SPEECH ENHANCEMENT BY COMBINING STATISTICAL ESTIMATORS OF SPEECH AND NOISE

Yang Lu and Philipos C. Loizou

Center for Robust Speech Systems
Department of Electrical Engineering
The University of Texas at Dallas, Richardson, Texas, 75080, USA
luyang1980@gmail.com, loizou@utdallas.edu

ABSTRACT
This paper presents a novel speech enhancement algorithm that can substantially improve the signal-to-residual spectrum ratio by combining statistical estimators of the spectral magnitude of the speech and noise. The noise spectral magnitude estimator is derived from the speech magnitude estimator, by appropriately transforming the a priori and the a posteriori SNR values. By expressing the signal-to-residual spectrum ratio as a function of the estimator’s gain function, we derive a hybrid strategy that can improve the signal-to-residual spectrum ratio when the a priori and the a posteriori SNR are detected to be lower than 0 dB. Experimental results showed that the signal-to-residual spectrum ratio as well as the PESQ scores can be improved substantially in stationary and quasi-stationary noise conditions with the proposed hybrid estimators. Informal listening tests revealed improved speech quality and no musical noise.

Index Terms— Statistical-model based speech enhancement, frequency-weighted SNR, SNR improvement

1. INTRODUCTION
A number of speech enhancement algorithms have been proposed in the past three decades [1, Ch.1–7]. The statistical-model based algorithms (e.g., [2][3][4]) have been found to work the best, among all other algorithms, when compared in terms of speech quality [5]. The majority of the statistical estimators apply a single gain function to the noisy speech magnitude spectrum to obtain the clean speech magnitude spectrum. The same gain function is applied regardless of the underly-

This research was supported in part by Grant No. R01 DC007527 from NIDCD/NIH.

ing SNR of each frequency bin. This can be problematic since some estimators can over-attenuate the signal when operating at extremely low SNR levels. On the other hand, other estimators (e.g., power spectrum estimators) tend to apply little attenuation, at the expense of introducing significant amounts of residual noise. It is not clear neither it is straightforward as to how best choose the right estimator for different SNR levels. For that, an objective criterion is needed to base the decision as to which estimator to use. Such a criterion needs to correlate highly with speech quality.

This paper proposes a new and simple strategy for speech enhancement by combining both the spectral magnitude estimators for speech and noise. The decision as to which estimator to use is based on the implicit maximization of the signal-to-residual spectrum ratio. This ratio has been found in [6][7] to correlate highly with both speech quality and speech intelligibility. The estimator for the spectral magnitude of the noise shares exactly the same form as its counterpart (e.g., see [8]), namely the speech spectral magnitude estimator. Once the magnitude spectrum of the noise is estimated, it is easy to compute the spectrum of the speech signal. These two estimators behave differently in various SNR levels, and this paper proposes a simple rule to combine both estimators to maximize the signal-to-residual spectrum ratio.

2. ESTIMATING THE SPEECH FROM THE NOISE ESTIMATE
Assuming that the observed noisy speech is corrupted by additive noise, the corrupted speech $Y$ is expressed in the short-term Fourier transform domain as $Y(\omega_k) = X(\omega_k) + D(\omega_k)$ and in polar form as $Y_k e^{j\theta_k} = X_k e^{j\theta_k} + D_k e^{j\theta_k}$, where $Y(\omega_k) = Y_k e^{j\theta_k}$, and $X_k$ and $D_k$ are the spectral magnitudes of the signal (clean speech) and the noise, respectively. The frequency-domain speech estimators could be expressed in a form of $\hat{X}_k = G_{x,m}^k (\xi^k, \gamma^k) Y_k$, where $G(\xi^k, \gamma^k)$ denotes the gain function, assumed to be a function of $\xi^k$ and $\gamma^k$, $k$ denotes the $k$-th frequency bin, and $m$ denotes the estimator used. The a priori and a posteriori SNRs are defined as $\xi^k \equiv \lambda^k_x / \lambda^k_d$ and $\gamma^k \equiv \lambda^k_y / \lambda^k_d$, where $\lambda^k_x$ and $\lambda^k_d$ are the variances of speech and noise respectively.

