euronoise

Analysis of short-time speech transmission index algorithms

Karen Payton and Mona Shrestha

ECE Dept., UMass Dartmouth, 285 Old Westport Rd., Dartmouth, MA 02747, USA
kpayton@umassd.edu

Various methods have been shown to compute the Speech Transmission Index (STI) using speech as a probe stimulus [Goldsworthy & Greenberg, *J. Acoust. Soc. Am.*, 116, 3679-3689, 2004]. Frequency-domain methods, while accurate at predicting the long-term STI, cannot predict short-term changes due to fluctuating backgrounds. Time-domain methods also work well on long speech segments and have the added potential to be used for short-time analysis. This study investigates the accuracy of two time-domain STI methods: Envelope Regression (ER) and Normalized Correlation (NC), as functions of window length, in various acoustically degraded environments with multiple talkers and speaking styles. Short-time STIs are compared with a short-time Theoretical STI, derived from octave-band signal-to-noise ratios. For windows as short as 0.3s, the ER and NC Methods track the short-time Theoretical STI and both the Theoretical and ER Methods converge to the long-term result for windows greater than 4s. Short-time STIs are also compared to intelligibility measurements on clear/conversational speech. Correlations between STI and intelligibility scores are high at the sentence and word levels and, consistent with the scores, short-time methods predict a higher average value of STI for clear than for conversational speech.

1 Introduction

The Speech Transmission Index (STI) is a physical metric demonstrated to be correlated with speech intelligibility [1]. A variety of methods have been proposed to compute the STI [2-12]. Some of these methods use speech as the probe stimulus rather than artificially modulated noise as originally proposed by Houtgast and Steeneken [13]. Of the speech-based methods, a subset have been shown to generate the same value as the theoretical STI, which is based on signal-to-noise ratio (SNR) and reverberation time (RT) [8, 12, 14]. To date, all speech-based approaches have used very long speech segments to generate a metric. Consequently, they have not been used to predict short-time changes or word-by-word intelligibility. The current work evaluates the Envelope Regression (ER) and Normalized Correlation (NC) speech-based methods using short windows (~1/3 s) to compute STI values in environments with stationary noise or multi-talker babble backgrounds for speech from multiple talkers speaking conversationally and clearly at normal rates (clear/norm) [15]. One set of speech materials was previously used in listening experiments and the performance of both methods are compared to the listeners' responses at both the sentence and word level. Additional analyses on the asymptotic behavior of the metrics in noise plus reverberation conditions can be found in Payton and Shrestha [16].

2 Methods

For both the ER and NC techniques, the clean and the degraded signals, digitized at 20 kHz using a 9.5 kHz antialiasing filter, were filtered by a bank of 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. For each band, i, the clean and the degraded signals were then squared and lowpass filtered with a cut off frequency of 50 Hz using a 10 ms Hamming window. The intensity envelope signals $x_i(t)$ and $y_i(t)$ were downsampled to 134 Hz to reduce computation time.

The modulation metric for each octave band, M_i, is computed from the envelope signals using Eq.(1) for the ER method

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

where μ_{xi} and μ_{yi} are the means of $x_i(t)$ and $y_i(t)$ respectively. M_i is computed using Eq.(2) for the NC method

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

[12]. The apparent signal-to-noise ratio (SNR) in each band, $aSNR_i$, is 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 and -15 dB. The $aSNR$ in each band is converted to a corresponding transmission index, TI_i, according to Eq.(4):

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

Finally, the overall STI (ranging from 0 to 1) is calculated as a weighted average 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 and β_i's correspond to the octave weighting and redundancy correction factors respectively given in the IEC standard [4].

2.1 Theoretical STI

In order to compare the short-time metrics with a standard, the theoretical STI is also calculated - over the same time windows as the speech-based STI. The clean speech and noise signals are passed through the same bank of octave

filters as before. The modulation index in each band, M_i, is then calculated [17] as

$$M_i = \left(1 + 10^{\frac{-S_i / N_i}{10}} \right)^{-1} . \quad (6)$$

where S_i and N_i are the signal and noise powers respectively. The theoretical STI is then computed using Eq.(3) through Eq.(5).

2.2 Stimuli

The stimuli used in this study are 50 nonsense sentences, either spoken conversationally by a male talker or both conversationally (conv) and clearly at normal rates (clear/norm) by both a male and a female talker. 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 contains four to six key words corresponding to the nouns, adjectives, verbs and adverbs in the sentence.

The clear/norm speaking style is not produced naturally. The talkers were trained to speak this way in order to remove a significant difference between clear and conversational speaking styles: the rate at which they are spoken. Normally, clear speech is produced at about half the speaking rate of conversational speech [15].

