Analysis of speech-based speech transmission index methods with implications for nonlinear operations

Ray L. Goldsworthy and Julie E. Greenberg

Citation: The Journal of the Acoustical Society of America **116**, 3679 (2004); doi: 10.1121/1.1804628
View online: https://doi.org/10.1121/1.1804628
View Table of Contents: https://asa.scitation.org/toc/jas/116/6
Published by the Acoustical Society of America

Analysis of speech-based speech transmission index methods with implications for nonlinear operations

Ray L. Goldsworthy and Julie E. Greenberg[a)]
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 and Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts 02139

(Received 1 October 2003; revised 16 April 2004; accepted 18 August 2004)

The Speech Transmission Index (STI) is a physical metric that is well correlated with the intelligibility of speech degraded by additive noise and reverberation. The traditional STI uses modulated noise as a probe signal and is valid for assessing degradations that result from linear operations on the speech signal. Researchers have attempted to extend the STI to predict the intelligibility of nonlinearly processed speech by proposing variations that use speech as a probe signal. This work considers four previously proposed speech-based STI methods and four novel methods, studied under conditions of additive noise, reverberation, and two nonlinear operations (envelope thresholding and spectral subtraction). Analyzing intermediate metrics in the STI calculation reveals why some methods fail for nonlinear operations. Results indicate that none of the previously proposed methods is adequate for all of the conditions considered, while four proposed methods produce qualitatively reasonable results and warrant further study. The discussion considers the relevance of this work to predicting the intelligibility of cochlear-implant processed speech. © *2004 Acoustical Society of America.* [DOI: 10.1121/1.1804628]

PACS numbers: 43.71.Gv, 43.60Wy, 43.71Ky [KWG] Pages: 3679–3689

I. INTRODUCTION

Early attempts to predict speech intelligibility led to the development of the articulation index (AI) (French and Steinberg, 1947; Kryter, 1962a, 1962b). A fundamental principle of the AI is that the intelligibility of speech depends on a weighted average of the signal to noise ratios (SNRs) in frequency bands spanning the speech spectrum. By accounting for the contribution of different regions of the spectrum to intelligibility, the AI successfully predicts the effects of additive noise and simple linear filters.

The Speech Transmission Index (STI) (Houtgast and Steeneken, 1971; Steeneken and Houtgast, 1980; IEC, 1998) is an intelligibility metric that differs from the AI by using reduction in signal modulation rather than band-specific SNRs. By including modulation reduction in the frequency band analysis, the STI can predict the effects of reverberation as well as additive noise. Calculation of the STI is based on changes in signal modulation when modulated probe stimuli are transmitted through a channel of interest. The responses to probe stimuli are measured in multiple frequency bands for a range of modulation frequencies relevant to speech. The STI successfully quantifies the effects of room acoustics and broadcast channels on speech intelligibility (Steeneken and Houtgast, 1982). The STI can also be adapted for use with hearing-impaired subjects (Humes *et al.*, 1986; Ludvigsen, 1987; Payton *et al.*, 1994).

Steeneken and Houtgast (1980) suggest that applying the STI to nonlinear operations requires more sophisticated probe signals than used in their original procedure. They introduced complex test signals that combine modulated noise with artificial speech-like signals, allowing the STI to predict the effects of automatic gain control and peak clipping. Other researchers have developed variations that use speech, rather than an artificial probe, to investigate nonlinear operations. These speech-based methods have been used to analyze dynamic amplitude compression (Hohmann and Kollmeier, 1995; Payton *et al.*, 2002; Drullman, 1995), spectral subtraction (Ludvigsen *et al.*, 1993), and envelope clipping (Drullman, 1995). In addition, speech-based STI methods have been used to investigate the intelligibility differences between clear and conversational speech (Payton *et al.*, 1994; Payton *et al.*, 1999).

The speech-based STI methods have generally failed to predict performance for nonlinear operations. In some studies, STI intelligibility predictions have been qualitatively inconsistent with performance results. A study of envelope expansion found that ''the prediction from STI is in the wrong direction for the expansion conditions'' (Van Buuren *et al.*, 1998). In an investigation of speech-based STI and spectral subtraction, researchers concluded ''STI, even in its modified version, is an unreliable predictor when non-linear processes are involved.'' (Ludvigsen *et al.*, 1993). Other researchers (Drullman, 1995; Payton *et al.*, 2002; Hohmann and Kollmeier, 1995) have also concluded that speech-based STI methods proposed thus far do not adequately predict the intelligibility of nonlinearly processed speech.

In this work, the various speech-based STI methods are analyzed to determine why they fail to predict intelligibility for nonlinear operations. Simple modifications are proposed to overcome problems with the existing speech-based STI methods. This results in four modified speech-based STI methods that are related to previously proposed methods. These modified STI methods are well correlated with the traditional STI for additive noise and reverberation and also

[a)]Author to whom correspondence should be addressed at: Massachusetts Institute of Technology, Building 36, Room 761, 77 Massachusetts Ave., Cambridge, MA 02139; electronic mail: jgreenbe@mit.edu

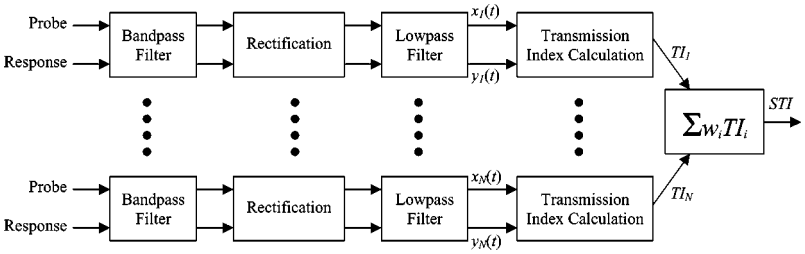

FIG. 1. General form of the STI calculation. For each frequency band, $i=1,...,N$, envelopes of the probe and response signals are compared to determine a transmission index (TI_i). The STI is a weighted average of TI_i values.

exhibit qualitatively reasonable behavior for selected nonlinear operations. As a result, the modified STI methods are promising candidates to predict intelligibility of nonlinearly processed speech.

II. BACKGROUND

Both the traditional and speech-based STI methods employ a frequency band analysis as illustrated in Fig. 1. A bank of bandpass filters splits the probe and response signals into frequency bands, where i indicates the frequency band number. Typically, octave bands with center frequencies from 125 to 8000 Hz are used. For each band, the probe and response envelope signals, $x_i(t)$ and $y_i(t)$, respectively, are computed by rectification and lowpass filtering and then compared to determine a transmission index, TI_i. The TI_i values are combined using a weighted average to determine the STI value. The various STI methods differ in how the envelope signals are computed and in how the TI_i values are computed from the envelopes.

