Effect of talker and speaking style on the Speech Transmission Index (L)

Sander J. van Wijngaarden[a] and Tammo Houtgast
TNO Human Factors, PO Box 23, 3769 ZG Soesterberg, The Netherlands

(Received 5 March 2003; revised 17 October 2003; accepted 30 October 2003)

The Speech Transmission Index (STI) is routinely applied for predicting the intelligibility of messages (sentences) in noise and reverberation. Despite clear evidence that the STI is capable of doing so accurately, recent results indicate that the STI sometimes underestimates the effect of reverberation on sentence intelligibility. To investigate the influence of talker and speaking style, the Speech Reception Threshold in noise and reverberation was measured for three talkers, differing in clarity of articulation and speaking style. For very clear speech, the standard STI yields accurate results. For more conversational speech by an untrained talker, the effect of reverberation is underestimated. Measurements of the envelope spectrum reveal that conversational speech has relatively stronger contributions by higher ($>$ 12.5 Hz) modulation frequencies. By modifying the STI calculation procedure to include modulations in the range 12.5–31.5 Hz, better results are obtained for conversational speech. A speaking-style-dependent choice for the STI modulation frequency range is proposed. © *2004 Acoustical Society of America.* [DOI: 10.1121/1.1635411]

PACS numbers: 43.71.Gv, 43.71.Hw [KWG] Pages: 38–41

I. INTRODUCTION

The Speech Transmission Index (IEC, 1998; Steeneken and Houtgast, 1980) is a physical measure for objectively predicting the intelligibility of speech. The Speech Transmission Index (STI) model uses modulation transfer functions (MTFs) to predict intelligibility under influence of a wide diversity of speech degradations, including degradations of a temporal nature, such as reverberation and echoes (Houtgast *et al.*, 1980). Through the modulation transfer function, these influences are translated into ‘‘equivalent speech-to-noise ratios,’’ and then treated in essentially the same way as additive noise.

The STI method was designed and optimized to yield representative and homogeneous intelligibility predictions across all kinds of speech degradation, including noise and reverberation. This was validated using consonant–vowel–consonant (CVC) words (Steeneken and Houtgast, 1980), and also found to be true for short, redundant sentences (Duquesnoy and Plomp, 1980).

However, recent experiences with SRT sentences based on more conversational speech (van Wijngaarden *et al.*, 2001) indicate a tendency for the STI to underestimate the effect of reverberation on sentence intelligibility. Indications for a mismatch between subjective intelligibility and the STI in combined ‘‘noise plus reverberation’’ conditions are also found in other studies (Payton *et al.*, 1994; Fig. 10, triangular data points on the left). The mismatch reported by Payton *et al.* (1994) seems to depend on speaking style, and is larger for a conversational than for a clear speaking style. This could suggest that the difference may be due to differences in speaking style.

In the next section, experiments along the lines of Duquesnoy and Plomp (1980) are described, in which the effect of noise and (simulated) reverberation on the STI is studied for talkers differing in speaking style.

The version of the STI method used throughout this letter is the revised STI ($\mathrm{STI_r}$), based on the most recent version of the standard available at the time this study was carried out (IEC, 1998).

II. SENTENCE INTELLIGIBILITY IN NOISE AND REVERBERATION

A. Method

The speech reception threshold (SRT; Plomp and Mimpen, 1979) is the speech-to-noise ratio at which 50% intelligibility of short, redundant sentences is realized. The original corpus of speech recordings made by Plomp and Mimpen has seen extensive application, and was also included in the present study.

A new, much larger, corpus of SRT test sentences is the ‘‘VU’’ corpus (Versfeld *et al.*, 2000). The sentences by the male talker of the VU corpus were used in this experiment. Versfeld *et al.* present the VU sentences as roughly equivalent to the Plomp and Mimpen sentences. However, the authors of this article perceive the adopted speaking style to be less clear.

A third corpus of SRT sentences is the multilingual SRT (ML-SRT) database (van Wijngaarden *et al.*, 2001; 2002). This corpus consists of material by many nonprofessional talkers in various languages. The single male Dutch talker used in the present study speaks less clearly than the VU talker, and certainly less clearly than the Plomp and Mimpen talker.