The statistical estimators of the speech and noise have a certain symmetry, in that one can derive the estimator for the noise signal based on the estimator of the speech signal by making the appropriate substitution of the constituent para-
meters $\xi_k$ and $\gamma_k$. More precisely, the statistical estimator of the noise can be derived as follows:

$$
\hat{D}_k = H_m^k \left( \xi_d^k, \gamma_d^k \right) Y_k,
$$

(1)

where the parameters $\xi_d^k$ and $\gamma_d^k$ are defined as:

$$
\xi_d^k \equiv \frac{\lambda_k^d}{\lambda_k^x} = \frac{1}{\xi_x}, \quad \gamma_d^k \equiv \frac{Y_k^2}{\xi_x^2} = \frac{\gamma_k^x}{\xi_x},
$$

(2)

The new gain function, denoted as $H_m^k$, applies the same rule as its counterpart $G_{x,m}^k$, and the parameters are now defined with respect to the noise. It is interesting to note that the $\xi_d$ and $\gamma_d$ parameters are simple transformations of $\xi_x$ and $\gamma_x$.

For example, the log-MMSE estimator [3] of the speech magnitude spectrum can be computed, based on the above, as follows:

$$
G_{x,\log\text{MMSE}}(\xi_x, \gamma_x) = \frac{\xi_x}{\xi_x + 1} \exp \left\{ \frac{1}{2} \int_{\nu_x}^{\infty} e^{-t} \frac{1}{t} dt \right\}
$$

(3)

where $\nu_x = \xi_x \cdot \gamma_x / (\xi_x + 1)$. The estimator of the noise magnitude spectrum can be computed, based on the above, as follows:

$$
H_{\log\text{MMSE}}(\xi_d, \gamma_d) = \frac{\xi_d}{\xi_d + 1} \exp \left\{ \frac{1}{2} \int_{\nu_d}^{\infty} e^{-t} \frac{1}{t} dt \right\}
$$

(4)

where $\nu_d = \gamma_x / \{ \xi_x (\xi_x + 1) \}$. Hence, with a simple parameter substitution (as defined in Eq. (2)), we can derive the noise magnitude estimator from the speech magnitude estimator.

Using the maximum likelihood estimation for the phase [2], i.e., $\theta_d = \theta_y$, we can easily compute the speech spectral component as follows:

$$
\hat{X}_k e^{j\theta_x} = Y_k e^{j\theta_y} - \hat{D}_k e^{j\theta_d}
$$

$$
= \left[ 1 - H_m^k \left( \xi_d^k, \gamma_d^k \right) \right] Y_k e^{j\theta_y}
$$

$$
= G_{x,m}^k \cdot Y_k e^{j\theta_y}
$$

(5)

where $G_{x,m}^k$ is the speech estimator derived from the noise spectrum estimate (Eq. (1)). Note that in the case of the Wiener estimator, due to its symmetry, the $G_{x,m}^k$ and $G_{d,m}^k$ gain functions are the same. This is not the case for other estimators such as the MMSE and logMMSE estimators.

### 3. Maximizing the Signal-to-Residual Noise Ratio Using Different Estimators

A frequency-weighted version of the signal-to-residual spectrum ratio (also referred to as the frequency-weighted segmental SNR measure [9][10]) has been found in [6][7] to correlate highly with speech quality and intelligibility. This ratio (denoted as $\text{SNR}_{ESI}$) is defined as:

$$
\text{SNR}_{ESI}(\xi_x^k) = \frac{E[|X(\omega_k)|^2]}{E[|X(\omega_k) - X(\omega_k)|^2]}
$$

$$
= \frac{E[|X(\omega_k)^2]}{E[|X(\omega_k)|^2 - G_k Y(\omega_k)|^2]}
$$

(6)

where $G_k$ is the gain function defining a statistical estimator. Given the high correlation of the $\text{SNR}_{ESI}$ ($\text{SNR}_{ESI}$ is similar to the frequency-weighted segmental SNR measure, $\text{fwSNRseg}$, used in [6], except for the weights) measure with speech quality [6] and speech intelligibility [7], we sought estimators, or combinations of estimators, that would maximize this measure.

Fig. 1 plots the $\text{SNR}_{ESI}$ measure as a function of $\xi_x$ for MMSE derived [2] estimators $G_x$ and $G_d$. The latter estima-
tor, $G_d$ was derived using the procedure outlined in Eq. 1-5. Figure 1 clearly reveals that when $\xi_x$ and $\gamma_x$ are less than 0 dB, the SNR$_{ESI}$ value of the $G_d$ estimator exceeds the corresponding SNR$_{ESI}$ value of the $G_c$ estimator. In this region, where both $\xi_x$ and $\gamma_x$ are smaller than 0 dB, the speech signal is typically weak or it falls within a speech-absent segment. Fig. 1 suggests that the $G_d$ estimator is more appropriate for low SNR regions, while the $G_c$ estimator is more appropriate for higher SNR regions. This becomes more evident when $\gamma_x < 0$ dB. A similar SNR$_{ESI}$ vs. $\xi_x$ pattern was observed with other estimators. Fig. 2 and Fig. 3 show the corresponding plots obtained with the logMMSE estimator [3] and the power-spectrum estimator (MMSE-PS)[4].

