Objective Measurement of the Speech Transmission Quality of Vocoders by Means of the Speech Transmission Index

Mr. Bastiaan J.C.M. van Gils
TNO Human Factors P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
Telephone +31 346 356 218, Telefax +31 346 353 977

vanGils@tm.tno.nl

Dr. Sander J. van Wijngaarden
TNO Human Factors P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
Telephone +31 346 356 230, Telefax +31 346 353 977

vanWijngaarden@tm.tno.nl

ABSTRACT

Nearly all types of military speech communication involve the use of so-called (narrow band) voice coders or vocoders. Usually the Speech Transmission Index (STI) uses artificial test signals, which can not be reproduced by vocoders with the usual fidelity. Therefore the STI is not able to evaluate vocoders at this time. Although it is theoretically feasible to measure the Speech Transmission Index with natural speech instead of the usual artificial signals, each of the various speech-based STI measurement methods proposed in the literature has its own shortcomings and inaccuracies. A new procedure is proposed for estimating a speech based modulation transfer function (MTF), on which the STI is based, that approaches the accuracy of conventional STI implementations based on artificial signals. The new method enables evaluation of vocoders by means of the STI. Applying the method to a voice coder database shows promising results, giving an average squared correlation coefficient R2 of 0.87 between the subjective CVC scores and the calculated STI for male speech.

INTRODUCTION

Nearly all types of military speech communication involve the use of so-called (narrow band) voice coders, or vocoders. These vocoders efficiently digitize human speech before transmission, making use of knowledge of the characteristics of the human voice and hearing. Speech communication equipment based on such vocoders allow more efficient use of the electromagnetic spectrum available for radio transmissions, while at the same time offering greatly increased possibilities for encryption and end-to-end secure transmission. Unfortunately, the Speech Transmission Index method, perhaps the most commonly used method for objective measurement of speech intelligibility, can not be applied when vocoders are involved. This method normally makes use of artificial test signals, which can not be reproduced by vocoders (which are optimized for natural speech) with the usual fidelity. As a consequence, speech intelligibility verifications of vocoders, as frequently required by NATO, depend on expensive and cumbersome subjective procedures involving listening panels. It is theoretically feasible to measure the Speech Transmission Index (STI) with natural speech instead of the usual artificial test signals. This has already been shown shortly after the introduction of the STI [1]. The speech-based approach would eliminate the problems as outlined above. However, each of the various speech-based STI measurement methods proposed in the literature [2] - [4] has its own shortcomings and inaccuracies. A new procedure is proposed for estimating the modulation transfer function (MTF), on which the STI is based, that approaches the accuracy of conventional STI implementations based on artificial signals.

van Gils, B.J.C.M.; van Wijngaarden, S.J. (2005) Objective Measurement of the Speech Transmission Quality of Vocoders by Means of the Speech Transmission Index. In *New Directions for Improving Audio Effectiveness* (pp. 12-1 – 12-6). Meeting Proceedings RTO-MP-HFM-123, Paper 12. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.aps.

METHODS AND DATA

A preliminary comparison of speech-based MTF (sMTF) estimation methods showed that the methods proposed by Steeneken en Houtgast [1] and Ludvigsen [2] do not perform as well as those by Drullman [3] and Payton [4]. Therefore we focused on the latter two methods. Mathematically the latter two methods are alike:

$$MTF_{drullman} \sim \frac{\Re(crossspectrum)}{\|autospectrum\|}$$

$$MTF_{payton} \sim \frac{\|crossspectrum\|}{\|autospectrum\|}$$

While the Drullman method uses the real part of the cross-correlation spectrum to estimate modulation transfer, the Payton method uses the cross-correlation spectrum magnitude. Difference between the methods is the weighting of out-of-phase transfer of modulations. Comparison of measurement results according to the Drullman and Payton methods indicates that the accuracy of the STI method may be improved by weighting out-of-phase modulations. Although the (almost implicit) phase-weighting approach adopted by Drullman is effective, its method and degree of phase-weighting is rather arbitrary.

