Evaluation of short-time speech-based intelligibility metrics

Karen L. Payton*, Mona Shrestha

University of Massachusetts Dartmouth, 285 Old Westport Rd., N. Dartmouth, MA 02747

*corresponding author: e-mail: kpayton@umassd.edu

INTRODUCTION

The Speech Transmission Index (STI) is based on acoustic measurements in environments and has been shown to be correlated with speech intelligibility under a wide range of acoustic conditions (Houtgast & Steeneken 1984). It is a weighted average of metrics derived from envelope signals in multiple frequency bands spanning the speech spectrum. A variety of methods have been proposed to compute the STI (Houtgast & Steeneken 1971; Steeneken & Houtgast 1980; Ludvigsen 1987; Drullman et al. 1994a, b; Payton et al. 1994; Drullman 1995; IEC 1998; Payton & Braida 1999; Payton et al. 2002; Goldsworthy & Greenberg 2004). Some of these methods use speech as the test stimulus rather than artificially modulated noise as originally proposed by Houtgast and Steeneken (1985). Many of the speech-based techniques have been shown to provide the same result as the traditional STI (Ludvigsen et al. 1990; Payton et al. 2002), which is based on modulation reductions in intensity-modulated noise and as a theoretically derived STI which is obtained from weighted signal-to-noise ratios (SNRs) in seven octave bands and room reverberation time (RT) (Houtgast & Steeneken 1985). To date, all speech-based approaches have used speech materials lasting at least a minute or two to generate metrics correlated with long-term speech intelligibility. Consequently, they have not been used to predict short-time changes in intelligibility due to time-varying environments such as fluctuating background noise. The current work investigates the ability of two speech-based methods to track short-term STI results by using speech segments of various lengths to compute results for environments with stationary speech-shaped noise, speech-shaped noise plus reverberation or multi-talker babble. The methods that will be evaluated are the Envelope Regression (ER) and the Normalized Correlation (NC) methods. The ER method is based on the speech-based STI method proposed by Ludvigsen et al. (1990). The NC method was proposed by Goldsworthy and Greenberg (2004) who also analyzed the long-term characteristics of both metrics.

METHODS

Figure 1 depicts a block diagram of the signal processing steps used to obtain the results for the speech-based algorithms. Specifically, for both the ER and NC techniques, the clean and the degraded signals, originally digitized at 20 kHz with a 9.5 kHz antialiasing filter, were digitally filtered using a bank of 6^{th} order octave-wide Butterworth band-pass filters with center frequencies from 125 Hz – 4 kHz and a 6^{th}-order Butterworth high-pass filter with a cutoff frequency of 6 kHz. For each band, i, the clean and the degraded signals were then squared and low-pass filtered with a cut off frequency of 50 Hz. The lowpass filter impulse response was a 10 ms Hamming window. The intensity envelopes, $x_i(t)$ and $y_i(t)$, were down-sampled to 134 Hz (a factor of 49) to reduce computation time without risking aliasing. Next, for each octave band, a modulation metric, M_i, was calculated from the intensity envelopes. Each approach used a different algorithm to compute this modulation metric.

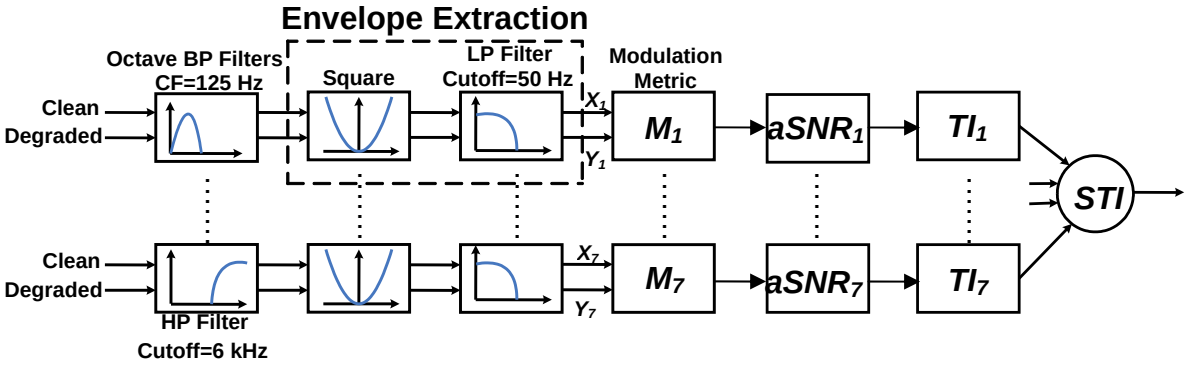

Figure 1: Block diagram of signal processing steps necessary to compute speech-based intelligibility metrics