Based on Fig. 1-3, we can conclude that in order to maximize the signal-to-residual spectrum ratio (and subsequently speech quality), the following rule needs to be adopted:

$$\hat{X}(\omega_k) = G_c^k \cdot Y(\omega_k) = \begin{cases} G_d^k \cdot Y(\omega_k) & \xi_x \leq 0 \text{dB} \\ G_x^k \cdot Y(\omega_k) & \text{otherwise} \end{cases},$$

(7)

where the subscript $c$ denotes the “combined” or hybrid estimator.

4. IMPLEMENTATION AND EXPERIMENTS

Speech was segmented into 20-ms frames and Hanning windowed with 50% overlap. The short-time Fourier transform was applied to each frame to obtain the noisy magnitude spectrum $Y_k$. An inverse Fourier transform was taken of $\hat{X}_k$ using the noisy speech phase spectrum to reconstruct the time-domain signal. The overlap-add method was used to obtain the enhanced signal. The “decision-directed” method was used to estimate the a priori SNR $\xi_x$, where the smoothing factor $\alpha$ was set to 0.98. The noise variance was estimated using the Minimum Controlled Recursive Average (MCRA) noise-estimation algorithm [11].

Fig. 3. Plot of SNR$_{ESI}$ values as a function of $\xi$ and fixed values of $\gamma$. The $G_c$ and $G_d$ estimators were derived using the MMSE-PS [4] estimator.

Fig. 4. $\Delta$fwSNRseg values obtained by various estimators in car, street, babble and white noise conditions at 0, 5, 10 and 15 dB SNR. Symbol ‘o’ denotes $\Delta$fwSNRseg$\text{MMSE}$, ‘x’ denotes $\Delta$fwSNRseg$\text{MMSE-PS}$, and ‘\n’ denotes $\Delta$fwSNRseg$\text{MMSE}$.

A total of 30 sentences taken from the publicly-available NOIZEUS database [5] was used to evaluate the performance of the proposed estimators. The sentences were corrupted by car, street, babble and white noise at 0 dB, 5 dB, 10 dB and 15 dB SNRs. Two measures were used to assess performance, the frequency-weighted segmented SNR (fwSNRseg), and the Perceptual Evaluation of Speech Quality (PESQ) [12] measure. Note that the fwSNRseg measure is a frequency-weighted version of the SNR$_{ESI}$ measure (Eq. (6)). To assess improvement in performance, relative to that attained when using the $G_x$ estimators alone, we used the following difference metric:

$$\Delta \text{fwSNRseg}_m = \text{fwSNRseg}_{G_c,m} - \text{fwSNRseg}_{G_x,m},$$

(8)

where $m$ represents the estimator used.

Three different statistical estimators were used to derive $G_d$ : (1). the MMSE estimator [2], (2). the log-MMSE estimator [3], and (3). the power-spectrum estimator (MMSE-PS) [4]. The results for the $\Delta$fwSNRseg measure are shown in 4 and the results in terms of PESQ values are shown in Table 1. The estimators labeled with the suffix “-Mod” are the proposed estimators implemented as per Eq. (7).

As can be seen from Figure 4, the values of $\Delta$fwSNRseg$_m$ obtained using the hybrid estimator $G_c$ are positive for all noise conditions. This suggests that the proposed estimator $G_c$ improved the fwSNRseg measure, relative to that obtained using the original $G_x$ estimators alone. The improvement ranged from 0.5 dB obtained with the logMMSE estimator to nearly 4 dB obtained with the MMSE-PS estimator.

