Modifying the normalized covariance metric measure to account for nonlinear distortions introduced by noise-reduction algorithms

Fei Chen
Division of Speech and Hearing Sciences, Prince Philip Dental Hospital, The University of Hong Kong, 34 Hospital Road, Hong Kong
feichen1@hku.hk

Yi Hu[a)]
Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee, 3200 North Cramer Street, Milwaukee, Wisconsin 53211
huy@uwm.edu

Abstract: In this study, two methods are proposed to modify the normalized covariance metric (NCM) measure to reduce the effects of gain-induced nonlinear distortions introduced by most noise-suppression algorithms. Considering that the gain-induced distortions behave differently dependent on the signal-to-noise ratio between the noise-reduced speech and the noise, the first approach introduces a penalty factor involving this ratio in the modified NCM measure. The second approach deemphasizes segments marked with amplification distortions that contribute less to intelligibility via adaptive thresholding. Significantly higher correlations with intelligibility scores were obtained from the modified NCM measures compared with the original NCM measures.

PACS numbers: 43.71.Gv, 43.71.Es, 43.71.An, 43.66.Ts [DO]
Date Received: February 24, 2013 **Date Accepted:** March 22, 2013

1. Introduction

The speech transmission index (STI) (Houtgast and Steeneken, 1971) is by far one of the most commonly used measures for predicting speech intelligibility in the presence of background noise. Although the STI-based intelligibility measures have been shown to successfully predict the effects of linear filtering, reverberation, room acoustics, and additive noise on speech intelligibility, they do not account for nonlinear distortions present in processed speech, and several modifications have been proposed to use speech or speech-like signals as probe signals in the computation of the STI measure (e.g., Steeneken and Houtgast, 1980). Despite these modifications, several studies have found that speech-based STI measures failed to predict the intelligibility of nonlinearly processed speech materials, especially those processed by noise-reduction algorithms (e.g., Goldsworthy and Greenberg, 2004; Ma *et al.*, 2009). To accommodate the effects of nonlinear processing by noise-reduction algorithms such as spectral subtraction methods, several variations such as those based on normalized covariance metric (NCM) measures were proposed by Goldsworthy and Greenberg (2004), but few of these modifications were validated using intelligibility scores obtained with normal-hearing (NH) human listeners in their studies. More recently, Taal *et al.* (2011) introduced a normalization-and-clipping approach involving short-time time-frequency analysis to improve the intelligibility prediction power of the STI-based index. The resulting short-time objective intelligibility (STOI) measure reflected two important contributions from Taal *et al.* (2011): a normalization procedure was applied to compensate for global level differences that affect speech intelligibility; and a clipping procedure

[a)]Author to whom correspondence should be addressed.

was implemented to ensure that the sensitivity of the intelligibility model was upper-bounded. In the study by Gómez *et al.* (2012), they took a significant initiative to improve the STI-based measures by considering the effects of nonlinear distortions on the subjective intelligibility assessment of noise-suppressed speech, and in their case, it was the detrimental effect of the negative spectral distortions introduced by unreliable noise estimation.

It is generally agreed that as a valuable tool to evaluate and optimize the performance of noise-reduction algorithms, objective intelligibility measures are of great interests to researchers. To further improve the predictive power of NCM measures for noise-suppressed speech, Ma *et al.* (2009) designed signal- and segment-dependent band-importance functions (BIFs) for predicting the intelligibility of noise-suppressed speech in situations where the target speech was corrupted by fluctuating maskers. Those BIFs were determined according to metrics that placed more weights to bands and segments presumably contributing more to speech intelligibility (e.g., spectral peak of vowels). In Chen and Loizou (2012), they assessed the contributions of various types of phonetic segments to predicting the intelligibility of noise-suppressed speech, and their modified NCM measure was implemented using segments selected using three segmentation methods. They found that higher correlation with intelligibility scores was obtained when segments were included containing consonant-vowel boundaries.