A. Traditional method of computing the STI

For the traditional method (Steeneken and Houtgast, 1980), the TI_i values are determined from an intermediate function called the modulation transfer function (MTF). The MTF is a function of modulation frequency, f, calculated individually for each value of f. For each frequency band, the probe signal consists of speech-shaped noise that has been bandpass filtered and then intensity modulated at a particular modulation frequency. The probe signal is passed through the system to be evaluated. The fractional change in modulation depth between probe and response intensity envelopes is quantified for that value of f, and the process is repeated for other modulation frequencies to determine the complete MTF for one frequency band. The MTF is typically characterized using modulation frequencies ranging from $f=0.63$ Hz to $f=12.7$ Hz in one-third octave intervals. As an alternative to artificial probe signals, Houtgast and Steeneken (1985) proposed determining the MTF for each frequency band from spectra of the intensity envelopes of running speech. Omitting the subscript i to simplify notation, this approach can be described as (Drullman, 1994b)

$$\mathrm{MTF}(f)=\alpha\frac{|Y(f)|}{|X(f)|}=\alpha\sqrt{\frac{S_{yy}(f)}{S_{xx}(f)}}, \tag{1}$$

where $\alpha=\mu_x/\mu_y$, $\mu_x=E\{x(t)\}$, $\mu_y=E\{y(t)\}$, and $E\{\cdot\}$ denotes expected value. $|X(f)|$ and $|Y(f)|$ are magnitude spectra, and $S_{xx}(f)$ and $S_{yy}(f)$ are power spectra, of the probe and response envelope signals, respectively.

The signal-to-noise ratio (SNR) in decibels as a function of f is calculated for each frequency band as

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

An overall apparent SNR (aSNR_i) for each frequency band is determined by clipping the $\mathrm{SNR}_i(f)$ values and then averaging across modulation frequencies, that is,

$$\mathrm{cSNR}_i(f)=\begin{cases}-15, & \mathrm{SNR}_i(f)<-15\\ \mathrm{SNR}_i(f), & -15\leqslant\mathrm{SNR}_i(f)\leqslant 15\\ 15, & \mathrm{SNR}_i(f)>15,\end{cases} \tag{3}$$

$$\mathrm{aSNR}_i=\mathrm{mean}(\mathrm{cSNR}_i(f)). \tag{4}$$

The transmission index is a linear function of the apparent SNR for each band, defined to be between zero and one,

$$\mathrm{TI}_i=\frac{\mathrm{aSNR}_i+15}{30}. \tag{5}$$

Finally, the overall STI value is calculated as a weighted average of the TI_i values,

$$\mathrm{STI}=\sum_i w_i\mathrm{TI}_i, \tag{6}$$

where w_i is a psycho-acoustically derived weighting. The weights, w_i, are defined to sum to one, thereby restricting the STI values to a range between zero and one.

B. Speech-based STI methods

This section summarizes four speech-based methods proposed in the literature. The first three speech-based methods use intensity envelopes calculated by squaring and then smoothing, while the fourth uses magnitude envelopes. For each method, the description focuses on the calculation of TI_i for one frequency band. To simplify notation, the subscript i is omitted for intermediate variables such as $\mathrm{MTF}(f)$ and aSNR.

1. Magnitude cross-power spectrum method

Payton and colleagues (2002) proposed a speech-based method where the MTF is based on the magnitude of the cross-power spectra as given by

$$\mathrm{MTF}(f)=\alpha\left|\frac{S_{xy}(f)}{S_{xx}(f)}\right|, \tag{7}$$

where $S_{xy}(f)$ is the cross-power spectrum of the probe and response envelopes. The MTF given by Eq. (7) is used in Eq. (2), and the STI is calculated from Eqs. (2) through (6).

2. Real cross-power spectrum method

Drullman and colleagues (1994b) introduced a phase-locked MTF in order to investigate the effects of reducing low-frequency modulations on the intelligibility of speech. The phase-locked MTF is defined as

$$\mathrm{MTF}(f)=\alpha\,\mathrm{Re}\left(\frac{S_{xy}(f)}{S_{xx}(f)}\right), \tag{8}$$

where $\mathrm{Re}(\cdot)$ denotes taking the real part of the complex-valued function. Although they did not propose a corresponding STI calculation procedure, the MTF in Eq. (8) could be used to calculate the STI in conjunction with Eqs. (2) through (6).

3. Envelope regression method

Ludvigsen and colleagues (1990) proposed a method where the probe envelope signal, $x(t)$, and the response envelope signal, $y(t)$, are compared using linear regression analysis. In this method, the apparent SNR for each frequency band is defined as

$$\mathrm{aSNR}=10\log_{10}\left(\frac{A\mu_x}{B}\right), \tag{9}$$

where A and B are the parameters that produce the best fit for the model $y(t)=Ax(t)+B$. This apparent SNR is clipped to values between ± 15 dB, and the STI is calculated via Eqs. (5) and (6).

4. Normalized covariance method

The normalized covariance method (Koch, 1992; Holube and Kollmeier, 1996) is based on the covariance between the probe and response envelope signals. For each frequency band, the apparent SNR is calculated as

$$\mathrm{aSNR}=10\log_{10}\left(\frac{r^2}{1-r^2}\right), \tag{10}$$

where r is the normalized covariance between $x(t)$ and $y(t)$ given by

$$r^2=\frac{\lambda_{xy}^2}{\lambda_x\lambda_y} \tag{11}$$

with

$$\lambda_{xy}=E\{(x(t)-\mu_x)(y(t)-\mu_y)\} \tag{12}$$

and

$$\lambda_x=E\{(x(t)-\mu_x)^2\}. \tag{13}$$

The apparent SNR of Eq. (10) is clipped to values between ± 15 dB and the STI is calculated via Eqs. (5) and (6).

5. Summary of speech-based methods

The above-described speech-based methods all compute the STI as a weighted sum of TI values determined from the envelopes of the probe and response signals in each frequency band. The key difference among the methods is how the TI values are calculated. Table I summarizes the intermediate modulation metrics used to calculate TI values for the different methods. In the case of the envelope regression method, the modulation metric in Table I is an alternate form that is derived in Appendix A. For the two cross-power spectrum methods, the modulation metric is a function of modulation frequency. For the other two methods there is a single value for each frequency band. The implications of this fundamental difference are discussed in Sec. VI A. In the following sections, these modulation metrics will be used to yield insight into the behavior of the speech-based STI methods.

TABLE I. Intermediate modulation metrics for speech-based STI methods proposed in the literature. These metrics use the normalization term $\alpha=\mu_x/\mu_y$. They are calculated for each frequency band and then combined to produce a single STI value as described in the text.

Magnitude cross-power spectrum	Real cross-power spectrum	Envelope regression	Normalized covariance
$\mathrm{MTF}(f)=\alpha\left\lvert\frac{S_{xy}(f)}{S_{xx}(f)}\right\rvert$	$\mathrm{MTF}(f)=\alpha\,\mathrm{Re}\left(\frac{S_{xy}(f)}{S_{xx}(f)}\right)$	$M=\alpha\frac{\lambda_{xy}}{\lambda_x}$	$r^2=\frac{\lambda_{xy}^2}{\lambda_x\lambda_y}$

III. PROPOSED METRICS

A. Normalization based on noise envelope

Both cross-power spectrum methods [Eqs. (7) and (8)] include the term α, which normalizes the envelopes to account for the power of the probe and response signals. The alternate form of the envelope regression method derived in Appendix A also depends on α; for this method the apparent SNR in Eq. (9) can be expressed as

$$\mathrm{aSNR}=10\log_{10}\left(\frac{M}{1-M}\right), \tag{14}$$

where M is a modulation metric defined as

$$M=\alpha\frac{\lambda_{xy}}{\lambda_x}. \tag{15}$$

Thus, the envelope regression method, as well as the two cross-power spectrum methods, include the normalization term α. This term successfully normalizes the envelopes for the cases of additive noise and reverberation; however, for a large class of operations this normalization ratio is not appropriate. In particular, when the processing reduces the overall amplitude of the response envelope, $y(t)$, α may increase without bound. As shown in Secs. V B and V C, this leads to invalid values of the intermediate modulation metrics listed in Table I.

