Generalized Parametric Spectral Subtraction Using Weighted Euclidean Distortion

Amit Das, John H. L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

amit.das@colorado.edu, john.hansen@utdallas.edu

Abstract

An improved version of the original parametric formulation of the generalized spectral subtraction method is presented in this study. The original formulation uses parameters that minimize the mean-square error (MSE) between the estimated and true speech spectral amplitudes. However, the MSE does not take into account any perceptual measure. We propose two new short-time spectral amplitude estimators based on a perceptual error criterion - the weighted Euclidean distortion. The error function is easily adaptable to penalize spectral peaks and valleys differently. Performance evaluations were performed using two noise types over four SNR levels and compared to the original parametric formulation. Results demonstrate that for most cases the proposed estimators achieve greater noise suppression without introducing speech distortion.

Index Terms: speech enhancement, generalized spectral subtraction

1. Introduction

The objective of any speech enhancement algorithm is to suppress background noise, improve perceived quality and intelligibility, reduce listener fatigue, and improve system performance. It is difficult to address all these objectives simultaneously in a single enhancement algorithm since this essentially means that noise should be suppressed in a way which does not introduce processing artifacts, musical noise, or speech distortions. In recent years, enhancement algorithms have evolved to incorporate perceptual information [1, 2, 3, 12]. For demanding real world applications, the challenge is to formulate simple yet effective enhancement algorithms that minimize processing time and memory on embedded devices such as mobile telephony. Spectral subtraction (SS) [4] has been widely used and accepted. The SS method estimates the noise spectrum, uses a simple subtraction technique, and makes use of a floor value from neighboring frames to estimate the clean speech spectral amplitudes. However, it comes with a drawback that it is likely to produce musical noise that is sometimes annoying to the listener.

In [5], a modified version of this scheme was proposed by suggesting the use of over-subtraction factor and spectral floor parameters. The over-subtraction factor determined the amount of signal distortion whereas the floor parameter was the key in balancing residual noise with musical noise. Later, in [6], the use of frequency band dependent over subtraction factor was advocated to take into account the narrowband nature of some real world noises. Sim et al. [7] introduced a short time spectral amplitude estimator of speech that minimizes the MSE between the true speech and estimated speech spectral amplitude by finding optimum weighting coefficients for both the noisy speech and noise. Improvement in SNR and spectral distance using this method were similar to that of Ephraim-Malah’s MMSE estimator [8]. The approach in [7] weights the estimation errors equally in spectral valleys and peaks. This does not exploit the effect of auditory masking where errors in formant peaks may not be perceptually noticeable due to the presence of strong masking components of speech [3]. Hansen et al. developed a generalized MMSE method [11] and later incorporated an auditory masking threshold formulation [12] for improved speech quality. In this study, we derive a more generalized version of the SS based spectral amplitude estimator that offers greater flexibility to penalize the errors in spectral valleys (or peaks) by simply tuning an exponent term.

2. Proposed Algorithm

Assuming that noise is additive and statistically independent of the speech signal, the representation of noise in the frequency domain can be given by,

\[ Y_k e^{j\phi_k} = S_k e^{j\phi_k} + D_k e^{j\phi_k}. \]  

Here, \( Y_k, S_k, D_k \) represent the spectral amplitudes and \( \phi_k \) represent the phases of the short-time Fourier transforms of noisy speech, clean speech and noise respectively at frequency bin \( k \).

Following the notation of Sim et al. in [7], the generalized parametric spectral subtraction [7] is represented as,

\[ \hat{S}_k = a_k Y_k^\alpha - b_k E[D_k^\alpha] \]  

where \( S_k^\alpha \) is the spectral amplitude of clean speech raised to the exponent \( \alpha \) and \( a_k, b_k \) are weighting parameters for frequency bin \( k \). Although the parameters are also functions of \( \alpha \), we have dropped \( \alpha \) from their subscript for notational simplicity. In [7], the parametric estimator in (2) is optimized by minimizing the MSE between \( S_k^\alpha \) and \( \hat{S}_k \). In the proposed solution, we modify the error function as:

