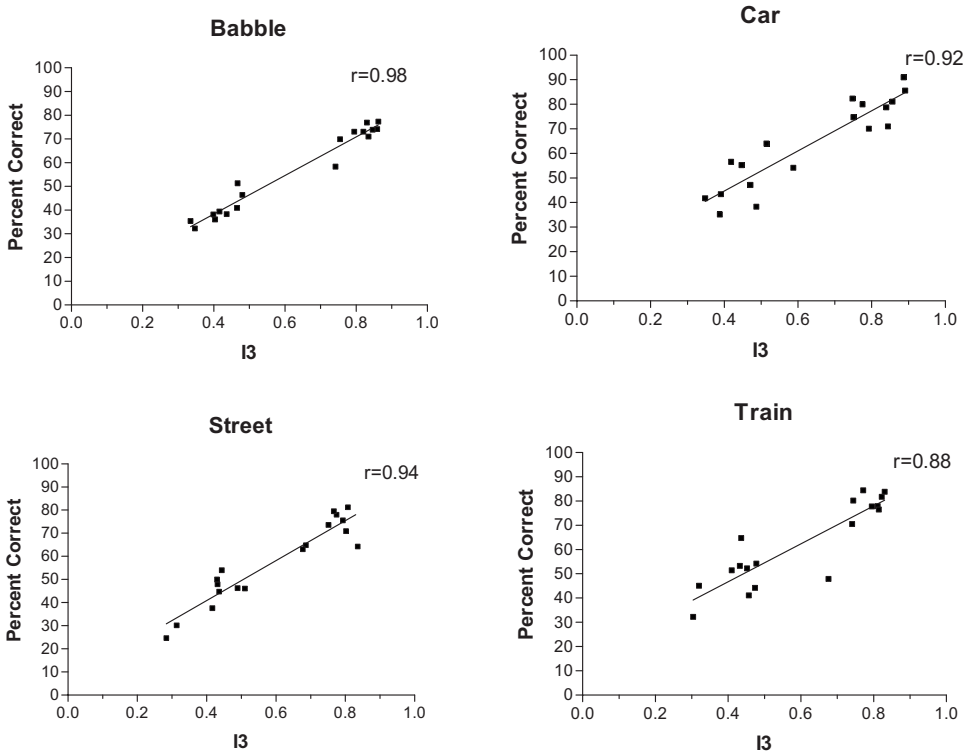

FIG. 4. Individual scatter plots of predicted I3 values against sentence recognition scores for the four types of maskers used.

scores are given in Tables III and IV, respectively. The *m*I3 measure, via the use of Eq. (26), improved the I3 correlation from 0.80 to 0.82, making it the highest correlation attained for consonants. For sentences, the improvement in performance, over that attained by the I3 measure, was marginal and not evident in Table IV due to the rounding of the correlation values to two decimal places. Further improvements in correlation were obtained with the three-level CSII measures for the sentence materials after applying the proposed signal- and phonetic-segment dependent band-importance functions given in Eq. (22). The correlation of the modified $CSII_{mid}$ measure improved from $r=0.92$ (7% prediction error) with ANSI (1997) weights to $r=0.94$ (6% prediction error) with the proposed BIF given in Eq. (22). The resulting correlation was higher than that attained with the I3 measure proposed by Kates and Arehart (2005), and it was the highest correlation obtained in the present study.

The next highest correlations were obtained with the modified NCM measure that used the BIF in Eq. (12). The resulting correlation coefficient for sentences was $r=0.89$

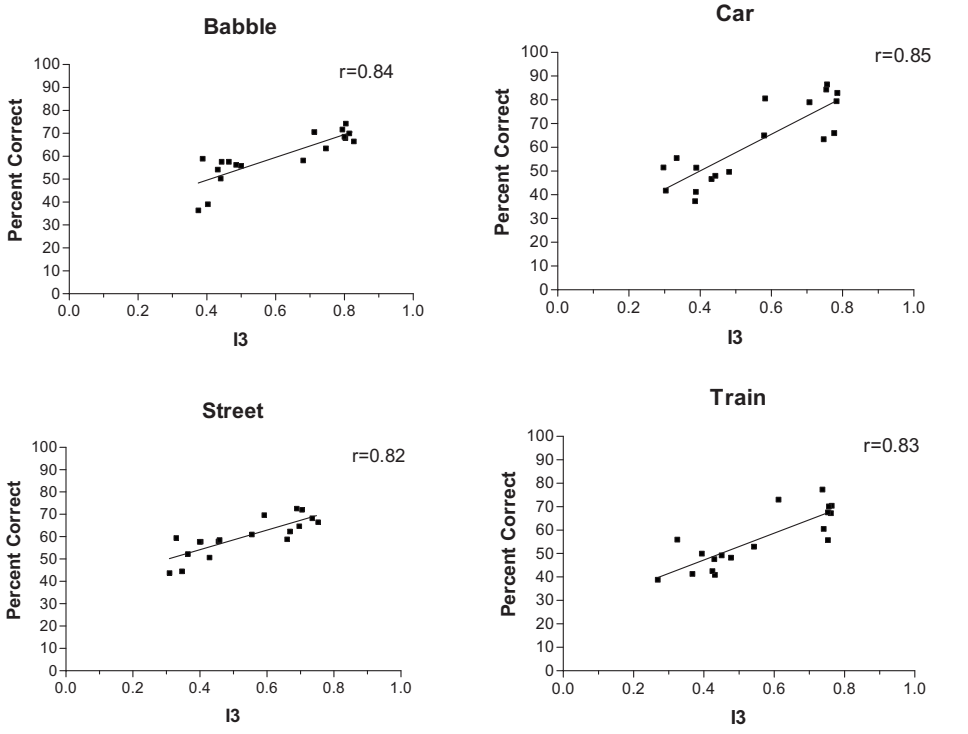

FIG. 5. Individual scatter plots of predicted I3 values against consonant recognition scores for the four types of maskers used.

