Comparison of a short-time speech-based intelligibility metric to the speech transmission index and intelligibility data[a)]

Karen L. Payton[b)] and Mona Shrestha
ECE Department, University of Massachusetts Dartmouth, 285 Old Westport Road, North Dartmouth, Massachusetts 02747

(Received 31 August 2012; revised 4 August 2013; accepted 30 August 2013)

Several algorithms have been shown to generate a metric corresponding to the Speech Transmission Index (STI) using speech as a probe stimulus [e.g., Goldsworthy and Greenberg, J. Acoust. Soc. Am. **116**, 3679–3689 (2004)]. The time-domain approaches work well on long speech segments and have the added potential to be used for short-time analysis. This study investigates the performance of the Envelope Regression (ER) time-domain STI method as a function of window length, in acoustically degraded environments with multiple talkers and speaking styles. The ER method is compared with a short-time Theoretical STI, derived from octave-band signal-to-noise ratios and reverberation times. For windows as short as 0.3 s, the ER method tracks short-time Theoretical STI changes in stationary speech-shaped noise, fluctuating restaurant babble and stationary noise plus reverberation. The metric is also compared to intelligibility scores on conversational speech and speech articulated clearly but at normal speaking rates (Clear/Norm) in stationary noise. Correlation between the metric and intelligibility scores is high and, consistent with the subject scores, the metrics are higher for Clear/Norm speech than for conversational speech and higher for the first word in a sentence than for the last word. [http://dx.doi.org/10.1121/1.4821216]

PACS number(s): 43.71.Gv, 43.55.Hy [AA] Pages: 3818–3827

I. INTRODUCTION

The Speech Transmission Index (STI) is a physical metric that has been developed to predict the intelligibility of speech in the presence of noise and/or reverberation (IEC, 1998, 2003, 2011). The STI is based on a weighted average of metrics derived from envelope signals in multiple frequency bands spanning the speech spectrum. The STI differs from other intelligibility metrics such as the Speech Intelligibility Index (SII) in that it measures reduction in signal envelope intensity modulations rather than band-specific signal-to-noise ratios (SNRs) to predict intelligibility (ANSI, 1997). One enhancement of the SII over its predecessor metric, the Articulation Index (AI), is that its estimate of the intelligibility in the presence of reverberation is very similar to the STI approach using room reverberation time (RT). The primary advantage of the STI is that, by using intensity envelope modulations, it can predict the effects of reverberation and distinct echoes in addition to interfering noise without requiring explicit knowledge about reverberation times or signal to noise ratios (Houtgast and Steeneken, 1973).

Several methods have been proposed to compute the STI. The traditional method, originally proposed by Houtgast and Steeneken and incorporated into the IEC standard, measures changes in envelope modulation depth using intensity-modulated noise as the excitation/source signal (Houtgast and Steeneken, 1973, 1980, 1985; IEC, 1998, 2003, 2011). Other methods have been proposed which use speech as the excitation signal rather than artificially modulated noise (Ludvigsen *et al.*, 1990; Ludvigsen, 1993; Ludvigsen *et al.*, 1993; Drullman *et al.*, 1994; Drullman, 1995; Payton and Braida, 1999; Payton *et al.*, 2002; Goldsworthy and Greenberg, 2004; Payton and Shrestha, 2008a,b). Using speech as an excitation signal allows one to compute the intelligibility of a talker during a live performance or other situation not amenable to modulated noise probe signals. Most of these speech-based techniques have been shown to provide nearly the same result as the traditional STI and as the "indirect method," described in the most recent IEC standard, in which the STI is obtained from weighted signal-to-noise ratios (SNRs) in seven octave bands and the Fourier transform of the squared room impulse response (IEC, 2011).

To date, almost all approaches to compute the STI have used excitation signals lasting at least a minute or two and generate a metric correlated with long-term speech intelligibility. For example, the Magnitude Cross-Power Spectrum (MCPS) method (Payton *et al.*, 2002), requires approximately 100 s of speech to estimate intelligibility due to acoustically degraded environments. Recently Payton and Shrestha (2008a,b) demonstrated the feasibility of a speech-based STI metric to predict changes in intelligibility over much shorter time intervals.

There have been efforts to develop other short-time speech intelligibility predictors. The Articulation Index (AI) and Speech Intelligibility Index (SII) have been modified to compute short-time intelligibility to help predict

[a)]Portions of this work were presented in "Analysis of short-time speech transmission index algorithms" Proceedings of Acoustics'08 Conference, Paris, France June 30, 2008 and "Evaluation of short-time speech-based intelligibility metrics" Proceedings of Int. Comm. Bio. Effects of Noise Conference, Foxwoods, CT, 23 July 2008.

[b)]Author to whom correspondence should be addressed. Electronic mail: kpayton@umassd.edu

performance of noise reduction algorithms (Kates, 1987) or effects of fluctuating noise (Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006). Those approaches used speech-shaped Gaussian noise instead of actual speech to approximate signal levels and used analysis windows as short as 6.4 ms. Boldt and Ellis (2009) developed a metric based on the correlation of a time-frequency binary mask of the clean speech envelope with a time-frequency representation of the degraded speech envelope to estimate intelligibility of speech in fluctuating and stationary background environments using 80 ms windows. Ma *et al.* (2009) studied the performance of several speech-based metrics, including the STI-based Normalized Covariance Method, using 20–30 ms frames of speech degraded by realistic environmental noises. Falk *et al.* (2010) used 256 ms windows to compute a metric based on modulation spectra vs filterbank center frequency of reverberant/processed speech to predict the performance of dereverberation algorithms. Taal *et al.* (2011) proposed a metric, similar to a short-time implementation of the Normalized Correlation metric described in Goldsworthy and Greenberg (2004), which used 384 ms windows to analyze nonlinearly processed speech. In 2012, Schlesinger (2012) created a "transient-based STI" using 400 ms analysis windows to predict the intelligibility of nonlinearly processed speech that built upon the preliminary results of Payton and Shrestha (2008a).

Clearly there is significant interest in short-time intelligibility metrics that can track speech intelligibility in time-varying environments, whether those environments are due to fluctuating backgrounds or speech enhancement algorithms. For a metric such as the STI, which has already been established as a standard, one necessary demonstration for any short-time version is that it approach the long-term results when averaged over many frames unless there is evidence that the long-term STI either over or under estimates intelligibility in a condition and the short-term version is a better fit to listener performance. In addition, a short-time metric should track speech intelligibility on the scale (window length) over which it is computed.

