PREDICTING THE INTELLIGIBILITY OF SPEECH CORRUPTED BY NONLINEAR DISTORTION

A.J. Brammer[1], G. Yu[1], E.R. Bernstein[1], M.G. Cherniack[1], J.B. Tufts[2], and D.R. Peterson[1]
[1]Ergonomic Technology Center, University of Connecticut Health Center, Farmington CT 06030, U.S.A.
[2]Department of Communication Sciences, University of Connecticut, Storrs CT 06269, U.S.A.

1. INTRODUCTION

Common methods for predicting the intelligibility of speech, the speech transmission index (*STI*) and the speech intelligibility index (*SII*), fail when the speech signal is corrupted by nonlinear distortion, e.g., center clipping (Steeneken & Houtgast, 2002). This limitation restricts the applicability of the metrics to many digital communication systems, where peak and center clipping are often unwanted byproducts of the signal processing. For distorted speech, the performance of the *SII* may be improved by calculating the speech signal-to-'noise' (or distortion) ratio from the coherence for three amplitude ranges, and combining the results to assess the intelligibility (Kates and Arehart, 2005).

In this paper we explore modifications to the *STI* to permit estimates of the intelligibility of speech corrupted within a communication system by center clipping. This type of distortion occurs when a signal rapidly changes polarity from a non-zero value, and is associated with quantization errors in digital signal processing. The development of models is first briefly described, followed by a summary of the psychophysical experiment employed to evaluate their performance. Results are presented for word intelligibility in speech-spectrum shaped noise, which is well predicted by the *STI*, and for speech subjected to center-clipping.

2. METHOD

2.1 Speech Transmission Index

The original *STI* was based on determining changes in the temporal envelope modulations of a test signal that substituted for real speech. The analysis involves computing an intensity modulation transfer function for seven octave bands from 125 to 8,000 Hz. A modulation transfer index is constructed for each octave band, MTI_k (k = 1, . .7), from the average intensity modulations within fourteen one-third octave bands with frequencies from 0.63 to 12.5 Hz, $TI_{k,f}$ (f = 1, . .14). The contributions to intelligibility are considered in two ways in the most recent version of the *STI*, the so-called revised *STI*, STI_r (Steeneken & Houtgast, 2002): 1) direct contributions from each octave band, which employ an empirically determined frequency dependent factor, α_k, and 2) indirect, or inter-band, contributions that are termed redundancy corrections, β_k, and arise from the observation that the energy of speech sounds in adjacent frequency bands may be correlated.

An alternative formulation consisting of a combination of the artificial test signal and an octave-band speech probe was necessary for predicting the intelligibility of speech corrupted by peak clipping, and noise. The use of solely speech as the probe signal and, in particular, the need to consider the coherence between speech and noise intensity modulations was discussed by Payton and Braida (1999). Our models build on this work, by employing a speech probe and formally introducing the coherence function between the original and corrupted speech, $\gamma^2_{k,f}$, to generalize the interaction between speech and interfering noise or distortion on the MTI_k, in the following way:

$$MTI_k = \frac{1}{14}\sum_{f=1}^{14} TI_{k,f} * \gamma^2_{k,f} \qquad (1)$$

The redundancy between speech information contained in different octave bands is treated formally by introducing the normalized cross-covariance function between intensity modulations in nearby bands, $\rho_{k,j}$ (j = k+1, . .7) In this paper we explore the consequence of including interactions between three adjacent octave bands (i.e., j = k+1, k+2), which coincides with the non-zero coefficients of the cross-correlation matrix observed for uncorrupted speech. The model for predicting intelligibility is then:

$$STI_{\rho-speech} = \sum_{k=1}^{7} \alpha_k MTI_k + \frac{1}{6}\sum_{k=1}^{6}\sum_{j=k+1}^{k+2}\sqrt{\rho^2_{k,j} * MTI_k * MTI_j} \qquad (2)$$

2.2 Modified Rhyme Test

The modified rhyme test (MRT) was used to characterize speech intelligibility (ANSI, 1989). A subject with normal hearing and confirmed understanding of American English was seated in an anechoic chamber. Speech was reproduced by a small, high-fidelity, low distortion loudspeaker located in front of the subject on the ear-nose plane, 2.4 m from the center-head position (Paradigm S1). Center-clipped speech was produced by removing the central 10% to 98% of the long-term histogram of the waveform of the carrier sentences and MRT words (Fig. 1). Speech-spectrum shaped noise was produced by four loudspeaker towers surrounding the subject (JVC SRX715Fs + SRX718Ss), with output processed to produce a flat, pseudo-diffuse field in the horizontal plane at the center-head position (±3 dB).

