Abstract

Although the speech transmission index (STI) has been shown to predict successfully the effects of linear distortions introduced by filtering and additive noise, it does not account for non-linear distortions present in noise-suppressed speech. In this study, the normalized covariance metric (NCM), a STI-based intelligibility measure, was modified to reduce the effects of non-linear distortions introduced by most noise-suppression algorithms for intelligibility prediction. This was done by designing a new definition of the output signal-to-noise ratio to compensate the biased estimation of the input SNR prior to the noise-suppression processing. The modified NCM measure was evaluated with intelligibility scores obtained by normal-hearing listeners in 72 noisy conditions involving noise-suppressed speech corrupted by four different maskers (babble, car, train and street interferences). Significantly higher correlation with intelligibility score was obtained from the modified NCM measure, in contrast to those from the original NCM measure.

Index Terms: Intelligibility prediction, speech-transmission index, noise-suppression algorithms.

1. Introduction

The speech transmission index (STI) [1-2] is by far one of the most commonly used measures for predicting speech intelligibility in the presence of background noise. Though the STI-based intelligibility indices have been shown to predict successfully the effects of linear filtering, reverberation, room acoustics, and additive noise on speech intelligibility (e.g., [1], [3]), they do not account for non-linear distortions present in processed speech. This is because the non-linear processing itself might introduce additional modulations which the STI measure interprets as increased signal-to-noise ratio (SNR) [4-6]. For that reason, several modifications have been proposed to use speech or speech-like signals as probe signals in the computation of the STI measure [6]. Despite these modifications, several studies have found that the speech-based STI methods failed to predict the intelligibility of nonlinearly-processed speech, especially that processed by noise-suppression algorithms [6-8].

The definition of the output SNR (or effective SNR) following the non-linear noise-suppression processing poses great challenges [7]. This is so because the definitions of the target and masker signals are no longer clear, and they both are affected following the non-linear noise-suppression processing. When using the target (clean) and processed signals to compute the output band SNR following the noise-suppression processing, we are actually estimating the output band SignAlto-RESidual distortion ratio (i.e., \( \text{SNR}_{\text{RES}} \)), rather than the input band SNR (see review in [9]). This \( \text{SNR}_{\text{RES}} \) metric is also known in the literature as the frequency-weighted segmental SNR [10]. Assuming independence of signal and masker, Lu and Loizou recently revealed that \( \text{SNR}_{\text{RES}} \) could be expressed in terms of the gain function \( G \) (see Fig. 1) as [11]:

\[
\text{SNR}_{\text{RES}}(\xi, G) = \frac{E[|X|^2]}{E[|X - \hat{X}|^2]} = \frac{E[|X-G\cdot(X+N)|^2]}{E[|X|^2]} = \frac{\sigma_x^2}{(1-G)^4 \sigma_x^2 + G^2 \sigma_{n}^2} = \frac{\xi}{(1-G)^2 \xi + G^2}.
\]
where $E$ is the expectation operator, $\hat{x}$ is the input SNR prior to the noise-suppression processing, $X, N,$ and $\hat{X} = G \cdot (X + N)$ are the clean, masker, and processed (via a noise-suppression algorithm) magnitude spectra, respectively (as shown in Fig. 1), and $\sigma^2$ and $\sigma^2_{\hat{x}}$ denote the variance of the target signal $x(t)$ and the masker signal $n(t)$, respectively. Eq. (1) gives the output (or effective) SNR following the non-linear noise-suppression processing, and can for $G=1$ be reduced to an estimation of the input SNR $\hat{x}$. Figure 2 shows the relation between the output SNR (i.e., $\text{SNR}_{\text{ESI}}$) and the input SNR $\hat{x}$ with gain value $G$ as a parameter from 0.01 to 1. It is observed from Eq. (1) and Fig. 2 that, when $G=1$ for a certain band, $\text{SNR}_{\text{ESI}} = \hat{x}$. On the other hand, a small gain will lead to a strongly reduced effect of the input SNR estimation. This indicates that the output SNR (i.e., $\text{SNR}_{\text{ESI}}$) used in the present intelligibility indices needs to be corrected to account for the effects of non-linear distortions introduced by noise-suppression algorithms. More precisely, the present estimation of the input SNR $\hat{x}$ in Eq. (1) needs to be compensated to reduce the estimation bias caused by the non-linear gain function $G$ [12].