The masking noise used in the SRT procedure was noise with the same long-term spectrum as speech by the corresponding talker. Noise was mixed with the target speech samples, after which this signal was convolved with suitable (synthetic) impulse responses to recreate reverberant speech

[a] Electronic mail: vanwijngaarden@tm.tno.nl

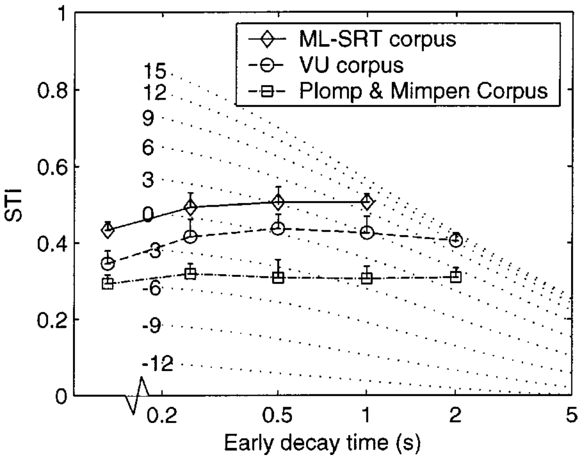

FIG. 1. STI at the SRT (native Dutch listeners), for conditions with and without synthetic reverberation. The dotted lines indicate the maximum STI at each EDT, as a function of the SNR. The error bars indicate the standard deviation (8 listeners, each 2 SRT measurements per condition). The leftmost data point of each line represents a condition without reverberation.

in noise. Use was made of synthetic instead of real impulse response, to exclude effects of nonsystematic differences in timbre for different early decay times (EDTs). In terms of the modulation transfer function, the resulting pseudo-reverberant conditions are identical to purely exponentially decaying naturally reverberant conditions of the same EDT.

B. Results

Duquesnoy and Plomp (1980) measured the SRT as a function of EDT, and then evaluated the STI at this speech-to-noise ratio (the "STI at the SRT"). They found that 50% sentence intelligibility always corresponds to the same STI, whether noise is the predominant speech degrading factor or reverberation. This experiment was essentially repeated, this time using not only the Plomp and Mimpen (1979) talker used by Duquesnoy and Plomp, but also the more conversational speech taken from the two other corpora described above.

Figure 1 shows STI at the SRT results for the three different talkers. First, the individual SRT was measured in a number of reverberation conditions. From this, the STI at the SRT (the STI corresponding to 50% sentence intelligibility) was calculated. If the STI model predicts effects of reverberation as accurately and unbiased as effects of noise (when related to sentence intelligibility), then the lines in Fig. 1 must be straight and horizontal.

The three talkers represented in Fig. 1 differ in terms of their average intelligibility; the three lines differ significantly. Figure 1 also clearly shows that 50% sentence intelligibility sometimes corresponds to a higher STI *with* than *without* reverberation. For the Plomp and Mimpen talker, the line in Fig. 1 follows the theoretical straight and horizontal line. There is a significant difference only between the STI without reverberation and the STI at $\mathrm{EDT}=0.25\ \mathrm{s}$, but this difference is relatively small. This essentially replicates the results found by Duquesnoy and Plomp (1980). For the VU and ML-SRT talkers, there is a mismatch; the STI without reverberation differs significantly ($p<0.05$) from the STI in any reverberation condition.

III. EXPLANATION FOR THE EFFECT OF SPEAKING STYLE

A. Trends observed in the data

For the upper two lines in Fig. 1, the STI at the SRT is clearly higher at a (relatively small) EDT of 0.2 s than in the absence of reverberation. This implies that even a small amount of reverberation may have an impact on intelligibility, to a degree not predicted by the STI model. This effect only appears for the two talkers adopting a more informal, conversational speaking style.

These observations can be explained by assuming that the relation between the envelope spectrum of speech and intelligibility depends on speaking style. The way that the STI model relates intelligibility to the modulation transfer function is apparently quite suitable for some talkers (and speech styles), but less so for others.

B. Between-corpus differences in the speech envelope spectrum

The STI model uses a fixed (logarithmic) set of 14 modulation frequencies ranging from 0.63 to 12.5 Hz, at 1/3-octave intervals. This represents, more or less, the modulation frequency range observed in natural speech. The envelope spectrum of speech normally shows a maximum around 3 Hz, and contains almost all of its energy in the range from 0–30 Hz.

The modulation frequency range in the STI model, and the choice to give each modulation frequency equal weight, are design choices, optimized to make the STI equally sensitive to all sorts of degradations in time and frequency domain. The chosen range was shown to be appropriate for CVC nonsense words (Steeneken and Houtgast, 1980). As shown above, this validation sometimes holds for short sentences, but apparently only for clear speech by a trained talker. If differences in clarity of articulation and speaking style translate into differences in the envelope spectrum, something may be said for adopting different modulation frequency weighting schemes for different speaking styles.