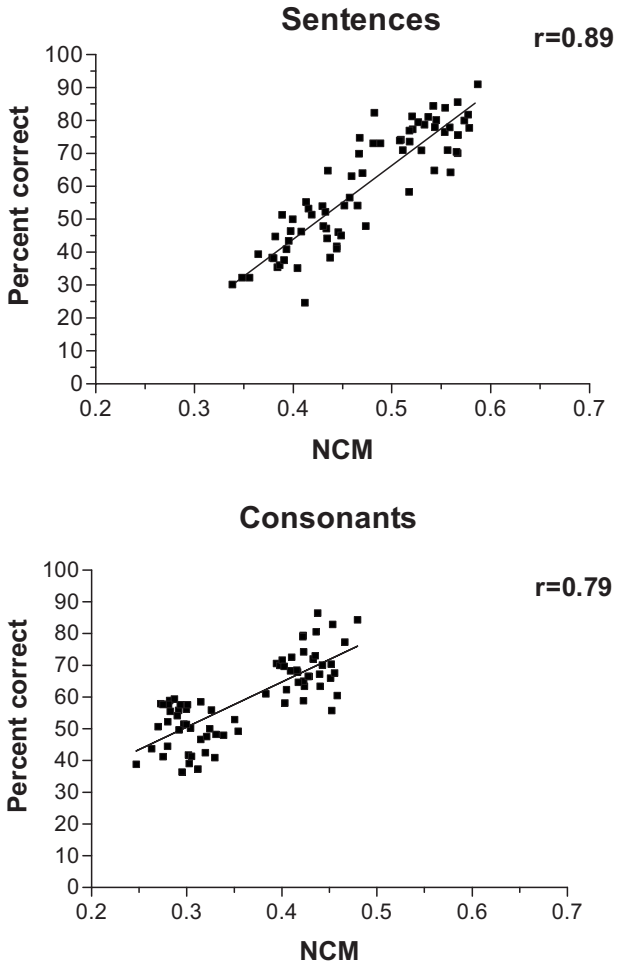

FIG. 6. Scatter plot of sentence recognition scores (top panel) and consonant recognition scores (bottom panel) against the predicted NCM values. In the implementation of the NCM metric, the SNR range was restricted to $[-10,35]$ dB and the BIF was set to that given in Eq. (12) with $p=1.5$ for the sentence materials and $p=0.25$ for the consonant materials.

(7% prediction error) and for consonants it was $r=0.79$ (8% error) when the $[-10,35]$ dB SNR range was used. Figure 6 shows the scatter plot of the predicted NCM scores against the listeners' speech recognition scores. Figures 7 and 8 show the individual scatter plots broken down by noise type for consonant and sentence recognition, respectively. A high correlation was maintained for all noise types, including modulated (e.g., train) and non-modulated (e.g., car) maskers. The correlations obtained with consonant recognition scores ranged from $r=0.75$ with babble to $r=0.89$ with train noise. The correlations obtained with sentence recognition scores ranged from $r=0.85$ with car noise to $r=0.94$ with babble.

As shown in Tables III and IV, performance was clearly influenced by the choice of the band-importance function. In all cases, the lowest correlation was obtained when the AI weights, taken from the ANSI (1997) standard, were used. This clearly demonstrates that the BIFs are material dependent, something that is already accounted for in the ANSI (1997) standard. Different sets of weights are provided for different speech materials (see Table B.1, ANSI, 1997). Complex procedures followed by lengthy experiments are needed to obtain the BIFs tabulated in the ANSI (1997) standard. In contrast, the proposed weighting functions, given in Eqs. (19)–(22), suggest an alternative and easier way for deriving the BIFs.

In the implementation of the NCM measure, we fixed the number of bands to 20, the speech dynamic range to $[-15,15]$ dB, and the range of modulation frequencies to 0–12.5 Hz. Additional experiments were run to assess the influence of the number of bands, range of modulation frequencies, and speech dynamic range on the prediction of speech intelligibility in noise. Note that the conventional STI measure uses seven 1/3-octave bands (Houtgast and