For the Envelope Regression (ER) method, the modulation metric for each band was computed from the envelope signals using Eqn (1):

$$M_i = \frac{\mu_{xi}}{\mu_{yi}} \frac{E\{(x_i(k) - \mu_{xi})(y_i(k) - \mu_{yi})\}}{E\{(x_i(k) - \mu_{xi})^2\}} \quad (1)$$

where μ_{xi} and μ_{yi} are the means of $x_i(t)$ and $y_i(t)$ respectively. For the Normalized Correlation (NC) method, M_i was computed using Eqn (2):

$$M_i = \frac{E\{x_i(k) \cdot y_i(k)\}^2}{E\{x_i^2(k)\} \cdot E\{y_i^2(k)\}} \quad (2)$$

(Goldsworthy & Greenberg 2004).

Once the modulation metrics were computed, the apparent signal-to-noise ratio in each band, $aSNR_i$, was computed as

$$aSNR_i = 10 \log_{10}\left(\frac{M_i}{1 - M_i}\right) \quad (3)$$

and then clipped to the range of -15 to +15 dB. The apparent SNR in each band was converted to a transmission index, TI_i, according to Eqn (4):

$$TI_i = \frac{aSNR_i + 15}{30} \quad (4)$$

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

$$STI = \sum_{i=1}^{7} \alpha_i TI_i - \sum_{i=1}^{6} \beta_i \sqrt{TI_i \times TI_{i+1}} \quad (5)$$

where the α_i's represent the octave weighting factors and the β_i's represent the redundancy correction factors given in the IEC standard (IEC 1998).

Short-Time Implementation Issues

For both the ER and NC methods, sample means of the windowed envelope signals were calculated. Correlations were calculated as biased estimates:

$$E\{x_i(k)y_i(k)\} = \frac{1}{N}\sum_{k=1}^{N}[x_i(k)y_i(k)] \text{ and } E\{x_i(k)^2\} = \frac{1}{N}\sum_{k=1}^{N}x_i(k)^2 \qquad (6)$$

where *N* was the window length (in samples). These correlation values were used directly in Eqn (2) for the NC method. The cross- and auto-covariances needed for the ER method were calculated from the correlation estimates of Eqns. (6) as

$$E\{(x_i(k)-\mu_{xi})(y_i(k)-\mu_{yi})\} = E\{x_i(k)y_i(k)\} - \mu_{xi}.\mu_{yi} \text{ and}$$
$$E\{(x_i(k)-\mu_{xi})^2\} = E\{x_i(k)^2\} - \mu_{xi}^2 \qquad (7)$$

and used in Eqn. (1). Window lengths were adjusted from 107 sec (length of 50 concatenated sentences) down to 78 ms for the analyses presented below. Windows were overlapped by 75 %.

Theoretical STI

In order to compare the short-time metrics with the "true" STI, the theoretical STI was also calculated over the same time windows as the short-time metrics. The speech and the noise (as opposed to the degraded speech) were separately passed through the octave-band filter bank shown in Figure 1 and within-band powers used to get signal to noise ratio (S_i/N_i in Eqn (8)) in each band. The modulation index in each band, $M_i(F)$, was then calculated as specified by Steeneken and Houtgast (1980):

$$M_i(F) = \left(\frac{1}{\sqrt{1+\left(\frac{2\pi FT}{13.8}\right)^2}}\right)\left(\frac{1}{1+10^{\frac{-S_i/N_i}{10}}}\right) \qquad (8)$$

The first term in Eqn (8) estimates the modulation reduction due to reverberation. The variable *F* corresponds to modulation frequency (between 0.63 and 25 Hz) and *T* corresponds to the reverberation time of the environment (T_{60}). The second term estimates the reduction due to additive noise. The theoretical STI was computed by substituting $M_i(F)$ for M_i in Eqn (3), the variable $aSNR_i(F)$ was averaged across *F* after clipping to obtain $aSNR_i$.