It is clear from these studies that understanding the mechanisms of noise-suppression algorithms and their impact on speech intelligibility is critical for designing improved intelligibility measures by accounting for nonlinear distortions introduced by noise-suppression algorithms. Most (if not all) noise-suppression algorithms involve a step in which the mixture envelope or spectrum is multiplied by a nonlinear gain function (i.e., G, taking values from 0 to 1) with the intention of suppressing background noise. The shape of the gain function varies across algorithms, but independent of its shape, when the gain function is applied to the mixture envelopes (or spectra), it introduces nonlinear distortions to the speech envelopes or spectra (i.e., the difference between the noise-reduced speech and the clean speech). Earlier studies (e.g., Loizou and Ma, 2011; Kim and Loizou, 2011; Gómez *et al.*, 2012) showed that it is beneficial to account for the effects of these nonlinear distortions in the objective intelligibility measures.

Measuring the effects of nonlinear distortions is certainly a challenging task considering that many types of distortions exist due to the varieties of noise-suppression algorithms. In an earlier effort to address this issue, Kim and Loizou (2011) simplified the classification of nonlinear distortions and classified the distortions into either amplification or attenuation distortions. The amplification distortion refers to the scenario that the envelope or magnitude spectrum of the noise-suppressed speech is larger than that of the clean speech; and the attenuation distortion refers to the scenario that the envelope or magnitude spectrum of the noise-suppressed speech is smaller than that of the clean speech. Kim and Loizou (2011) made a compelling case that the effects of the gain-induced nonlinear distortions on the intelligibility of noise-suppressed speech are different depending on the gain function values inducing amplification or attenuation distortions. They found that amplification distortion was practically harmful to speech intelligibility, and in contrast, attenuation distortion did not impair speech intelligibility. This implies that, while amplification and attenuation distortions co-exist in noise-suppressed speech, these two types of distortions need to be treated differently for speech intelligibility prediction. It is worthwhile to note that this type of characterization of nonlinear distortions only applies to those introduced by noise-suppression algorithms.

The aim of the present study is to improve the predictive power of the NCM intelligibility measures by accounting for nonlinear distortions introduced by noise-suppression algorithms. To this end, two approaches are proposed to modify the implementations of the NCM measure in order to reflect the insight from the mechanisms of noise-suppression algorithms. In the first approach, we base our proposed modification on the observation that depending on the signal-to-noise ratio (SNR) between the noise-reduced speech signal and the masker signal, the nonlinear distortions behave dramatically differently in terms of their distribution. The rationale is that in the segments where these distortions distribute widely (i.e., covering a wide range from large attenuation distortions to large amplification distortions, see

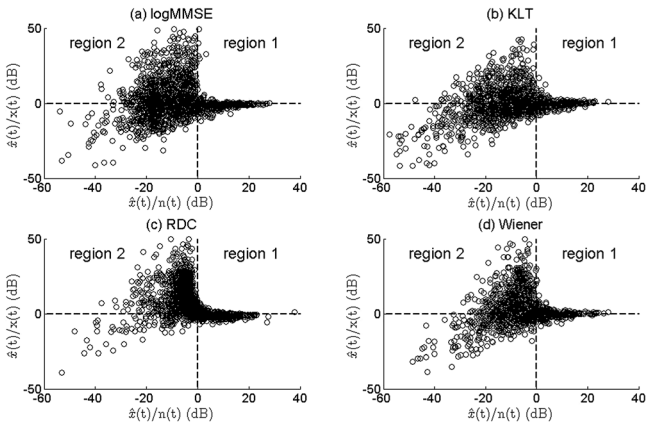

Fig. 1. Example scatter plots of the nonlinear distortion $[\hat{x}(t)/x(t)]$ against $[\hat{x}(t)/n(t)]$ for a noise-suppressed sentence originally corrupted by babble masker at 0 and 5 dB SNR levels and processed by four types of noise-suppression algorithms.

an illustration in Fig. 1), unreliable noise estimation usually dominates, and therefore a penalty factor needs to be applied to reduce the contribution of these sample values to the similarity measurement (i.e., the normalized covariance metric in this study) between the noise-reduced speech and the clean speech. In the second approach, by recognizing that it is critical for the intelligibility measures to treat the two types of distortions following noise-suppression differently, we identify the segments carrying more intelligibility information (i.e., those marked with attenuation distortion) from noise-suppressed speech, and only include these segments into the computation of NCM measures. The underlying hypothesis is that by emphasizing those intelligibility-information-carrying segments in the computation of intelligibility index, the correlation with human listener's intelligibility scores improves.