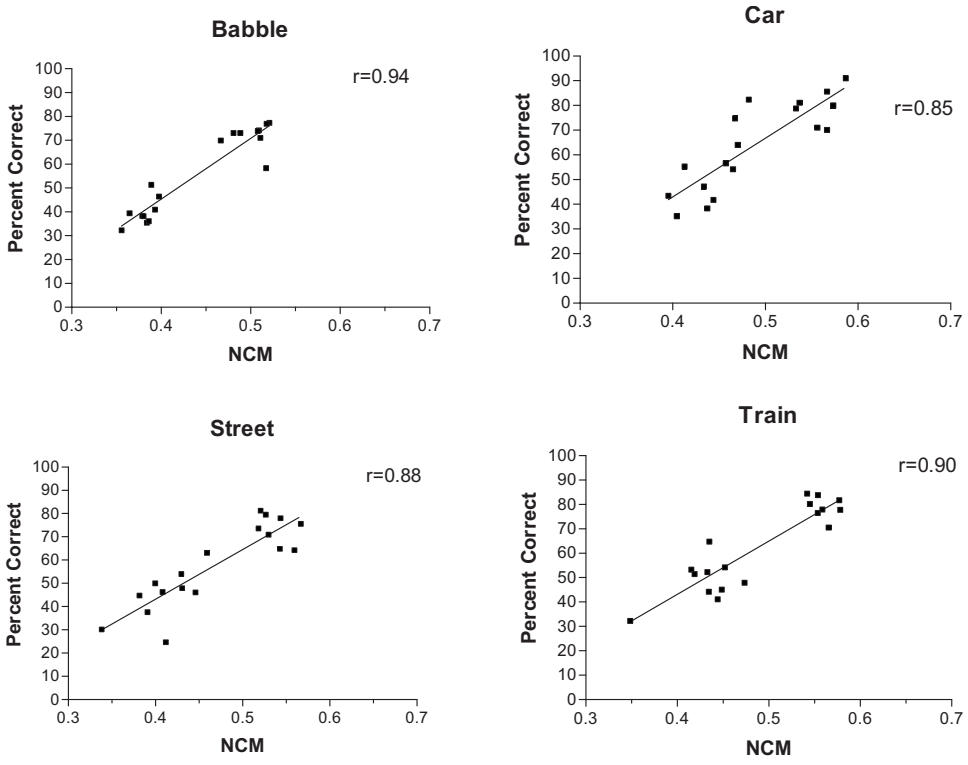

FIG. 7. Individual scatter plots of predicted NCM values against sentence recognition scores for the four types of maskers used.

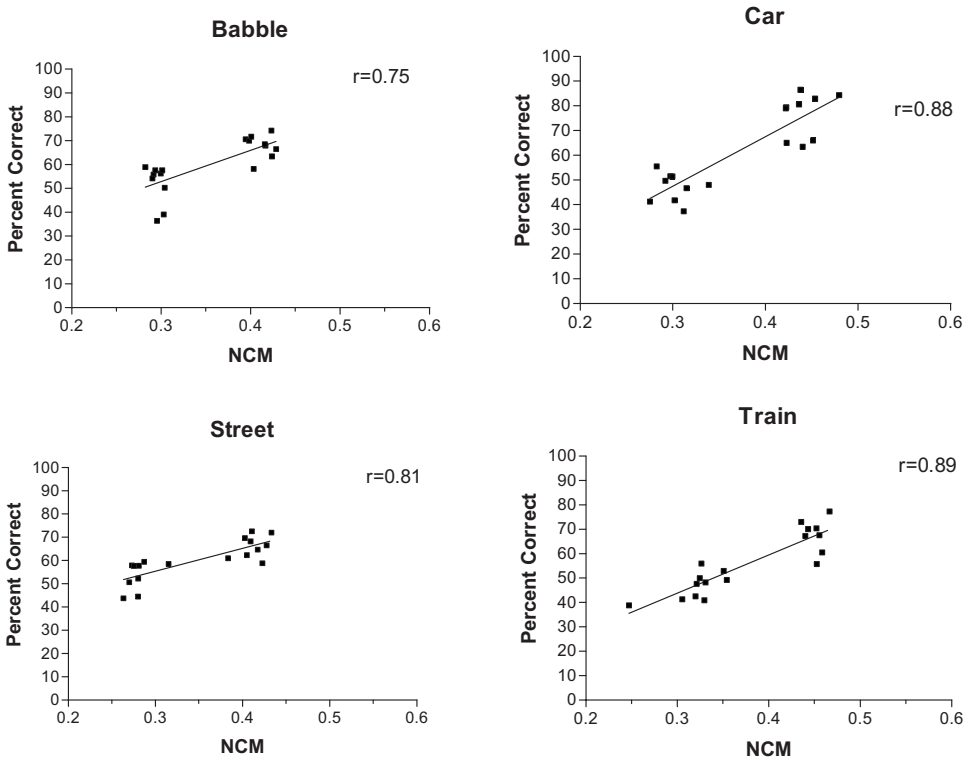

FIG. 8. Individual scatter plots of predicted NCM values against consonant recognition scores for the four types of maskers used.

Steeneken, 1985). To assess the influence of the number of bands on the computation of the NCM measure, we varied the number of bands from 7 to 20. The band center frequencies were logarithmically spaced in the 300–3400 Hz bandwidth. The weighting function given in Eq. (12) with $p=1.5$ was used in all conditions. The resulting correlation coefficients are given in Table V. As can be seen, there is a small, but non-significant, improvement in the correlation as the number of bands increases. Hence, the number of bands used in the computation of the NCM measure does not influence significantly its prediction power.

The implementation of the STI measure typically uses a set of 14 modulation frequencies ranging from 0.63 to 12.5 Hz (Houtgast and Steeneken, 1985). To further assess whether including higher (>12.5 Hz) modulation frequencies would improve the correlation of the NCM measure, we tested two additional implementations that included modulation frequencies up to 20 Hz and up to 31 Hz. The results obtained for different SNR ranges and different ranges of modulation frequencies are tabulated in Table VI. As can be seen there is no improvement for sentences, but a small improvement for consonants. The small improvement obtained with consonants might reflect a difference in the speaking style between the production of consonants vs. sentences (van Wijngaarden and Houtgast, 2004) used in the present study. The sentences used in the present study (taken from Loizou, 2007) were produced with a clear, rather than conversational, speaking style.

TABLE V. Correlation coefficients, r, and standard deviations of the error, σ_e, between sentence recognition scores and the NCM measure as a function of the number of bands used.