Stimuli

The stimuli used in this study were 50 concatenated nonsense sentences, spoken conversationally by a male talker totaling 107 s of speech (Payton et al. 1994). These nonsense sentences are grammatically correct but do not provide any semantic context to help word identification, e.g., "His guests could teach his turnpike". Each sentence consists of four to eight key words (underlined in example) where the key words consist of the nouns, adjectives, verbs and adverbs in the sentence.

Degradation Conditions

Three environmental degradations were evaluated: stationary speech-shaped noise, stationary noise plus simulated reverberation and multi-talker babble. The speech-shaped noise was generated by filtering white Gaussian noise to approximate the average long-term spectra of speech (Payton et al. 1994). The noise was added to the speech at an average SNR of 0dB. For the noise plus reverberation condition, speech plus noise at 0 dB SNR was convolved with a simulated conference room impulse response (Peterson 1986; Payton et al. 1994). The multi-talker babble was taken from a recording of restaurant noise. The babble also was added to the speech at 0 dB SNR.

RESULTS

Results from both the ER and NC methods were compared with the theoretical STI for each degradation condition as functions of window length. Linear regression analyses also were carried out for the metrics and theoretical STI results. For the regression analyses, results for two window lengths are presented. The 0.3 s window results are typical of all the longer windows. The 78 ms window is presented to show a window for which the metrics deviate from the theoretical STI during silent intervals.

Zero dB SNR with Stationary Speech-Shaped Noise

The results for each method over the length of one sentence are plotted as functions of time in Figure 2.

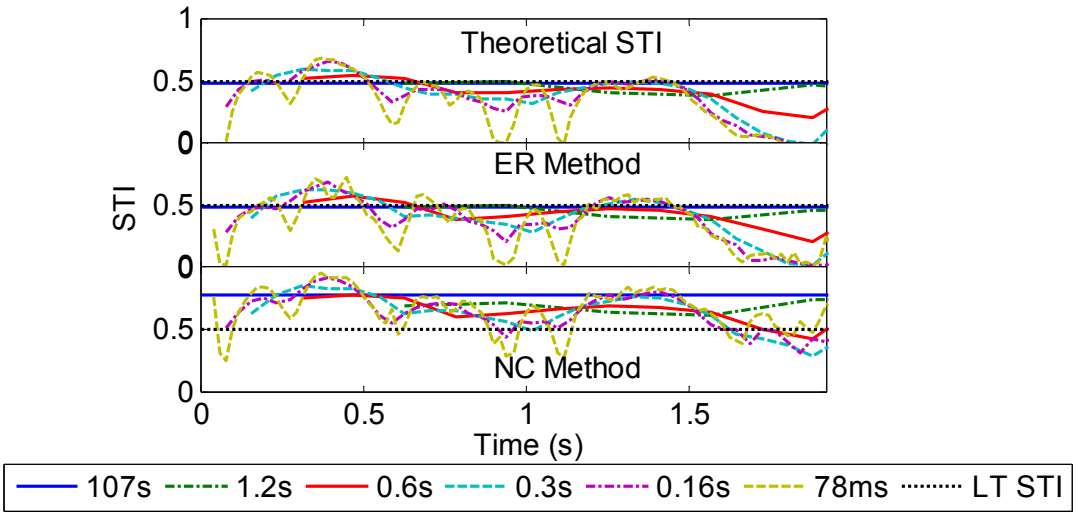

Figure 2: Metric results vs. window length (top) theoretical STI (center) ER method and (bottom) NC method for 0 dB SNR stationary speech-shaped noise condition. Different curve types represent results with different window lengths as given in the legend. The black dotted line in each plot represents the long-term STI.

For visual reference, an SNR of 0 dB corresponds to an STI value of about 0.5 (the exact value depends on the spectral characteristics of the speech and noise). Both the ER and NC metrics (center and bottom plots respectively) generally matched local fluctuations in the theoretical STI (top plot) for each window length and the ER result for entire corpus (blue line in center plot) matched the long-term STI (black dotted line) exactly. The ER method tracked the theoretical STI more closely than the NC method for all window lengths analyzed. For all window lengths, the NC method predicted slightly higher values than either the ER method or the theoretical STI in agreement with long-term results of Goldsworthy and Greenberg (2004).

Once window length was decreased to 78 ms (tan dashed lines), both the ER and NC methods deviated greatly from the theoretical STI at the beginnings and ends of sentences. Where the theoretical STI was zero because only noise was present (SNR = -∞ dB) both metrics often generated non-zero results.