2. Methods

2.1 The normalized covariance metric

The NCM measure is similar to the STI measure in that it computes a weighted sum of transmission index (TI) values determined from the envelopes of the probe and response signals in each frequency band (Goldsworthy and Greenberg, 2004). Unlike the traditional STI measure, however, the NCM measure is based on the covariance between the probe (input) and response (i.e., processed by a noise-suppression algorithm) envelope signals computed in each band of speech signals. The implementation of the NCM measure and more details on its computation can be found in Ma *et al.* (2009).

It can be shown that if the envelope of the processed signal $\hat{x}_j(t)$ is generated by applying a linear processing to the envelope of the clean signal $x_j(t)$ (e.g., by additive background noise), the output SNR in each spectral band is essentially equal to the SNR prior to the processing. For this reason, the NCM measure is accurate in predicting the intelligibility of unprocessed noisy speech even in fluctuating maskers, and this is partially confirmed by our small data set consisting of eight noisy conditions (see Sec. 2.4) that shows a correlation value of 0.92 between the predicted NCM measures and the NH listeners' intelligibility scores.

2.2 Modifying the NCM measure by introducing a penalty factor

For linear distortions, the output band SNR is equal to the input SNR prior to the linear processing assuming that the envelope of the additive noise signal in the jth band [i.e., $n_j(t)$] is uncorrelated to that of the speech; however, for nonlinear distortions $r_j(t) = \hat{x}_j(t) - x_j(t)$, this is in general not the case due to the fact that $r_j(t)$ contains both speech distortion and residual noise that are nonlinearly related to $x_j(t)$ depending on

the types of gain functions applied to the mixture envelope or spectrum. For this reason, different noise-suppression algorithms may generate various types of nonlinear distortions at the same input SNR, consequently resulting in different impact on speech intelligibility. Clearly, quantifying the effects of various types of nonlinear distortion on speech intelligibility is a formidable task; as a first step, Kim and Loizou (2011) used a binary classification to discriminate the nonlinear distortions induced by the gain functions of the noise-reduction algorithms, namely, the amplification distortion [i.e., $\hat{x}_j(t) > x_j(t)$], and the attenuation distortion [i.e., $\hat{x}_j(t) < x_j(t)$]. They showed that these two types of distortions have dissimilar effects on speech intelligibility, and different treatment is needed to differentiate their effects. Following this idea, Loizou and Ma (2011) extended the articulation index to improve its predictive power for materials processed by noise-suppression algorithms.

In the present study, we adopt a similar methodology that provides a simplified characterization of the nonlinear distortions in order to make the task manageable, and our proposed approach is based on the observation that the distributions of the nonlinear distortions $r_j(t) = \hat{x}_j(t) - x_j(t)$ vary widely dependent on the $\overline{\mathrm{SNR}}_j(t)$ [i.e., SNR defined between the envelope of the noise-reduced speech signal $\hat{x}_j(t)$ and that of the masker signal $n_j(t)$]. Note that the same SNR definition [i.e., between the noise-reduced speech signal $\hat{x}_j(t)$ and the masker signal $n_j(t)$] has been previously used as output band SNR after noise-suppression processing in Ma *et al.* (2009) and Loizou and Ma (2011). Figure 1 shows example scatter plots for a noise-suppressed sentence originally corrupted by babble masker at 0 and 5 dB SNR levels and processed by four types of noise-suppression algorithms (see more on noise-suppression algorithms in Hu and Loizou, 2007), with the x axis being the SNR in dB scale between the envelopes of the noise-reduced speech signal and the masker signal, and the y axis being the ratio in dB scale between the envelopes of the noise-reduced speech signal and the clean speech signal. It can be seen from Fig. 1 that when the $\overline{\mathrm{SNR}}_j(t)$ value is above 0 dB (labeled as region 1), the distortions are usually small distributing in a narrow dynamic range; and when the $\overline{\mathrm{SNR}}_j(t)$ value is below 0 dB (labeled as region 2), the distortions are usually nonuniformly distributed in a much larger dynamic range. We hypothesize that distortions in region 1 are more important to subjective intelligibility assessment than those in region 2. The cause for this outcome is the unreliable noise estimation for the segments in region 2 resulting in inaccurate gain function values applied to suppress the noise (Hu and Loizou, 2007; Gómez *et al.*, 2012). It can be shown that the normalized covariance ρ_j between $x_j(t)$ and $\hat{x}_j(t)$ measures the degree of similarity between the change of the clean speech envelope and that of the envelope of the processed speech (an absolute value of 1 indicates a perfect linear relationship); and it does not take into account the observation that the nonlinear distortions depend on the SNR between the noise-reduced speech signal and the masker signal. Note that although not shown in Fig. 1 (due to space limitation), similar patterns to those in Fig. 1 were also found in the analyses of the other three types of maskers (i.e., car, street, and train). In other words, the $\overline{\mathrm{SNR}}_j(t)$-dependent distributions of nonlinear distortions could be generalized to the various maskers and noise-reduction algorithms used in this study.