An alternative normalization term is proposed here. The noise envelope is defined as

$$z(t)=|y(t)-x(t)|, \tag{16}$$

and a new normalization term is defined as

$$\beta=\frac{\mu_x}{\mu_x+\mu_z}. \tag{17}$$

For cases when $y(t)>x(t)$ for all t (as is typically the case for additive noise and reverberation) then $\mu_z=\mu_y-\mu_x$ and, consequently, $\beta=\alpha$. Thus, for certain operations, the pro-

TABLE II. Intermediate modulation metrics for speech-based STI methods proposed in this work. These metrics use the normalization term β as defined in Eq. (17). They are calculated for each frequency band and then combined to produce a single STI value as described in the text.

Magnitude cross-power spectrum	Real cross-power spectrum	Envelope regression	Normalized correlation
$\mathrm{MTF}(f)=\beta\left\lvert\frac{S_{xy}(f)}{S_{xx}(f)}\right\rvert$	$\mathrm{MTF}(f)=\beta\,\mathrm{Re}\left(\frac{S_{xy}(f)}{S_{xx}(f)}\right)$	$M=\beta\frac{\lambda_{xy}}{\lambda_x}$	$\rho^2=\frac{\phi_{xy}^2}{\phi_x\phi_y}$

posed normalization term equals the original.

When the processing reduces the response envelope so that $y(t)<x(t)$ for some values of t, then μ_y decreases, causing α to increase. In some cases, high values of α may result in erroneously high values of apparent SNR for that frequency band. Since $\mu_z+\mu_x$ is always greater than μ_x, β will avoid characterizing reduced response envelopes as improved SNR.

B. Normalized correlation

We hypothesize that the normalized covariance method (Sec. II B 4) is well suited to nonlinear operations. The normalized covariance defined in Eq. (11) is a metric that necessarily falls between zero and one, with a value of unity achieved only when the envelopes are identical. These constraints ensure that the method always produces valid values of the intermediate metric. For the other speech-based methods, the intermediate metrics in Table I are not restricted to values between zero and one, and operations that increase the modulation depth may cause the intermediate metrics to take on invalid values greater than one, as demonstrated in Secs. V B and V C.

As a variation on the normalized covariance method, we consider the normalized correlation,[1] ρ, where

$$\rho^2=\frac{\phi_{xy}^2}{\phi_x\phi_y} \tag{18}$$

with $\phi_{xy}=E\{x(t)y(t)\}$ and $\phi_x=E\{x^2(t)\}$. The STI is subsequently calculated by substituting ρ for r in Eq. (10), clipping to values between ± 15 dB, and applying Eqs. (5) and (6). The normalized correlation method differs from the normalized covariance method only in that the envelope means are included in the correlation terms.

Table II summarizes the intermediate modulation metrics for the proposed speech-based methods. Comparing Table II to Table I reveals the key differences between the methods proposed in this work and those proposed previously.

IV. METHODS

This section describes the calculation of the various speech-based STI methods for three sets of processing conditions: acoustic degradation, envelope thresholding, and spectral subtraction. For the acoustic degradation conditions, speech-based STI values are compared to the traditional STI. For the envelope thresholding and spectral subtraction conditions, the speech-based STI methods are characterized by intermediate modulation metrics for a single frequency band.

A. Common elements

For all speech-based STI methods, the probe stimulus was a 120 s speech signal formed by concatenating 42 phonetically balanced sentences (IEEE, 1969). For the traditional method, the probe stimulus was based on a 60 s noise sequence with the same long-term spectrum as the speech. In both cases the sampling rate was $F_s=22050$ Hz.

The bandpass filters were seven octave-band filters with center frequencies ranging from 125 Hz to 8 kHz. All filters were eighth-order Butterworths. Intensity envelopes were calculated by squaring the bandpass-filtered signals and lowpass filtering. Magnitude envelopes were calculated by full-wave rectification of the bandpass-filtered signals followed by lowpass filtering. In both cases the lowpass filter was an eighth-order Butterworth with 50 Hz cutoff frequency. Envelopes were downsampled to 200 Hz before calculating the various metrics. This resulted in discrete-time probe and response envelope signals, $x[n]$ and $y[n]$, that were $N=24\,000$ samples long for the speech sequence and $N=12\,000$ samples long for the noise sequence.

The octave band weighting function used in Eq. (6) was taken from Houtgast and Steeneken (1985). All processing was performed in MATLAB® on a personal computer with a Pentium III processor.

B. Metric calculation

1. Traditional method

The traditional STI was calculated using fourteen modulation frequencies ranging from $f=0.63$ to 12.7 Hz in one-third-octave increments. Because it requires the use of a probe noise sequence, it was only practical to compute the traditional STI for the acoustic degradation conditions. For each modulation frequency, the noise sequence described in Sec. IV A was amplitude modulated by $\sqrt{1+\cos(2\pi(f/F_s)n)}$ to form the probe signal. The response signal consisted of the probe signal combined with additive noise and/or reverberation. Both the probe and response signals were bandpass filtered into octave bands and intensity envelopes were computed by squaring followed by lowpass filtering. The modulation depth of each envelope was measured as the maximum value of the cross-covariance between the envelope and the function $\cos(2\pi(f/F_s)n)$ normalized by the envelope mean. The MTF value was determined from the ratio of the response envelope's modulation depth to the probe envelope's modulation depth.

2. Cross-power spectrum methods

Both the magnitude cross-power spectrum method and the real cross-power spectrum method use intensity envelopes. Sample envelope means were calculated from the average of the envelope signals. The MTF for the two cross-power spectrum methods requires estimating the auto- and cross-power spectra. This was accomplished using the periodogram method with 4096-point Hanning windows, 4096-point FFTs, and 50% overlap. The resulting 0.05 Hz fre-

quency bins were averaged to produce values in one-third octave intervals (Payton *et al.*, 1999) centered from 0.63 to 12.7 Hz. This resulted in averaging of three bins for the lowest modulation frequency and 60 bins for the highest modulation frequency. These quantities were used in Eqs. (7) and (8) for the original methods, and with β [Eq. (17)] in place of α for the proposed methods. Then STI was calculated via Eqs. (2) through (6).

3. Envelope regression method

The envelope regression method was calculated from the intensity envelopes using the alternate form derived in Appendix A. Sample envelope means were computed from the average of the envelope signals and the covariance was calculated as an unbiased estimate, that is,

$$\lambda_{xy}=E\{(x[n]-\mu_x)(y[n]-\mu_y)\}$$
$$\approx\left(\frac{1}{N-1}\right)\sum_{i=1}^{N}(x[i]-\mu_x)(y[i]-\mu_y). \tag{19}$$

For each frequency band, the modulation metric, M, was calculated using Eq. (15) for the existing method and with β in place of α for the proposed method. The apparent SNR was then calculated from Eq. (14), clipped to values between ± 15 dB, and used in Eqs. (5) and (6).