\[ C_k(S_k^\alpha, \hat{S}_k) = (S_k^\alpha - \hat{S}_k)^T W (S_k^\alpha - \hat{S}_k) \]  

where \( W = \text{diag}(S_1^\alpha, S_2^\alpha, ..., S_K^\alpha) \), \( K \) is the size of the short-time Fourier transform representation of the analysis frame and \( P \) is a constant exponent term. The error function in (3) can be treated as the weighted Euclidean distortion measure as in [3].
It should be noted that when $P > 0$, the error function penalizes the errors in the spectral peaks more heavily than spectral valleys. When $P < 0$, the errors in the spectral valleys are penalized more than those in spectral peaks. Since MSE weighs the errors equally in all regions of the spectrum, the value of $P$ in the weighted Euclidean distortion offers flexibility in modifying the error function. The musical noise present in spectral subtraction approaches can be attributed to the random occurrence of sinusoidal peaks along the spectrum. In an additive white noise scenario, musical noise can be considered predominant in the region of spectral valleys since these are regions of low SNR. Therefore, in this paper, we focus on values where $P < 0$. For any frequency bin $k$, where $1 \leq k \leq K$, (3) reduces to,

$$
C_k(S_k^a, S_k^b) = S_k^p(S_k^a - S_k^b)^2.
$$

(4)

The true spectral amplitude can be written as $S_k^a = Y_k^a - D_k^a$. Substituting the true spectral amplitude and (2) in (4), and finding the expectation of the estimation error results in,

$$
E[C_k] = (1 - a_k)^2 \mu_{S_k^a}^2 + a_k^2 \mu_{S_k^p}^2 \mu_{D_k^a}^2 + (b_k^2 - 2a_kb_k) \mu_{S_k^p}^2 \mu_{D_k^a}^2 + 2(1 - a_k)(b_k - a_k) \mu_{S_k^p} \mu_{D_k^a}^2
$$

(5)

where $\mu(\cdot)$ represents the expected value of the random variable in (.)

The optimum values of $a_k, b_k$ that minimize the error in (4) are found by differentiating (5) with respect to $a_k, b_k$ separately and setting them to zero. After simplification, the generalized solutions for the optimum values are given by,

$$
a_k = \frac{(\mu_{S_k^p}^2 + \mu_{S_k^p}^2 \mu_{D_k^a}^2)}{\mu_{S_k^a}^2 (\mu_{S_k^p}^2 + \mu_{S_k^p}^2 \mu_{D_k^a}^2)}
$$

(6)

$$
b_k = a_k - (1 - a_k) \frac{\mu_{S_k^p}^2 + \mu_{S_k^p}^2 \mu_{D_k^a}^2}{\mu_{S_k^a}^2 (\mu_{S_k^p}^2 + \mu_{S_k^p}^2 \mu_{D_k^a}^2)}
$$

(7)

Assuming the case where real and imaginary parts of the clean speech and noise DFTs to be Gaussian distributed with zero means in the analysis frame, their spectral amplitudes will be Rayleigh distributed. Substituting the distribution and its simplification ([7],[11]-[14]) in (6) and (7) further simplifies the parameters to,

$$
a_k = \frac{\xi_k^a \theta_1}{\xi_k^a \theta_1 + \theta_2}
$$

(8)

$$
b_k = \frac{\xi_k^a \theta_1 \Gamma(\alpha/2 + 1) - \xi_k^a \theta_2 \Gamma(\alpha/2 + 1)}{\Gamma(\alpha/2 + 1)} \frac{1}{\xi_k^a \theta_1 + \theta_2}
$$

(9)

where $\xi_k$ is the a priori SNR [8] at the $k^{th}$ frequency component given by,

$$
\xi_k = \frac{\mu_{S_k^2}}{\mu_{D_k^a}^2}
$$

(10)

and constants $\theta_1, \theta_2, \theta_3$ are given by,

$$
\theta_1 = \Gamma(\alpha + P/2 + 1) \Gamma(P/2 + 1) - \Gamma^2(\alpha/2 + P/2 + 1)
$$