The current work examines both issues. It compares short-time predictions of one speech-based STI technique, the Envelope Regression (ER) metric, to short-term theoretical STI results. The ER metric is based on the speech-based STI method proposed by Ludvigsen *et al.* (1990) then reformulated by Goldsworthy and Greenberg (2004) who also demonstrated that the long-term characteristics of the metric were almost identical to the long-term STI for noisy and/or reverberant environments. The short-time theoretical STI is calculated using a frame-based application of the IEC indirect method, herein referred to as the Theoretical Method. Frame-level and averaged results are evaluated as a function of window length for environments degraded by stationary speech-shaped noise, fluctuating restaurant babble or stationary speech-shaped noise plus reverberation. Linear regression analyses of the short-time ER results vs the Theoretical Method are performed where computations for both metrics are made over the same time frame. The averaged results are also compared to long-term STI, as determined by the MCPS technique (Payton *et al.*, 2002). Finally, the ER metric results are compared to listener intelligibility scores for two talkers, speaking both conversationally and clearly at normal speaking rates in the presence of stationary speech-shaped noise.

II. METHODS

A. Stimuli

The stimuli used in this study were three sets of 50 nonsense sentences from the corpus of Picheny *et al.* (1985) that had been digitized at a 20 kHz sampling rate. Nonsense sentences are grammatically correct but do not provide any semantic context to aid word identification, e.g., "His *guests could teach* his *turnpike*." Each sentence consisted of four to six key words (italicized in example) consisting of the nouns, adjectives, and verbs in the sentence. The set used in Secs. III A to III C was spoken conversationally by a male talker (Payton *et al.*, 1994). The second and third sets, used in Sec. III D, were spoken both conversationally (Conv) and articulated clearly but at normal speaking rates (Clear/Norm) by a different male talker and by a female talker, respectively (Krause, 2001; Krause and Braida, 2002, 2004).

B. Degradation conditions

Three types of environmental degradations were considered: Stationary speech-shaped noise, speech-shaped noise plus simulated reverberation and fluctuating restaurant babble. The speech-shaped noise was generated by filtering white Gaussian noise sequences to approximate the average long-term spectra of speech. The restaurant babble consisted of multi-talker babble plus random impulsive dish and utensil noises sampled at 20 kHz.

In Sec. III A, the speech-shaped noise was added to the speech at an average SNR of 0 dB. For the noise plus reverberation condition in Sec. III B, speech plus noise at 0 dB SNR was convolved with a simulated conference room impulse response ($T_{60} = 0.6$ s) (Peterson, 1986; Payton *et al.*, 1994). The restaurant babble was added to the speech at 0 dB SNR in Sec. III C. In Sec. III D, stationary speech-shaped noise was added to the speech at an average SNR of −1.8 dB to match the listening conditions reported in Krause (2001; Krause and Braida, 2002, 2004).

C. Processing steps

Figure 1 depicts a block diagram of the signal processing steps used to obtain the speech-based STI values. The clean and degraded signals were separately filtered using a bank of six sixth-order octave-wide Butterworth band-pass filters with center frequencies from 125 Hz to 4 kHz and a sixth-order Butterworth high-pass filter with a cutoff frequency of 6 kHz to extract the 8 kHz band. For each frequency band, i, intensity envelopes were extracted from the clean and the degraded signals by squaring and low-pass filtering with a cutoff frequency of 50 Hz. The lowpass filter was implemented using a 10 ms (200 point) Hamming window as the impulse response in order to have a constant group delay and to avoid negative-valued envelopes. Frequencies below 50 Hz were attenuated by the filter less

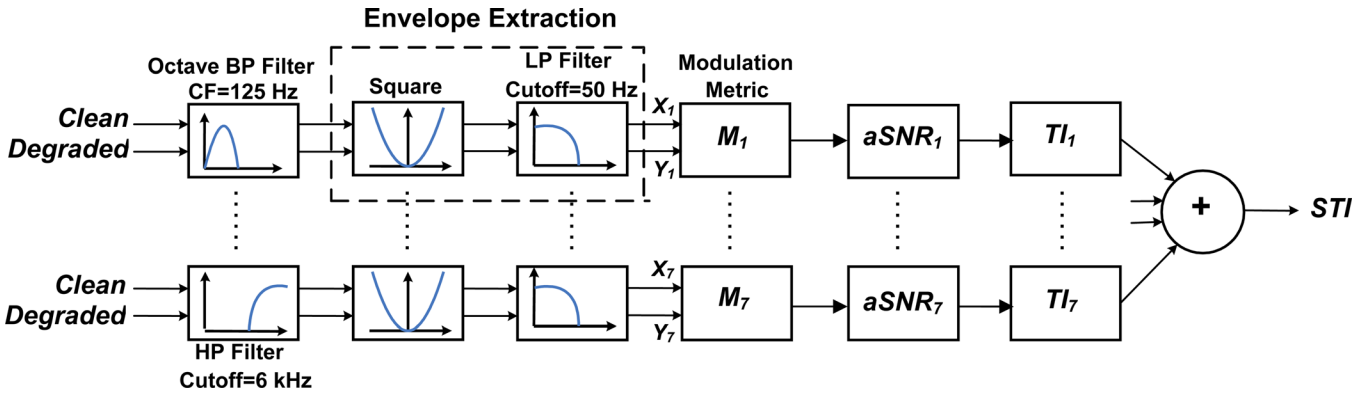

FIG. 1. (Color online) Block diagram of signal processing steps necessary to compute the speech-based intelligibility metric.

than 2 dB, the first null occurred at 203 Hz and the side lobe peaks were at least −42 dB below the main lobe peak.

The clean and degraded intensity envelopes, $x_i(k)$ and $y_i(k)$, respectively, were down-sampled by a factor of 49 to 408 Hz. The new sampling rate was selected to decrease computation time without attenuating envelope frequencies below 50 Hz due to the anti-aliasing filter. Next, for each octave band, a modulation metric, M_i, was calculated based on a comparison of the clean and degraded intensity envelopes.

For the Envelope Regression (ER) method, each band's modulation metric, M_i, was computed from the clean and degraded envelope signals using Eq. (1):

$$M_i = \frac{\mu_{xi}}{\mu_{yi}} \frac{\frac{1}{N}\sum_{k=1}^{N}[x_i(k)y_i(k)] - \mu_{xi}\mu_{yi}}{\frac{1}{N}\sum_{k=1}^{N}[x_i(k)]^2 - (\mu_{xi})^2} \tag{1}$$

where μ_{xi} and μ_{yi} are the means of the clean and degraded intensity envelopes $x_i(k)$ and $y_i(k)$, respectively. This equation is a short-time implementation of the ER algorithm proposed in Goldsworthy and Greenberg (2004). The variable N corresponds to the rectangular window length used, from 43 656 (107 s, the length of 50 concatenated sentences) down to 32 (78 ms) for the analyses presented in Secs. III A to III C below. Except for the 107 s window, windows were overlapped by 75%. The analyses in Sec. III D, with conversational and Clear/Norm sentences, were performed with windows equal to either individual sentence lengths or keyword lengths.