4. Normalized covariance and normalized correlation methods

The normalized covariance and normalized correlation methods were calculated based on magnitude envelopes. For each frequency band, the normalized covariance, r, was calculated from Eq. (11), with estimates of the variance and covariance calculated as in Eq. (19). The normalized correlation, ρ, was calculated according to Eq. (18) with the correlation estimated as

$$\phi_{xy}=E\{x[n]y[n]\}\approx\left(\frac{1}{N-1}\right)\sum_{i=1}^{N}(x[i]\cdot y[i]). \tag{20}$$

The apparent SNRs were calculated from Eq. (10) (replacing r with ρ for the normalized correlation method), clipped to values between ± 15 dB, and used in Eqs. (5) and (6).

C. Acoustic degradations

For the acoustic degradation conditions, speech-shaped noise was added to the probe stimulus and the resulting signal was convolved with a reverberant impulse response. The speech-shaped noise had the same long-term spectrum as the probe stimulus. Two-second-long reverberant impulse responses were generated using a room simulation based on the image method (Allen and Berkley, 1979). The speech-shaped noise was scaled to produce SNRs between -15 and 30 dB in 3 dB increments as well as a no-noise condition. Reverberation times (T_{60}) ranged from 0 to 1.5 s in 0.3 s increments. The traditional and speech-based STIs were computed for all combinations of SNR and reverberation time.

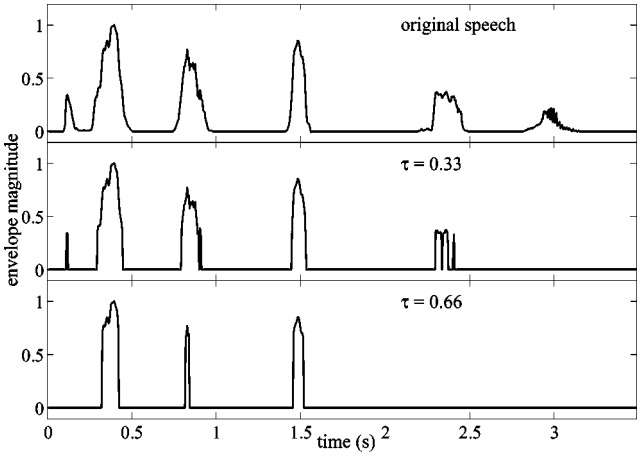

FIG. 2. Effect of envelope thresholding on a speech envelope for the octave band centered at 1 kHz, shown for two values of the fractional threshold, τ.

D. Envelope thresholding

Envelope thresholding is a nonlinear operation that consists of setting to zero any samples of the original envelope that are below a threshold, that is

$$y[n]=\begin{cases} x[n], & x[n]\geqslant\tau\max(|x[n]|) \\ 0, & x[n]<\tau\max(|x[n]|), \end{cases} \tag{21}$$

where $x[n]$ and $y[n]$ are the probe and response envelopes, respectively, and τ is a fractional threshold relative to the maximum value of the probe envelope. Figure 2 illustrates the effect of the envelope thresholding on a speech envelope and shows that increasing the value of the threshold results in greater levels of modulation and increasingly distorted envelopes. Intermediate modulation metrics were calculated for all speech-based STI methods for values of τ ranging from 0 to 1 in increments of 0.02.

E. Spectral subtraction

Spectral subtraction attempts to reduce background noise by subtracting a spectral estimate of the noise from short-time spectra of the noisy signal. Generalized spectral subtraction (Lim and Oppenheim, 1979) scales the noise spectral estimate by a constant factor, that is,

$$|P(F)|=|D(F)|-\kappa|\hat{N}(F)|, \tag{22}$$

where $D(F)$ is a short-time spectrum of the input signal, $\hat{N}(F)$ is the spectral estimate of the noise, $P(F)$ is the processed spectrum, and κ is a parameter that scales the noise estimate. $|P(F)|$ is multiplied by the phase of the original input signal and short-time reconstruction is performed to produce the time-domain output signal.

Figure 3 illustrates the effects of spectral subtraction on speech envelopes. For $\kappa=1$, the noise component of the envelope is suppressed with relatively little effect on the speech envelope. For $\kappa=8$, the noise is suppressed, but the speech envelope is highly distorted. Spectral subtraction with large values of κ is similar to envelope thresholding in that it distorts the envelope and increases the level of modulation.

The speech signal was degraded by noise with the same long-term spectrum as the probe stimulus (0 dB SNR) and

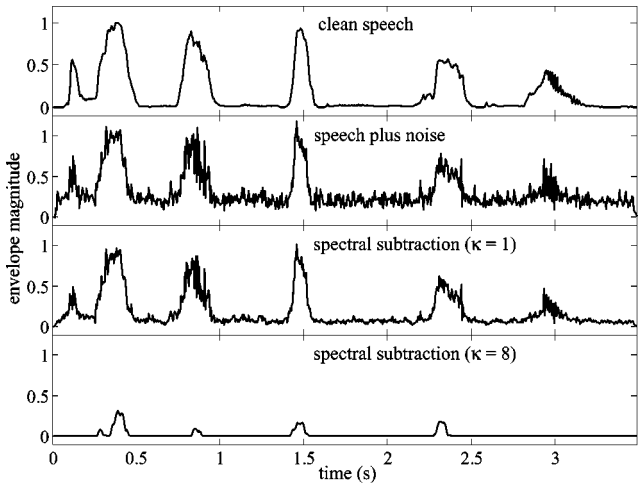

FIG. 3. Effect of spectral subtraction on the envelope of noisy speech for the octave band centered at 1 kHz, shown for two values of the control parameter, κ.

then processed by the spectral subtraction algorithm using the overlap-add method with 25 ms Hamming windows with 50% overlap. Intermediate modulation metrics were calculated for all speech-based STI methods for values of κ ranging from zero to eight in increments of 0.25. A value of $\kappa = 0$ corresponds to no spectral subtraction processing and a value of $\kappa = 1$ corresponds to standard spectral subtraction. A value of $\kappa = 8$ corresponds to an extreme version where the spectral subtraction processing eliminates all but the highest spectral peaks.

V. RESULTS

A. Acoustic degradation

Since the traditional STI method is well established as an accurate predictor of speech intelligibility for additive stationary noise and reverberation, any proposed speech-based method must produce similar values of STI under these conditions. Figure 4 compares the speech-based STI methods to the traditional STI for the acoustic degradation conditions of additive noise and reverberation described in Sec. IV C. Figures 4(a)–(d) show the four previously proposed speech-based methods described in Sec. II B, while Figs. 4(e)–(h) show the methods proposed in Sec. III. Each curve represents STI values calculated over the 45 dB range of SNRs for one level of reverberation.

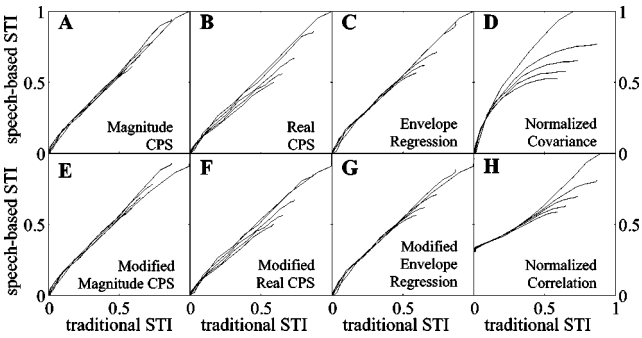

FIG. 4. Comparison of speech-based STI methods to the traditional STI. Each plot shows the relationship between one speech-based method and the traditional STI. Each curve corresponds to the 45 dB range of SNR values for one level of reverberation. More reverberant conditions terminate at lower values of the traditional STI.