2.3 Evaluation Conditions

For the first corpus of speech materials, results for two conditions will be shown: concatenated sentences plus stationary speech-shaped noise at 0 dB SNR and sentences plus multi-talker babble at 0 dB SNR. For the second speech corpus, stationary speech-shaped noise at -1.4 dB was added to individual sentences to match the stimuli presented to listeners.

3 Results

3.1 Comparison of Short-Time Metrics to Theoretical STI

Figure 1 plots regression analyses for speech in stationary noise at 0 dB SNR versus the Theoretical STI computed over the same time window.

The left plot corresponds to the ER method and the right plot corresponds to the NC method. Each point represents the STIs for a windowed segment of the speech. A 0.3 s analysis window was used for these plots. A total of about 1.75 minutes of speech were analyzed.

Note that although the nominal SNR was 0 dB, individual windowed STI values vary from 0 to 0.7. For both metrics, the regression "goodness of fit" term, R^2, is very high: 0.99 for the ER method and 0.96 for the NC method. The most significant difference between the two metrics is that the NC method results are shifted up relative to the ER method and the theoretical STI.

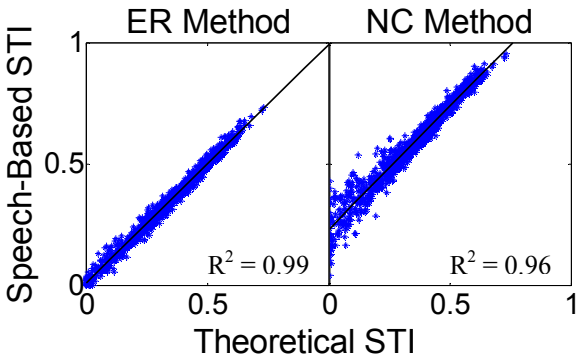

Fig. 1 Regression Analysis - 0dB Speech-Shaped Noise Comparison of ER (left plot) and NC (right plot) methods with theoretical STI computed over 0.3 s windows.

Figure 2 presents the regression analyses for speech plus multi-talker babble at 0 dB SNR for the two metrics. While the R^2 values are slightly less: 0.93 for both, the fits are still extremely good.

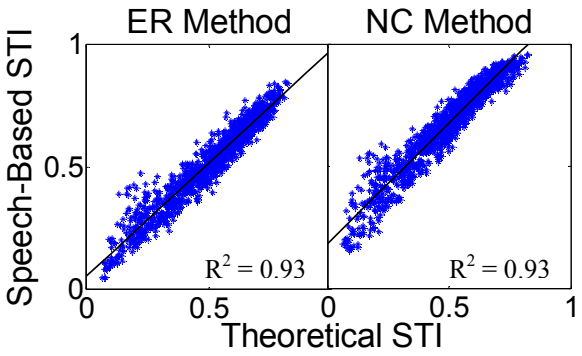

Fig. 2 Regression Analysis - 0dB Multi-Talker Babble Comparison of ER (left plot) and NC (right plot) methods with theoretical STI computed over 0.3 s windows.

For both types of noise, when window lengths were reduced below 0.3 s, the correlations of the ER and NC metrics to the Theoretical STI were less robust. In fact, for windows in which the theoretical STI is zero (e.g. silent intervals), both the ER and NC metrics often generated values greater than zero – as high as 0.4 for the ER method and 0.8 for the NC method. Reasons for this behavior are being investigated.

3.2 Comparison of Metrics to Subject Intelligibility Scores

A second corpus of nonsense sentences, spoken either conversationally (conv) or clearly at normal rates (clear/norm) by a female (RG) and a male (SA) talker in the presence of -1.4 dB speech-shaped noise was also analyzed [15]. The purpose of this analysis was three-fold. First, we wished to verify metric performance on more than one voice. Second, since subject intelligibility data was available for these speech materials and condition, we were

interested in seeing how well the metrics would compare to subject performance at the sentence and word levels. Third, we were interested in how the short-time metrics would perform on different speaking styles for which we have previously demonstrated significant intelligibility differences [6, 15].

To demonstrate the first goal, Fig. 3 plots regression analyses of STI for keywords from talkers RG (top row) and SA (bottom row) for the two metrics (ER on the left and NC on the right).

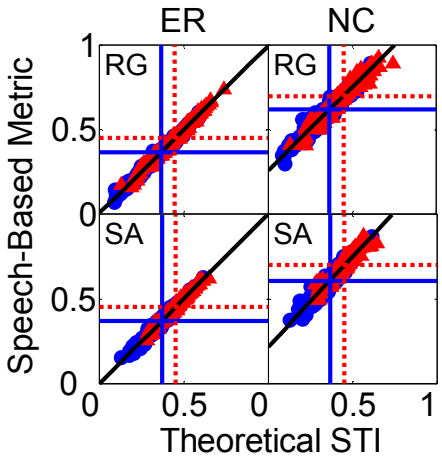

Fig.3 Speech-based Metrics vs. Theoretical STI
Regression analyses for speech based STIs vs. theoretical STI, using windows equal to keyword lengths. Top row: talker RG, bottom row: talker SA. Circles represent conv keyword STIs and triangles represent conv/clear keyword STIs. Solid crosshairs: mean STIs for conv words; dotted crosshairs: mean STIs for clear/norm words.

