Intelligibility of speech corrupted by nonlinear distortion

A.J. Brammer[1,2], G. Yu[1], E.R. Bernstein[1], M.G. Cherniack[1], J.B. Tufts[3], D.R. Peterson[1]

[1] Ergonomic Technology Center, University of Connecticut Health Center, 263, Farmington Ave., Farmington CT 06030, USA, brammer@uchc.edu

[2] Envir-O-Health Solutions, 4792, Massey Lane, Ottawa, ON K1J 8W9, Canada

[3] Dept. of Communication Sciences, University of Connecticut, Storrs, CT 06269, USA

INTRODUCTION

Attempts to predict the intelligibility of speech transmitted by a communication system have led to numerous models (see, for example, ANSI 1997; IEC 2003; Christiansen et al. 2010; Elhilali et al. 2003; Kates & Arehart 2005; Payton & Braida 1999; Steeneken & Houtgast 2002a; Yu et al. 2010). Of these, the speech intelligibility index (SII) (ANSI 1997) and the speech transmission index (STI) (IEC 2003) have received most attention. Both provide an index of intelligibility from 0 to 1 based on the speech signal-to-noise ratio in discrete frequency bands. The frequency bands of the SII were originally chosen to reflect the psychoacoustic masking of test sounds by noise (critical bands). The method was later standardized with the speech spectrum alternatively broken down into fewer, broader frequency bands for convenience of calculation (one-third octave, and octave bands from 125 Hz to 8 kHz). The test signals are those naturally occurring in the communication system (i.e., speech and noise, the levels of which need to be separately determined). The STI focuses on the temporal modulation of speech sounds and adopted octave bands as the basis for calculating the modulation spectrum (Steeneken & Houtgast 2002b). It replaced speech by a probe signal to ensure that the modulation could be determined in each modulation frequency band, which have frequencies from 0.63 to 12.5 Hz in the international standard.

In modern communication systems the speech signal is often corrupted by the signal processing and electronic circuitry, as well as by noise, which in some circumstances introduces audible distortion and may degrade intelligibility. The SII and STI have been shown to predict speech intelligibility for a range of conditions in which speech understanding is impeded by continuous noise, but fail when the speech signal is corrupted by nonlinear distortion such as center clipping. In these circumstances the performance of the SII has been improved by calculating the speech signal-to-'noise' (or distortion) ratio from the coherence, which needs to be determined for different amplitude ranges of the speech signal in order to assess the intelligibility (Kates & Arehart 2005). We have explored replacing the test signal of the STI by speech and adjusting the metric for the coherence between the original and corrupted speech, as a means for determining when the observed modulations are due to speech rather than 'noise' (Payton & Braida 1999; Goldsworthy & Greenberg 2004). Also, the contributions to intelligibility from speech information in nearby frequency bands is known not to be independent, and so cannot be simply summed as in some models (e.g., ANSI 1997), resulting in the need to estimate inter-band redundancy (Steeneken & Houtgast 1999; Brammer et al. 2010).

In this paper we briefly describe our models and their application to speech-spectrum shaped noise and center clipping. The latter occurs when a signal within a communication channel rapidly changes polarity from a non-zero value. An example is given

in Figure 1, where the time history of a short segment (0.1 s) of a speech sound is shown (above) as well as a corrupted waveform in which 75 % of the amplitude distribution of the speech sounds has been removed (below).