In Fig. 4, complete agreement between the traditional STI method and a speech-based STI method would appear as a straight line from the bottom left to the top right of a particular plot. As seen in Figs. 4(a), (b), and (c), the original cross-power spectrum methods and the original envelope regression method all provide a reasonable match to the traditional method, although the real cross-power spectrum method is slightly less well-matched to the traditional than the other two.

Comparing Figs. 4(a), (b), and (c) to Figs. 4(e), (f), and (g) shows that for these acoustic degradation conditions, the modified methods using β as the normalization term are equivalent to the original methods using α. As described in Sec. III A, this equivalence is expected because the acoustic degradations increase the response envelopes relative to the probe envelopes.

The normalized covariance method [Fig. 4(d)] and the proposed normalized correlation method [Fig. 4(h)] are distinctly different from the other speech-based methods. The normalized covariance method does not exhibit a one-to-one relationship with the traditional method. The curves for different levels of reverberation are not superimposed, indicating that the normalized covariance method is not consistent with the traditional method in accounting for reverberation. Given the success of the traditional STI, this implies that the normalized covariance method will not be a good predictor of intelligibility for additive noise and reverberation. The normalized correlation method comes closer to having a one-to-one relationship to the traditional method, with some divergence at high SNRs. This implies that the normalized correlation method may perform poorly when accounting for the effects of reverberation in quiet and low-noise environments.

While the relationship between the normalized correlation method and the traditional STI is approximately one-to-one, they are not equivalent metrics. In other words, some mapping is required to transform the values produced by the normalized correlation method to values corresponding to the traditional STI. To the extent that a unique mapping does exist for these conditions, the new metric will retain the predictive power of the traditional STI for additive noise and reverberation.

B. Envelope thresholding

Figure 5 shows the effect of envelope thresholding on intermediate modulation metrics used to compute the various speech-based STI methods. Investigating these metrics, rather than the final STI values, is necessary to identify methods that produce invalid results. All of the intermediate modulation metrics have a valid range from zero to one, where zero indicates no preservation of the envelope modulations and one indicates perfect preservation. Values of the intermediate metric greater than one indicate a failure of the corresponding method.

Figures 5(a), (b), and (c) reveal that the original cross-power spectrum methods and the original envelope regression method fail for envelope thresholding. In all three plots,

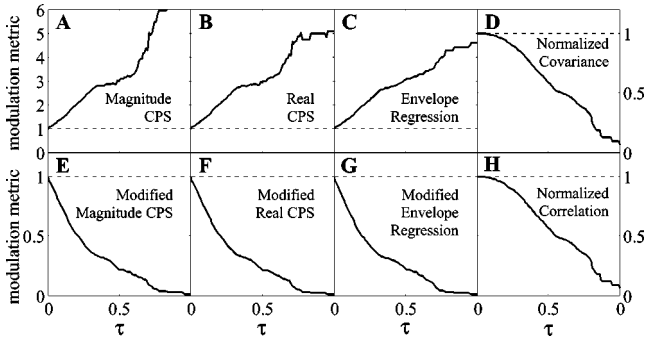

FIG. 5. Intermediate modulation metrics of speech-based STI methods for envelope thresholding as a function of threshold, τ. For the cross-power spectrum (CPS) methods, the intermediate metrics are the MTFs from Eqs. (7) and (8) averaged over modulation frequency. For the envelope regression methods, the intermediate metric is M [Eq. (15)]. For the normalized covariance and normalized correlation methods, the intermediate metrics are r [Eq. (11)] and ρ [Eq. (18)], respectively. All results are for the octave band centered at 1 kHz. The dotted line indicates unity, the maximum valid value for all metrics.

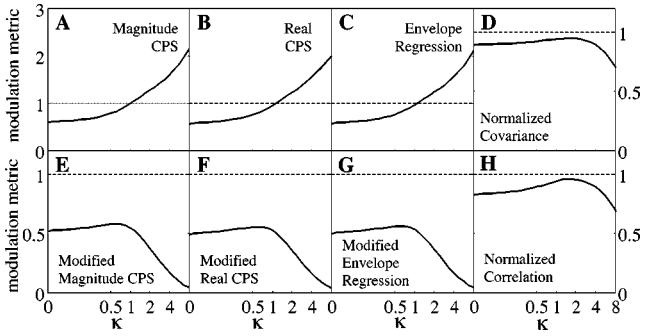

FIG. 6. Intermediate modulation metrics of speech-based STI methods (as in Fig. 5) for spectral subtraction as a function of control parameter, κ. All results are for the octave band centered at 1 kHz. The dotted line indicates unity, the maximum valid value for all metrics.

the modulation metrics increase above one as the threshold increases. These invalid values of the intermediate metrics indicate that these methods are not applicable to the nonlinear operation of envelope thresholding. The remaining five plots reveal that all of the proposed methods [Figs. 5(e)–(h)], as well as the normalized covariance method [Fig. 5(d)], produce valid values of the intermediate metrics. As the threshold increases, all of the intermediate metrics monotonically decrease from an initial value of one.

The general effect of envelope thresholding is to emphasize peaks in the envelope by setting low-amplitude samples of the envelope to zero. As the threshold increases, more samples are set to zero. Because this increases the modulation depth of the envelope, most of the previously proposed speech-based STI methods erroneously interpret this operation as increasing intelligibility beyond the initial value of one for speech in quiet. These methods fail because envelope thresholding reduces the mean of the response envelope, μ_y. Since it is the denominator of the normalization term, α, small values of μ_y can lead to extremely large values of α. Although envelope thresholding also reduces the cross-power spectrum, $S_{xy}(f)$, and cross-covariance, λ_{xy} [which contribute to the numerator of the modulation metrics in Eqs. (7), (8), and (15)], empirical observations indicate that as the threshold increases, these terms decrease more gradually than μ_y, leading to invalid values of the modulation metrics.

The modified methods that use β as the normalization term do not fail in this way because, for envelope thresholding, μ_z varies from zero to μ_x as the threshold goes from 0 to 1, corresponding to values of β ranging from 1 to 0.5 for the full range of envelope thresholding. This causes the intermediate metrics to decrease with increasing threshold.

The results for the three modified methods, as well as the normalized correlation and normalized covariance methods, are qualitatively consistent with the expected effect of envelope thresholding on the intelligibility of speech in quiet. The effect of increasing the threshold is to increase the distortion of the processed signal, thereby making it less intelligible. Increasing the threshold of a slightly different envelope manipulation has been shown to decrease intelligibility (Drullman, 1995). Therefore, the methods that account for envelope thresholding by decreasing as the threshold increases are viable candidates for speech-based STI.

C. Spectral subtraction

Figure 6 shows the effects of spectral subtraction on intermediate modulation metrics used to compute the various speech-based STI methods. Figures 6(a), (b), and (c) reveal that the original cross-power spectrum methods and the original envelope regression method fail for spectral subtraction. In all three plots, the modulation metrics increase monotonically as the control parameter, κ, increases and eventually reach invalid values greater than one. This indicates that these methods are not applicable to spectral subtraction. The remaining five plots reveal that all of the proposed methods [Figs. 6(e)–(h)], as well as the normalized covariance method [Fig. 6(d)], produce valid values of the intermediate metrics. As the control parameter increases, all of the intermediate metrics initially increase to a maximum and then decrease.