These figures include data on both speaking styles. The circles correspond to conv speech and the triangles correspond to clear/norm speech and the cross-hairs denote the means with the solid lines corresponding to the conv means and the dotted lines indicating the clear/norm means. Note that, despite the substantial overlap of STIs for the two speaking styles, both talkers' means for clear/norm words are greater for both the theoretical and speech based STIs. The R^2 statistic is 0.97 for the ER method and 0.93 for the NC method for both talkers (fit was made over words from both speaking styles). As was seen for the previous speech materials, the NC method results are shifted upward relative to the ER method results and the theoretical STI.

Next the speech-based STI methods were compared with intelligibility scores for sentences averaged across subjects. In Fig. 4, speech-based STIs are plotted on the horizontal axes and average percent correct values are plotted on the vertical axes.

The R^2 statistics for the data in Fig. 4, 0.34 (ER method) and 0.28 (NC method) for talker RG and 0.45 (ER method) and 0.42 (NC method) for talker SA, are much lower than for the regressions of the metrics against the theoretical STI for word-length windows. The reasons are twofold. First, the goodness of fit statistic is not as meaningful if the data is spread vertically around the mean, as is somewhat true of the sentence data since all sentences were presented at -1.4 dB SNR. Second, there are many more data points in the word regression than there are in the sentence regression analysis (4 to 6 words per sentence).

It should be pointed out that, for both the word-level metric vs. theoretical STI and sentence-level metric vs. intelligibility, the averages for clear/norm are always greater than the averages for conv speech. This means that the metrics are able to capture some aspects of the clear/norm speech that contribute to its higher intelligibility.

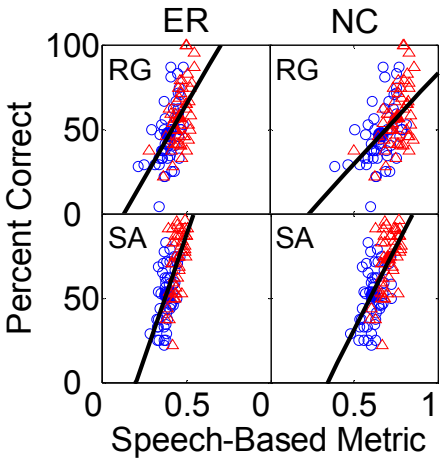

Fig. 4 Speech-based Metrics vs. Percent Correct
Regression analyses for speech-based STI vs. percent correct using windows equal to sentence lengths. Top row: talker RG; bottom row: talker SA. Left column: ER method; right column NC method. Triangles correspond to clear/norm sentences and circles correspond to conv sentences.

In an effort to compare speaking style results more directly, we analyzed the difference in intelligibility between clear/norm and conv sentence pairs vs. difference in the metrics. Figure 5 shows the results for the two methods.

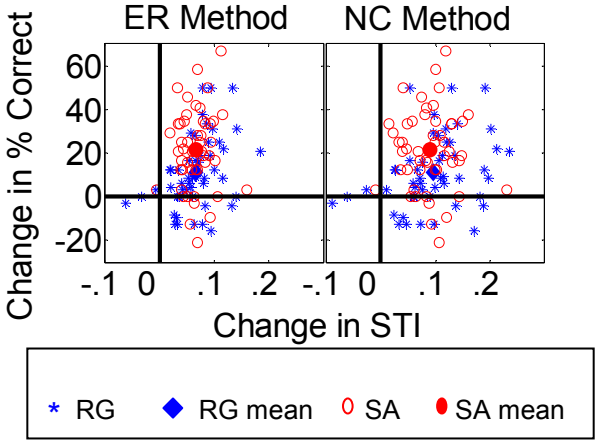

Fig. 5 Change in STI vs Change in Percent Correct Due to Change in Speaking Style.
Stars represent sentences spoken by RG. Circles represent sentences spoken by SA. Solid horizontal and vertical lines mark zero change in percent correct and STI respectively.

The vertical lines in each plot correspond to no change in STI and the horizontal lines correspond to no change in percent correct. Each symbol corresponds to a sentence with the stars corresponding to talker RG and the circles corresponding to talker SA. Symbols in the first and third quadrant correspond to sentence results accurately predicted by the metrics, i.e. clear/norm sentences which

are more intelligible than their conv counterparts and have higher STI values or clear/norm sentences that are less intelligible and have lower STI values. Most sentence pairs and the means for both talkers fall in the first quadrant. A few sentence pairs fall in the third quadrant. The fourth quadrant, with the second largest cluster of sentence pairs - mostly spoken by RG - corresponds to clear/norm sentences which are less intelligible than their conv counterparts but have STI values that are higher. These are the sentences for which the STI metrics changed in the opposite direction from the subject data. These sentences have been analyzed further and it has been determined that the metric values are driven by voiced sounds in the words such as vowels and, despite strong vowels, some key words have low probability of correct identification.