(11)

$$
\theta_2 = \Gamma^2(P/2 + 1) \{ \Gamma(\alpha + 1) - \Gamma^2(\alpha/2 + 1) \}
$$

(12)

$$
\theta_3 = \frac{\Gamma(\alpha/2 + 1) \Gamma(\alpha + 1)}{\Gamma(P/2 + 1)}
$$

(13)

These constants are functions of $\alpha$ and $P$. Substituting (8), (9) in (2), the gain function $G_k = S_k/Y_k$ can be written as,

$$
G_k = \left\{ \frac{\xi_k^a \theta_1}{\xi_k^a \theta_1 + \theta_2} \right\}^{\gamma_k - \alpha/2} \left\{ \theta_3 - (\theta_1 \Gamma(\alpha/2 + 1) - \theta_2) \right\}^{\gamma_k - \alpha/2}
$$

(14)

where $\gamma_k = Y_k^2/\mu_{D_{k^2}}$ is the a posteriori SNR [8] at the $k^{th}$ frequency component. Following the nomenclature in [7], (14) may be considered as the P-unconstrained parametric estimator. Applying constraints (i.e., setting $a_k = b_k$ in (2)), the value of $a_k$ that minimizes (5) is,

$$
a_k = \frac{\xi_k^a}{\xi_k^a + \beta_{\alpha, P}}
$$

(15)

where,

$$
\beta_{\alpha, P} = \frac{\Gamma(\alpha + 1)}{\Gamma(P/2 + 1)} \{ \Gamma(\alpha + 1) - \Gamma^2(\alpha/2 + 1) \}
$$

(16)

Finally, the gain function for the P-constrained parametric estimator can be represented as,

$$
G_k = \left\{ \frac{\xi_k^a}{\xi_k^a + \beta_{\alpha, P}} \right\}^{\gamma_k - \alpha/2} \left\{ 1 - (\Gamma(\alpha/2 + 1) \gamma_k - \alpha/2) \right\}^{\gamma_k - \alpha/2}
$$

(17)

It is easy to see that if we set $P = 0$, then the gain functions of the proposed P-unconstrained and P-constrained estimators become the gain functions of the estimators introduced in ([7], [32], [33]).

3. Gain Characteristics

The characteristics of the proposed estimators can be studied from their gain curves. From (14) and (17), the gains are functions of $\xi_k$, $\gamma_k$, $\alpha$, and $P$. In Fig.1 and Fig.2, the gain functions are plotted using a fixed value of $\alpha$ and a priori SNR $\xi_k$ of 5dB. Each curve represents the gain behavior at a particular value of $P$ across a wide range of instantaneous SNRs $\gamma_k - 1$. At $P = 0$, it becomes the estimator derived in [7]. For the P-unconstrained estimator, $\alpha$ is set to 1, whereas for the P-constrained estimator $\alpha$ is set to a conservative value of 2 (larger $\alpha$ corresponds...
to higher gain) due to its stronger suppression characteristics than the unconstrained estimator. From Fig.1, it is evident that at lower values of $\gamma_k$, the P-unconstrained estimator achieves a greater degree of noise suppression when $-1.9 < P < 0$. However, when $P > 0$, noise is under suppressed compared to the original unconstrained estimator. At higher $\gamma_k$ (possibly regions of strong speech components), the attenuation is low (higher gain) when $-1 < P < 0$. This may be noted that for the case of constrained estimator in Fig.2, the gain functions for $-1.5 < P < 0$ do not offer significant improvement compared to the original constrained estimator. This might prompt us to choose $P$ such that $P < -1.5$. However, in this case, we run the risk of severely attenuating most of the signals including those in the regions of high $\gamma_k$. The gain characteristics are similar in nature for other values of $\xi_k$. We highlight that we used a data of $P = -1.9$ instead of $-2.0$ since $S_k^c$ is Rayleigh distributed and $\mu_{S_k^c}$ exists only if $P > -2$ [9, (6.631.1)]. This is the same limitation encountered in [3].