Once the modulation metrics were computed, the apparent signal-to-noise ratio (SNR) in each band, $a\text{SNR}_i$, was computed based on the IEC standard (IEC, 1998, 2003, 2011) as

$$a\text{SNR}_i = 10\log_{10}\left(\frac{M_i}{1-M_i}\right) \tag{2}$$

where the results were clipped to ±15 dB then converted to a transmission index, TI_i:

$$\text{TI}_i = \frac{a\text{SNR}_i + 15}{30}. \tag{3}$$

Finally, the overall STI value (ranging from 0 to 1) was calculated as a weighted sum of the TI_i values:

$$\text{STI} = \sum_{i=1}^{7}\alpha_i\text{TI}_i - \sum_{i=1}^{6}\beta_i\sqrt{\text{TI}_i \times \text{TI}_{i+1}} \tag{4}$$

where the α_i's represent the octave band weighting factors and the β_i's represent the redundancy correction factors specified in the IEC standard (IEC, 1998, 2003, 2011).

D. Reference methods

In order to compare the ER method with the "true" STI, the long-term STI, was computed using the speech-based Magnitude Cross Power Spectrum (MCPS) method (Payton *et al.*, 2002). Also a short-term Theoretical Method based on the IEC standard's indirect method (IEC, 2011) was also calculated.

The first set of 50 sentences were concatenated in order to compute the long-term STI. A degraded version was created for each condition in Secs. III A to III C as described in Sec. II B. Clean and degraded envelopes were computed as described above. Auto-power spectra and cross-power spectra for each band, i, were estimated using Welch's averaged, modified periodogram method with 4096-point FFTs using Hamming windows and 50% overlap. The resulting 0.1 Hz wide frequency bins of each spectrum were summed across one-third octave intervals centered from 0.315 to 25 Hz. Equation (5) was used to calculate the Modulation Transfer Function (MTF) as a function of interval frequency, F,

$$M_i(F) = \frac{\mu_{xi}}{\mu_{yi}}\frac{|S_{xyi}(F)|}{|S_{xxi}(F)|} \tag{5}$$

where μ_{xi}, μ_{yi}, $x_i(k)$ and $y_i(k)$ are defined above and $S_{xyi}(F)$ and $S_{xxi}(F)$ are the third-octave cross- and auto-power spectra. The long-term STI was computed by substituting $M_i(F)$ for M_i in Eq. (2). The resulting variable, $a\text{SNR}_i(F)$, was averaged across F after clipping to ±15 dB to obtain $a\text{SNR}_i$. The long-term STI then was computed using Eqs. (3) and (4).

The short-term Theoretical Method STI was calculated over the same window lengths as the ER metric. The speech and the noise (as opposed to the degraded speech) were

passed separately through the octave-band filter bank shown in Fig. 1 and, rather than extracting intensity envelopes, within-band powers were used to obtain the signal-to-noise ratio [SNR_i in Eq. (6)] in each band. The modulation index in each band, $M_i(F)$, was then calculated as specified by the most recent version of the IEC standard for the "indirect method" (IEC, 2011):

$$M_i(F) = \left(\frac{\left| \sum_{0}^{N-1} h_i(n)^2 e^{-j2\pi F n} \right|}{\sum_{0}^{N-1} h_i(n)^2} \right) (1 + 10^{(\mathrm{SNR}_i)/10})^{-1}. \quad (6)$$

The first parenthetical term in Eq. (6) estimates the modulation reduction due to reverberation in band i and uses the simulated room impulse response, $h(n)$, of length N, filtered by the ith octave filter. The variable F corresponds to modulation frequency (0.63 to 25 Hz). The second parenthetical term in Eq. (6) estimates the modulation reduction due to additive noise in the analysis window in band i where SNR_i is the signal-to-noise ratio in the ith band (in dB). The Theoretical Method for each window was computed by substituting $M_i(F)$ for M_i in Eq. (2) then following the steps described for the long-term STI above.

III. RESULTS

In Secs. III A to III C, the ER metric results are compared to the Theoretical Method for three degradation condition as functions of window length. Linear regression analyses for the metric vs the Theoretical Method results are examined. Results for two window lengths are presented for the linear regression analyses. The 0.3 s window results are typical of all the longer windows. The 78 ms window results demonstrate metric behavior for window lengths that are too short to track the Theoretical Method, particularly during silent intervals. In addition, ER and Theoretical Method averages when different window lengths are used are compared to the long term STI.

In Sec. III D, regression analyses for the ER metric results vs the Theoretical Method are presented for window lengths matched to each sentence or to each keyword for the two talkers who spoke both conversationally (Conv) and clearly at normal speaking rates (Clear/Norm). Also, the resulting metrics are compared to average listener intelligibility scores at both the sentence and word level. Finally, changes in metric results and intelligibility due to speaking style are evaluated at the sentence level and at the word level, based on word position within sentences.

A. Zero dB SNR with stationary speech-shaped noise

In the first experiment, stationary speech-shaped noise was added to concatenated sentences at 0 dB SNR. Figure 2 depicts the Theoretical Method and ER metric as functions of time for 2 s of speech mixed with noise. Each panel corresponds to the metrics calculated using the indicated window length.

As depicted by the dashed line in each panel, an SNR of 0 dB corresponds to a long-term STI value of approximately 0.5 (the exact value depends on the spectral characteristics

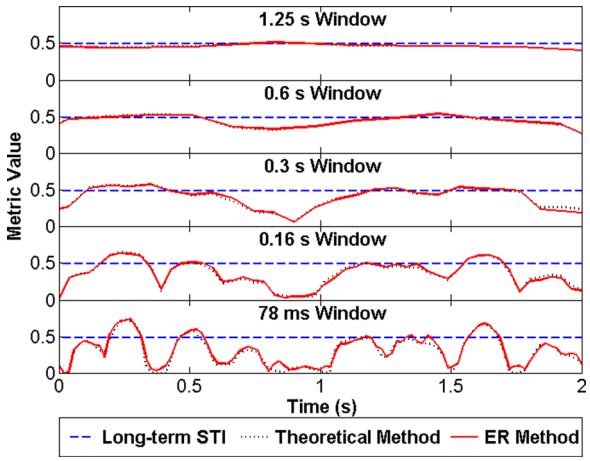

FIG. 2. (Color online) Metrics vs time for several window lengths in the 0 dB SNR stationary speech-shaped noise condition. Predictions for 2 s of speech (~2 sentences) are shown. Different curve types represent: The long-term STI determined by the MCPS method (dashed), Theoretical method (dotted), and ER method (solid).

of the speech and noise). For windows equal to the entire corpus of 107 s (not shown) both the Theoretical and ER methods matched the long-term STI value of 0.49, computed using MCPS method. As window length was shortened, the Theoretical Method tracked the short-term fluctuations in SNR. The ER method generally matched local fluctuations in the Theoretical Method for each window length.

