PHONEME SELECTIVE SPEECH ENHANCEMENT USING THE GENERALIZED PARAMETRIC SPECTRAL SUBTRACTION ESTIMATOR

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

In this study, the generalized parametric spectral subtraction estimator is employed in the context of a ROVER speech enhancement framework to develop a robust phoneme class selective enhancement algorithm. The parametric estimator is derived by a) optimizing the weighted Euclidean distortion cost function and b) by modeling clean speech spectral magnitudes as Rayleigh distributed priors. A set of enhanced utterances are generated from a single noisy utterance by tuning the parameters of the parametric estimator for different phoneme classes. The speech and non-speech segments are segregated using a voice activity detector. Thereafter, the mixture maximum model is used to make soft decisions on these segments to determine their phoneme class weights. The segments from the enhanced utterances are weighted by these decisions and combined to form the final composite utterance. Using segmental SNR and Itakura-Saito metrics over two noise types and four SNR levels, it was demonstrated that the composite utterance exhibited better phoneme class improvement than the individual utterances enhanced from the parametric estimator.

Index Terms—generalized spectral subtraction, phoneme selective speech enhancement.

1. INTRODUCTION

Each phoneme class can undergo different levels of degradation in the same noise condition. This is due to its frequency content, and the degree of the influence of the noise type based on the stationarity and the bandwidth of the noise ([1], Chapter 2). Hansen and Arslan [2] used hidden Markov models (HMM) to create 13 phoneme class models. Using the forward algorithm scoring procedure, conditional probabilities \( p(X_i|\lambda_i) \), \( i = 1,2,\ldots,13 \), were obtained where \( X_i \) represents the observation vector from noisy speech, and \( \lambda_i \) is the noisy speech HMM model for phoneme class \( i \). The difference of the top two class scores was used for phoneme enhancement. In this paper, a class selective enhancement approach is presented based on the generalized spectral subtraction (GSS) estimator derived by Sim et al. [3]. For this purpose, the phonemes are split into three broad phoneme classes (BPCs)—sonorants, obstruents, and silence. In the GSS algorithm, the parameters of the gain functions are fixed over the entire utterance. In this study, the parameters of the gain functions are varied by using the weighted Euclidean distortion (WED) as the cost function and by assuming Rayleigh distributed priors for the clean speech spectral magnitudes. It is shown that these WED optimized gain functions obtain better enhancement levels per phoneme class than the baseline GSS estimator of Sim et al. However, even with the WED optimized gains, no single set of parameter exists that can attain equivalent enhancement potential across all classes. To overcome this limitation, a ROVER (Recognizer Output Voting Error Reduction) based solution is employed. In this solution, for a given noisy utterance, three different enhanced utterances are generated from the parametric estimators with each estimator tuned for a specific BPC. The utterance is partitioned into speech and non-speech regions using a voice activity detector (VAD). The ROVER solution applies soft decisions using the mixture maximum (MIXMAX) model, first formulated in [5] and later used for class independent MMSE speech enhancement [6], to weigh and combine the phoneme segments obtained from the three enhanced utterances. The remainder of the paper is organized as follows. In Section 2, derivations of the parametric versions of GSS estimators are outlined and the application of the MIXMAX model is briefly reviewed. Objective quality evaluations are performed in Section 3 and the conclusion of the paper is drawn in Section 4.

2. ENHANCEMENT MODEL

2.1. Generalized Spectral Subtraction

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

\[
Z_k e^{j\phi_{z,k}} = X_k e^{j\phi_{x,k}} + D_k e^{j\phi_{d,k}}.
\]

(1)

Here, \( Z_k, X_k, D_k \) represent the spectral magnitudes and \( \phi_{z,k}, \phi_{x,k}, \phi_{d,k} \) represent the phases of the discrete Fourier transforms (DFT) of noisy speech, clean speech and noise respectively at frequency bin \( k \). If \( X_k \) is the spectral magnitude estimate of clean speech, then using the GSS [3] estimator this can be represented as,

\[
X_k^\alpha = a_k Z_k^\alpha - b_k E|D_k^\alpha|,
\]