The aim of the present work is to improve the prediction power of the speech-based STI to account for non-linear distortions introduced by noise-suppression algorithms. More specifically, we will propose a new definition of the effective output SNR following noise-suppression by incorporating information from the masker signal to compensate the bias effect in estimating the input SNR by the present $\text{SNR}_{\text{ESI}}$ metric, and introduce this new output SNR into the implementation of the STI measure.

2. The Normalized Covariance Metric

From the various speech-based STI measures proposed, this study chose the normalized covariance metric (NCM) [6]. The NCM measure computes a weighted sum of transmission index (TI) values determined from the envelopes of the probe (input) and response (output) signals in each frequency band [6]. In this study, the stimuli are first decomposed into $K=20$ bands spanning the signal bandwidth (300–3400 Hz). The speech decomposition is implemented with a series of fourth-order Butterworth filters [6, 8]. Let $x(t)$ and $\hat{x}(t)$ be the envelopes of the clean and processed signals, respectively, in the $j$-th band. The normalized covariance in the $j$-th band is computed as:

$$\rho_j^2 = \frac{\sum \left( x_j(t) - \mu_j \right) \left( \hat{x}_j(t) - \nu_j \right)}{\sqrt{\sum \left( x_j(t) - \mu_j \right)^2} \cdot \sqrt{\sum \left( \hat{x}_j(t) - \nu_j \right)^2}},$$

where $\mu_j$ and $\nu_j$ are the mean values of $x_j(t)$ and $\hat{x}_j(t)$, respectively. The apparent output SNR in the $j$-th band is defined as:

$$\text{SNR}(j) = 10 \log_{10} \left( \frac{\rho_j^2}{1 - \rho_j^2} \right),$$

and subsequently limited to the range of $[-15, 15]$ dB. The transmission index in each band is computed by linearly mapping the output SNR values between 0 and 1, as:

$$\text{TI}_j = \frac{\text{SNR}(j) + 15}{30}.$$

Finally, the transmission indices are averaged across all frequency bands to produce the NCM index:

$$\text{NCM} = \frac{1}{\sum_{j=1}^{K} W_j} \cdot \sum_{j=1}^{K} \text{TI}_j,$$

The most commonly-used weight $W_j$ is ANSI articulation index [8].

Note that the output SNR defined in Eq. (3) is not the input SNR prior to the noise-suppression processing, but rather the output band signal-to-residual distortion ratio [i.e., $\text{SNR}_{\text{ESI}}$ in Eq. (1)]. This is so because the noise term in Eq. (3) is the residual noise between processed and clean signals. The following paragraph will prove that Eq. (3) gives the output band signal-to-residual distortion ratio [$\text{SNR}_{\text{ESI}} = \sigma^2 / \sigma^2_{\hat{x}}$], or Eq. (1), where $\sigma^2$ denotes the variance of the distortion signal $r(t) = \hat{x}(t) - x(t)$, rather than the input SNR ($\xi = \sigma^2 / \sigma^2_{\hat{x}}$) prior to the noise-suppression processing.

We assume that the target clean signal $x(t)$ and distortion signal $r(t)$ have zero mean, and they are uncorrelated, i.e., $E[x(t) \cdot r(t)] = 0$. After dropping the band index $j$ for convenience, we compute the normalized covariance between $x(t)$ and $\hat{x}(t)$ [i.e., Eq. (2)] as follows:

$$\rho_j^2 = \frac{E[x(t) \cdot \hat{x}(t)]^2}{\sigma^2_j \cdot \sigma_{\hat{x}}^2_j},$$

where $\sigma_j^2$ denotes the variance of the processed signal $\hat{x}(t)$. Substituting $\hat{x}(t) = x(t) + r(t)$ into the above equation yields:

$$\rho_j^2 = \frac{E[x(t) \cdot (x(t) + r(t))]^2}{\sigma_j^2 \cdot (\sigma_j^2 + \sigma^2_r)} = \frac{E[x(t)]^2}{\sigma_j^2 \cdot (\sigma_j^2 + \sigma^2_r)} = \frac{\sigma_j^2}{\sigma_j^2 + \sigma^2_r}.$$

It is easy to get $\rho_j^2 = \frac{\sigma_j^2}{\sigma_j^2 + \sigma^2_r} = \text{SNR}_{\text{ESI}}$, indicating that Eq. (3) gives the output band signal-to-residual distortion ratio in Eq. (1), but not the input SNR prior to the noise-suppression processing.