No. of bands	r	σ_e
7	0.88	0.08
12	0.88	0.08
16	0.89	0.08
20	0.89	0.08

The correlations obtained with the NCM measure after varying the SNR dynamic range from [−15,15] to [−15,35] dB are shown in Table VII. Performance improved on the consonant recognition task. The correlation coefficient, for instance, obtained with the NCM measure improved from 0.77 to 0.79 when the speech dynamic range increased from 30 to 45 dB. No improvement was noted for the sentence recognition task, at least for the indicated band-importance function. Table VIII shows in more detail the correlations obtained with other band-importance functions and with the SNR dynamic range set to [−10,35] dB. Overall, correlations improved for both consonants and sentences when a wider dynamic range was used.

The performance obtained with the AI-ST measure was quite poor ($r=0.39$ for consonants and $r=0.33$ for sentences) when the AI AI weights were used, but improved considerably when the proposed BIFs were used ($r=0.68$ for consonants and $r=0.80$ for sentences). Compared to the SII implementation (ANSI, 1997) which incorporates upward-spread of masking effects, the AI-ST implementation is rather simplistic. In addition, the averaging of the individual frame AI-ST values in Eq. (16) implicitly assumes that all short (phonetic) segments should be weighted uniformly, i.e., that equal emphasis should be placed on consonant segments, steady-state vowels, and/or vowel-consonant transitions. Furthermore, it is assumed that the same weighting function should be applied to vowels and consonants. Further work is thus needed to develop weighting functions specific to consonants and vowels.

TABLE VI. Correlation coefficients obtained with the NCM measure for different modulation bandwidths and different SNR dynamic ranges. For the sentence materials, the W_1 band-importance function was used with $p=1.5$ and for the consonant materials the W_1 function was used with $p=0.25$.

Material	Modulation bandwidth (Hz)	SNR dynamic range (dB)			
		[−15,20]	[−15,25]	[−15,30]	[−15,35]
Sentences	20.0	0.89	0.89	0.89	0.89
	31.5	0.88	0.88	0.88	0.88
Consonants	20.0	0.74	0.74	0.74	0.74
	31.5	0.77	0.77	0.77	0.77

In the computation of the SII index, the time interval over which the noise and signal are integrated is 125 ms (ANSI, 1997). Within this integration time, the distribution of the speech rms values is approximately linear within a 30 dB dynamic range (Dunn and White, 1940), which is the range adopted for the computation of the SII and STI measures. Several studies have argued, however, that this estimate of speech dynamic range is conservative (e.g., Boothroyd *et al.*, 1994; Studebaker and Sherbecoe, 2002). Studebaker and Sherbecoe (2002), for instance, reported that the dynamic range of BIFs (derived for monosyllabic words) ranged from 36 to 44 dB, with an average value of about 40 dB. Hence, we considered varying the speech dynamic range for both the AI-based and fwSNRseg measures. The resulting correlation coefficients obtained with the wider dynamic range are given in Table VII. As can be seen, the larger dynamic range seemed to influence the performance of the AI-ST measure, but not the fwSNRseg and NCM measures.

Unlike the SII standard (ANSI, 1997) which uses a 125-ms integration window, a 30-ms integration window was used in our present study for the implementation of the AI-ST measure. To assess the influence of window duration, we varied the window duration from 30 to 125 ms. The resulting correlation coefficients are tabulated in Tables IX and X for consonants and sentences, respectively. As can be seen from these tables, performance was influenced by both the weighting function and window duration used. Small improvements were obtained in the prediction of consonant recognition when the window duration increased (Table IX), and considerably larger improvements were obtained in the prediction of sentence recognition (Table X).

V. DISCUSSION

The PESQ measure, which was originally designed to predict quality of speech transmitted over IP networks (ITU-T, 2000), performed modestly well ($r=0.77-0.79$) on predicting intelligibility of consonants and sentences in noise. This was surprising at first, given that this measure assesses overall loudness differences between the input (clean) and processed speech signals, and as such it is more appropriate for predicting subjective quality ratings (Bladon and Lindblom, 1981) than intelligibility. The PESQ measure has been shown in Hu and Loizou (2007) to correlate well ($r=0.81$) with subjective ratings of speech distortion introduced by noise-suppression algorithms. Hence, on this regard it is reasonable to expect that a measure that assesses accurately speech distortion (and overall quality) should also be suitable for assessing speech intelligibility. This is based on the premise (and expectation) that the distortion often introduced by noise-suppression algorithms (e.g., spectral attenuation near formant regions) and imparted on the speech signal, should degrade speech intelligibility. Indeed, the intelligibility study by Hu and Loizou (2007) showed that some noise-suppression algorithms may degrade speech intelligibility in noisy conditions.

Among all objective measures examined in the present study, the modified CSII and NCM measures incorporating

TABLE VII. Correlation coefficients obtained for different SNR dynamic ranges. The band-importance function (BIF) used is given in the third column.

Material	Objective measure	BIF	SNR dynamic range (dB)					
			[−15,15]	[−15,20]	[−15,25]	[−15,30]	[−15,35]	[−15,35]
Sentences	fwSNRseg	$p=2$, Eq. (6)	0.81	0.79	0.78	0.77	0.77	0.80
	NCM	W_1, $p=1.5$, Eq. (12)	0.89	0.89	0.89	0.89	0.89	0.89
	AI-ST	W_2, $p=3$, Eq. (20)	0.80	0.81	0.82	0.83	0.83	0.83
Consonants	fwSNRseg	$p=2.5$, Eq. (6)	0.68	0.65	0.65	0.64	0.64	0.64
	NCM	W_1	0.77	0.78	0.78	0.78	0.78	0.79
	AI-ST	W_3, $p=4$, Eq. (21)	0.67	0.68	0.69	0.69	0.69	0.69

TABLE VIII. Correlation coefficients, r, and standard deviations of the error, σ_e, between sentence/consonant recognition scores and the various objective measures examined. The BIF are given in the third column. The SNR was restricted in the range of [−10,35] dB.