Envelope spectra were calculated from the recorded SRT sentences. The method for calculating envelope spectra essentially follows the procedure originally proposed in the context of the STI model (Houtgast *et al.*, 1980), but is implemented in digital algorithms rather than analog hardware. The speech (sampled at 44 100 Hz) is band filtered into the seven audio-frequency octave bands used by the STI model. Next, the modulation spectrum is derived from the squared signal by means of a discrete Fourier transform. The obtained line spectrum is normalized by its dc component to allow interpretation in terms of modulation indices, and binned into 1/3-octave bands in the range from 0.40 to 31.5 Hz. This gives a separate modulation spectrum for each of the audio-frequency octave bands.

As shown in Fig. 2, the envelope spectra for the three different speech materials all have the usual maximum around 3–4 Hz, but show differences in magnitude. Results

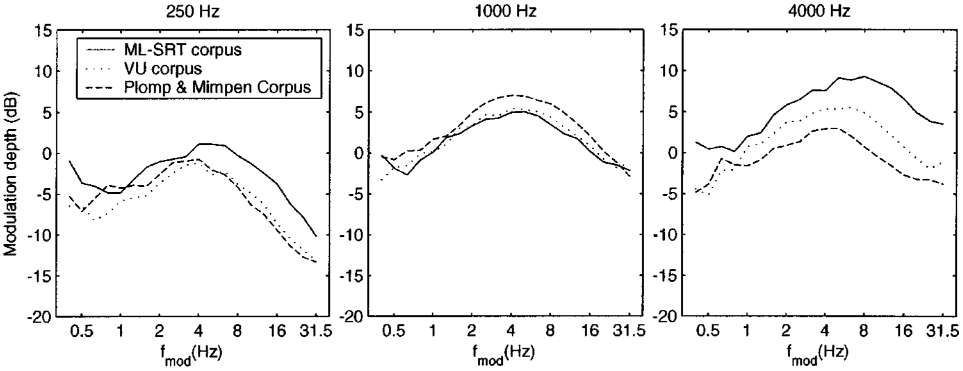

FIG. 2. Averaged envelope spectra (250-, 1000-, and 4000-Hz audio frequency bands) of speech by three different talkers.

appear different for each individual audio-frequency octave band, making it difficult to detect systematic differences due to the speech material.

By inspecting envelope spectra such as Fig. 2, the only (subtle) trend that may be observed is that for the clearer Plomp and Mimpen sentences, the energy in the envelope spectrum appears to be concentrated more around the maximum at 3 Hz. It spreads a smaller fraction of its total energy to higher modulation frequencies.

To investigate whether this is a systematic effect, frequency-integrated versions of the envelope spectra are calculated, averaged across audio frequency, and normalized by dividing through their cumulative maximum (making the value at 31.5 Hz, the highest measured modulation frequency, equal to 1). Figure 3 shows these integrated (or cumulative) spectra for the three different speech materials, integrated from 1 Hz upward.

The tendency in Fig. 3 appears to be that the envelope spectrum of clearer speech shows relatively smaller contributions by the higher modulation frequencies. The modulation frequencies in Fig. 3 not taken into account by the STI model (>12.5 Hz) represent only a small portion of the total energy for the Plomp and Mimpen corpus, but are of greater importance for the ML-SRT and VU material.

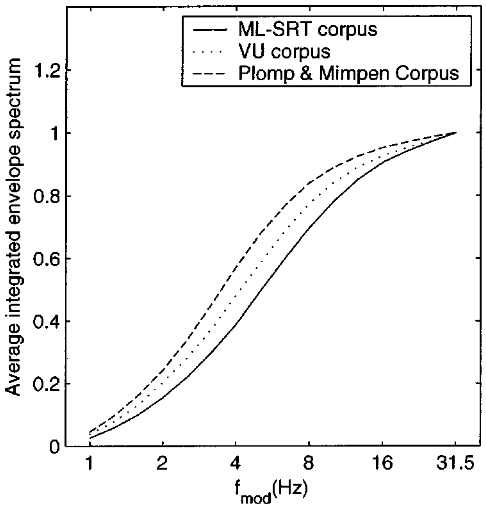

FIG. 3. Integrated (cumulative) envelope spectra of speech by three different talkers. The square of modulation index *m* was integrated from 1 Hz upward, and averaged across the audio frequency octave bands 125–8000 Hz.