3. Define Output Signal-to-Noise-Ratio With Masker Information

Let $z(t) = \hat{x}(t) + n(t)$ be the synthetic signal involving both the processed and masker signals. We first modify the computation of the normalized covariance in the $j$-th band by replacing the processed signal $\hat{x}(t)$ in Eq. (2) with the synthetic signal $z(t)$, as:

$$\rho_j^2 = \frac{\sum (x_j(t) - \mu_j) \cdot (z_j(t) - \nu_j)}{\sqrt{\sum (x_j(t) - \mu_j)^2} \cdot \sqrt{\sum (z_j(t) - \nu_j)^2}},$$

where $w_j$ is the mean value of $z(t)$. We further modify the apparent output SNR in the $j$-th band as:

$$\text{SNR}_{\text{masker}}(j) = 10 \log_{10} \left( \frac{(1 + a_j) \cdot \rho_j^2}{1 - \rho_j^2} \right),$$

where

$$a_j = \frac{\sum (z_j(t) - x_j(t))^2}{\sum |z_j(t)|^2} \cdot \frac{\sum |x_j(t)|^2}{\sum |z_j(t)|^2}.$$

The $\text{SNR}_{\text{masker}}(j)$ in Eq. (9) is subsequently limited to the range of $[-15, 15]$ dB, linearly mapped to the values between 0 and 1, and finally summed across all bands with ANSI weights to yield to the modified NCM measure, i.e., $\text{NCM}_{\text{masker}}$.

The following will prove that the modified SNR defined in Eq. (9) leads to an unbiased estimation of the input SNR prior to the noise-suppression processing. We assume that $x(t)$ and $n(t)$
Table 1. Correlation coefficients (r) and standard deviation of the error (σe) between sentence recognition scores and the NCM-based measures. Asterisk denotes that the difference of correlation coefficients between the NCM measure and its modified measure is significant (α=0.05).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Babble</th>
<th>Car</th>
<th>Street</th>
<th>Train</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCM</td>
<td>0.91</td>
<td>0.82</td>
<td>0.78</td>
<td>0.85</td>
</tr>
<tr>
<td>NCMmasker</td>
<td>0.95</td>
<td>0.90</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>STOI</td>
<td>0.93</td>
<td>0.86</td>
<td>0.81</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 2. Correlation coefficients (r) between sentence recognition scores and the NCM-based measures for four types of maskers. Asterisk denotes that the difference of correlation coefficients between the NCM measure and its modified measure is significant (α=0.05).

<table>
<thead>
<tr>
<th>Masker</th>
<th>NCM</th>
<th>NCMmasker</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babble</td>
<td>0.82</td>
<td>0.89 *</td>
<td>0.87 *</td>
</tr>
<tr>
<td>Car</td>
<td>0.88</td>
<td>0.88 *</td>
<td>0.88</td>
</tr>
<tr>
<td>Street</td>
<td>0.85</td>
<td>0.87</td>
<td></td>
</tr>
</tbody>
</table>

4. Speech Intelligibility Data

The speech intelligibility data was taken from the intelligibility evaluation of noise-corrupted speech processed through eight different noise-suppression algorithms by a total of 40 normal-hearing (NH) listeners [13]. IEEE sentences were used as test material, and all sentences were produced by a male talker. The sentences were originally sampled at 25 kHz and down-sampled to 8 kHz. The masker signals included the following real-world recordings from different places: 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. Twenty IEEE sentences were used for each condition, and none of the sentences were repeated. The intelligibility scores were obtained from NH listeners in a total of 72 conditions (=4 maskers × 2 SNR levels × 8 recordings + 4 maskers × 2 noisy references (i.e., 2 noise-corrupted conditions at 2 SNR 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. The eight different noise-reduction algorithms included the generalized subspace approach, the perceptually-based subspace approach, the log minimum mean square error (logMMSE) algorithm, the logMMSE algorithm with speech-presence uncertainty, the spectral subtraction algorithm based on reduced-delay convolution, the multiband spectral-subtractive algorithm, the Wiener filtering algorithm based on wavelet-thresholded multistaper spectra, and the traditional Wiener algorithm. More details about the noise-suppression algorithms and the procedure used to collect the intelligibility data can be found in [8, 13].