Material	Objective measure	Band-importance function	r	σ_e
Consonants	fwSNRseg	ANSI (Table I)	0.60	0.10
	fwSNRseg	Eq. (6), p=2.5	0.64	0.09
	NCM_{LF}	W_i=1	0.69	0.09
	NCM_{LF}	$W_i^{(1)}$, p=2, Eq. (12)	0.74	0.08
	NCM	ANSI (Table II)	0.73	0.08
	NCM	$W_i^{(1)}$, p=0.25, Eq. (12)	0.79	0.08
	NCM	$W_i^{(2)}$, p=0.25, Eq. (13)	0.76	0.08
	AI-ST	ANSI (Table I)	0.42	0.11
	AI-ST	W_1, Eq. (19)	0.57	0.10
	AI-ST	W_2, p=4, Eq. (20)	0.69	0.09
	AI-ST	W_3, p=4, Eq. (21)	0.68	0.09
	AI-ST	W_4, p=4, Eq. (22)	0.62	0.10
Sentences	fwSNRseg	ANSI (Table I)	0.78	0.11
	fwSNRseg	Eq. (6), p=2	0.80	0.10
	NCM_{LF}	W_i=1	0.81	0.10
	NCM_{LF}	$W_i^{(1)}$, p=1.5, Eq. (12)	0.87	0.09
	NCM	ANSI (Table II)	0.84	0.09
	NCM	$W_i^{(1)}$, p=1.5, Eq. (12)	0.89	0.08
	NCM	$W_i^{(2)}$, p=0.25, Eq. (13)	0.86	0.08
	AI-ST	ANSI (Table I)	0.43	0.16
	AI-ST	W_1, Eq. (19)	0.66	0.13
	AI-ST	W_2, p=3, Eq. (20)	0.83	0.10
	AI-ST	W_3, p=3, Eq. (21)	0.83	0.10
	AI-ST	W_4, p=4, Eq. (22)	0.73	0.11

signal-specific weighting information have been found to perform the best in terms of predicting speech intelligibility in noise. The CSII measures have been found previously to correlate highly with both speech intelligibility (Kates and Arehart, 2005) and speech quality (Arehart *et al.*, 2007), at least for sentence materials subjected to hearing-aid type of distortions (e.g., clipping). On this regard, the present study extends the utility of the CSII measures for the prediction of the intelligibility of noise-suppressed speech. The proposed band-importance functions [Eq. (22)] had a big influence on the performance of the modified CSII measures, particularly for the prediction of sentence intelligibility scores. The correlation coefficient of the $CSII_{mid}$ measure with sentence recognition scores, in particular, improved from r=0.92 to r=0.94 after using the proposed BIF given in Eq. (22). Similar improvement was noted for consonants, but only for the $CSII_{low}$ measure. The lack of improvement for the $CSII_{mid}$ measure can be attributed to the non-uniform, and perhaps skewed, distribution of segments falling in the three regions, at least for the consonant materials used in this study (note that for sentences, a roughly equal number of segments fall in the three regions). Only a small percentage (<16%) of segments were found to be classified as mid-level, suggesting that perhaps different regions need to be considered for consonants. Further work is thus warranted to optimize the selection of regions for isolated vowel-consonant–vowel syllables.

High performance was expected of the NCM measure as it belongs to the speech-based STI measures, which have been shown in many studies to correlate highly with the intelligibility of nonsense syllables (e.g., Steeneken and Houtgast, 1982; Houtgast and Steeneken, 1985). The speech-based STI measures (Goldsworthy and Greenberg, 2004) generally assess the amount of reduction in temporal-envelope modulations incurred when the input signal goes through a sound transmission system. In our case, the NCM measure [Eq. (8)] assesses the fraction of the processed envelope signal that is linearly dependent on the input (clean) envelope signal at each frequency band. This measure accounts for the average envelope power in each band as well as for the low-frequency (<12.5 Hz) envelope modulations, which are known to carry critically important information about speech (e.g., Drullman *et al.*, 1994a, 1994b; Arai *et al.*, 1996). Compared to the conventional NCM measure (Hollube and Kollmeier, 1996) which uses fixed (for all speech stimuli) weights, the modified NCM measure uses signal-dependent weighting functions and performed substantially better. Overall, the proposed BIFs [Eqs. (12) and (13)] had a big influence on the performance of the modified NCM measure. The correlation coefficient obtained with the consonant materials improved from 0.66 when fixed ANSI (1997) weights were used to 0.77 when the signal-dependent weighting function given in Eq. (12) was used (Table III). Similar improvements were also noted on the sentence recognition task (Table IV). Aside from the use of the proposed BIFs, the use of a wider speech dynamic range (45 dB) improved slightly the performance of the NCM measure (see

TABLE IX. Correlation coefficients between consonant recognition scores and the AI-ST measure as a function of window duration (in milliseconds), SNR range, and BIF.

SNR range	Band-importance function	Window duration			
		30 ms	60 ms	100 ms	125 ms
[−15,15] dB	W_1, Eq. (19)	0.56	0.56	0.59	0.59
	W_2, p=1, Eq. (20)	064	0.64	0.66	0.68
	W_3, p=2, Eq. (21)	0.65	0.64	0.66	0.66
	W_4, p=2, Eq. (22)	0.51	0.53	0.56	0.59
[−10,35] dB	W_1, Eq. (19)	0.57	0.55	0.57	0.58
	W_2, p=1, Eq. (20)	0.66	0.65	0.66	0.67
	W_3, p=2, Eq. (21)	0.67	0.65	0.67	0.67
	W_4, p=2, Eq. (22)	0.60	0.61	0.63	0.64

TABLE X. Correlation coefficients between sentence recognition scores and the AI-ST measure as a function of window duration (in milliseconds), SNR dynamic range, and BIF.