Table 1 shows the PESQ comparison. Improvement in the PESQ scores was obtained using the proposed $G_c$ estima-
This paper presented a novel speech enhancement algorithm that can substantially improve the $\text{SNR}_{\text{ESI}}$ measure by combining the estimators of the spectral magnitude of the speech and noise. The two estimators are combined so as to maximize the $\text{SNR}_{\text{ESI}}$ measure, which was found to be highly correlated with speech quality [6]. Experimental results demonstrated a significant improvement in the frequency-weighted segmental SNR (a measure similar to $\text{SNR}_{\text{ESI}}$) ranging from 0.5 to 4 dB. The PESQ scores were also found to be higher, compared to that obtained by the original estimators (e.g., MMSE, logMMSE), particularly in the high-SNR conditions (5 and 10 dB). Informal listening tests revealed better background noise suppression with no noticeable musical noise.

5. CONCLUSIONS

6. REFERENCES


Table 1. PESQ performance comparison.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Method</th>
<th>15dB</th>
<th>10dB</th>
<th>5dB</th>
<th>0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>MMSE</td>
<td>2.90</td>
<td>2.61</td>
<td>2.25</td>
<td>1.94</td>
</tr>
<tr>
<td></td>
<td>MMSE-Mod</td>
<td>3.00</td>
<td>2.67</td>
<td>2.29</td>
<td>1.94</td>
</tr>
<tr>
<td></td>
<td>LogMMSE</td>
<td>2.94</td>
<td>2.64</td>
<td>2.28</td>
<td>1.93</td>
</tr>
<tr>
<td></td>
<td>LogMMSE-Mod</td>
<td>2.98</td>
<td>2.64</td>
<td>2.26</td>
<td>1.91</td>
</tr>
<tr>
<td></td>
<td>MMSE-PS</td>
<td>2.73</td>
<td>2.42</td>
<td>2.09</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>MMSE-PS-Mod</td>
<td>2.98</td>
<td>2.67</td>
<td>2.29</td>
<td>1.95</td>
</tr>
<tr>
<td>Street</td>
<td>MMSE</td>
<td>2.81</td>
<td>2.52</td>
<td>2.19</td>
<td>1.87</td>
</tr>
<tr>
<td></td>
<td>MMSE-Mod</td>
<td>2.87</td>
<td>2.57</td>
<td>2.12</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>LogMMSE</td>
<td>2.85</td>
<td>2.56</td>
<td>2.22</td>
<td>1.87</td>
</tr>
<tr>
<td></td>
<td>LogMMSE-Mod</td>
<td>2.88</td>
<td>2.57</td>
<td>2.24</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td>MMSE-PS</td>
<td>2.68</td>
<td>2.40</td>
<td>2.07</td>
<td>1.78</td>
</tr>
<tr>
<td></td>
<td>MMSE-PS-Mod</td>
<td>2.85</td>
<td>2.54</td>
<td>2.19</td>
<td>1.84</td>
</tr>
<tr>
<td>Babble</td>
<td>MMSE</td>
<td>2.90</td>
<td>2.55</td>
<td>2.21</td>
<td>1.81</td>
</tr>
<tr>
<td></td>
<td>MMSE-Mod</td>
<td>2.94</td>
<td>2.56</td>
<td>2.18</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>LogMMSE</td>
<td>2.94</td>
<td>2.57</td>
<td>2.21</td>
<td>1.86</td>
</tr>
<tr>
<td></td>
<td>LogMMSE-Mod</td>
<td>2.95</td>
<td>2.56</td>
<td>2.19</td>
<td>1.83</td>
</tr>
<tr>
<td></td>
<td>MMSE-PS</td>
<td>2.79</td>
<td>2.46</td>
<td>2.15</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>MMSE-PS-Mod</td>
<td>2.91</td>
<td>2.54</td>
<td>2.16</td>
<td>1.79</td>
</tr>
<tr>
<td>White</td>
<td>MMSE</td>
<td>2.92</td>
<td>2.64</td>
<td>2.30</td>
<td>1.92</td>
</tr>
<tr>
<td></td>
<td>MMSE-Mod</td>
<td>3.05</td>
<td>2.74</td>
<td>2.38</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td>LogMMSE</td>
<td>2.98</td>
<td>2.69</td>
<td>2.35</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>LogMMSE-Mod</td>
<td>3.02</td>
<td>2.70</td>
<td>2.34</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td>MMSE-PS</td>
<td>2.70</td>
<td>2.40</td>
<td>2.05</td>
<td>1.74</td>
</tr>
<tr>
<td></td>
<td>MMSE-PS-Mod</td>
<td>3.03</td>
<td>2.75</td>
<td>2.39</td>
<td>2.02</td>
</tr>
</tbody>
</table>

The algorithms indicated with “-Mod” are based on Eq. (7).