Figure 3 plots linear regression analyses of metric results versus theoretical STI for two window lengths: 0.3 s (top row) and 78 ms (bottom row). Each data point represents the results for a single window. Regression lines and the goodness of fit (R^2) statistics are also shown for each window length. As can be seen from the figure, the ER method results (left column) closely match the theoretical STI for the 0.3 s window, indicated by the R^2 statistic of 0.99. The results are also close for the 78 ms window (R^2=0.91). However, for the 78 ms window, some of the ER results were above zero on the y-axis which means that, during the silent intervals, when the theoretical STI was zero the ER method sometimes generated values greater than zero (up to 0.4).

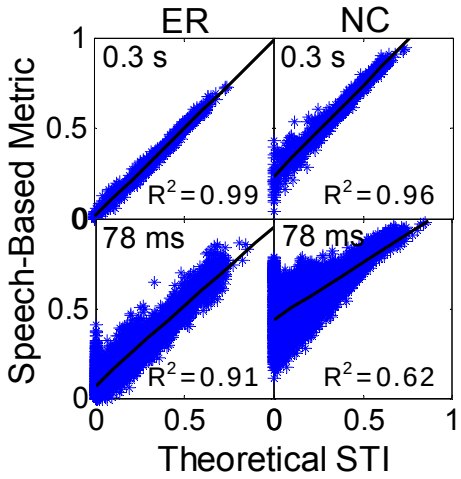

Figure 3: Metrics computed from ER (left column) and NC (right column) methods vs. theoretical STI for 0 dB SNR using 0.3 s windows (top row) and 78 ms windows (bottom row). The solid lines represent best linear fits to the data.

The NC method regression analysis results are shown in the right column of Figure 3. This method predicted higher values than the theoretical STI for all window lengths as can be seen by the upward shift of the linear regression lines from the main diagonal. The R^2 statistic of 0.96 for 0.3 s window shows that, despite this shift, the NC method followed the theoretical STI quite closely. For the 78 ms window, the metric did not perform as well. The R^2 statistic is also reduced (0.62) in part because, when the theoretical STI was zero, the NC method generated values ranging from 0.1 to 0.8.

In order to study how well, on average, the short-time metrics match the long-term theoretical STI over the range of window lengths, the metrics and theoretical STI were averaged over the entire speech corpus (107 s) for each window length. The averages are plotted in Figure 4 as functions of window length. In the figure, the solid red line represents ER method averages, the blue dash-dot line represents the NC method averages and the black dotted line represents the theoretical STI.

It can be seen that ER method produced the same average value as the theoretical STI over virtually the entire window range studied. The averages for all metrics decreased as the window was decreased. This is because voiced speech segments dominated the metric results and when the windows were shortened to the point that some windows contained primarily unvoiced and/or silent intervals then the results for those windows were significantly reduced. The leftmost data points are for the 78 ms window. For that window length, the ER did not decrease quite as much as the theoretical STI and the NC method actually increased slightly.

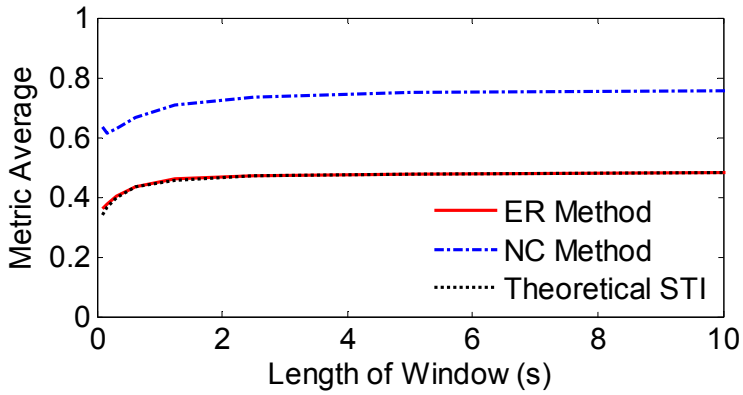

Figure 4: Metric averages computed over entire speech corpus for speech in 0 dB stationary speech-shaped noise, as functions of window lengths.