SNR range	Band-importance function	Window duration			
		30 ms	60 ms	100 ms	125 ms
[−15, 15] dB	W_1, Eq. (19)	0.66	0.71	0.75	0.76
	W_2, $p=1$, Eq. (20)	0.77	0.81	0.84	0.85
	W_3, $p=2$, Eq. (21)	0.79	0.82	0.85	0.86
	W_4, $p=2$, Eq. (22)	0.67	0.68	0.71	0.73
[−10, 35] dB	W_1, Eq. (19)	0.66	0.71	0.74	0.75
	W_2, $p=1$, Eq. (20)	0.80	0.83	0.85	0.86
	W_3, $p=2$, Eq. (21)	0.82	0.84	0.86	0.86
	W_4, $p=2$, Eq. (22)	0.71	0.73	0.75	0.77

Table VII). However, neither the use of a wider range of modulation frequencies (see Table VI) nor the use of smaller number of channels (see Table V) influenced significantly the performance of the NCM measure. The power exponent (p) used in the BIFs can be clearly optimized for different speech materials, but only a slight dependence on the specific value of the power exponent was observed (see Table XI), at least for the NCM measure.

The performance of the proposed low-frequency (100–1000 Hz) version of the NCM measure [see Eq. (15)] was comparable to that of the NCM measure. This suggests that the low-frequency region of the spectrum carries critically important information about speech. The low-frequency region of the spectrum is known to carry F1 and voicing information, which in turn provides listeners with access to low-frequency acoustic landmarks of the signal (Li and Loizou, 2008). These landmarks, often blurred in noisy conditions, are critically important for understanding speech in noise as it aids listeners to better determine syllable structure and word boundaries (Stevens, 2002; Li and Loizou, 2008).

The performance of the AI-ST measure was modest and comparable to that obtained with the PESQ measure. Higher performance was expected with the AI-ST measure, at least for predicting consonant recognition in noise, given the success of the AI index in predicting the intelligibility of nonsense syllables (e.g., Kryter, 1962b). Our implementation, however, was rather simplistic as it did not incorporate upward spread of masking or any other non-linear auditory effects modeled in the ANSI (1997) standard. Furthermore, the AI-ST measure operates at a short, segmental (phonetic) level, while the SII measure operates on the average long-term spectra of the target and masker signals. Operating at a short-term (segmental) level was found necessary in the present study in order to capture the changing temporal/spectral characteristics of fluctuating maskers (e.g., train), but it imposes some limitations on the AI-ST measure that are difficult to overcome. For one, the segmental AI-ST values were averaged over all segments to produce one value. In doing so, it is implicitly assumed that all short (phonetic) segments should be weighted uniformly, i.e., that equal emphasis should be placed on consonant segments, steady-state vowels, and/or vowel-consonant transitions. Since our knowledge is limited as to how normal-hearing listeners integrate over time vowel and consonant information for sentence recognition, one can consider devising separate BIFs that are more appropriate for vowels and consonants. A better temporal weighting function, perhaps one derived psychoacoustically and incorporating forward/backward masking effects (e.g., Rhebergen *et al.*, 2006), might be needed to improve further the performance of the AI-ST measure.

The performance of the AI-ST measure on the prediction of sentence intelligibility in noise was higher than that on consonant intelligibility. This was surprising since the AI-ST measure as well as the other measures examined in this study do not model contextual or any other high-level (involving central processes) effects, which are known to play a significant role on sentence recognition. We speculate that this was accomplished, or perhaps compensated, by the use of signal-dependent BIFs. In the absence of those functions, the performance of the AI-ST measure on the sentence recognition task was found to be poor ($r<0.4$).

The data shown in Tables III and IV clearly demonstrate that the performance of the AI-ST measure depends largely on the choice of the BIF. The BIF given in Eq. (20), in particular, was found to work the best on both consonant and sentence recognition tasks. The performance, for instance, of the AI-ST measure when applied to sentence recognition improved from $r=0.33$ with ANSI (1997) weights to $r=0.80$ with the proposed BIF given in Eq. (20). The results from the present study clearly suggest that the traditional SII index (ANSI, 1997), as well as the STI index, could benefit from

TABLE XI. Correlation coefficients, r, and standard deviations of the error, σ_e, between sentence recognition scores and the NCM measure as a function of the power exponent, p, used in the BIF in Eq. (12).

Power exponent, p	r	σ_e
0.12	0.85	0.09
0.25	0.87	0.08
0.50	0.89	0.08
0.62	0.89	0.08
0.75	0.89	0.08
1.00	0.89	0.08
1.50	0.89	0.07

the use of signal-dependent BIFs, such as those given in Eqs. (19)–(22).

VI. CONCLUSIONS

The present study evaluated the performance of traditional (e.g., SNRseg) as well as new objective measures in terms of predicting speech intelligibility in realistic noisy conditions. The objective measures were tested in a total of 72 noisy conditions which included processed sentences and nonsense syllables corrupted by four real-world types of noise (car, babble, train, and street). The distinct contributions of the present work include the following:

(1) An AI-ST measure was proposed operating on short-term (30 ms) segments. This measure was found to predict modestly (r=0.68–0.83) well the intelligibility of speech embedded in fluctuating maskers when the proposed BIFs were used. The performance of the AI-based measure was quite poor (r=0.33) when the ANSI (1997) AI weights were used, but improved to r=0.83 when the proposed (segment-dependent) BIFs were used.