This observation led us to propose a new speech-based STI phase-weighting method. This method is basically a generalization of the methods by Payton and Drullman, using an explicit and adjustable phase-weighting of the cross-correlation spectrum.

$$MTF_{phaseweight} \sim \frac{\|crossspectrum\| \cdot f(\angle(crossspectrum))}{\|autospectrum\|}$$

To test the phase-weighting method we first compared the results to a theoretical STI, calculated for simulated noise and reverberation conditions [5]. From the results we selected the best performing variant of the method.

This variant was applied to material of a narrow-band voice coder (NBVC) database [8] – this database was also used in Beerends & van Wijngaarden, this volume, consisting of speech signals, as processed through a set of 9 different military vocoders (including legacy systems as well as the current state-of-the-art, at bit rates ranging from 1200 – 4800 bps) in 12 channel conditions. Each of the 108 combinations was tested using 8 talkers (4 male – 4 female). Speech intelligibility was measured using standard subjective assessment methods (CVC and SRT). The database also entails the audio of the original CVC word lists and the input and output signals. We extracted the mean CVC score in 216 different vocoder – condition – gender combinations and calculated the mean STI of each combination. The results were compared to the well established relation between STI and CVC [6].

RESULTS

We simulated 14 speech intelligibility reducing conditions (Noise: 7 SNR ratios / Reverberation: 7 reverberation times) and calculated the theoretical STI value. The simulated speech files were processed. Since time delay between input and output signals influences the results of the phase-weighting method, we aligned the input and output signals in time, based on maximum correlation of the input and output speech envelope.

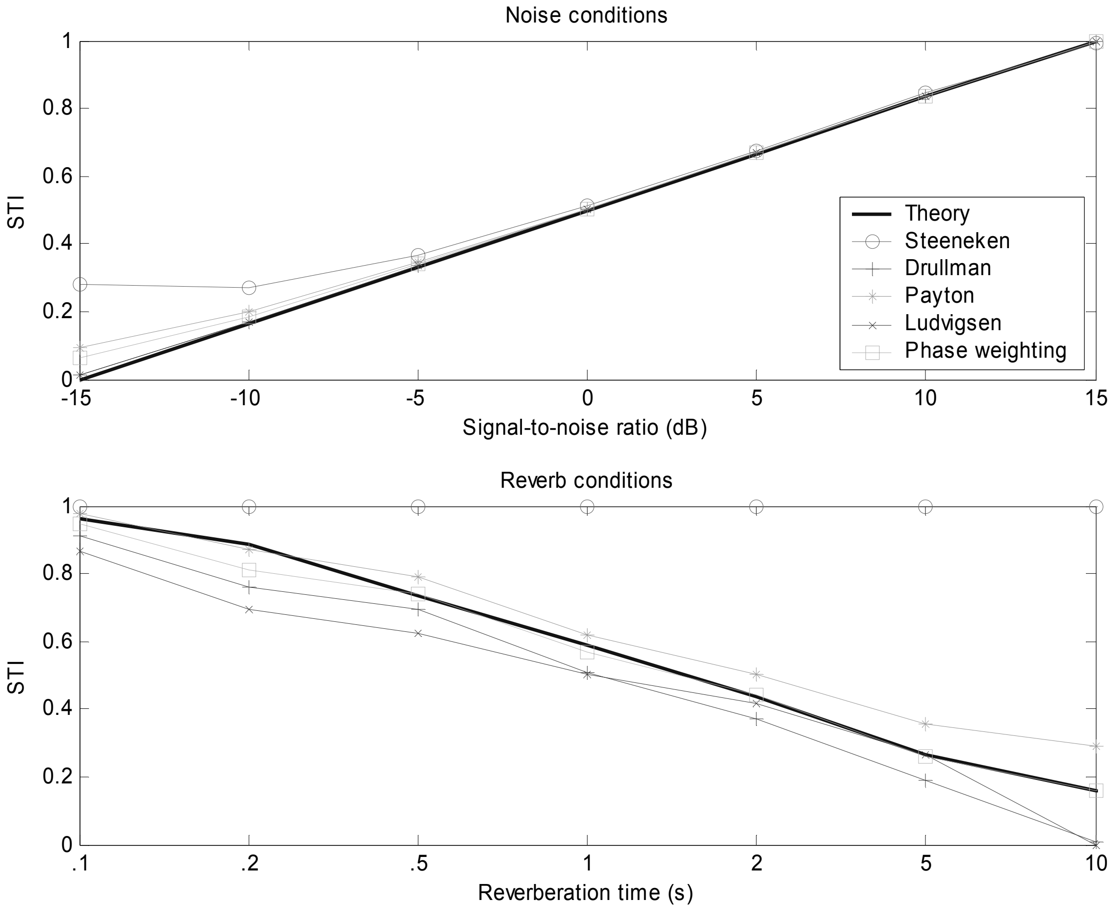

Figure 1. Comparison of speech-based STI methods with theoretically computed STI in conditions with added noise and reverberation.