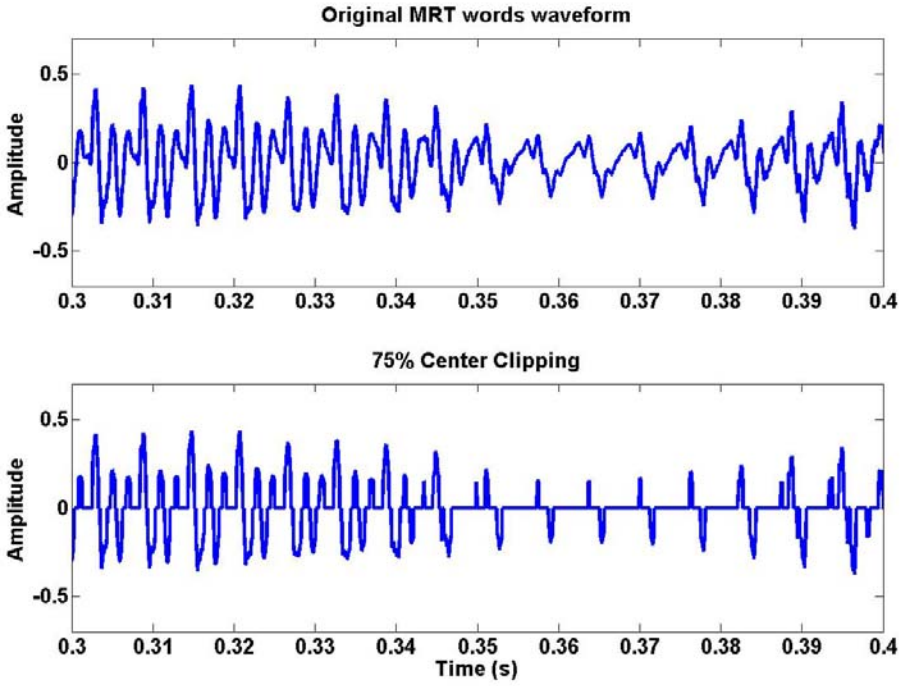

Figure 1: Example of center clipping for a 0.1 s time segment of a speech waveform

METHODS

Speech Transmission Index

The models employ running speech as the input to the communication system, contrary to the conventional STI method specified in IEC 2003, and are described in detail elsewhere. Other essential components of the traditional STI model are, however, retained, viz. establish changes in the temporal modulation of speech sounds within octave bands with center frequencies from 125 Hz to 8 kHz. Specifically, the ratio of the input to output modulations is first determined in one-third octave bands (with center frequencies, f, from 0.63 to 12.5 Hz) from the intensity envelopes of the signals in each octave band (k). The magnitude squared coherence between the input and output modulation envelopes is then computed for each modulation frequency band, $\gamma_{k,f}$. The coherence is used to correct the measured modulation reduction, $mi_{k,f}$, to account for situations in which the measured modulation reduction is influenced by the competing noise or distortion (i.e., when the coherence will be reduced):

$$mc_{k,f} = mi_{k,f} * \gamma_{k,f}^2 \quad (1)$$

The corrected modulation reduction, $mc_{k,f}$, is converted to a signal-to-noise (or distortion) ratio, which in turn is adjusted to range from 0 to 1 so that a transfer index can be computed for each modulation frequency band (within each octave band). A modulation transfer index, MTI_k, is then computed for each octave band from the arithmetic mean of the transfer indices for that octave band, and used to construct a speech-based STI, according to:

$$STI_{speech} = \sum_{k=1}^{7} \alpha_k MTI_k - \sum_{k=1}^{6} \beta_k \sqrt{MTI_k * MTI_{k+1}} \quad (2)$$

In this equation, α_k are numerical coefficients that represent the contribution to speech intelligibility from sounds in each octave band. They are equivalent to the 'importance functions' employed in calculations of the SII and are empirically determined from speech tests with human subjects. The second term was introduced with empirical coefficients (β_k) to account for inter-band contributions to intelligibility arising from speech information common to adjacent octave bands (Steeneken & Houtgast 1999).