The proposed methods as well as the existing normalized covariance method exhibit behavior that is qualitatively consistent with a hypothetical trade-off between noise reduction and signal distortion. For each of these methods, the modulation metric initially increases, predicting slight improvements in intelligibility due to moderate levels of spectral subtraction ($\kappa \approx 1$), and then decreases, predicting degradations in intelligibility for more severe processing ($\kappa > 2$). The modified cross-power spectrum methods and the modified envelope regression method predict the most benefit from spectral subtraction with $\kappa = 0.6$, while the normalized covariance and normalized correlation method predict an optimum value of $\kappa = 1.4$. Further studies are required to determine if the proposed methods predict the intelligibility of speech processed by spectral subtraction and if they account for the effects of musical noise, an unpleasant artifact introduced by spectral subtraction (Goh *et al.*, 1998).

These results imply that spectral subtraction may improve the intelligibility of speech degraded by additive noise. A number of studies have shown that spectral subtraction does not improve the intelligibility of speech for normal-hearing listeners (Lim and Oppenheim, 1979). However,

spectral subtraction has been shown to improve intelligibility for cochlear implant listeners (Weiss, 1993; Hochberg *et al.*, 1992). This is discussed in more detail in Sec. VI B.

VI. DISCUSSION

A. Candidate speech-based STI methods

The results presented in the previous section indicate the suitability of the various speech-based STI methods for predicting intelligibility under conditions of acoustic degradation, envelope thresholding, and spectral subtraction. The long-term goal is to identify and validate a speech-based STI method that accurately predicts intelligibility of speech processed by a wide variety of linear and nonlinear operations. The immediate goal of this study is to identify speech-based STI methods that maintain a one-to-one relationship with the traditional STI for acoustic degradations while also producing qualitatively reasonable results for selected nonlinear operations.

Of the four original methods, only the normalized covariance method exhibited qualitatively reasonable behavior for the nonlinear operations considered in this study. However, this method does not have a one-to-one correspondence to the traditional STI for acoustic degradations. The other three previously proposed methods produce invalid results for the nonlinear operations considered. Therefore, we conclude that none of the four original methods are suitable for both conventional acoustic degradations and nonlinear operations.

The four proposed speech-based STI methods exhibit a one-to-one relationship with the traditional STI for acoustic degradations and produce qualitatively reasonable results for the nonlinear operations. However, the normalized correlation method may be less accurate for predicting the intelligibility of reverberant speech in quiet. Even so, all of the proposed methods are potential candidates to extend the STI to nonlinear operations while retaining its applicability to acoustic degradations. Additional work is required to determine if any of the proposed methods accurately predict speech intelligibility for these and other nonlinear operations.

Substantial differences exist among the four proposed methods. The two cross-power spectrum methods are computed using a modulation transfer function as the intermediate variable for each frequency band, and these MTFs are computed as a function of modulation frequency. In contrast, for the envelope regression and normalized correlation methods, the intermediate metrics consist of a single value for each frequency band and are not computed as functions of modulation frequency. However, it is shown in Appendix B that the normalized correlation method can be expressed as the energy-weighted average of an alternate MTF. The weights applied to the alternate MTF represent the proportion of the total energy in the probe envelope at each modulation frequency. A similar derivation can be performed for the envelope regression method but is complicated by the fact that the intermediate metric is based on covariance rather than correlation.

This interpretation of the normalized correlation and the envelope regression methods as the energy-weighted average of a MTF facilitates comparison with the cross-power spectrum methods. One area of concern relates to nonlinear operations that alter envelope spectra at modulation frequencies above 15 Hz. Such operations will not affect the STI values produced by the cross-power spectrum methods, because those methods only include modulation frequencies up to the one-third-octave band centered at 12.7 Hz. Indeed, there is evidence that modulation frequencies above 16 Hz provide only a marginal contribution to intelligibility (Drullman, 1994a). Because the envelope regression and normalized correlation methods use intermediate metrics that incorporate all frequencies in the envelopes (up to 50 Hz in the current implementation), one might expect these metrics to produce vastly different predictions of intelligibility for alterations in the envelope spectra above 15 Hz. However, since the intermediate metrics can be expressed as the energy-weighted average of a MTF, we must consider how much energy is present at higher modulation frequencies. For typical speech signals, less than 5% of the envelope energy occurs above 15 Hz. As a result, alterations to the envelope spectra above 15 Hz have only minor effects on the STI values produced by the envelope regression and normalized correlation methods.

The normalized correlation method and envelope regression methods can be calculated efficiently because they require estimates of envelope means and variances, which can be computed using running averages or windows of various lengths. The cross-power spectrum methods that calculate the MTF explicitly require at least several seconds of speech in order to estimate power spectra and cross-power spectra with a resolution less than 1 Hz, and calculating these spectra is computationally more intensive than calculating means and variances. Finally, because Figs. 4–6 illustrate that the behavior of the envelope regression method is similar to that of the cross-power spectrum methods, we conclude that the envelope regression method is a more practical choice than the two cross-power spectrum methods.

The normalized correlation method presents a substantial deviation from the traditional STI. The other proposed methods are equivalent to the traditional STI, that is, the speech-base STI values correspond directly to traditional STI values. However, as seen in Fig. 4, the normalized correlation method is not equivalent to the traditional STI, or is it a linear transformation of traditional STI. A (nonlinear) function is required to map the normalized correlation STI values to the traditional STI.

Another difference, illustrated in Figs. 5 and 6, is that the qualitative behavior of the normalized correlation method is substantially different from the other three proposed methods. As mentioned above, additional work is required to determine if any of the proposed methods accurately predict speech intelligibility. Note that although the normalized correlation method uses magnitude envelopes rather than the intensity envelopes used in the other methods, the major differences in qualitative behavior cannot be attributed to this difference in envelope computation. The normalized correlation metric is admittedly a departure from many of the principles of the traditional STI, and it may be preferable to consider it a new intelligibility metric distinct from the STI

except for the common elements of using frequency-band envelopes.

B. Predicting intelligibility of cochlear-implant processed speech

The STI has already been adapted for use with hearing-impaired subjects (Humes *et al.*, 1986; Payton *et al.*, 1994), and it is a good candidate for predicting intelligibility of speech processed by cochlear-implant (CI) speech processors. This expectation is based primarily on similarities between the STI calculation procedure and CI processing strategies; both the STI and conventional CI processing strategies use information from the envelopes in a number of frequency bands and neglect the fine structure. The STI calculation procedures can be tailored to match a particular CI speech processor by matching the frequency bands and method of envelope calculation.

Although the absolute performance of subjects listening to CI-processed speech differs from that of subjects listening to unprocessed speech, additive noise has relatively similar effects in both cases (Hochberg, 1992). Therefore, the STI methods that accurately predict the relative intelligibility among conditions of speech with additive noise (Fig. 4) should also be valid for CI-processed speech with additive noise, although an alternate mapping from STI to percent correct scores may be required for CI-processed speech. It is expected that the same trends will exist for reverberant conditions, although there has been relatively little research assessing the intelligibility of CI-processed speech in reverberation.