To differentiate the distortions distributing in a narrow range and those in a much larger range in computing the NCM measure, we propose in this study introducing a penalty factor $\alpha_j(t)$ to the computation of ρ_j in order to penalize the contribution of the $\hat{x}_j(t)$ values in the region with smaller $\overline{\mathrm{SNR}}_j(t)$ values. To achieve this goal, a simple approach is to replace $\hat{x}_j(t)$ with $\hat{x}_j(t)\,\alpha_j(t)$, where $\alpha_j(t)$ is defined as

$$\alpha_j(t) = \begin{cases} \left[1 + \dfrac{1}{\overline{\mathrm{SNR}}_j(t)}\right] & \text{if } 0 < \overline{\mathrm{SNR}}_j(t) < 1 \\ 1, & \text{if } \overline{\mathrm{SNR}}_j(t) \geq 1. \end{cases} \tag{1}$$

The penalty factor in Eq. (1) takes different definitions according to the range of $\overline{\mathrm{SNR}}_j(t)$. It is intuitive to see that the lower the $\overline{\mathrm{SNR}}_j(t)$ is in region 2, the higher $\alpha_j(t)$ and $\hat{x}_j(t)\alpha_j(t)$

will be, resulting in a smaller ρ_j value than obtained before. Note that a smaller ρ_j value indicates that we will deemphasize the effects of those envelope values $\hat{x}_j(t)$ in region 2 with lower $\overline{\mathrm{SNR}}_j(t)$ on predicting the intelligibility of noise-suppressed speech, as they cover almost all of the large distortions exemplified in Fig. 1. The modified ρ_j values were then used to compute the modified NCM measure (i.e., $\mathrm{NCM}_{\mathrm{masker}}$).

2.3 Separating segments of two types of distortions by thresholding processing

Kim and Loizou (2011) showed that in cases when amplification distortion and attenuation distortion co-exist, segments marked with attenuation distortions (via noise-suppression algorithms) contribute more to speech intelligibility because segments with amplification distortions are usually those that are discarded in the ideal binary masking (IdBM) processing that maximizes a simplified form of articulation index (Loizou and Kim, 2011).

In order to minimize the influences of amplification distortions, an amplitude thresholding procedure could be deployed for intelligibility prediction of the noise-suppressed speech. Specifically, we could alternatively remove those frames from the computation of intelligibility indices by using a channel-specific adaptive thresholding processing, and the adaptive threshold is defined as

$$\mathrm{Thr}_j = \min[x_j(t)] + b\{\max[x_j(t)] - \min[x_j(t)]\}, \tag{2}$$

where b is a constant ($b = 0.1$ in this study). The envelope segments with magnitudes larger than the threshold in Eq. (2) are then concatenated into one composite envelope in each frequency band. Note that the thresholding procedure is only applied to clean envelopes, and for the envelopes of the noise-suppressed speech, the same segments as those from the clean envelopes are selected. Finally, the corresponding composite (concatenated) clean and processed envelopes are used to compute the NCM measure. By using the new composite envelopes formed by concatenating the attenuation-distortion dominated segments, the modified NCM measure (i.e., $\mathrm{NCM}_{\mathrm{thr}}$) computes a perceptually more relevant SNR in each band, since only intelligibility-information-bearing segments are included in its computation (Loizou and Kim, 2011). Note that the threshold value b in Eq. (2) was empirically determined by analyzing the peak and valley values of envelope waveform in each band, and further study is needed to investigate how to select the optimal value of constant b.