The information common to the speech signal in different octave bands may be formally identified by calculating the normalized cross-covariance between intensity modulations in different octave bands, $\rho_{k,j}$ (j=k+1, …7). For the results reported in this paper we have included interactions between three adjacent octave bands based on observations of real speech (i.e., j=k+1, k+2), and computed the mean cross covariance. The model for predicting intelligibility then becomes:

$$STI_{\rho-speech} = \sum_{k=1}^{7} \alpha_k MTI_k - \frac{1}{6} \sum_{k=1}^{6} \sum_{j=k+1}^{k+2} \left[\rho_{k,j}^2 * MTI_k * MTI_j \right]^{1/2} \quad (3)$$

STI values were computed from sound pressures levels (SPLs) recorded at the center-head position in the absence of the subject.

Modified rhyme test

The modified rhyme test (MRT) was used to characterize the intelligibility of individual words in a six-alternative forced choice paradigm (ANSI 1989). The word lists were those standardized for American English and were a commercial recording by a male talker. Speech sounds were reproduced in an anechoic chamber by a small, high fidelity, low distortion loudspeaker positioned 2.4 m in front of the subject with tweeter at ear height (Paradigm Signature S1 v2) (see Figure 2).

Speech-spectrum shaped noise was produced by four loudspeaker towers surrounding the subject at a distance of almost 2.0 m (each tower consisting of a JVC SRX715F woofer and tweeter, and SRX718 sub-woofer). The signals driving the loudspeaker towers were processed to yield a pseudo-diffuse sound field in the horizontal plane at the center-head position (Yamaha SPX2000 and SP2060). The sound levels were equalized to produce a flat frequency response from 80 Hz to 10 kHz (± 3 dB).

Center-clipped speech was obtained by removing the central 10% to 98% of the long-term histogram of the speech waveform, which consisted of the concatenated MRT words and carrier phrases with the silence between test phrases removed. The distorted signals were adjusted to the same SPL as the unperturbed speech, and replayed to the subject at 65 dBA as measured at the center-head position (in the absence of the subject). The speech-spectrum shaped noise was also reproduced at 65 dBA, and the SPL of the MRT word and carrier phrase were adjusted to the desired speech signal-to-noise ratio (from -17 to +10 dB).

Subjects sat with a computer-controlled touch screen at a convenient height in front of them (see Figure 2) and initiated each test sentence by pressing the 'play' button,

whereupon one of six words displayed on the screen was randomly presented within the carrier phrase. Subjects were instructed to choose one of the six words by touching the screen, and then initiate the next trial when they were ready. There were 50 trials in each measurement block from which a word score was derived for a given speech signal-to-noise or distortion. Each measurement was repeated on three separate occasions. Word scores from the three replications were first averaged for each subject before the mean scores were combined across subjects.

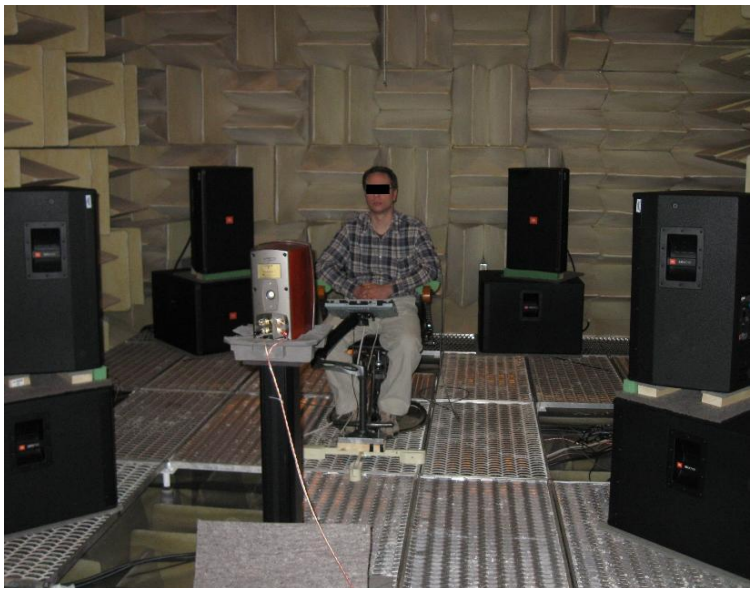

Figure 2: Determination of word intelligibility

Subjects

Six healthy subjects ranging in age from 25 to 45 years with normal hearing (3 male and 3 female) participated in all the experiments. Hearing thresholds were determined at study induction and were better than 20 dB HL at audiometric frequencies from 500 Hz to 8 kHz: in addition, the differences in HLs between ears were less than 10 dB. All were native speakers of American English and were paid for their time.