The selection of envelope thresholding as a nonlinear operation was guided by our interest in CI-processed speech. Some CI processors use N-of-M processing, coding only a subset (N) of the total (M) frequency-band envelopes during each stimulation cycle (Loizou, 1998). The stimulation cycle is relatively short (a few milliseconds) compared to the STI analysis frame (typically several seconds). The effect of N-of-M processing is comparable to setting the remaining M-N envelopes to zero during intervals when the envelope is not selected. Although this is not identical to envelope thresholding, it has a similar effect on the shape of the envelope, preserving the envelope in intervals where its amplitude is relatively high and eliminating the envelope in intervals where its amplitude is low.

The envelope thresholding results in Fig. 5 indicate that the four proposed methods are potential candidates for predicting the effect of N-of-M processing. If a frequency band is selected all of the time (equivalent to a threshold of 0%), then the intermediate modulation metric is one, contributing a transmission index value (TI_i) of one for that band. If a frequency band is never selected (equivalent to a threshold of 100%), then the intermediate modulation metric is zero and $\mathrm{TI}_i = 0$. If a frequency band is selected intermittently, then the corresponding modulation metric will fall between zero and one, producing a transmission index that reflects that band's partial contribution to intelligibility. While all of the proposed methods are qualitatively correct in that they decrease monotonically from one to zero with increasing threshold, additional work is required to determine which methods, if any, are quantitatively accurate in predicting the effects of envelope thresholding and N-of-M processing on intelligibility.

While research indicates that spectral subtraction does not improve intelligibility for normal-hearing listeners (Lim and Oppenheim, 1979), it has been demonstrated to improve intelligibility for CI users (Weiss, 1993; Hochberg *et al.*, 1992). We hypothesize that this may be related to the effective spectral resolution of the listeners; normal-hearing listeners have relatively fine spectral resolution that permits perception of narrow spectral peaks that rise above the background noise, while CI users are restricted to the relatively broad frequency bands used by their speech processors and therefore cannot perceive spectral peaks within a wider band of noise. As a result, normal-hearing listeners do not benefit from spectral subtraction, since they are already able to listen in relatively narrow bands. On the other hand, CI users benefit from spectral subtraction algorithms that operate in frequency bins substantially narrower than the broader bands used by their speech processors. A related interpretation is that by suppressing narrow frequency bands with low SNR, spectral subtraction removes noise from the broadband temporal envelope, an improvement that provides greater benefit to CI users than to normal hearing listeners. The spectral subtraction results in Fig. 6 indicate that the four proposed methods are potential candidates for predicting the effect of spectral subtraction on CI-processed speech. The intermediate metrics indicate that the proposed STI methods will predict an improvement for speech processed with spectral subtraction algorithms using moderate values of the control parameter, κ. It appears that an appropriate speech-based STI may predict the effect of spectral subtraction on intelligibility more accurately for CI-users than for normal-hearing listeners precisely because it uses a broad frequency-band analysis similar to that used by CI speech processors. In fact, the success of the traditional STI for normal-hearing listeners may be due to the historic focus on broadband distortion such as reverberation and additive broadband noise. For example, consider the case of speech corrupted by a pure tone. This specialized interference would have little or no effect on intelligibility for normal-hearing listeners, but would have a detrimental effect on intelligibility when passed through a CI-speech processor. In computing the STI, the effect of the pure tone would also show up in the apparent SNR for the corresponding frequency band, so that the STI would better predict the effect on intelligibility for CI-processed speech than for a normal-hearing listener.

C. Alternate intelligibility metrics

Because these quantities can be calculated on arbitrarily small speech segments, this raises the possibility of calculating the STI on phoneme-length segments. Traditionally, STI has focused on long-term effects; however, focusing on short segments could prove useful in a number of areas. For example, researchers have studied the effect of mutual independence of adjacent frequency bands based on long-term averages (Steeneken and Houtgast, 1999). However, mutual information may be modeled more accurately using short time segments that carry information concerning the fluctu-

ating short-term SNR. Incorporating short-term averages could potentially lead to speech-based STI metrics that use mutual information from neighboring frequency bands on a phonemic level rather than a global level.

Another approach to combining spectral and temporal information is the physiologically motivated spectro-temporal modulation index (STMI; Elhilali *et al.*, 2003). The STMI is based on an auditory model (Chi *et al.*, 1999) and quantifies the difference in the auditory model output between clean and degraded speech. It operates along spectral and temporal dimensions jointly and explicitly accounts for changes in spectro-temporal modulations. The STMI has been shown to be comparable to the traditional STI for additive noise and reverberation. In addition, for nonlinear distortions consisting of phase jitter or phase shifts, the STMI tracks subject performance on intelligibility tests, while the traditional STI does not. Both the STMI and the methods proposed in this work seek to extend the traditional STI to nonlinear operations. In order to compare these two approaches, future investigations should assess the ability of both the STMI and the proposed metrics to capture the effects of a wide variety of nonlinear operations that includes envelope thresholding, spectral subtraction, phase jitter, and phase shifts.

VII. CONCLUSIONS

The main conclusions of this study follow.

(1) None of the four original speech-based STI methods are suitable for both conventional acoustic degradations and nonlinear operations.
(2) All four of the proposed speech-based STI methods produce reasonable results for conventional acoustic degradations, although preliminary evidence suggests that the normalized correlation method may predict intelligibility less accurately than the other methods for reverberant speech in quiet. All four proposed methods produce qualitatively reasonable results for the nonlinear operations considered in this study. Additional work is required to determine if any of the proposed methods accurately predict speech intelligibility for these and other nonlinear operations.
(3) The normalized correlation and envelope regression methods are computationally less complex than the two cross-power spectrum methods and therefore offer the possibility of computing STI on a short-term (phonemic) level. The envelope regression method is preferred over the two cross-power spectrum methods, because it produces comparable results with less computational complexity.
(4) Of the proposed methods, the normalized correlation method represents the most substantial deviation from the traditional STI. Because it produces results that are qualitatively different from the other methods, it provides an important alternative for fitting data from future speech intelligibility studies.
(5) The proposed speech-based STI methods offer the potential to predict the intelligibility of CI-processed speech.

ACKNOWLEDGMENTS

This work was supported by the National Institute of Deafness and Other Communicative Disorders under Grant Nos. 1-R01-DC00117 and 5-T32-DC0038. The authors are grateful to Karen Payton for her helpful comments on an earlier version of this paper.