Once window length was decreased to 78 ms (bottom panel), the ER method frequently deviated from the Theoretical Method, particularly during low SNR intervals, such as pauses in the speech and occasionally during higher SNR intervals. In windows where the Theoretical Method was zero because only noise was present (around 0.4, 0.8, 1, and 1.75 s), the ER method often generated non-zero results. In some cases, the ER metric was higher during pauses than it was during windows with higher Theoretical values (e.g., at 0.8 s the ER method is higher than it is during the interval between 0.8 and 1 s).

Regression analyses were performed to compare the short-time ER results to the Theoretical Method on a window-by-window basis for two window lengths: 0.3 s and 78 ms. Figure 3 summarizes the regression results for the three degradation conditions evaluated in this manner. Each row represents a different acoustic condition. A separate data point was plotted for the metrics computed on each windowed segment. This resulted in 1364 points in the left panels (0.3 s windows) and 5467 points in the right panels (78 ms windows). Regression lines, standard deviation bars, and the goodness-of-fit (R^2) statistics are also shown for each window length. The standard deviations were calculated after subtracting the linear regression values. The R^2 statistic measures the proportion of the observed variations around the mean that can be explained by the regression fit. The closer R^2 is to 1, the greater the degree of association between the two metrics. If all of the variation can be explained by the mean, then $R^2 = 0$.

Figure 3(a) depicts the correspondence of the ER method to the Theoretical method for speech plus stationary

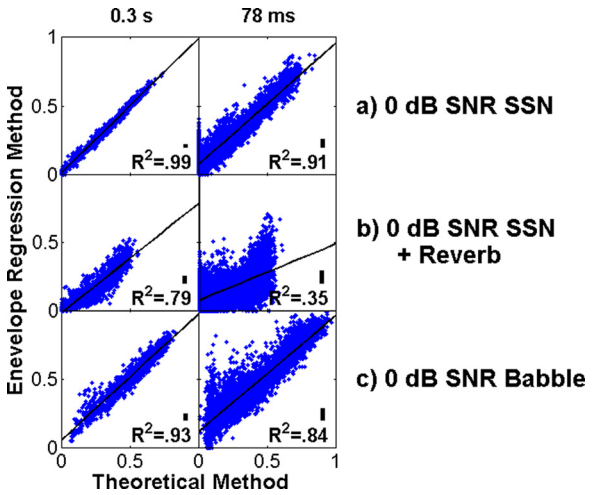

FIG. 3. (Color online) ER Metrics computed from 0.3 s windows (left column) and 78 ms windows (right column) vs Theoretical Method for (a) 0 dB SNR stationary speech-shaped noise, (b) 0 dB SNR stationary speech-shaped noise plus reverberation, and (c) 0 dB SNR fluctuating restaurant babble. When the window length was 0.3 s, 1364 window frames were available for analysis. When the window length was 78 ms, there were 5467 window frames. The solid lines represent best linear fits to the data, the vertical bars to the right in each square indicate 1 standard deviation from the best linear fit and the R^2 statistics indicate the goodness of fit.

speech-shaped noise (SSN) at 0 dB SNR. The R^2 statistic of 0.99 and standard deviation of 0.02 in the left frame indicates the ER metric values are highly predictive of the corresponding Theoretical Method values for the 0.3 s window segments. The slope and intercept of the best-fit line are 0.985 and 0.01, respectively, indicating the two metrics are, on average, almost identical. The fit of the data to a straight line is also very good for the 78 ms windowed segments ($R^2 = 0.91$ and standard deviation of 0.06) whose data are presented in the right frame. Note though, for this shorter window, when the Theoretical Method is zero there are many ER values greater than zero along the y axis. The measured slope and intercept of 0.89 and 0.07, respectively, reflect this asymmetry. The data points along the y axis correspond to the intervals mentioned above in reference to Fig. 2 where the ER values were greater than the Theoretical method during pauses. The Theoretical Method will be zero whenever the talker pauses. The fact that all the corresponding ER values are not zero during the pauses indicates the ER method does not track the Theoretical Method well during the silent intervals when the window is this short. Preliminary analysis indicates that the ER value for these short windows tracks the noise envelope in the window when the clean signal is very small or zero.

In order to study how well the averaged short-time metrics correspond to the long-term STI, the ER method and the Theoretical Method for a given window length were averaged over the entire speech corpus (107 s). The averages are plotted in Fig. 4(a) as functions of window length. For comparison, the long-term STI is also plotted for this condition.

It can be seen that ER method produced the same average value as the Theoretical Method for window lengths greater than about 0.3 s and that both agree with the long-term STI for windows longer than 10 s. The ER averages decreased noticeably relative to both the long-term and

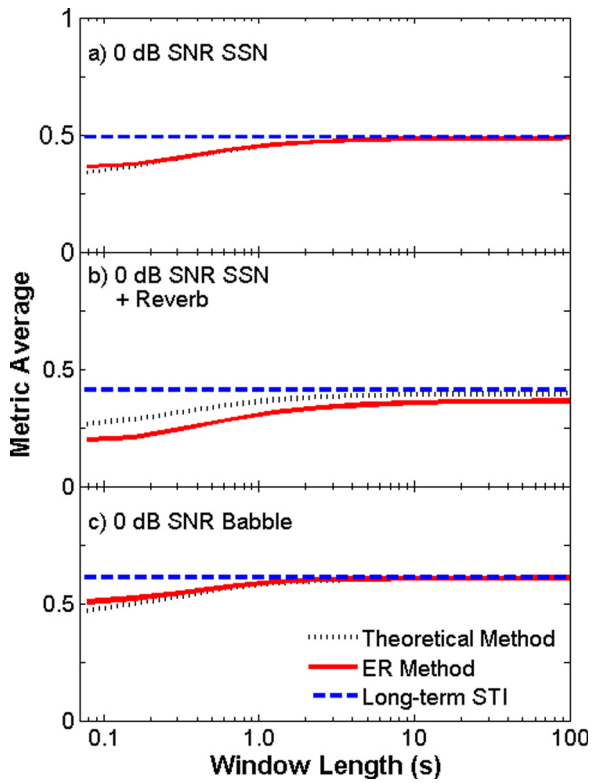

FIG. 4. (Color online) Metric averages computed over entire speech corpus, plotted as functions of analysis-window length for speech in (a) 0 dB SNR stationary speech-shaped noise, (b) 0 dB SNR stationary speech-shaped noise plus reverberation, and (c) 0 dB SNR restaurant babble. The dotted line corresponds to the Theoretical Method average, the solid line to the ER method and the dashed line to the long-term STI as calculated using the MCPS method.