The present study assessed the prediction power of the proposed NCM measure (i.e., NCMmasker) involving the new definition of the output SNR in Eq. (9). For the purpose of comparison, we also computed the intelligibility indices by the original NCM measure [6] and the STOI measure. Taal et al. introduced a normalization-and-clipping method to improve the intelligibility prediction power of the STI-based index [14]. They implemented the STI measure in short-time frequency unit, and termed their method as a short-time objective intelligibility measure (STOI). The normalization procedure was claimed as compensating for global level differences which should not have a strong effect on speech intelligibility; while the clipping procedure was implemented to ensure that the sensitivity of the intelligibility model was upper-bounded.

5. Results

The average intelligibility scores obtained by NH listeners were subjected to correlation analysis with the corresponding values obtained by the NCM-based measures (i.e., NCM, NCMmasker and STOI). The Pearson’s correlation coefficient (r) was used to assess the performance of the intelligibility measures to predict intelligibility scores. A larger value of correlation coefficient (r) indicates that the index performs better for predicting speech intelligibility score. The standard deviation of the error (σe) was also used for performance evaluation, as σe = σi√1−r², where σi is the standard deviation of the speech intelligibility scores. A smaller value of σi indicates that the measure performs better in predicting intelligibility scores.

Table 1 shows the correlation coefficients of the NCM, NCMmasker and STOI measures with sentences intelligibility scores. It is seen in Table 1 that, when introducing the masker information into the output SNR definition as per Eq. (9), the NCMmasker measure predicts the intelligibility scores much better than the NCM measures, with correlation coefficient improved from r=0.82 to 0.89. Statistical analysis reveals that this correlation coefficient improvement is significant [15]. This shows that the proposed new output SNR definition [i.e., Eq. (9)] can be used as an alternative and simpler method for improving the intelligibility prediction power of the NCM measure. Figure 3 gives the scatter plots of the predicted NCM and NCMmasker values against listeners’ sentence recognition scores. It is noted that the correlation coefficient computed with the NCMmasker measure is slightly higher than that with the STOI measure (i.e.,
Figure 3. Scatter plots of the predicted (a) NCM and (b) NCMmasker values against listeners’ sentence recognition scores.

0.89 vs. 0.87). The values of standard deviation of the error (σ_e) are also shown in Table 1. The NCMmasker and STOI measures give smaller σ_e values than the NCM measure in predicting intelligibility score.

Table 2 lists the NCMmasker based correlation coefficients broken down by masker type. As can be seen, high correlation is maintained for all noise types, with a maximum correlation of r = 0.95 achieved in babble noise condition. The higher correlation coefficients for all noise types are observed relative to those obtained from the NCM and STOI measures in Table 2.

6. Discussion and Conclusion

Recently, many methods have been developed for improving envelope-based speech intelligibility [e.g., 16]. The present work modified the computation of the existing speech transmission index (i.e., the NCM measure) to account for nonlinear distortions introduced by noise-reduction algorithms. The modification was motivated by the impact of the gain-induced distortions caused by most noise-suppression algorithms on speech intelligibility [9]. The present output band SNR computation following the noise-suppression processing was actually the signal-to-residual distortion ratio (SNRsi). and it gave a biased estimation of the input SNR prior to noise-suppression processing [see Eq. (1)]. In addition, although not shown in this paper, our work also found that the above-mentioned modification could be incorporated into the computation of other intelligibility measure, e.g., coherence-based speech intelligibility index (CSII) [17], and led to improved intelligibility prediction performance. Note that the present work only investigated the noise-corruption conditions with additive noises. It is unclear how well the proposed method performs for convolutional noises, which warrants further study in the future.

In conclusion, the present work showed that the output SNR (i.e., SNRsi) computed in the present speech transmission index (i.e., NCM) represented a biased estimation of the input SNR prior to the noise-suppression processing, and was not the best choice accounting for the intelligibility variance of noise-suppressed speech. With the additional information of masker signal [Eq. (9)], the definition of the output SNR for noise-suppressed speech was modified to reliably estimate the input SNR, and to compensate the biased input SNR estimation due to the gain-induced nonlinear distortions introduced by noise-suppression algorithms. The proposed output SNR computation could be easily integrated into the NCM measure, and the modified NCM measures achieved high correlation with intelligibility scores of the noise-suppressed speech relative to the traditional NCM measure.

7. References