All subjects gave their informed consent to participate in the study, which was conducted according to the provisions of the university's ethics review board.

RESULTS AND DISCUSSION

Results are presented in Figure 3 for the two STI models described in this paper (STI_{speech} and $STI_{\rho\text{-}speech}$), and for the most recent version of the conventional STI incorporated in the IEC standard (IEC 2003), the so-called STI_r (shown by the line for speech-shaped noise).

The first model, STI_{speech}, employs a speech probe signal in which the change in the intensity modulation from input to output has been corrected for the magnitude squared coherence between the original and corrupted speech signals, according to equation 1. The coefficients for the contributions to intelligibility from each octave frequency band (α_k) as well as the redundancy corrections (β_k) are those contained in the IEC standard (IEC 2003), and the metric has been constructed according to equation 2. MRT scores and STI values for this model are shown by the open symbols in Figure 3.

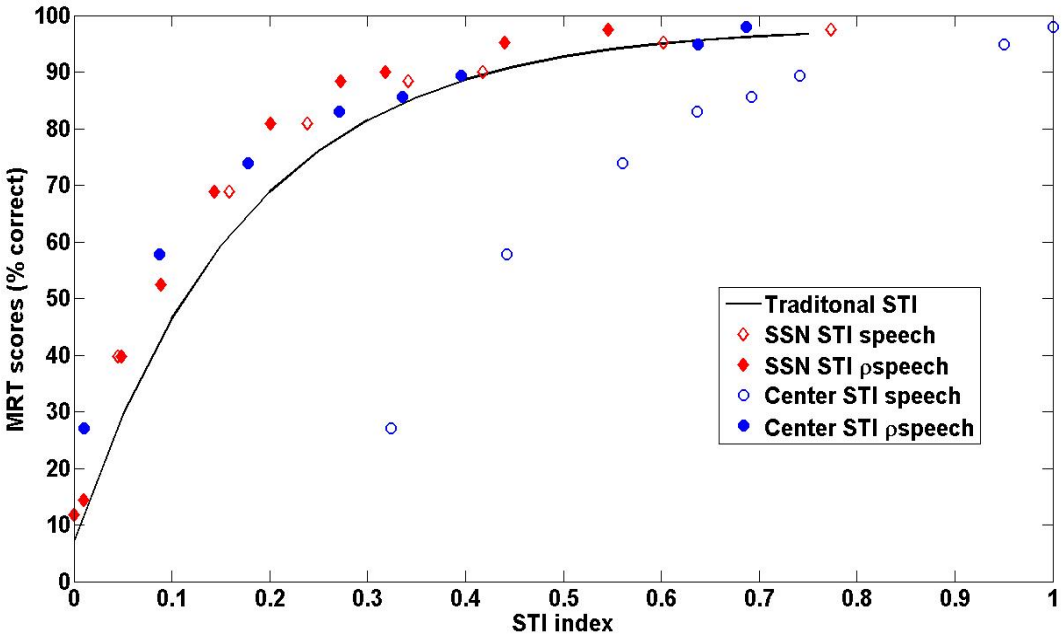

Figure 3: Mean MRT scores and predictions using three models for the STI: open symbols - STI_{speech}, filled symbols - $STI_{\rho\text{-}speech}$, and line - STI_r; 'SSN' - speech-spectrum shaped noise, and 'Center' - center clipping

The second model, $STI_{\rho\text{-}speech}$, also employs a speech probe signal in which the change in the intensity modulation from input to output has been corrected for the magnitude squared coherence between the original and corrupted speech signals. However, in addition, the empirically determined redundancy factors of the IEC standard (IEC 2003) have been replaced by the measured normalized cross covariance of the intensity modulations between adjacent and next nearest neighbor octave bands, according to equation 3. The coefficients for the contributions to intelligibility from each octave frequency band (α_k) are again those contained in the IEC standard (IEC 2003). MRT scores and STI values for this model are shown by the filled symbols in Figure 3.