The next step was to consider how SNR changed from the beginning to the end of the sentences. The hypothesis was that some of the clear/norm advantage was due to a maintenance of SNR for later words in the sentences when compared to a reduction of SNR for those words in the conv sentences. As shown in Fig. 6, this hypothesis is supported by the data.

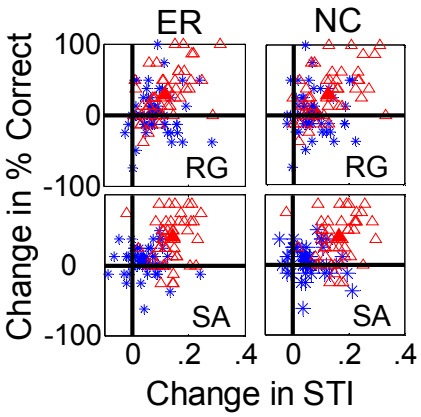

Fig. 6 Change in Percent Correct vs. Change in STI for First and Last Words Due to Change in Speaking Style
Stars correspond to first word pairs. Triangles correspond to last word pairs. Solid horizontal and vertical lines mark zero change in percent correct and STI respectively.

The stars represent the first word changes and the triangles represent the last word changes in STI and percent correct when the talkers change speaking style. The last words have the greatest increase in both STI and percent correct going from conv to clear/norm. This is true for both talkers although the result is more noticeable for talker SA. The main reason for the increase is that most talkers let their voice level fall off as they speak conversationally but maintain a more stable level when they speak clearly, even when speaking clearly at normal speaking rates.

4 Conclusions

The work reported herein has demonstrated two important results. First it has been demonstrated that the short-time STI methods considered can generate accurate STI values down to time scales on the order of 1/3 s in both stationary and fluctuating noise backgrounds. Second, these methods successfully predict intelligibility differences due to speaking style at both the sentence and word level, specifically tracking differences in acoustic features such as SNR word by word through a sentence. Clearly work needs to be done to more thoroughly investigate the limitations of these new methods but they represent promising new ways to objectively predict speech intelligibility in a variety of acoustic environments.

Acknowledgments

The authors wish to thank Dr. Jeanie Krause for providing the stimuli and subject response data presented in Sect. 3.2. We also thank Mr. Kenneth Schutte and Dr. Louis Braida for their comments and suggestions. This work was supported by NIDCD grant -RO1-DC007152-01A2.

References

[1] T. Houtgast, H. J. M. Steeneken, "A multi-language evaluation of the RASTI-method for estimating speech intelligibility in auditoria," *Acustica*, 54, 185-199 (1984).

[2] T. Houtgast, H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, 25, 355-367 (1971).

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

[4] IEC, "Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index," Internat. Electrotech. Commiss. (1998).

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

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

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

[8] K. L. Payton, L. D. Braida, S. Chen, P. Rosengard, R. L. Goldsworthy, "Computing the STI using speech as a probe stimulus," in *Past, Present and Future of the Speech Transmission Index*, TNO, 97-110 (2002).

[9] R. Drullman, J. M. Festen, R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, 95, 1053-1064 (1994).

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

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

[12] R. L. Goldsworthy, J. E. Greenberg, "Analysis of speech-based speech transmission index methods with

implications for nonlinear operations," *J. Acoust. Soc. Am.*, 116, 3679-3689 (2004).

[13] T. Houtgast, H. J. M. Steeneken, "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 (1985).

[14] C. Ludvigsen, C. Elberling, G. Keidser, T. Poulsen, "Prediction of intelligibility of non-linearly processed speech," *Acta Octolaryngol. Suppl.*, 469, 190-195 (1990).

[15] J. C. Krause, "Properties of Naturally Produced Clear Speech at Normal Rates and Implications for Intelligibility Enhancement," Ph.D., Dept. Elec. Eng. Comp. Sci., Mass. Inst. Tech., Cambridge, MA (2001).

[16] K. L. Payton, M. Shresha, "Evaluation of short-time speech-based intelligibility metrics," to be presented at *Proc. Internat. Commiss. Biol. Effects Noise*, Foxwoods Resort, CT (2008).

[17] T. Houtgast, H. J. M. Steeneken, R. Plomp, "Predicting speech intelligibility in rooms from the Modulation Transfer Function I. General room acoustics," *Acustica*, 46, 60-72 (1980).