The resulting STI from the existing sMTF methods and the phase-weighting method were compared with the theoretical value. Figure 1 shows the results of the existing sMTF methods, including the best variant of the phase-weighting method. In noise conditions (top) the methods perform equally well in SNR ratios higher than -5dB. In reverberation conditions (bottom) however, the differences are larger and the phase-weighting method performs the best.

In a second step we used the phase-weighting method to process the NBVC vocoder database. Figure 2 shows that for male speech, although there is an offset, the correlation between the reference curve and the calculated data is high for the older vocoder methods (CVSD, LPC-10 and CELP). The high correlation for the CVSD data can be explained by the fact that CVSD is a general waveform coder instead of a vocoder, and does not affect the temporal fine structure of the signal as much.

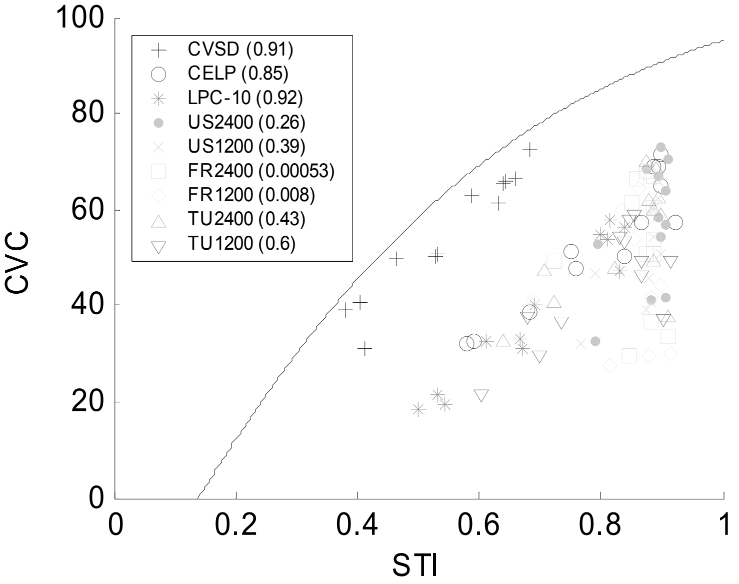

Figure 2. Scatter plot of speech-based STI with corresponding CVC scores. Solid line: STI-CVC reference curve. Numerical values: squared correlation coefficient of each data set.

The results for state-of-the-art vocoders (indicated here by their nation of origin: US, FR and TU) are not as good. Analysis of the results showed that deviation from the reference curve was highly dependent on the SNR at the vocoder input. This result can be explained by the noise suppression accomplished by the vocoders.

Vocoders function as, or explicitly implement, a noise suppression algorithm. It has been known that noise suppression does, at best, provide a modest improvement in speech intelligibility. A reference experiment however, showed that the STI does increase considerably after noise suppression. The noise reduction restores the signal envelope, while the intelligibility remains approximately the same. We conclude that the STI is blind to reduction of intelligibility due to noise addition and suppression.