(2)

where the term \( \alpha \) is an exponent term, and \( a_k, b_k \) are weighting parameters at frequency bin \( k \). Although these weighting parameters are functions of \( \alpha \), we have dropped \( \alpha \) from their subscripts for notational simplicity. In [3], the estimator in (2) is optimized by minimizing the mean square error (MSE) between \( X_k^\alpha \) and \( X_k^\phi \). In this study, we minimize the generalized weighted Euclidean distortion (WED) between the clean speech and estimate of clean speech spectral magnitude in the context of GSS [4] as shown in (3). If the clean
speech spectral magnitude vector is \( \mathbf{X} = [X_1, X_2, X_3, \ldots, X_K]^T \) in a short-time analysis frame of speech, then the WED error between clean speech and estimate of clean speech spectral magnitude vector is:

\[
C_r(\mathbf{X}^a, \hat{\mathbf{X}}^a) = (\mathbf{X}^a - \hat{\mathbf{X}}^a)^TW(\mathbf{X}^a - \hat{\mathbf{X}}^a),
\]

(3)

where \( W = \text{diag}(X_1^a, X_2^a, \ldots, X_K^a) \), \( K \) is the length of the DFT of the analysis frame, and \( \alpha, \beta \) are constant exponent terms. It should be noted that when \( \beta > 0 \), the error function penalizes the errors in the spectral peaks more heavily than spectral valleys. When \( \beta < 0 \), 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 \( \beta \) in WED error offers flexibility in modifying the error function. The musical noise present in spectral subtraction approaches are predominant in the region of spectral valleys since these are the regions of low SNR. Therefore, we focus on values when \( \beta < 0 \). For any frequency bin \( k, 1 \leq k \leq K \), the “ideal” spectral magnitude [3, eq.6] can be written as \( X_k^a = Z_k^2 - D_k^2 \).

Substituting this and (2) in (3), and taking the expectation results in:

\[
E[C_r] = E[X_k^2]\{(1 - a_k)X_k^a - a_kD_k^2 + b_kE[D_k^2]\} = (1 - a_k)^2\mu_{X_k^{\beta+2}} + a_k^2\mu_{X_k^{\beta}D_k^2} + b_k(b_k - 2a_k)\mu_{X_k^{\beta}}\mu_{D_k^2}^2 + 2(1 - a_k)(b_k - a_k)\mu_{X_k^{\beta+2}}\mu_{D_k^2}.
\]

(4)

where \( \mu(.) \) represents the expectation operator on the random variable in (\( . \)). Assuming mutual independence between frequency components, we attempt to minimize the WED error at each frequency bin independently to minimize the overall WED error in (4). The optimum values of \( a_k, b_k \) that minimize the WED error in (4) can be obtained by partially differentiating (4) with respect to \( a_k \) and \( b_k \) separately, setting them to zero, and solving for \( a_k, b_k \). The optimum values are given by:

\[
a_k = \frac{(\mu_{X_k^{\beta+2}}) - \mu_{X_k^{\beta}}^2}{\mu_{X_k^{\beta+2}}\mu_{X_k^{\beta}} - \mu_{X_k^{\beta+2}}^2},
b_k = a_k - (1 - a_k)\frac{\mu_{X_k^{\beta}}\mu_{D_k^2}}{\mu_{X_k^{\beta}}^2\mu_{D_k^2}}.
\]

(5)

Assuming the case where real and imaginary parts of the clean speech and noise DFTs in the analysis frame are independent and Gaussian distributed with zero means, their spectral magnitudes will be Rayleigh distributed. The \( r^{th} \) moment of a Rayleigh distribution \( f(X_k) \) can be simplified as:

\[
E[X_k^r] = \frac{1}{\Gamma(r/2) + 1}E[X_k^{2r/2}].
\]

(6)

The exponent of \( X_k \) in (6) must satisfy the condition \( r > -2 \) for (6) to exist. Substituting (6) in (5) further simplifies the optimum values of parameters to:

\[
a_k = \frac{\xi_k^\alpha\theta_1}{\xi_k^\alpha\theta_1 + \theta_2}, \quad b_k = \frac{\xi_k^\alpha\theta_1\Gamma(\alpha/2 + 1) - \xi_k^{\alpha/2}\theta_3}{\Gamma(\alpha/2 + 1) \{\xi_k^\alpha\theta_1 + \theta_2\}},
\]

(7)

where \( \Gamma(\cdot) \) represents the complete Gamma function and \( \xi_k \) is the \( \alpha \) priori SNR at the \( k^{th} \) frequency component given by \( \xi_k = \frac{\lambda_X(k)}{\lambda_D(k)} \).