2.4 Speech intelligibility data

More details of the speech intelligibility data and the noise-suppression algorithms involved can be found in Hu and Loizou (2007). In summary, the intelligibility evaluation by a total of 40 NH listeners was conducted using noise-corrupted speech sentences processed through eight different noise-suppression algorithms, and the intelligibility scores were obtained from NH listeners in a total of 72 conditions (=4 maskers $\times$ 2 SNR levels $\times$ 8 algorithms + 4 maskers $\times$ 2 noisy references). Twenty IEEE sentences were used for each condition, and none of the sentences were repeated across testing conditions. The masker signals included four types of noises recorded from real-world environments: babble, car, street, and train. The maskers were added to the speech signals at SNR levels of 0 and 5 dB. The processed speech sentence files, along with the noisy speech files, were presented monaurally to the listeners in a double-walled sound-proof booth via Sennheiser's HD 250 Linear II circumaural headphones at comfortable listening levels. The percentage intelligibility score for each condition was calculated by dividing the number of words correctly identified by the total number of words in a particular testing condition.

3. Results

The average intelligibility scores obtained by NH listeners were subjected to correlation analysis with the corresponding values obtained by the two modified NCM measures (i.e., $\mathrm{NCM}_{\mathrm{masker}}$ and $\mathrm{NCM}_{\mathrm{thr}}$). More specifically, correlation analysis was performed between the mean (across all subjects) intelligibility scores obtained in each of the testing conditions and the corresponding mean (computed across the 20 sentences used in

Table 1. Correlation coefficients (r) between sentence recognition scores and the NCM-based measures.

	r	
Intelligibility measure	noisy + noise-suppressed (72 conditions)	noise-suppressed (64 conditions)
NCM	0.82	0.84
NCM_{masker}	0.91[a]	0.92[a]
NCM_{thr}	0.87[a]	0.87[a]
STOI	0.86	0.87

[a]The difference of correlation coefficients between the NCM measure and its modified measure is significant ($\alpha = 0.05$).

each condition) intelligibility index values obtained in each condition. The Pearson's correlation coefficient (r) was used to assess the performance of the intelligibility measures to predict intelligibility scores.

Table 1 shows the correlation of the two modified NCM measures (i.e., NCM_{masker} and NCM_{thr}) with the intelligibility scores from NH listeners. For the purpose of comparison, the correlation results of the original NCM measure (i.e., NCM, Goldsworthy and Greenberg, 2004) and the STOI measure (Taal *et al.*, 2011) are also included using all 72 noisy and noise-suppressed conditions. It is seen in Table 1 that, when introducing the $\overline{SNR}_j(t)$-dependent penalty factor, the NCM_{masker} measure predicted the intelligibility much better than the original NCM measure, with the correlation coefficient being improved from 0.82 to 0.91; similarly NCM_{thr} yielded a higher correlation coefficient 0.87 as well. Statistical analysis reveals that these improvements are significant (Steiger, 1980). Figure 2 shows the scatter plots of speech recognition scores against the predicted NCM, NCM_{masker} and NCM_{thr} values. Analysis also shows that the correlation coefficient of the NCM_{masker} measure is significantly higher than that of the STOI measure, i.e., $r = 0.91$ vs 0.86 in Table 1. Table 1 also shows the correlation results computed only with the 64 noise-suppressed conditions involving nonlinear distortions, and the rationale is that the other eight noisy conditions do not have nonlinear distortions targeted by the proposed approach. When introducing the $\overline{SNR}_j(t)$-dependent penalty factor, the NCM_{masker} and NCM_{thr} measures predicted the intelligibility significantly better than the original NCM measure.

4. Summary and conclusion

The present study modified the computation of the existing NCM measure to account for nonlinear distortions introduced by noise-reduction algorithms. The modification was motivated by the different effects of gain-induced distortions by most noise-suppression algorithms on speech intelligibility. Taking this important observation into account, this study proposed two approaches to modify the conventional NCM measure, and it was done by (1) treating differently the nonlinear distortions in the region with smaller $\overline{SNR}_j(t)$ values and those in the region with larger $\overline{SNR}_j(t)$ values; and (2) reducing the influence of the amplification distortions contained in the processed envelope. In summary, the contributions of the present work include the following.