(2) A low-frequency version of the NCM measure was proposed that incorporates only low-frequency (100–1000 Hz) envelope information in its computation. The correlation obtained with this measure for predicting sentence recognition scores was high (r=0.87) and nearly as good as that obtained with the full-bandwidth (300–3400 Hz) NCM measure (r=0.89). This outcome provides additional support for the importance of low-frequency (<1000 Hz) acoustic landmarks on speech recognition (Li and Loizou, 2008).

(3) The conventional SNRseg measure, which is widely used for assessing performance of noise-suppression and speaker-separation algorithms, predicted poorly (r=0.40–0.46) the intelligibility of consonants and sentences.

(4) The PESQ measure, which was originally designed to predict speech quality, performed modestly well (r=0.77–0.79) on predicting speech intelligibility in noise. Of all the conventional subjective quality measures tested, the fwSNRseg and PESQ measures performed modestly well in terms of predicting *both* quality and intelligibility.

(5) The influence of speech dynamic range (varying from 30 to 50 dB), integration window (varying from 30 to 125 ms), number of bands (varying from 7 to 20 bands), and range of modulation frequencies (varying from 12.5 to 30 Hz) was assessed on the performance of the AI-based and STI-based (i.e., NCM) measures. Of all these parameters, only the use of a wider dynamic range (45–50 dB) improved somewhat the correlation of the NCM and AI-ST measures. Increasing the window duration also improved the correlation of the AI-ST measure in predicting sentence recognition (Table X).

(6) Of all parameters examined in this study, the BIFs influenced the performance of the AI-based, STI-based (NCM), and coherence-based (CSII) measures the most. The proposed signal and phonetic-segment dependent BIFs [Eqs. (19)–(22)] were found to be suitable for predicting the intelligibility of speech in fluctuating maskers. Additional flexibility is built in the proposed band-importance functions for emphasizing spectral peaks and/or spectral valleys. The proposed BIFs improved consistently the performance of all three sets of measures. This outcome clearly suggests that the traditional SII index (ANSI, 1997) as well as the STI index could benefit from the use of signal-dependent band-importance functions, such as those proposed in Eqs. (19)–(22).

(7) Among all objective measures examined in the present study, the modified CSII and NCM measures incorporating signal-specific weighting information have been found to perform the best in terms of predicting speech intelligibility in noise. The modified CSII_{mid} measure, in particular, which only includes vowel/consonant transitions and weak consonants in its computation, yielded the highest correlation (r=0.94) with sentence recognition scores. This outcome further corroborates the large contribution of weak consonants on speech recognition in noise (Li and Loizou, 2008).

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC007527 (P.C.L.) and R03 DC008887 (Y.H.) from the National Institute of Deafness and other Communication Disorders, NIH.

APPENDIX

The MSC function is given by

$$\text{MSC}(\omega) = \frac{|S_{XY}(\omega)|^2}{S_{XX}(\omega)S_{YY}(\omega)}. \tag{A1}$$

Let the MTF at frequency ω be given by (Drullman *et al.*, 1994b)

$$\text{MTF}(\omega) = \alpha\sqrt{\frac{S_{YY}(\omega)}{S_{XX}(\omega)}}, \tag{A2}$$

where α is a normalization factor, and let $W(\omega)$ be the following weighting function at frequency ω:

$$W(\omega) = \frac{1}{\alpha}\frac{|S_{XY}(\omega)|^2}{\sqrt{S_{XX}(\omega)}(S_{YY}(\omega))^{3/2}}. \tag{A3}$$

Then, the MSC function can be written as a weighted MTF, i.e.,

$$\text{MSC}(\omega) = W(\omega)\cdot\text{MTF}(\omega). \tag{A4}$$

Allen, J. B. (**1994**). "How do humans process and recognize speech," IEEE Trans. Speech Audio Process. **2**, 567–577.

Anderson, W. B., and Kalb, J. T. (**1987**). "English verification of STI method for estimating speech intelligibility of a communications channel," J. Acoust. Soc. Am. **81**, 1982–1985.

ANSI (**1997**). "Methods for calculation of the speech intelligibility index," S3.5–1997 (American National Standards Institute, New York).

Arai, T., Pavel, M., Hermansky, H., and Avendano, C. (**1996**). "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proceedings of the ICSLP*, pp. 2490–2493.