Theoretical STI as the window was decreased below about 1 s. This is because voiced speech segments dominate the metric results for longer windows. When the windows are shortened such that some contain primarily unvoiced and/or silent intervals, the metric results for those windows are much closer to zero, pulling the average down. The leftmost data points are for the 78 ms window. For that window length, the ER average did not decrease quite as much as the Theoretical Method average relative to the next window length of 0.157 s.

B. Zero dB SNR stationary speech-shaped noise plus reverberation

When the noisy speech was convolved with the reverberant room impulse response ($T_{60} = 0.6$ s), the ER method generated values that varied more widely when compared to the Theoretical Method. In Fig. 3(b), ER results are plotted versus the Theoretical Method for the two window lengths. As before, the 0.3 s window results are plotted in the left column and the 78 ms results in the right column and each symbol corresponds to a single window result, linear regression lines are overlaid on the data and both standard deviations and goodness-of-fit statistics (R^2) are shown.

It can be seen from Fig. 3(b) that the ER method tracks the Theoretical Method fairly well using the 0.3 s window for this condition. The corresponding R^2 statistic is 0.79, and

the standard deviation is 0.04. The ER method predicts values that are, on average, slightly lower than the Theoretical Method across the range as indicated by the linear regression line always lying below the main diagonal in the left frame (slope = 0.80, intercept = −0.02). It should be noted that this condition was a more severe degradation than speech plus noise at 0 dB SNR, as evidenced by the fact that the Theoretical method never exceeded 0.56 for the 0.3 s windows or 0.61 for the 78 ms windows whereas, in the 0 dB SNR condition, the Theoretical Method had maxima of 0.73 and 0.85 for the 0.3 s and 78 ms windows, respectively.

For the 78 ms window analysis, the ER metric results for noise plus reverberation demonstrate even more variability relative to the Theoretical Method ($R^2 = 0.35$, standard deviation = 0.09) indicating a very poor fit to the linear regression line (slope = 0.42 and intercept = 0.07). Looking at specific trends, when the Theoretical Method was zero, the ER metric varied over a wide range (0 to 0.4). It also appears that, in the reverberant condition, the ER metric values varied more at the highest Theoretical Method values than at the lowest, in contrast to the other conditions analyzed.

The ER metric and the Theoretical Method were also averaged across windows. The averages are plotted vs window length in Fig. 4(b), along with the long-term STI for comparison. The Theoretical Method asymptotes to a value just slightly less than the long-term STI (0.394 vs 0.41) for this condition when analysis windows are greater than 5 s. The ER average asymptotes at 0.363, 0.031 less than the Theoretical Method average. For windows less than 5 s, both the Theoretical Method and ER averages decrease, similar to what was observed in the stationary speech-shaped noise condition. The ER method averages parallel, but are consistently less than, the Theoretical Method for all window lengths (a difference of 0.03 for the longest windows and 0.08 for 0.3 s windows).

C. Zero dB SNR with fluctuating restaurant noise

The third condition analyzed was speech degraded by restaurant babble at 0 dB SNR. As for the prior two conditions, the window-by-window metric results in Fig. 3(c) are plotted versus the corresponding Theoretical Method and a linear regression analysis is presented for each plot. It can be seen from the left plot that the ER results are highly correlated with the Theoretical Method for the 0.3 s window where $R^2 = 0.93$ and the standard deviation is 0.04. The slope of 0.92 and intercept of 0.05 indicate the ER is very close in value to the Theoretical Method. For the 78 ms window, the data is more scattered, with $R^2 = 0.84$ and standard deviation equal to 0.09. The regression line has a slope of 0.85 and intercept equal to 0.11. One can see that the Theoretical Method is rarely zero for either window length but, as was observed for the other conditions, when the Theoretical Method produced values less than 0.1 using the shorter analysis window, the ER values spanned a wide range, in this case from 0 to 0.75.

Figure 4(c) demonstrates that the averaged behavior of the metrics when speech is degraded by restaurant babble is very similar to that when speech is degraded by stationary speech-shaped noise. The primary difference is that the Theoretical Method and ER method asymptote at 0.6 rather than 0.5 as they had for the speech-shaped noise condition.

D. Different speaking styles at −1.8 dB SNR with stationary noise

For the next investigation, nonsense sentences were analyzed that had been spoken either conversationally (Conv) or clearly at normal rates (Clear/Norm) by a female and a different male talker (T4 and T5, respectively) in the presence of stationary speech-shaped noise presented at −1.8 dB SNR (Krause and Braida, 2002). The purpose of this analysis was three-fold. First, it was intended to verify metric performance on additional voices. Second, it was of interest to see how well the metrics would compare to subject performance at the sentence and word levels since subject intelligibility data for these sentences and keywords at this signal-to-noise ratio were provided by J. Krause (personal communication). This would be the first instance when an STI metric for individual words could be compared to subject responses to those words. Third, it was not known how the short-time metric would perform as a function of the two speaking styles for which significant intelligibility differences have been demonstrated (Payton *et al.*, 1994; Krause, 2001; Krause and Braida, 2002).

To demonstrate the first goal, a regression analysis of the ER method vs the Theoretical Method was performed on each talker's Conv and Clear/Norm keywords. The window length was the same as the length of the corresponding word. Virtually all the words analyzed were longer than 0.3 s which was the window duration established in Sec. III A as sufficient for the ER method to track the Theoretical Method for speech-shaped noise degradation. Out of 342 words analyzed (170 by T4 and 172 by T5), only 18 (10 by T4 and 8 by T5) were less than 0.3 s in duration (minimum word length was 0.2 s). The results for each talker and speaking style are not plotted because they were almost identical to the 0.3 s window 0 dB SNR results presented in Fig. 3(a). For both talkers, the best-fit lines had slopes of 1.0. The "goodness-of-fit" statistic (R^2) was 0.98 for T4, 0.99 for T5 and there was no difference in the quality of fit based on speaking style. The means of both talkers' Conv words were 0.37 for both the ER and Theoretical methods while the Clear/Norm word means were 0.45 for each talker and metric. Individual Conv word metric values ranged from 0.07 to 0.61 for T4 and from 0.13 to 0.63 for T5 while the Clear/Norm word metrics ranged from 0.13 to 0.74 for T4 and from 0.25 to 0.65 for T5.