Fig.3 and Fig.4 illustrate the comparison of the gain functions of the proposed estimators with the original estimators. In both figures, $\alpha = 2$ and $P = -1$ remain constant for all the regions have similar gain values, and for low energy regions the gains of the proposed and original estimators are nearly the same, suggesting the proposed estimator does well in preserving the speech signal components. Although even noise-only components are still retained due to lower attenuation, they are masked by the presence of strong speech components. As $\gamma_k$ decreases (maintaining $\xi_k > -5\,\text{dB}$), stronger attenuation is applied but mostly invariant for values of $\gamma_k < -5\,\text{dB}$. As noted in [7], this is a region of low probability in the context of speech enhancement and is of least interest. The next region where both $\gamma_k < 0\,\text{dB}$ and $\xi_k < -5\,\text{dB}$ is likely to be the noise-only region. It is clear that the proposed estimator offers greater noise suppression than its original counterpart. Finally, the region of $\gamma_k > 0\,\text{dB}$ and $\xi_k < -5\,\text{dB}$ is most vulnerable to musical noise. The presence of high instantaneous $a posteriori$ SNR $\gamma_k$ when average $a priori$ SNR $\xi_k$ low suggests the presence of random spectral peaks [10]. The proposed estimator offers greater suppression of noise in this region. For the case of the constrained estimator in Fig.4, high energy regions have similar gain values, and for low energy regions the proposed estimator offers greater attenuation.

4. Results

A set of 16 phonetically balanced test utterances from the TIMIT corpus were used for objective quality evaluations. The test corpus spans speakers from 8 dialect regions with 1 male and 1 female speaker per region. The corpus was sampled at 8kHz and degraded with two noise types - flat communications channel noise (FLN) (mostly stationary), and large crowd noise (LCR) (mostly non-stationary) at global SNRs of -5, 0, 5, and 10 dB. The quality of enhanced speech was assessed using objective speech quality measures such as the Itakura-Saito (IS) distortion (lower the better), segmental SNR (SegSNR) (higher the better). In all the experiments, the $a priori$ SNR $\xi_k$ in (10) was evaluated using the decision-directed approach [8], (51)). A value of 0.98 was used as the smoothing constant to weigh the
SNR of the previous frame. Further, we maintain the same noise estimation method used by Sim et al. ([7], (24)). The smoothing constant (\( \gamma \)) for noise estimation was set to 0.90.

A summary of the performance of the estimators for the case of flat communications channel noise are tabulated in Table 1. The IS values in the table represent the percentage improvement which is calculated as the ratio of the difference in the IS distortion between noisy and enhanced speech and the IS distortion of the noisy speech expressed in percentage. The values indicated in bold correspond to the best estimator for a given SNR. For relatively lower SNRs, the original constrained estimator of [7] has the best performance. The original constrained estimator uses an additional spectral gain floor for uniformity in our assessment. The most notable improvement is in the rise of segmental SNR by about 3dB from the IS improvement (comparable to original constrained estimator). How- ever, the P-constrained estimator still performs better than the original constrained estimator and P-unconstrained estimator. It achieves this without under- going severe attenuation of the speech signal as is evident from the IS improvement (comparable to original constrained estimator). Moreover, the P-unconstrained estimator performs the best in terms of noise suppression. However, this requires further investigation of empirical setting of the spectral gain floor value and smoothing mechanism of the estimated spectral amplitudes. In our experiments, we used the same value for gain floor (0.1) as the original constrained estimator. How- ever, the P-constrained estimator still performs better than the original constrained estimator in terms of degree of noise suppression, but did not perform as well as the P-unconstrained estimator.

5. Conclusions

Two short time spectral amplitude estimators, P-unconstrained and P-constrained, are derived using parametric coefficients that minimize the weighted Euclidean distortion between the clean speech and estimated speech spectral amplitudes. For negative values of \( P \), the proposed estimators emphasize the spectral valley errors more than peaks and achieve more noise suppression than the original estimators. From our experimental evaluations, the P-unconstrained estimator performed better than the P-constrained estimator and the original estimators.