APPENDIX A: ALTERNATE FORM OF THE ENVELOPE REGRESSION METHOD

The following is a stochastic reformulation of the envelope regression method (Sec. II B 3) that facilitates comparison with other methods. It begins with the assumption that the linear regression of the sampled response envelope, $y[n]$, onto the sampled probe envelope, $x[n]$, is performed using a minimum mean square error criterion (Ross, 1998). In this case, the optimal fit is

$$y_{\mathrm{MMSE}}[n]=\mu_y+\frac{\lambda_{xy}}{\lambda_x}(x[n]-\mu_x), \tag{A1}$$

where λ_{xy} and λ_x are defined in Eqs. (12) and (13). Thus, the slope (A) and the y-intercept (B) calculated using a minimum mean square error criterion are

$$A=\frac{\lambda_{xy}}{\lambda_x}, \tag{A2}$$

and

$$B=\mu_y-\frac{\lambda_{xy}}{\lambda_x}\mu_x. \tag{A3}$$

Substituting Eqs. (A2) and (A3) into Eq. (9) and rearranging allows the apparent SNR to be expressed as

$$\mathrm{aSNR}=10\log_{10}\left(\frac{M}{1-M}\right), \tag{A4}$$

where M is a modulation metric defined as

$$M=\frac{\mu_x}{\mu_y}\frac{\lambda_{xy}}{\lambda_x}. \tag{A5}$$

APPENDIX B: NORMALIZED CORRELATION METHOD EXPRESSED AS AN ENERGY-WEIGHTED MTF

Equation (18) defines the normalized correlation as

$$\rho=\frac{\phi_{xy}}{\sqrt{\phi_x\phi_y}}. \tag{B1}$$

Using the relationship between the cross-correlation function, $R_{xy}[k]$, and the cross-power spectrum, $S_{xy}(f)$, (Papoulis, 1984) together with the observation that ϕ_{xy} equals the cross-correlation function computed at zero lag, yields

$$\phi_{xy}=R_{xy}[0]=\int_{f=-1/2}^{1/2}S_{xy}(f)\,df, \tag{B2}$$

where $\phi_{xy}\triangleq E\{x[n]y[n]\}$ and $R_{xy}[k]\triangleq E\{x[n]y[n-k]\}$. The normalized correlation can then be expressed as

$$\rho=\frac{1}{\sqrt{\phi_x\phi_y}}\int_{f=-1/2}^{1/2}S_{xy}(f)\,df. \tag{B3}$$

Bringing the denominator inside the integral and multiplying numerator and denominator by the same terms yields

$$\rho = \int_{f=-1/2}^{1/2} \sqrt{\frac{\phi_x}{\phi_y}} \left[\frac{S_{xy}(f)}{S_{xx}(f)} \right] \left[\frac{S_{xx}(f)}{\phi_x} \right] df. \tag{B4}$$

Defining a new MTF,

$$\mathrm{MTF}_\rho(f) \triangleq \sqrt{\frac{\phi_x}{\phi_y}} \frac{S_{xy}(f)}{S_{xx}(f)}, \tag{B5}$$

and a weighting function,

$$W(f) \triangleq \frac{S_{xx}(f)}{\phi_x}, \tag{B6}$$

allows describing ρ as an energy-weighted average of this new MTF, that is,

$$\rho = \int_{f=-1/2}^{1/2} \mathrm{MTF}_\rho(f) W(f) df. \tag{B7}$$

The weighting function, $W(f)$, is the ratio of the power of the probe envelope at each modulation frequency to the total power in the probe envelope.

The MTF defined in Eq. (B5) is similar in form to the MTFs defined for the cross-power spectrum methods in Eqs. (7) and (8). All three MTFs are based on the normalized ratio of the cross-power spectrum between probe and response envelopes to the power spectrum of the probe envelope. The main differences are the factor used for normalization ($\sqrt{\phi_x/\phi_y}$ rather than $\alpha = \mu_x/\mu_y$) and the fact that in Eq. (B5) the MTF is complex-valued. However, since $S_{xx}(f)$ is real and symmetric, and $S_{xy}(f)$ is complex-conjugate symmetric, the integral over equal ranges of positive and negative frequencies will be real-valued.

[1]Motivation for considering the normalized correlation comes in part from studies of binaural detection (Bernstein and Trahiotis, 1996), which have shown that the normalized correlation, ρ, is a better indicator of performance than the normalized covariance, r. By including the envelope means, the metric accounts for the average envelope power as well as the envelope fluctuations. While binaural detection is clearly different than speech intelligibility, it is possible that in both cases the auditory system utilizes the additional information about average envelope power provided by the normalized correlation.

Allen, J. B., and Berkley, D. A. (**1979**). "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943–950.

Bernstein, L. R., and Trahiotis, C. (**1996**). "The normalized correlation: Accounting for binaural detection across center frequency," J. Acoust. Soc. Am. **100**, 3774–3784.

Chi, T., Gao, M., Guyton, M. C., Ru, P., and Shamma, S. A. (**1999**). "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc. Am. **106**, 2719–2732.

Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592.

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

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

Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun. **41**, 331–348.

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

Goh, Z., Tan, K. C., and Tan, B. T. G. (**1998**). "Postprocessing method for suppressing musical noise generated by spectral subtraction," IEEE Trans. Speech Audio Process **6**, 287–292.

Hochberg, I., Boothroyd, A., Weiss, M., and Hellman, S. (**1992**). "Effects of noise and noise suppression on speech perception by cochlear implant users," Ear Hear. **13**, 263–271.

Hohmann, V., and Kollmeier, B. (**1995**). "The effect of multichannel dynamic compression on speech intelligibility," J. Acoust. Soc. Am. **97**, 1191–1195.

Holube, 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. J. M. (**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.

Humes, L. E., Dirks, D. D., Bell, T. S., Ahlstrom, C., and Kincaid, G. E. (**1986**). "Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners," J. Speech Hear. Res. **29**, 447–462.

IEC (**1998**). *Sound System Equipment—Part 16: Objective rating of speech intelligibility by Speech Transmission Index*, 2nd ed., International Standard No. 60268-16.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE, NY.

Koch, R. (**1992**). "Gehörgerechte Schallanalyse zur Vorhersage und Verbesserung der Sprachverständlichkeit," ("Auditory sound analysis for the prediction and improvement of speech intelligibility"), Dissertation, Universität Göttingen.

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.

Lim, J. S., and Oppenheim, A. V. (**1979**). "Enhancement and bandwidth compression of noisy speech," Proc. IEEE **67**, 1586–1604.

Loizou, P. C. (**1998**). "Mimicking the human ear," IEEE Signal Process. Mag. **15**, 101–130.

Ludvigsen, C. (**1987**). "Prediction of speech intelligibility for normal-hearing and cochlear hearing-impaired listeners," J. Acoust. Soc. Am. **82**, 1162–1171.

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.

Papoulis, A. (**1984**). *Probability, Random Variables, and Stochastic Processes* (McGraw–Hill, New York), pp. 263–293.

Payton, K. L., and Braida, L. D. (**1999**). "A method to determine the speech transmission index from speech waveforms," J. Acoust. Soc. Am. **106**, 3637–3648.

Payton, K. L., Braida, L. D., Chen, S., Rosengard, P., and Goldsworthy, R. (**2002**). "Computing the STI using speech as a probe stimulus," *Past, Present and Future of the Speech Transmission Index* (TNO Human Factors, Soesterberg, The Netherlands), pp. 125–138.

Payton, K. L., Uchanski, R. M., and Braida, L. D. (**1994**). "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," J. Acoust. Soc. Am. **95**, 1581–1592.

Ross, S. (**1998**). *A First Course in Probability*, 5th ed. (Prentice Hall, Englewood Cliffs, NJ), pp. 350–354.

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

Steeneken, H. J. M., and Houtgast, T. (**1982**). "Some applications of the speech transmission index (STI) in auditoria," Acustica **51**, 229–234.

Steeneken, H. J. M., and Houtgast, T. (**1999**). "Mutual dependence of the octave-band weights in predicting speech intelligibility," Speech Commun. **28**, 109–123.

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

Weiss, M. R. (**1993**). "Effects of noise and noise reduction processing on the operation of the Nucleus-22 cochlear implant processor," J. Rehabil. R. D**30**, 117–128.