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.
Beerends, J., Larsen, E., Lyer, N., and van Vugt, J. (**2004**). "Measurement of speech intelligibility based on the PESQ approach," in Proceedings of the Workshop Measurement of Speech and Audio Quality in Networks (MESAQIN), Prague, Czech Republic.
Beerends, J., van Wijngaarden, S., and van Buuren, R. (**2005**). "Extension of ITU-T recommendation P.862 PESQ towards measuring speech intelligibility with vocoders," in *New Directions for Improving Audio Effectiveness*, Proceedings of the RT0-MP-HFM-123, Neuilly-sur-Seine, France, pp. 10-1–10.6.
Bladon, R., and Lindblom, B. (**1981**). "Modeling the judgment of vowel quality differences," J. Acoust. Soc. Am. **69**, 1414–1422.
Boothroyd, A., Erickson, F. N., and Medwetsky, L. (**1994**). "The hearing aid input: A phonemic approach to assessing the spectral distribution of speech," Ear Hear. **6**, 432–442.
Brachmanski, S. (**2004**). "Estimation of logatom intelligibility with STI method for polish speech transmitted via communication channels," Arch. Acoust. **29**, 555–562.
Carter, C., Knapp, C., and Nuttall, A. (**1973**). "Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing," IEEE Trans. Audio Electroacoust. **AU-21**, 337–344.
Cohen, I., and Berdugo, B. (**2002**). "Noise estimation by minima controlled recursive averaging for robust speech enhancement," IEEE Signal Process. Lett. **9**, 12–15.
Drullman, R., Festen, J., and Plomp, R. (**1994a**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.
Drullman, R., Festen, J., and Plomp, R., (**1994b**). "Effect of reducing slow temporal modulations on speech reception" J. Acoust. Soc. Am. **95**, 2670–2680.
Dunn, H., and White, S. (**1940**). "Statistical measurements on conversational speech," J. Acoust. Soc. Am. **11**, 278–288.
Ephraim, Y., and Malah, D. (**1985**). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-33**, 443–445.
Fletcher, H., and Galt, R. H. (**1950**). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89–151.
French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.
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.
Gustafsson, H., Nordholm, S., and Claesson, I. (**2001**). "Spectral subtraction using reduced delay convolution and adaptive averaging," IEEE Trans. Speech Audio Process. **9**, 799–807.
Hansen, J., and Pellom, B. (**1998**). "An effective quality evaluation protocol for speech enhancement algorithms," in Proceedings of the International Conference on Spoken Language Processing, Vol. **7**, pp. 2819–2822.
Hirsch, H., and Pearce, D. (**2000**). "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA Tutorial and Research Workshop ASR2000*, Paris, France.
Hohmann, V., and Kollmeier, B. (**1995**). "The effect of multichannel dynamic compression on speech intelligibility," J. Acoust. Soc. Am. **97**, 1191–1195.
Hollube, I., and Kollmeier, K. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**, 1703–1715.
Houtgast, T., and Steeneken, H. J. M. (**1971**). "Evaluation of speech transmission channels by using artificial signals," Acustica **25**, 355–367.
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.
Hu, Y., and Loizou, P. C. (**2003**). "A generalized subspace approach for enhancing speech corrupted by colored noise," IEEE Trans. Speech Audio Process. **11**, 334–341.
Hu, Y., and Loizou, P. C. (**2004**). "Speech enhancement based on wavelet thresholding the multitaper spectrum," IEEE Trans. Speech Audio Process. **12**, 59–67.
Hu, Y., and Loizou, P. C., (**2007**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1786.
Hu, Y., and Loizou, P. C. (**2008**). "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Audio, Speech, Lang. Process. **16**, 229–238.
IEC 60268-16 (**2003**). "Sound system equipment—Part 16: Objective rating of speech intelligibility by speech transmission index," Ed. 3 (International Electrotechnical Commission, Geneva, Switzerland).
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
ITU-T (**2000**). "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation P. 862.
Jabloun, F., and Champagne, B. (**2003**). "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," IEEE Trans. Speech Audio Process. **11**, 700–708.
Kamath, S., and Loizou, P. C. (**2002**). "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL.
Kates, J. (**1987**). "The short-time articulation index," J. Rehabil. Res. Dev. **24**, 271–276.
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.
Kitawaki, N., Nagabuchi, H., and Itoh, K. (**1988**). "Objective quality evaluation for low bit-rate speech coding systems," IEEE J. Sel. Areas Commun. **6**, pp. 262–273.
Klatt, D. H. (**1982**). "Prediction of perceived phonetic distance from critical-band spectra: A first step," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. **2**, pp. 1278–1281.
Kryter, K. D. (**1962a**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.
Kryter, K. D. (**1962b**). "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698–1706.
Larm, P., and Hongisto, V. (**2006**). "Experimental comparison between speech transmission index, rapid speech transmission index, and speech intelligibility index," J. Acoust. Soc. Am. **119**, 1106–1117.
Li, N., and Loizou, P. (**2008**). "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," J. Acoust. Soc. Am. **124**, 498–509.
Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice* (CRC, Boca Raton, FL).
Ludvigsen, C., Elberling, C., and Keidser, G. (**1993**). "Evaluation of a noise reduction method—Comparison of observed scores and scores predicted from STI," Scand. Audiol. Suppl. **38**, 50–55.
Ludvigsen, C., Elberling, C., Keidser, G., and Poulsen, T. (**1990**). "Prediction of intelligibility of non-linearly processed speech," Acta Oto-Laryngol., Suppl. **469**, 190–195.
Mapp, P. (**2002**). "A comparison between STI and RASTI speech intelligibility measurement systems," in The 111th AES Convention, Los Angeles, CA, Preprint No. 5668.
Moore, B., and Glasberg, B. (**1993**). "Suggested formulas for calculation auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.
Pavlovic, C. V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**, 413–422.
Quackenbush, S. R., Barnwell, T. P., and Clements, M. A. (**1988**). *Objective Measures of Speech Quality*, (Prentice-Hall, Englewood Cliffs, NJ).
Rhebergen, K. S., and 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.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988–3997.
Rix, A., Beerends, J., Hollier, M., and Hekstra, A. (**2001**). "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. **2**, pp. 749–752.
Scalart, P., and Filho, J. (**1996**). "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 629–632.
Steeneken, H., and Houtgast, T. (**1980**). "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Steeneken, H., and Houtgast, T. (**1982**). "Some applications of the speech

transmission index (STI) in auditoria," Acustica **51**, 229–234.
Stevens, K. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891.
Studebaker, G., and Sherbecoe, R. (**2002**). "Intensity-importance functions for bandlimited monosyllabic words," J. Acoust. Soc. Am. **111**, 1422–1436.
van Buuren, R., Festen, J., and Houtgast, T. (**1999**). "Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality," J. Acoust. Soc. Am. **105**, 2903–2913.
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.