C. Adapting the STI method by using a wider modulation frequency range

A straightforward first step in trying to adapt the STI model for more conversational speech would be to extend the modulation frequency range to 31.5 Hz, maintaining equal weight for all modulation frequencies. The modulation frequencies remain separated by 1/3 octave, so the extension to 31.5 Hz increases the number of modulation frequencies from 14 to 18.[1]

Figure 4 is based on the same SRT data as Fig. 1, this time with the modulation frequency range for the STI calculation extended to 31.5 Hz. The ML-SRT data in Fig. 4 show a much closer resemblance to the expected horizontal line than in Fig. 1. The same is true for the VU data, even if some dependence of the STI on the EDT is still observed (the STI at EDT=0.50 differs significantly from the STI without reverberation). Only for the Plomp and Mimpen data, Fig. 1 fits the expected horizontal line better. This confirms the expectations based on the modulation spectra of Fig. 3.

D. Envelope spectra for a larger population of talkers

Given the differences in modulation spectra for the three SRT talkers, the question arises what variations may be expected for a greater population of (arbitrarily selected) talkers.

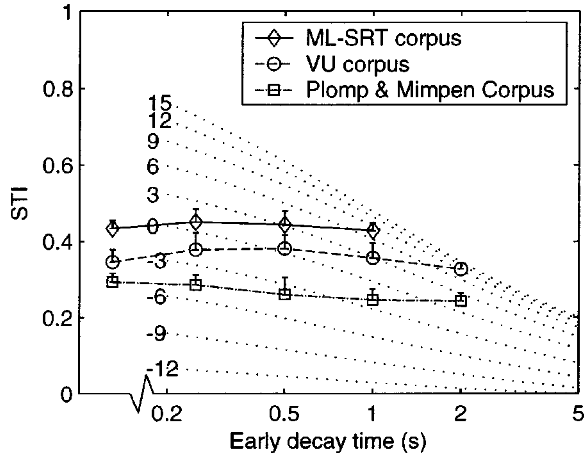

FIG. 4. STI at the SRT results, based on the same data as Fig. 1, but with a wider modulation frequency range (0.63–31.5 Hz). The dotted reference lines (STI vs EDT as a function of SNR) are also based on this extended modulation frequency range.

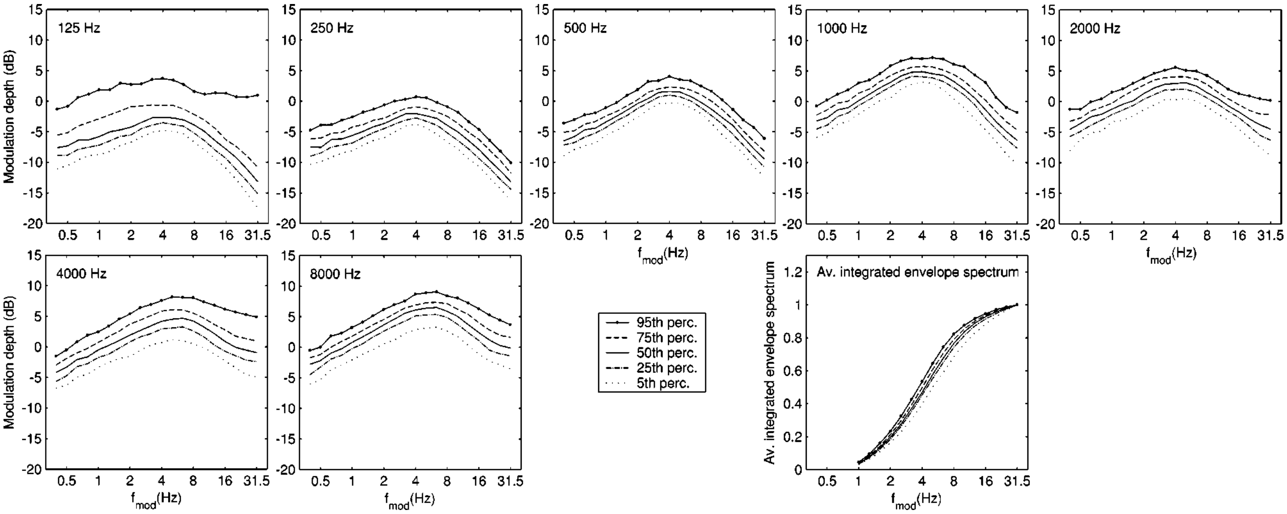

FIG. 5. Envelope spectra (125–8000-Hz audio frequency octave bands) of speech by 134 different talkers. The data are represented by the 5th, 25th, 50th, 75th, and 95th percentile. The corresponding integrated (cumulative) envelope spectra (the square of modulation index *m* integrated from 1 Hz upward and averaged across all audio frequency octave bands) are also given.