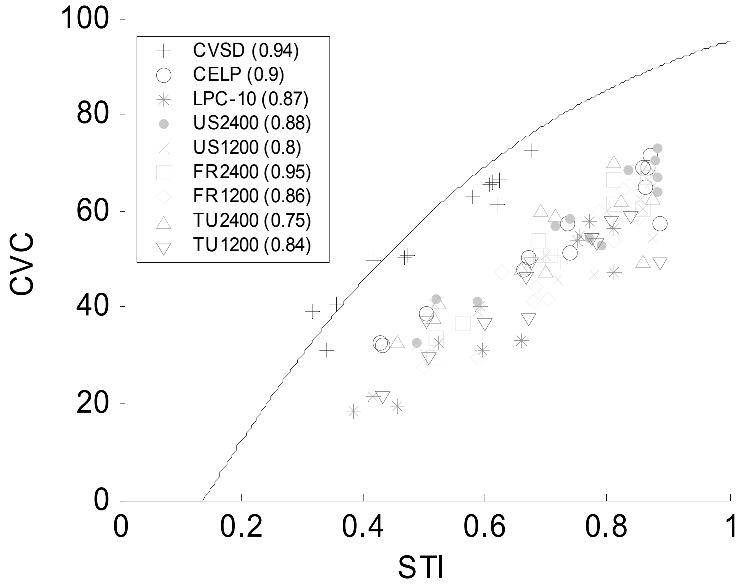

Figure 3. Scatter plot of speech-based STI using noise suppression correction with corresponding CVC scores. Solid line and numerical values as in Figure 2.

The blindness to noise suppression can be brought into the results externally. Figure 3 shows the results for male speech after combining the sMTF between the CVC source and vocoder output and the sMTF of the transfer between CVC source and vocoder input [7]. This procedure improves the correlation of all vocoders greatly as shown in Table 1.

The remaining offset however, seems to be constant among the tested vocoder types, having a value of about 0.3 STI. It may be attributed to the fact that vocoders erode the spectral content of speech fine structure, while the signal envelope remains unaffected. Again speech intelligibility is reduced without any effect noticed in the STI, which is only observing changes in signal envelope.

Not shown in this paper, the same procedure was used to estimate the STI for female speech. These results are less convincing; the squared correlation coefficient varies from .63 to .93. One reason could be that vocoder algorithms are mostly optimized for male speech.

Coder	Squared correlation coefficient R^2	Coder	Squared correlation coefficient R^2
CVSD	0.94	FR 2400	0.86
CELP	0.90	FR 1200	0.95
LPC-10	0.87	TU 2400	0.75
US 2400	0.88	TU 1200	0.84
US 1200	0.80	Average	0.87

Table 1. Squared correlation coefficient R2 of the data sets and the STI-CVC reference curve.

CONCLUSIONS AND DISCUSSION

A new method was proposed to measure the STI with real speech instead of an artificial test signal. On conventional channels (without vocoders), this procedure yields more accurate results than existing methods to measure the STI with real speech. Initial results obtained with a database of speech processed through vocoders show that the speech-based STI can indeed be used to predict speech intelligibility of vocoders, using a correction for noise suppression effects, but that additional work must be done to improve accuracy.

[1] Steeneken, H.J.M., Houtgast, T. (1983). The temporal envelope spectrum of speech and its significance in room acoustics. Proceedings 11th ICA Congress, Paris 1983. Vol. 7, 85-88.

[2] Ludvigsen, C., Elberling, C., Keidser, G. and Poulsen, T. (1990). Prediction of intelligibility of non-linearly processed speech. Acta Otolaryngol Suppl., 469, 190-5.

[3] Drullman, R., Festen, J.M. and Plomp, R. (1994) Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am., 95, 2670-80.

[4] Payton, K.L., Braida, L.D., Chen, S. Rosengard, P. and Goldsworthy, R. (2002) Computing the STI using speech as a probe stimulus. In Past, Present and Future of the Speech Transmission Index., (ed. van Wijngaarden), TNO Human Factors: Soesterberg, the Netherlands, 125-37.

[5] Houtgast, T., Steeneken, H.J.M. and Plomp, R. (1980) Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics. Acustica 46, 59-72.

[6] Steeneken, H.J.M. (1992) On measuring and predicting speech intelligibility. Soesterberg: TNO Institute for Perception, ISBN 90-6743-209-1.

[7] van Wijngaarden, S.J. and Verhave, J. Prediction of speech intelligibility for public address systems in traffic tunnels, submitted.

[8] Tardelli, J.D., van Wijngaarden, S. J., Hassanein, H. and Collura, J.S. (2002) A Precision-weighted Rank-ordering Procedure for the Combination of Voice Coder Evaluation Results. IEEE Speech Coding Workshop, Tsukuba, Ibaraki, Japan, October 6-9 2002, pp 99-101