Here, \( \lambda_X(k) = \mu_{X_k^2} \) and \( \lambda_D(k) = \mu_{D_k^2} \) are the variances of clean speech and noise respectively. The constants \( \theta_1, \theta_2, \theta_3, \theta_4 \) are given by:

\[
\begin{align*}
\theta_1 &= \Gamma(\alpha + \beta/2 + 1)\Gamma(\beta/2 + 1) - \Gamma^2(\alpha/2 + \beta/2 + 1), \\
\theta_2 &= \Gamma^2(\beta/2 + 1)\{\Gamma(\alpha + 1) - \Gamma^2(\alpha/2 + 1)\}, \\
\theta_3 &= \theta_2 \Gamma(\alpha/2 + \beta/2 + 1) \Gamma(\beta/2 + 1), \\
\theta_4 &= \frac{\Gamma(\beta/2 + 1)\{\Gamma(\alpha + 1) - \Gamma^2(\alpha/2 + 1)\}}{\Gamma(\alpha + 3/2 + 1)}.
\end{align*}
\]

(8)

which are functions of \( \alpha \) and \( \beta \). Substituting (7)-(8) in (2), the gain function \( G_k = X_k/Y_k \) can be written as:

\[
G_k = \sqrt{\frac{\xi_k^\alpha\theta_1 + \theta_2}{\xi_k^\alpha\theta_1 + \theta_4}}\left(\theta_1 - (\theta_1\Gamma(\alpha/2 + 1) - \theta_4\xi_k^{\alpha/2}\gamma_k^{-\alpha/2})\right),
\]

(9)

where \( \gamma_k = Z_k^2/\mu_{D_k^2} \) is the \( \alpha \) posteriori SNR at the \( k^{th} \) frequency component and \( \alpha, \beta \) are tunable parameters. The gain equation of (9) may be considered as the GSS \( \beta \)-unconstrained (GBU) parametric estimator. To arrive at the constrained estimator, the same optimization procedure is followed as in GBU after applying the constraint \( a_k = b_k \) in (2). Then the GSS \( \beta \)-constrained (GBC) estimator can be represented as:

\[
G_k = \sqrt{\frac{\xi_k^\alpha\theta_1}{\xi_k^\alpha\theta_1 + \theta_4}}\left(1 - \Gamma(\alpha/2 + 1)\gamma_k^{-\alpha/2}\right).
\]

(10)

It is easy to see that if we set \( \beta = 0 \) in (9) or (10), the gain functions collapse to the gain functions of the GSS estimators derived by Sim et al. [3, eqs.(32),(33)].

### 2.2. Application of the MIXMAX Model

The application of the MIXMAX model to phoneme selective speech enhancement [7] is briefly reviewed in this section. For a given noisy utterance, three utterances may be generated from (9) or (10), using parameters for each BPC. We change the notation slightly from the previous sections. Let \( \mathbf{X} \) denote the random vector representing the Mel frequency cepstral coefficients (MFCC) of clean speech sonorants with the \( k^{th} \) component being \( X_k \), where \( k = 1, \ldots, K \) with \( K \) being the length of the MFCC vector. The probability density function of \( \mathbf{X} \) can be modeled by a Gaussian mixture model (GMM) with each mixture consisting of \( K \) components and a diagonal covariance matrix. This can be represented by:

\[
f(x) = \sum_i c_i(x) \prod_k f_{i,k}(x),
\]

(11)

where, \( f_{i,k}(x) \sim N(\mu_{i,k}, D_{i,k}) \) and \( c_i(x) \) is the weight of the \( i^{th} \) mixture. Similarly, let \( Y, D, Z \) represent the MFCC vectors of obstruents in clean speech, noise, and noisy speech respectively. As in \( f_{i,k}(x) \), the \( k^{th} \) component of the \( j^{th} \) mixture of the GMM representing \( Y \) can be given by the p.d.f. \( t_{j,k}(y) \sim N(\mu_{j,k}, D_{j,k}) \) with mixture weight \( c_j(y) \).

Noise is modeled using a single mixture of \( K \) dimensional Gaussian represented by \( g(d) = \prod_k g_k(d) \) where \( g_k(d) \sim N(0, D_{j,k}) \).