6. References


Table 1: Itakura-Saito, SegSNR results of the original and proposed estimators using \( \alpha = 1.2 \) and \( P = -1.5 \) for the case of flat communications channel noise

<table>
<thead>
<tr>
<th>Global SNR</th>
<th>-5dB</th>
<th>0dB</th>
<th>5dB</th>
<th>10dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Itakura-Saito Distortion (%)</td>
<td>8.13</td>
<td>7.33</td>
<td>6.3</td>
<td>5.7</td>
</tr>
<tr>
<td>Orig Unconstrained</td>
<td>59.94</td>
<td>63.08</td>
<td>66.67</td>
<td>70.21</td>
</tr>
<tr>
<td>Orig Constrained</td>
<td>63.33</td>
<td>66.96</td>
<td>68.03</td>
<td>71.84</td>
</tr>
<tr>
<td>P-Unconstrained</td>
<td>62.19</td>
<td>65.50</td>
<td>69.46</td>
<td>73.44</td>
</tr>
<tr>
<td>P-Constrained</td>
<td>63.20</td>
<td>65.73</td>
<td>68.73</td>
<td>70.98</td>
</tr>
<tr>
<td>SegSNR (dB)</td>
<td>-11.28</td>
<td>-7.85</td>
<td>-3.74</td>
<td>0.81</td>
</tr>
<tr>
<td>Orig Unconstrained</td>
<td>2.69</td>
<td>2.95</td>
<td>2.62</td>
<td>1.64</td>
</tr>
<tr>
<td>Orig Constrained</td>
<td>4.86</td>
<td>4.98</td>
<td>4.20</td>
<td>2.61</td>
</tr>
<tr>
<td>P-Unconstrained</td>
<td>8.26</td>
<td>8.25</td>
<td>7.05</td>
<td>5.33</td>
</tr>
<tr>
<td>P-Constrained</td>
<td>6.96</td>
<td>7.25</td>
<td>6.23</td>
<td>4.52</td>
</tr>
</tbody>
</table>

Table 2: Itakura-Saito, SegSNR results of the original and proposed estimators using \( \alpha = 1.1 \) and \( P = -1.3 \) for the case of large crowd noise

<table>
<thead>
<tr>
<th>Global SNR</th>
<th>-5dB</th>
<th>0dB</th>
<th>5dB</th>
<th>10dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Itakura-Saito Distortion (%)</td>
<td>6.30</td>
<td>6.03</td>
<td>5.34</td>
<td>4.7</td>
</tr>
<tr>
<td>Orig Unconstrained</td>
<td>55.47</td>
<td>56.99</td>
<td>59.25</td>
<td>61.11</td>
</tr>
<tr>
<td>Orig Constrained</td>
<td>59.40</td>
<td>62.87</td>
<td>64.12</td>
<td>64.54</td>
</tr>
<tr>
<td>P-Unconstrained</td>
<td>59.37</td>
<td>62.17</td>
<td>66.70</td>
<td>70.87</td>
</tr>
<tr>
<td>P-Constrained</td>
<td>60.43</td>
<td>65.18</td>
<td>67.30</td>
<td>68.68</td>
</tr>
<tr>
<td>SegSNR (dB)</td>
<td>-9.27</td>
<td>-6.99</td>
<td>-2.85</td>
<td>1.73</td>
</tr>
<tr>
<td>Orig Unconstrained</td>
<td>2.96</td>
<td>4.00</td>
<td>3.70</td>
<td>2.89</td>
</tr>
<tr>
<td>Orig Constrained</td>
<td>4.65</td>
<td>5.32</td>
<td>4.61</td>
<td>3.31</td>
</tr>
<tr>
<td>P-Unconstrained</td>
<td>6.44</td>
<td>6.47</td>
<td>5.52</td>
<td>4.12</td>
</tr>
<tr>
<td>P-Constrained</td>
<td>5.25</td>
<td>5.87</td>
<td>4.95</td>
<td>3.45</td>
</tr>
</tbody>
</table>