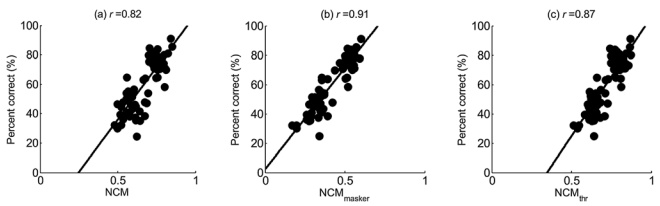

Fig. 2. Scatter plots of the 72 sentence recognition scores against values of the predicted (a) NCM measures, (b) NCM_{masker} measures, and (c) NCM_{thr} measures.

(1) As shown by Kim and Loizou (2011) and Loizou and Kim (2011), the noise-suppressed speech segments showing amplification or attenuation distortions affect speech intelligibility differently. A similar methodology is proposed in the present study that treats the nonlinear distortions differently dependent on $\overline{\mathrm{SNR}}_j(t)$ which is the SNR ratio between the noise-reduced speech and the masker signal, as the segments with smaller $\overline{\mathrm{SNR}}_j(t)$ are usually those with unreliable noise estimation that leads to widely distributed distortions, and the contribution of the speech segments containing these distortions to the computation of the NCM measure is reduced by introducing the penalty factor based on $\overline{\mathrm{SNR}}_j(t)$. The modified NCM measure (i.e., $\mathrm{NCM}_{\mathrm{masker}}$) achieved a much higher correlation with intelligibility scores. It is worthwhile to note that the penalty factor $\alpha_j(t)$ in Eq. (1) can take other forms, and it is beneficial to integrate a similar clipping procedure to that of Taal *et al.* (2011) in order to make sure that the sensitivity of the model in those envelope samples severely degraded is upper bounded; we will conduct further investigations on these issues in our future studies.

(2) A simple adaptive thresholding processing was proposed to separate the two types of envelope distortions (i.e., amplification and attenuation distortions), and improved intelligibility prediction was obtained. This result is consistent with recent outcomes on the contributions of different phonetic segments for predicting the intelligibility of noisy and noise-suppressed speech. For instance, Chen and Loizou (2012) showed that sonorant segments (e.g., vowels) contributed more to the intelligibility prediction than obstruent segments (e.g., stops and fricatives). When tested on the same dataset, the correlation of sonorant-based NCM measures with intelligibility scores was $r = 0.84$, while that of the obstruent-based NCM measures was $r = 0.64$. In addition, the advantage of using the adaptive-thresholding based segmentation is that it does not need explicit phonetic segmentation, which is extremely challenging to implement in practice even with using the most sophisticated phoneme detection algorithms.

References and links

Chen, F., and Loizou, P. (**2012**). "Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise," J. Acoust. Soc. Am. **131**, 4104–4113.
Goldsworthy, R., and Greenberg, J. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679–3689.
Gómez, A. M., Schwerin, B., and Paliwal, K. (**2012**). "Improving objective intelligibility prediction by combining correlation and coherence based methods with a measure based on the negative distortion ratio," Speech Commun. **54**, 503–515.
Houtgast, T., and Steeneken, H. J. M. (**1971**). "Evaluation of speech transmission channels by using artificial signals," Acustica **25**, 355–367.
Hu, Y., and Loizou, P. (**2007**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1786.
Kim, G., and Loizou, P. (**2011**). "Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms," J. Acoust. Soc. Am. **130**, 1581–1596.
Loizou, P., and Kim, G. (**2011**). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," IEEE Trans. Speech, Audio Proc. **19**, 47–56.
Loizou, P., and Ma, J. (**2011**). "Extending the articulation index to account for nonlinear distortions introduced by noise-suppression algorithms," J. Acoust. Soc. Am. **130**, 985–995.
Ma, J., Hu, Y., and Loizou, P. (**2009**). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am. **125**, 3387–3405.
Steeneken, H., and Houtgast, T. (**1980**). "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Steiger, J. H. (**1980**). "Tests for comparing elements of a correlation matrix," Psychol. Bull. **87**, 245–251.
Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**, 2125–2136.