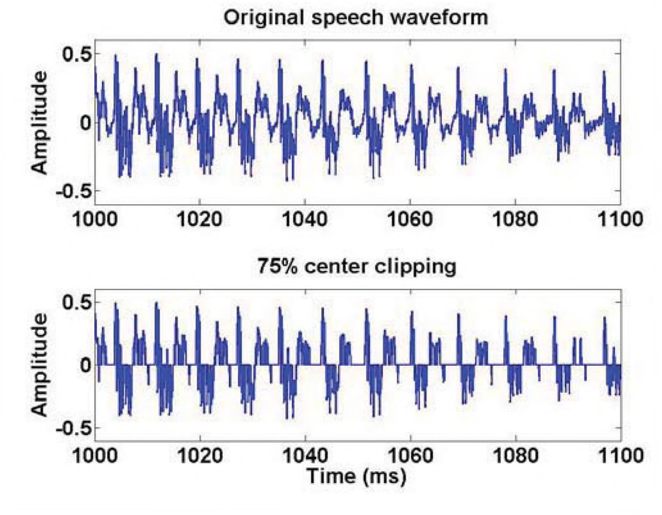

Fig. 1. Example of 75% center clipping shown for a 10 ms duration segment from a speech waveform.

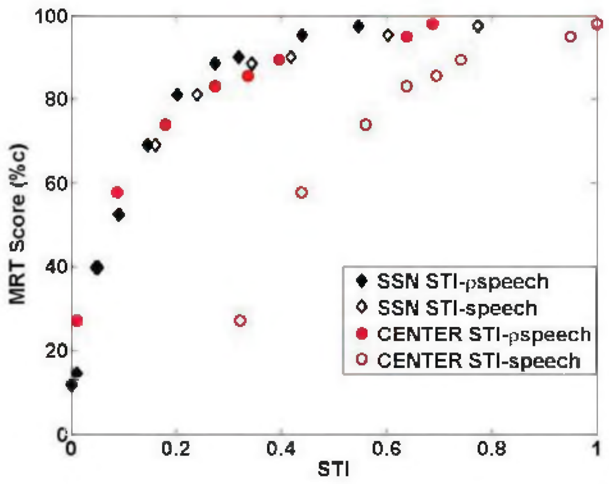

Fig. 2. Mean MRT scores and predictions using two *STI* models for: speech-spectrum shaped noise – diamonds, and; center clipping – circles.

All subjects (3 male, 3 female) gave their informed consent to participate in the study according to the provisions of the University's institutional review board.

3. RESULTS AND DISCUSSION

Mean MRT scores are presented in Fig. 2 for three replications of each experimental condition by each subject, for two *STI* models. The first model, *STI-speech*, consists of a speech probe in which values of the MTI_k are calculated from the transmission indices and the coherence between the original and modified speech according to eqn. 1. The transmission indices are computed following the procedure described in IEC 60628-16, 2003, and the contributions of each octave band to the intelligibility employ the values of α_k and β_k contained in the standard. The MRT scores are shown by unfilled diamonds for speech in speech-spectrum shaped noise, and unfilled circles for speech corrupted by center clipping. The second model, *STIρ-speech*, also employs a speech probe and again introduces the coherence to calculate MTI_k as in the first model. However, the second model replaces the empirically determined redundancy factors (β_k) of Steeneken and Houtgast, and the international standard, by the measured cross-covariance of the intensity modulations between adjacent and the next nearest neighbor octave bands, as given by eqn.2. For this model the MRT scores are shown by filled diamonds for speech in speech-spectrum shaped noise, and filled circles for speech corrupted by center clipping.

Inspection of Fig. 2 reveals that the model employing speech as the probe signal but not accounting for the correlated energy in adjacent (octave) frequency bands (*STI-speech*) yields *STI* values that depend on the competing or distorting sounds interfering with word intelligibility (i.e., the unfilled symbols). In contrast, the model incorporating a measure of the correlation between sounds in adjacent octave bands (*STIρ-speech*) yields *STI* values that are substantially the same irrespective of whether the interfering sounds are speech-spectrum shaped noise or the distortions introduced by center clipping (i.e., compare filled symbols).

It thus appears that a model accounting for the redundancy between energy in adjacent octave bands is required to predict the intelligibility of speech corrupted by center clipping using the *STI*. The formulation proposed here (*STIρ-speech*) is one of a family of models constructed using generalizations of eqn. 2 to introduce varying degrees of interaction between the speech information in different frequency bands. It should be noted that the results for both models described employ only seven adjustable parameters, the values of α_k, in contrast to the 22 adjustable parameters introduced to model speech corrupted by nonlinear distortion using the *SII* (Kates and Arehart, 2005).

REFERENCES

ANSI S3.2-1989 (1989). American National Standard: Methods for the Intelligibility of Speech over Communication Systems (American National Standards Institute, New York).

IEC 60628-16 (2003). Sound System Equipment, Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index (International Electrotechnical Commission, Geneva).

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

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.

Steeneken, H.J.M., and Houtgast, T. (2002). Validation of the revised STI$_\mathrm{r}$ method. Speech Comm., 38, 413-425.

ACKNOWLEDGEMENTS

This work was supported by NIOSH research grant R01 OH008669.