Speech material (Dutch newspaper sentences) read aloud by 134 (male and female) native Dutch talkers, taken from the NRC corpus (van Leeuwen and Orr, 2000), was subjected to the same modulation spectrum calculations as the SRT sentences. The NRC corpus consists of high-quality recordings of untrained talkers, screened for impairments, but otherwise randomly selected.

Figure 5 shows that the maximum of the envelope spectrum shifts slightly (from approximately 3 to 4 Hz) for higher audio frequencies. The statistical spread is considerable, especially for the higher frequency bands. The percentile curves were calculated separately for each band. The percentile belonging to a certain talker in a certain band was found to have no predictive value for the percentile this talker would correspond to in other bands; no systematic correlations were found. The bottom right panel of Fig. 5 shows 5th–95th-percentile versions of the integrated envelope spectrum, derived from the individual talker data rather than by integrating the curves shown in the other panels.

The 5th- and 95th-percentile curves in this panel are close to the ML-SRT and Plomp and Mimpen data, respectively, in Fig. 3. This indicates that the Plomp and Mimpen and ML-SRT talkers represent the extremes of the talker population on which Fig. 5 is based.

IV. CONCLUSIONS AND DISCUSSION

Depending on the talker and the adopted speaking style, the standardized STI calculation procedure (IEC, 1998) may give inaccurate predictions of sentence intelligibility in reverberant conditions. Based on the data presented in this paper, we propose to apply a wider range of modulation frequencies (0.63–31.5 Hz instead of 0.63–12.5 Hz) for predicting the intelligibility of conversational speech. For clear speech, the standard modulation frequency range remains more appropriate.

Further fine-tuning of the STI model may be possible through the application of modulation frequency weighting functions that depend on speaking style. For the limited range of variations in speaking style and voice quality addressed in this study, such a refined and complex approach would not be justified. However, more extreme variations in speaking style (including true conversations, where the interaction between the communicators becomes important) may require this more refined approach.

[1] Extension of the range to higher modulation frequencies is one of the ways in which the STI model can be made more sensitive to reverberation. Another possibility would have been to maintain 14 modulation frequencies, but shift the entire range upward. It has been verified that, for the present data, this leads to essentially similar results. However, this would also affect the relation between the STI and intelligibility for conditions affecting the low-frequency end of the envelope spectrum, such as AGC (automatic gain control).

Duquesnoy, A. J. H. M., and Plomp, R. (**1980**). ‘‘Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis,’’ J. Acoust. Soc. Am. **68**, 537–544.

Houtgast, T., Steeneken, H. J. M., and Plomp, R. (**1980**). ‘‘Predicting speechintelligibility in rooms from the modulation transfer function. I. General room acoustics,’’ Acustica **46**, 60–72.

IEC (**1998**). IEC 60268-16 2nd edition, ‘‘Sound system equipment. Part 16: Objective rating of speech intelligibility by speech transmission index’’ (International Electrotechnical Commission, Geneva, Switzerland).

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.

Plomp, R., and Mimpen, A. M. (**1979**). ‘‘Improving the reliability of testing the speech reception threshold for sentences,’’ Audiology **18**, 43–52.

Steeneken, H. J. M., and Houtgast, T. (**1980**). ‘‘A physical method for measuring speech transmission quality,’’ J. Acoust. Soc. Am. **67**, 318–326.

van Leeuwen, D. A., and Orr, R. (**2000**). ‘‘Speech recognition of non-native speech using native and non-native acoustic models,’’ in Proceedings of the RTO workshop MIST, RTO-MP-28 AC/323(IST)TP/4, Neuilly-sur-Seine, France.

van Wijngaarden, S. J., Steeneken, H. J. M., and Houtgast, T. (**2001**). ‘‘Methods and models for quantitative assessment of speech intelligibility in cross-language communication,’’ in Proceedings of the RTO Workshop on Multi-lingual Speech and Language Processing, Aalborg, Denmark.

van Wijngaarden, S. J., Steeneken, H. J. M., and Houtgast, T. (**2002**). ‘‘Quantifying the intelligibility of speech in noise for nonnative listeners,’’ J. Acoust. Soc. Am. **111**(4), 1906–1916.

Versfeld, N. J., Daalder, J., Festen, J. M., and Houtgast, T. (**2000**). ‘‘Method for the selection of sentence materials for efficient measurement of the speech reception threshold,’’ J. Acoust. Soc. Am. **107**, 1671–1684.