Assuming zero means and mutual statistical independence between \( X, Y, D \), the objective is to find the maximum component among sonorants, obstruents and noise from the noisy speech modeled using the MIXMAX model \( Z \approx \max(X, Y, D) \). As mentioned earlier, three utterances, each enhanced for a specific BPC from (9) or (10), are present prior to the MIXMAX evaluation. Let \( Z_X, Z_Y, Z_D \) be the MFCC vectors obtained from the parametric estimators of sonorants, obstruents, and silence BPCs respectively. For a given segment boundary, if the segment is truly a sonorant, then \( Z_X \) should ideally generate a higher
likelihood score than the MFCC vectors obtained from obstruents or silence customized utterances. Hence, under noisy conditions, the probability that $X$ is the maximum is given by
\[
p(X = z_X | X = z_X) = \prod_{k=1} p(X_k = z_X | Y_k = z_X).
\]

A similar equation can be written for the case of obstruents using $p(Y = z_Y | Z = z_Y)$. Denoting the log probabilities of $p(X = z_X | Z = z_X)$ as $\phi_X$ and $p(Y = z_Y | Z = z_Y)$ as $\phi_Y$, the resulting MFCC vector is given by,
\[
\{X \text{ or } Y\} = \frac{e^{\phi_X}}{e^{\phi_X} + e^{\phi_Y}} z_X + \frac{e^{\phi_Y}}{e^{\phi_X} + e^{\phi_Y}} z_Y.
\]

which is converted back into magnitude spectrum. Using this estimate of magnitude spectrum and noisy phase, the composite speech signal is reconstructed by the standard overlap-add method. Therefore, the ROVER enhancement framework uses (9) or (10) along with (13) to form the composite utterance.

### 3. EXPERIMENTAL RESULTS

A set of 32 (16 females, 16 males) phonetically balanced utterances from the TIMIT test corpus, downsampled from 16kHz to 8kHz, was used for objective quality evaluations. The test utterances were 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, and 10 dB. The quality of enhanced speech was assessed using objective quality metrics such as the segmental SNR (SegSNR), and the Itakura-Saito (IS) distortion. Higher values of SegSNR and lower values of IS distortion represent better speech quality. Segmental SNR evaluations were limited to the range of -10 dB to 35 dB. The clean speech GMMs for BPCs used in the MIXMAX model were constructed using 300 utterances taken from the TIMIT training corpus. We used 39 dimensional MFCC vectors for training these GMMs. We found that generating these MFCC vectors from a small frame size of 10ms and skip rate of 5ms resulted in the best performance. This is because with a frame size of 10ms the likelihood of missing capturing phone-to-phone boundaries is lower than with a larger frame size. Hence, we maintained the same frame size and skip rate even while making MIXMAX decisions. We used 16 mixtures to model the GMMs for sonorants and obstruents and 1 mixture for noise GMM. We updated the noise p.d.f. $g(d)$ in noise only regions estimated by the VAD. In the final composite utterance, we replaced these regions with frames obtained from the parametric estimators GBU/GBC customized for silence. For the remaining regions, we used the MIXMAX decision of (13). The corresponding ROVER implementations of GBU and GBC estimators are denoted by - ROVER GBU (RGBU), and ROVER GBC (RGBC). Similarly, the baseline estimators [3, eqs.(31),(33)] are denoted by - GSS unconstrained (GU), and GSS constrained (GC). The SegSNR and IS distortion values for clean speech BPCs degraded by FLN and LCR noises at different global SNRs are given in Table 1 for reference.

A comparison of improvement in objective quality metrics of ROVER enhancement versus its corresponding parametric and baseline estimators across different BPCs is tabulated in Table 2 for the case of speech degraded by FLN noise at global SNR of 0 dB. The table represents a) improvement in SegSNR as a measure of increase in SegSNR over noisy speech values in Table 1, and b) improvement in IS distortion as a measure of decrease in IS distortion from noisy speech values in Table 1. Hence, in both metrics, higher the improvement greater is the objective quality of speech. In addition, these tables also mention the values of the tunable parameters used in the parametric estimators of (9) or (10) to generate the best objective quality for a particular BPC. The tunable parameters used in the algorithms are indicated as $(\alpha, \beta)$. The subscripts - S(sonorants), O(obstruents), N(silence), Ovl(overall) - denote that the parameters $(\alpha, \beta)$ were tuned to generate the best quality for that BPC while overlooking the quality of other BPCs. For example, focusing only on sonorants in Table 2, the SegSNR improvement for GBU$_S$(4.85) is higher than GBU$_O$(3.94), GBU$_N$(-2.70), or GBU$_{Ovl}$(3.98). The quality of BPCs other than sonorants in GBU$_S$ is compromised solely to generate the best quality of sonorant based segments. The rise in SegSNR for sonorants in GBU$_S$, obstruents in GBU$_O$, silence regions in GBU$_N$, and overall in GBU$_{Ovl}$ are higher than those in the baseline estimator GU. Although this demonstrates the efficacy of GBU over GU, none among GBU$_S$/GBU$_O$/GBU$_N$ provide improved levels of enhancement across all BPCs. GBU$_O$(7.77) and GBU$_N$(7.45) do well in obstruents and silence regions but the rise in SegSNR of their sonorants is much lower than the base-
Table 3. SegSNR and IS improvements of ROVER vs baseline algorithms across all BPCs.