Next, the ER metric was computed for each of 50 sentences common to both the Conv and Clear/Norm corpuses with the window length set to the sentence length (i.e., one value was obtained for each sentence). These values were compared to average subject intelligibility scores for the corresponding sentences' keywords. In Fig. 5, sentence ER values are plotted on the horizontal axes (T4 on the left and T5 on the right) and average percent-correct values are plotted on the vertical axes.

Note that the Clear/Norm averages for each talker, indicated by the intersection of the dotted crosshairs, are to the

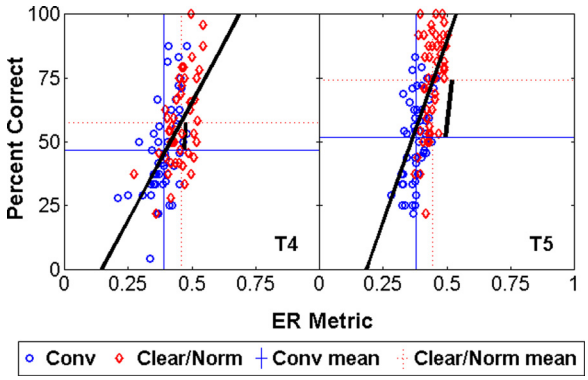

FIG. 5. (Color online) Regression analyses for ER metric vs percent correct on sentences using windows equal to sentence lengths. (left) Talker T4; (right) talker T5. Circles correspond to Conv sentences and diamonds correspond to Clear/Norm sentences. The solid crosshairs indicate the means for conversational sentences and the dotted crosshairs indicate the means for clear sentences. The thick lines indicate the best linear fit to the combined data. The line segments between the percent-correct lines indicate the change in long-term STI for the two data sets [from Krause and Braida (2004)].

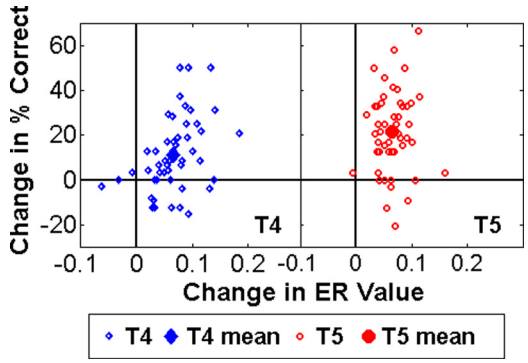

FIG. 6. (Color online) Change in ER metric vs change in percent correct for sentences due to change in speaking style from Conv to Clear/Norm. Diamonds represent sentences spoken by T4. Circles represent sentences spoken by T5. The larger filled symbols represent talker averages. The solid horizontal lines denote zero change in percent correct and solid vertical lines indicate zero change in ER value.

right and above the Conv averages which are indicated by the intersection of the solid crosshairs. This implies that the metric is able to capture some aspects of Clear/Norm speech that contributed to its higher intelligibility even though it was presented at the same SNR as Conv speech. These results are consistent with those reported by Krause and Braida (2004) who demonstrated in their Fig. 4 that the long-term STI for these two talkers increased slightly when speaking style was changed. To allow a comparison, line segments have been included in Fig. 5 to indicate their long-term STI changes for these speech materials. The segment for T4 is almost vertical indicating very little change in STI for the two speaking styles. The segment for T5 demonstrates a greater change in STI due to speaking style but still not as large as that demonstrated by the ER metric.

The R^2 statistics for the data in Fig. 5 are much lower than for the regressions of the ER metric against the Theoretical Method for word-length windows. For these sentences, $R^2 = 0.34$ for talker T4 and 0.45 for talker T5. The R^2 statistic indicates the extent to which the regression line is a better fit to the data than just the mean value. The regression line is not very meaningful if the data are spread vertically about the mean. This is the situation for much of the sentence data since each sentence was presented at -1.8 dB SNR and the resulting intelligibility metrics varied over a range of only 0.28 for T4 and 0.14 for T5. The only difference in the metric values appears to be a shift of the mean due to speaking style from an average of 0.39 for T4 and 0.38 for T5 on Conv sentences to 0.46 for T4 and 0.45 for T6 on Clear/Norm sentences. On the other hand, subject intelligibility averages varied from about 20% correct to nearly 100% correct on the sentences. A word-level regression analysis is not shown because subject intelligibility scores were highly quantized (8 subjects, each either getting the word correct or not) resulting in even lower correlations.

Since the Conv and Clear/Norm experiments included a sentence list in common to the two speaking styles, it was possible to compare the intelligibility differences due to speaking style on specific keywords with the corresponding changes in metric value. The change in intelligibility between Clear/Norm and Conv sentence pairs vs the change in ER value for the same pair are presented in Fig. 6. A positive change corresponds to a higher score for the Clear/Norm speech than the Conv speech.

Symbols in the first and third quadrant correspond to subject performance trends accurately predicted by the metric, i.e., Clear/Norm sentences which are more intelligible than their Conv counterparts with corresponding positive metric changes or Clear/Norm sentences less intelligible with negative metric changes. Most sentence pairs and the means for both talkers fell in the first quadrant. A couple sentence pairs fell in the third quadrant for T4. The fourth quadrant, with the second largest cluster of sentence pairs (mostly spoken by T4), corresponds to Clear/Norm sentences less intelligible than their Conv counterparts but with higher metric values. These are sentences for which the subjects had decreased intelligibility for Clear/Norm sentences over the corresponding Conv sentences but the metric still predicted improved intelligibility. They were analyzed further and it was determined that the metric predominately tracked voiced sounds such as vowels and, despite strong vowels, some Clear/Norm key words had low probability of correct identification.

Next, the metric was evaluated on its ability to predict intelligibility changes due to speaking style for keywords based on their position in the sentence. The hypothesis was that talkers tend to speak more softly toward the ends of sentences when speaking conversationally, reducing the effective SNR for those words. Some of the Clear/Norm intelligibility advantage could be due to talkers maintaining a more constant SNR throughout the sentence, improving the SNR for sentence-ending words when compared to the SNR for these words in Conv sentences. Therefore, as shown in Fig. 7, changes in intelligibility and ER values of first keywords were compared to changes in intelligibility and ER values of final keywords.