Zero dB SNR Plus Reverberation

When reverberation was added to the noisy speech, the metrics generated values that varied more widely when compared to the theoretical STI. In Figure 5, metrics are plotted (ER on the left and NC on the right) versus the theoretical STI for the two window lengths. The 0.3 s window results are plotted in the top row and the 78 ms results in the bottom row. As before, each symbol corresponds to a single window result, linear regression lines are overlaid on the data and the goodness of fit statistics are shown.

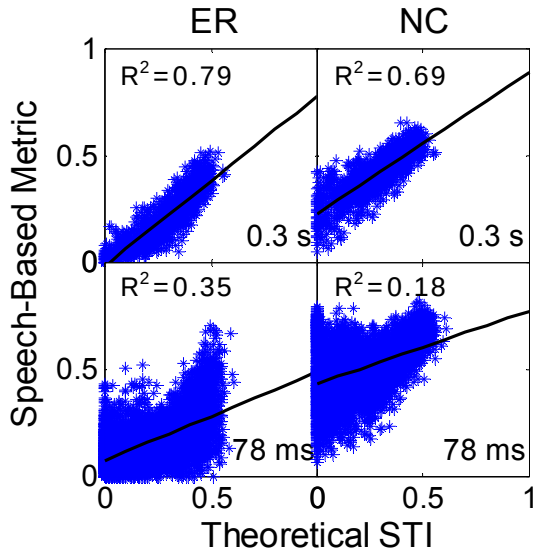

Figure 5: Metrics computed from ER (left column) and NC (right column) methods vs. theoretical STI for 0 dB SNR plus reverberation using 0.3 s windows (top row) and 78 ms windows (bottom row). The solid lines represent best linear fits to the data.

It can be seen from Figure 5 that, for the 0.3 s window, the results from both methods tracked the theoretical STI fairly closely although the ER method predicted values that were, on average, slightly lower than the theoretical STI across the range. The NC method predicted higher values than the theoretical at the low STI end and lower values at the high STI end. The corresponding R^2 statistics are 0.79 and 0.69 for the ER and NC methods respectively. For the 78 ms window, the results are much more divergent (R^2=0.35 and 0.18 respectively indicating very poor correlations). In particular, when the theoretical STI was zero, both metrics generated results that varied over a wide range (0 to 0.4 for the ER method and 0.1 to 0.8 for the NC method). Furthermore, there appears to be a nonlinear relation such that the metric values deviated from the linear regression line more at the higher STI values.

Averages for both methods and the theoretical STI as functions of window length are given in Figure 6. The solid red line plots the ER method averages, the blue dash-dot line shows the NC method and the black dotted line represents the theoretical STI.

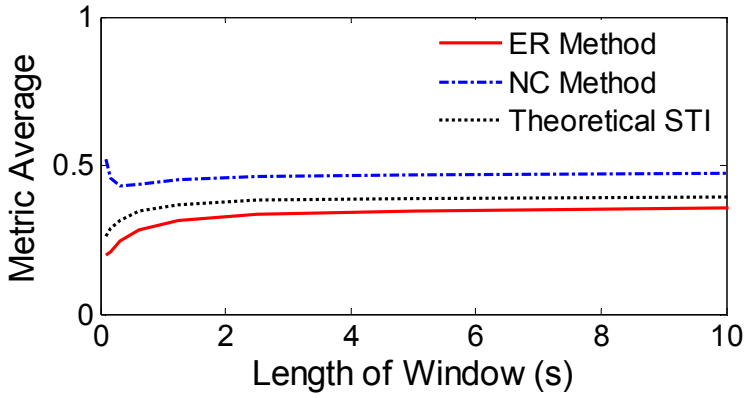

Figure 6: Metric averages computed over entire speech corpus for speech in 0 dB stationary speech-shaped noise plus reverberation, as functions of window length.

It can be seen from Figure 6 that, for the noise plus reverberation condition, the ER method generated values that paralleled but were consistently less than the theoretical STI for all window lengths. It should also be noted that, as for the speech plus noise condition, the NC method actually increased for the shortest windows while the ER and theoretical STI continued to decrease.

Zero dB SNR with Multi-Talker Babble

As for the prior two conditions, metric results are plotted against the theoretical STI in Figure 7 and a linear regression analysis is performed for each plot. It can be seen from the left column in the figure that the STI from ER method is highly correlated with the theoretical STI for the 0.3 s window where R^2=0.93 while data is much more scattered for the 78 ms window for which R^2=0.84. As was observed for the other conditions, when the theoretical STI produced values near zero, the ER values covered a wide range, in this case from 0 to 0.8.