Reference to Figure 3 shows that both models produce similar results and a similar relationship between model values and MRT word scores for speech masked by speech-spectrum shaped noise (the diamonds). The relationship between MRT scores and the two STI indices differ somewhat from the traditional STI_r, shown by the line, with STI_{speech} and $STI_{\rho\text{-}speech}$ predicting higher word scores for a given value of the index. Equally important, for a given word score, equations 2 and 3 produce values for the indices that are less than that produced by the traditional STI_r, suggesting that the balance between the two terms in equations 2 and 3 may need adjusting.

For speech corrupted by center clipping, it is evident from Figure 3 that only the model incorporating a measure of the correlation between sounds in adjacent and next nearest neighbor octave bands can produce values of the metric that are essentially the same as those for speech-spectrum shape noise. This may be seen by comparing the filled symbols. In contrast, the values of the metric produced by the other model (STI_{speech}), shown by open circles, deviate substantially from the values for

speech-spectrum shaped noise. A prediction of intelligibility for nonlinear distortion cannot be obtained by the traditional STI method.

It should be noted that the model for speech corrupted by nonlinear distortion developed here contains only seven adjustable parameters, the α_k, which have been assigned the values in the IEC standard (IEC 2003). Thus no attempt has been made to optimize the fit of the model to the word scores. An equivalent model for the SII would require 22 fitted parameters to characterize speech corrupted by center clipping (Kates & Arehart 2005).

CONCLUSIONS

It thus appears that a STI model accounting for the redundancy between speech information contained in nearby octave bands is required to predict the intelligibility of speech corrupted by center clipping. The model proposed here, $STI_{\rho\text{-}speech}$, which includes the mean of the measured contributions to the normalized cross covariance from nearest neighbor and next nearest neighbor bands according to equation 3, can account for center clipping from 10 to 98 % of the speech amplitude-time histogram.

[Work supported by the National Institute for Occupational Safety and Health under grant 5 R01 OH008669-05]

REFERENCES

ANSI S3.2-1989 (1989). Method for measuring the intelligibility of speech over communication systems. , New York: American National Standards Institute.

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

Brammer AJ, Yu G, Bernstein ER et al. (2010). Predicting the intelligibility of speech corrupted by nonlinear distortion. Can Acoust 38: 80-81.

Christiansen C, Pedersen MS, Dau T (2010). Prediction of speech intelligibility based on an auditory preprocessing model. Speech Commun 52: 678–692.

Elhilali M, Chi T, Shamma SA (2003). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Commun 41: 331-348.

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

IEC 60268-16 (2003). Sound system equipment—Part 16: Objective rating of speech intelligibility by speech transmission index. (3rd ed.) Geneva: International Electro-technical Commission.

Kates J, Arehart K (2005). Coherence and the speech intelligibility index. J Acoust Soc Am 117: 2224–2237.

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

Steeneken HJM, Houtgast T (1999). Mutual dpendence of the octave-band weights in predicting speech intelligibility. Speech Commun 28: 109-123.

Steeneken HJM, Houtgast T (2002a). Validation of the revised STIr method. Speech Commun 38: 413-425.

Steeneken HJM, Houtgast T (2002b). Basics of the STI measuring method. In: Houtgast T, Steeneken H (eds): Past, present and future of the Speech Transmission Index (pp 13-43). Soesterberg: TNO Human Factors.

Yu G, Brammer AJ, Swan K et al. (2010). Relationships between the modified rhyme test and objective metrics of speech intelligibility. J Acoust Soc Am 127:1903.