As was the case in Fig. 6, positive changes indicate higher intelligibility and/or ER value for Clear/norm

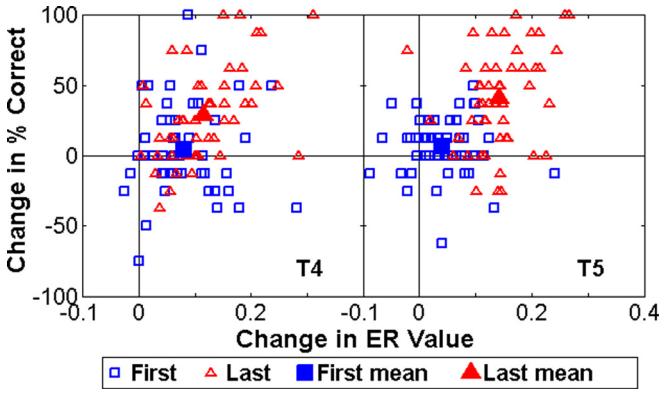

FIG. 7. (Color online) Change in percent correct vs change in metric for first and last keywords due to change in speaking style for each talker. Squares correspond to first word pairs. Triangles correspond to last word pairs. The large filled symbols indicate averages. Solid horizontal and vertical lines mark zero change in percent correct and ER values, respectively.

pronunciation of the words. While there is some tendency for the ER value to shift to the right for both initial and final keywords, the average change in intelligibility for the initial keywords is nearly zero for both talkers. Nineteen of T4s and 11 of T5s 50 initial keywords were less intelligible when spoken in the Clear/Norm style. The final keyword changes are more varied but tend to be farther from the zero-change axes and in the first quadrant. Only six of T4s and four of T5s final keywords were less intelligible when spoken in the Clear/Norm style than in the Conv style and only one final keyword (spoken by T5) had a lower ER value. On average, sentence-final keywords have a greater increase in both metric values and percent correct going from Conv to Clear/Norm than sentence-initial keywords for both talkers although the result is more noticeable for talker T5. The hypothesis therefore appears to be supported by the data: The short-term metrics decrease from beginnings to ends of conversational sentences but maintain a more stable level during clear speech, even clear speech produced at normal speaking rates.

IV. DISCUSSION

Previous efforts to use speech as a probe stimulus in STI calculations, rather than modulated noise, as specified by the IEC standard, required very long speech segments to generate an accurate metric (e.g., Payton *et al.*, 2002). This was due primarily to the fact that they were derived from modulation spectra and needed signals of sufficient duration to provide the necessary spectral resolution. Goldsworthy and Greenberg (2004) demonstrated that the time-domain ER and other STI-based methods, computed using speech as a probe stimulus, were highly correlated with the long-term STI. They used speech durations comparable to the spectral methods to validate the metrics. The current work has demonstrated that the ER metric tracks the Theoretical Method using analysis windows as short as 0.3 s in a variety of acoustic degradation conditions. In addition, the algorithm to compute the ER metric is computationally efficient and can be implemented in a real-time system. The results presented indicate that the averaged ER method equals the averaged Theoretical Method when the analysis window is as short as 1 s and also the long-term STI when the analysis window is as short as 8–10 s for the two noise conditions.

For the noise plus reverberation condition, there is evidence in the literature that the long-term STI over-predicts intelligibility for noise plus reverberation. The intelligibility scores reported by Payton *et al.* (1994) were lower for speech degraded by noise plus reverberation than the STI would have predicted. For example, in two conditions that differed only in whether reverberation was part of the acoustic degradation, the authors reported a difference of only 0.05 in STI while their subjects' intelligibility scores dropped substantially due to the added reverberation. The average intelligibility scores dropped 18 percentage points, from 79% to 61% correct, for the Clear talker and 24 percentage points, from 51% to 27% correct, for the Conv talker. If a 0.3 s window ER metric were used and averaged, the noise-plus-reverberation STIs would be reduced by 0.17. This would move the STI predictions for that condition from being outliers to close to the third-order curve fits reported by Payton *et al.*

Some points should be made about the Theoretical Method as computed in this study. First, while it has been demonstrated to match the long-term STI over a wide range of analysis window lengths, it requires access to knowledge of both the signal power and noise power in each octave band and is therefore not a reasonable choice for applications in which the SNR is not known. Second, when the condition is noise plus reverberation, the theoretical Modulation Transfer Function for speech plus noise is adjusted by a multiplicative term derived from the room impulse response intensity envelope spectrum [Eq. (6)] to obtain the modulation index. In practice, the room impulse response is usually not known. Third, even if the impulse response were known, the reverberation adjustment in Eq. (6) is not likely to be valid for analysis windows shorter than the reverberation time (the reverberation adjustment is not a function of analysis window length). This issue should be investigated further; it may explain why the scatterplot comparing the ER method to the Theoretical Method for the noise plus reverberation condition was so much worse than those for the stationary speech-shaped and the fluctuating restaurant babble conditions (Fig. 3). Both the 0.3 s and the 78 ms analysis windows are shorter than the 0.6 s reverberation time. George *et al.* (2008) took a hybrid approach to model speech intelligibility in the presence of non-stationary background noise plus reverberation. They computed the Modulation Transfer Function to account for the reverberation and combined it with the extended SII (ESII), proposed by Rhebergen and Versfeld (2005), which was computed using analysis frames that varied from 35 ms down to 9.4 ms depending on the filter band. In the future, it may make more sense to compute the Theoretical Method value by treating the portion of the reverberant speech due to reflections as a masking speech interferer like the fluctuating noise condition.

As indicated above, further work must be done to thoroughly investigate the limitations of this short-time speech-based STI method. Using the current computational technique, analysis windows shorter than 0.3 s result in ER metric values that deviate from the Theoretical Method. Doubbelboer and

Houtgast (2007) point out that, in addition to the "systematic lift" of degraded speech envelopes due to the mean noise intensity, there are stochastic envelope fluctuations that are not captured by the Theoretical Method or by the long-term STI. Those stochastic envelope fluctuations are what cause the ER metric to deviate from the Theoretical Method.