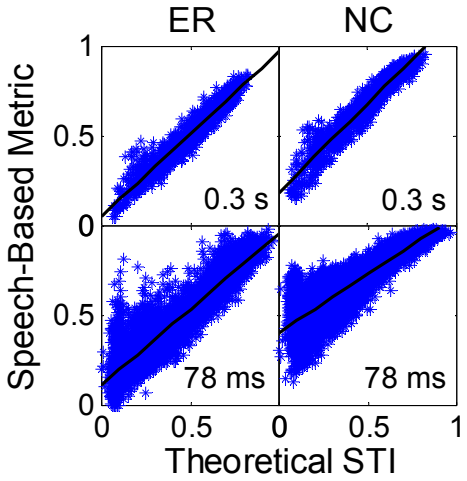

Figure 7: Metrics computed from ER (left column) and NC (right column) methods vs. theoretical STI for 0 dB SNR multi-talker babble using 0.3 s windows (top row) and 78 ms windows (bottom row). The solid lines represent best linear fits to the data.

Regression analysis results for the NC method are shown in right column of Figure 7. The R^2 statistic of 0.93 for the 0.3 s window indicates that the NC method followed the theoretical STI fairly closely although the values it generated were consistently greater than the theoretical STI. For the 78 ms window, when the theoretical STI generated values below 0.1, the NC method results ranged from 0.1 up to 0.8 and R^2=0.74. When the asymptotic behavior of the metrics was analyzed for speech plus multi-talker babble, the plots were identical in shape to Figure 4, just shifted up slightly to asymptote at 0.6 for the theoretical STI and ER method and 0.8 for the NC method (plot not shown due to space constraints).

CONCLUSIONS

The data presented have demonstrated the ability of two short-time, speech-based, metrics to accurately track short-term fluctuations in STI down to window lengths of 0.3 s for two different noise environments and a noise plus reverberation environment. Because these metrics are speech based, they have the potential to be used in a wide variety of settings to estimate speech intelligibility under conditions not ame-

nable to standard intelligibility measurement techniques such as during live performances. Further investigation is underway to analyze the 78 ms window results more thoroughly.

ACKNOWLEDGEMENTS

This work was supported by NIDCD grant RO1-DC007152.

REFERENCES

Drullman R (1995). Temporal envelope and fine structure cues for speech intelligibility. J Acoust Soc Am 97: 585-592.

Drullman R, Festen JM, Plomp R (1994a). Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am 95: 2670-2680.

Drullman R, Festen JM, Plomp R (1994b). Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 95: 1053-1064.

Goldsworthy RL, Greenberg JE (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. J Acoust Soc Am 116: 3679-3689.

Houtgast T, Steeneken HJM (1971). Evaluation of speech transmission channels by using artificial signals. Acustica 25: 355-367.

Houtgast T, Steeneken HJM (1984). A multi-language evaluation of the RASTI-method for estimating speech intelligibility in auditoria. Acustica 54: 185-199.

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

Houtgast T, Steeneken HJM, Plomp R (1980). Predicting speech intelligibility in rooms from the Modulation Transfer Function I. General room acoustics. Acustica 46: 60-72.

IEC (1998). Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index. 2nd Ed, Internat. Standard No. 60268-16, International Electrotechnical Commission.

Ludvigsen C (1987). Prediction of speech intelligibility for normal-hearing and cochlearly hearing-impaired listeners. J Acoust Soc Am 82: 1162-1171.

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

Payton KL, Braida LD (1999). A method to determine the speech transmission index from speech waveforms. J Acoust Soc Am 106: 3637-3648.

Payton KL, Uchanski RM, Braida LD (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.

Payton KL, Braida LD, Chen S, Rosengard P, Goldsworthy R (2002). Computing the STI using speech as a probe stimulus. In: v. Wijngaarden, SJ (ed.): Past, present and future of the SpeechTtransmission Index (pp 97-119). TNO Human Factors, The Netherlands.

Peterson PM (1986). Simulating the response of multiple microphones to a single acoustic source in a reverberant room. J Acoust Soc Am 80: 1527-1529.

Steeneken HJM, Houtgast T (1980). A physical method for measuring speech-transmission quality. J Acoust Soc Am 67: 318-326.