<table>
<thead>
<tr>
<th>BPC</th>
<th>Algo</th>
<th>Rise in SegSNR (FLN)</th>
<th>Fall in IS Distortion (FLN)</th>
<th>Rise in SegSNR (LCR)</th>
<th>Fall in IS Distortion (LCR)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-5 dB</td>
<td>0 dB</td>
<td>5 dB</td>
<td>10 dB</td>
</tr>
<tr>
<td>Son</td>
<td>GU</td>
<td>4.13</td>
<td>4.39</td>
<td>3.71</td>
<td>2.90</td>
</tr>
<tr>
<td></td>
<td>GC</td>
<td>4.87</td>
<td>4.75</td>
<td>3.86</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>RGBU</td>
<td>5.88</td>
<td>5.19</td>
<td>4.06</td>
<td>2.69</td>
</tr>
<tr>
<td></td>
<td>RGBC</td>
<td>5.54</td>
<td>5.11</td>
<td>4.06</td>
<td>2.80</td>
</tr>
<tr>
<td>Obs</td>
<td>GU</td>
<td>1.80</td>
<td>3.36</td>
<td>4.59</td>
<td>4.81</td>
</tr>
<tr>
<td></td>
<td>GC</td>
<td>2.70</td>
<td>4.19</td>
<td>5.14</td>
<td>5.01</td>
</tr>
<tr>
<td></td>
<td>RGBU</td>
<td>6.95</td>
<td>7.89</td>
<td>7.72</td>
<td>6.49</td>
</tr>
<tr>
<td></td>
<td>RGBC</td>
<td>4.78</td>
<td>6.35</td>
<td>5.74</td>
<td>5.77</td>
</tr>
<tr>
<td>SIl</td>
<td>GU</td>
<td>0.05</td>
<td>0.22</td>
<td>1.00</td>
<td>2.69</td>
</tr>
<tr>
<td></td>
<td>GC</td>
<td>0.13</td>
<td>0.53</td>
<td>1.64</td>
<td>3.52</td>
</tr>
<tr>
<td></td>
<td>RGBU</td>
<td>2.06</td>
<td>4.99</td>
<td>7.16</td>
<td>8.38</td>
</tr>
<tr>
<td></td>
<td>RGBC</td>
<td>0.62</td>
<td>1.91</td>
<td>4.10</td>
<td>6.04</td>
</tr>
<tr>
<td>Ovl</td>
<td>GU</td>
<td>3.14</td>
<td>3.86</td>
<td>3.96</td>
<td>3.67</td>
</tr>
<tr>
<td></td>
<td>GC</td>
<td>3.83</td>
<td>4.35</td>
<td>4.27</td>
<td>3.86</td>
</tr>
<tr>
<td></td>
<td>RGBU</td>
<td>6.30</td>
<td>6.59</td>
<td>6.07</td>
<td>5.02</td>
</tr>
<tr>
<td></td>
<td>RGBC</td>
<td>5.30</td>
<td>5.76</td>
<td>5.48</td>
<td>4.68</td>
</tr>
</tbody>
</table>

The ROVER algorithms outperformed their corresponding baselines at all global SNRs. c) Comparing RGBU vs RGBC, RGBU was the better estimator for obstruents, silence, and overall cases across both noise types. d) For the case of sonorants in LCR noise, RGBC performed better than RGBU in most global SNRs. For the case of sonorants in FLN noise, RGBU was the better performer.

4. CONCLUSIONS

A phoneme class selective speech enhancement algorithm was presented in this paper. This was done in two stages. In the first stage, short-time spectral magnitude generalized spectral subtraction β-unconstrained and β-constrained parametric estimators were derived using coefficients that minimize the weighted Euclidean distortion between the clean speech and estimate of clean speech spectral magnitudes. In the second stage, three enhanced utterances from these parametric estimators, each customized for a specific phoneme class, were generated. Using the mixture maximum model, phoneme classification was done using probabilistic decisions and these decisions were used as weights to combine the phoneme segments from the enhanced utterances to form one single composite utterance.

5. REFERENCES