Another question to be answered is: What is the right length analysis window to use? Following the example of Rhebergen and Versfeld, analysis windows as short as 9.4 ms might be needed to accurately predict speech intelligibility in fluctuating backgrounds. On the other hand, the envelope signal extracted by an analysis window this short would always be a constant since only frequencies greater than 106 Hz (period equal to 9.4 ms) would appear as time-varying fluctuations within the window but the envelopes are lowpass filtered at 50 Hz. Also, Gallun and Souza (2008) demonstrated that consonant confusions by listeners were highly correlated with modulation spectra spanning modulation frequencies from 1 to 32 Hz. Even a 0.3 s window only "sees" frequencies greater than 3.3 Hz which is already near the syllable rate of speech (~4 Hz). Clearly more work needs to be done to clarify this issue.

Even with its limitations, the short-time ER metric represents a promising new way to objectively predict speech intelligibility in a variety of acoustic environments. In particular, it might be possible to predict differences in intelligibility between stationary and modulated noise attributed to masking release if the metric can be computed using appropriate window lengths. The current work has shown that, in the presence of a fluctuating background (restaurant babble), the short-term Theoretical Method and ER method both generated slightly higher values than they did for stationary noise (0.6 rather than 0.5).

One reason an STI metric that uses speech as its probe stimulus is important is it opens up the range of environments under which the STI can be measured. Using the ER method, speech intelligibility can be predicted during a lecture or other live performance in a populated auditorium or classroom. The existing tools for measuring the STI require the presentation of intensity-modulated noises that most audiences might find rather annoying to listen to. A speech-based metric might also be applied to a time-varying situation such as that resulting from an amplitude-compression hearing aid. This would also be a situation where short-term windows might be more appropriate than a long-term analysis so that distortions during gain transitions do not necessarily skew predicted intelligibility during steady-state intervals.

V. CONCLUSIONS

The work reported herein has demonstrated two important results. First, a short-time, speech-based, metric is able to track the short-term fluctuations in STI accurately down to window lengths of 0.3 s for two different noise environments and a noise plus reverberation environment. Second, this short-time metric has successfully predicted intelligibility differences due to speaking style at both the sentence and word level. It tracked differences in acoustic features such as SNR word by word through a sentence.

ACKNOWLEDGMENTS

This work was supported by NIDCD grant 1-RO1-DC007152. The authors wish to thank Dr. Jeanie Krause for providing the stimuli and subject response data presented in Sec. III D and for her comments on the manuscript. We also thank Mr. Kenneth Schutte and Dr. Louis Braida for their comments and suggestions.

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

Boldt, J. B., and Ellis, D. P. W. (**2009**). "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proc. 17th European Sig. Process. Conf.* (EURASIP, Lausanne, Switzerland), pp. 1849–1853.

Doubbelboer, F., and Houtgast, T. (**2007**). "A detailed study on the effects of noise on speech intelligibility," J. Acoust. Soc. Am. **122**, 2865–2871.

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

Falk, T. H., Zheng, C., and Chan, W.-Y. (**2010**). "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," IEEE Trans. Audio Speech Lang. Process. **18**, 1766–1774.

Gallun, F., and Souza, P. (**2008**). "Exploring the role of the modulation spectrum in phoneme recognition," Ear Hear. **29**, 800–813.

George, E. L. J., Festen, J. M., and Houtgast, T. (**2008**). "The combined effects of reverberation and nonstationary noise on sentence intelligibility," J. Acoust. Soc. Am. **124**, 1269–1277.

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

Houtgast, T., and Steeneken, H. J. M. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica **28**, 66–73.

Houtgast, T., and Steeneken, H. J. M. (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**, 60–72.

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.

IEC (**1998**). "Part 16: Objective rating of speech intelligibility by speech transmission index (2nd edition)," in *IEC 60268 Sound System Equipment* (Int. Electrotech. Commiss., Geneva, Switzerland).

IEC (**2003**). "Part 16: Objective rating of speech intelligibility by speech transmission index (3rd edition)," in *IEC 60268 Sound System Equipment* (Int. Electrotech. Commiss., Geneva, Switzerland).

IEC (**2011**). "Part 16: Objective rating of speech intelligibility by speech transmission index (4th edition)," in *IEC 60268 Sound System Equipment* (Int. Electrotech. Commiss., Geneva, Switzerland).

Kates, J. M. (**1987**). "The short-time articulation index," J. Rehab. Res. Develop. **24**, 271–276.

Krause, J. C. (**2001**). "Properties of naturally produced clear speech at normal rates and implications for intelligibility enhancement," Ph.D. dissertation, Dep. Electon. Eng. Comput. Sci. (Mass. Inst. Technol., Cambridge, MA).

Krause, J. C., and Braida, L. D. (**2002**). "Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility," J. Acoust. Soc. Am. **112**, 2165–2172.

Krause, J. C., and Braida, L. D. (**2004**). "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **115**, 362–378.

Ludvigsen, C. (**1993**). "The use of objective methods to predict the intelligibility of hearing aid processing speech," in *Recent Developments in Hearing Instrument Technology*, edited by J. Beilin and G. R. Jensen (Danavox/Stougaard Jensen, Copenhagen, Denmark), pp. 81–94.

Ludvigsen, C., Elberling, C., and Keidser, G. (**1993**). "Evaluation of a noise reduction method-comparison between observed scores and scores predicted from STI," Scand. Audiol. **22**, 50–55.

Ludvigsen, C., Elberling, C., Keidser, G., and Poulsen, T. (**1990**). "Prediction of intelligibility of non-linearly processed speech," Acta Otolaryngol. Suppl. **469**, 190–195.

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

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," in *Past, Present and Future of the Speech Transmission Index*, edited by S. J. v. Wijngaarden (TNO Human Factors, Soesterburg, The Netherlands), Chap. 11, pp. 125–138.

Payton, K. L., and Shrestha, M. (**2008a**). "Analysis of short-time speech transmission index algorithms," in *Acoustics'08 Paris*, 633–638.

Payton, K. L., and Shrestha, M. (**2008b**). "Evaluation of short-time speech-based intelligibility metrics," in *9th Internat. Congress on Noise as a Public Health Problem*, edited by B. Griefahn (IfADo, Dortmund, Germany), pp. 243–251.

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.

Peterson, P. M. (**1986**). "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," J. Acoust. Soc. Am. **80**, 1527–1529.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1985**). "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," J. Speech Hear. Res. **28**, 96–103.

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. A. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988–3997.

Schlesinger, A. (**2012**). "Transient-based speech transmission index for predicting intelligibility in nonlinear speech enhancement processors," in *Proc. IEEE Internat. Conf. Acoust. Speech Sig. Process.* (IEEE, Piscataway, NJ), pp. 3993–3996.

Taal, C. H., Hendriks, R. C., Heudens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**, 2125–2136.