Speech Enhancement Minimizing Generalized Euclidean Distortion Using Supergaussian Priors

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

We introduce short time spectral estimators which minimize the weighted Euclidean distortion (WED) between the clean and estimated speech spectral components when clean speech is degraded by additive noise. The traditional minimum mean square error (MMSE) estimator does not take into account sufficient perceptual measure during enhancement of noisy speech. However, the new estimators discussed in this paper provide greater flexibility to improve speech quality. We explore the cases when clean speech spectral magnitudes and discrete Fourier transform (DFT) coefficients are modeled by super-Gaussian priors like Chi and bilateral Gamma distributions respectively. We also present the joint maximum a posteriori (MAP) estimators of the Chi distributed spectral magnitudes and uniform phase. Performance evaluations over two noise types and three SNR levels demonstrate improved results of the proposed estimators.

Index Terms: MMSE estimation, joint MAP estimation, weighted Euclidean distortion, Chi prior, two-sided Gamma prior, super-Gaussian prior

1. Introduction

In speech enhancement, a widely accepted approach is split-ting a segment of noisy speech into smaller overlapping analysis frames (usually 30ms), applying DFT to convert the time domain samples to frequency domain, and then multiplying the noisy spectral components with gain estimates at each frequency bin. The multiplication yields clean speech spectral estimates which are reverted back to the time domain using inverse DFT. Therefore, the key lies in estimating the spectral gains as accurately as possible.

The Ephraim-Malah algorithm [1] adopts a Bayesian approach to minimize the mean square error between the clean and estimated speech spectral magnitudes making an assumption that clean speech and noise DFT coefficients are Gaussian distributed priors. Later, Martin [2] derived complex spectrum estimators from DFT coefficients that are better modeled by Gamma and Laplacian distributed priors. Lotter and Vary [3] obtained MAP estimates of clean speech spectral magnitudes observing that they could be fit into a generalized prior to represent a Gamma or Laplacian function by a simple change in parameter. In all these algorithms, the priors were changed to improve the accuracy of the gain estimate. However, there are other approaches where the Bayesian cost error function was penalized more heavily than those in spectral peaks. Penalizing the errors in the spectral valleys is more pronounced when \( \alpha > 0 \), and this effect is more pronounced when \( \beta > 0 \). Penalizing the errors in the spectral peaks happens when \( \beta < 0 \) but this effect is more pronounced when \( \alpha > 0 \).

In [4], the square of the log power spectral error was minimized. Hansen et al. [5] minimized the squared error between arbitrary powers of clean and estimated speech power spectral densities and You et al. [6] explored the case for spectral magnitudes. Natarajan et al. [7] incorporated an auditory masking threshold formulation to the estimator in [5]. Later, Loizou [8] obtained estimators for several perceptual cost functions. In this study, we derive estimators using super-Gaussian priors and WED as the cost function since they offer greater flexibility to penalize the errors in spectral valleys (or peaks) by simply tuning two exponent terms.

The rest of the paper is organized as follows. In Section 2, we derive the generalized parametric WED estimator based on Chi and Gamma priors, and the joint MAP using the Chi prior. In Section 3, we report the analysis of our experimental evaluations.

2. Statistical Model

2.1. MMSE Estimation

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

\[
Y_t e^{j\phi_{y,k}} = X_t e^{j\phi_{x,k}} + D_t e^{j\phi_{d,k}}.
\]

Here, \( Y_t, X_t, D_t \) represent the spectral magnitudes and \( \phi_{y,k}, \phi_{x,k}, \phi_{d,k} \) represent the phases of the short-time Fourier transforms (STFT) of noisy speech, clean speech and noise respectively at frequency bin \( k \). In the proposed solution, we modify the error function as:

\[
C_e(X^\alpha, \hat{X}^\alpha) = (X^\alpha - \hat{X}^\alpha)^T W (X^\alpha - \hat{X}^\alpha)
\]

where \( W = \text{diag}(X_1^{\beta}, X_2^{\beta}, ..., X_K^{\beta}) \), \( K \) is the length of the STFT of the analysis frame and hence the length of the vector \( \hat{X} \), \( \hat{X}_k \) is the estimate of clean speech spectral magnitude and \( (\alpha, \beta) \) are constant exponent terms. The error function in (2) can be treated as the generalized version of weighted Euclidean distortion in [8] where the term \( \alpha \) was set to 1. It should be noted that when \( \beta < 0 \), the errors in the spectral valleys are penalized more heavily than those in spectral peaks. Penalizing the spectral peaks happens when \( \beta > 0 \) but this effect is more pronounced when \( \alpha > 0 \).

Letting \( Y(\omega_k) = Y_t e^{j\phi_{y,k}} \), the MMSE estimator of the
spectral magnitude that minimizes the error in (2) is,

\[ X_k = \left\{ \frac{E[X_k^{\alpha+\beta} | Y(\omega_k)]}{E[X_k^\alpha | Y(\omega_k)]} \right\}^{\frac{1}{\beta}} = \left\{ \int_{-\infty}^{\infty} X_k^{\alpha+\beta} p(Y(\omega_k)|X_k) p(X_k) \, dX_k \right\}^{\frac{1}{\beta}} \]

after canceling out the common term \( p(Y(\omega_k)) \) in the denominator. The resulting estimator in (3) depends on the choice of the distribution of the prior \( p(X_k) \). In this paper, we investigate the performance using the Chi and two-sided Gamma priors. If the speech DFT coefficients are Gaussian distributed, the resulting speech spectral magnitude pdf \( p(X_k) \) is Rayleigh distributed [1]. Next, the Chi distribution is represented by [9]

\[ p(X_k) = \frac{2 X_k^{2a-1}}{\lambda_x(k) Y(a)} \exp \left( - \frac{X_k^2}{\lambda_x(k)} \right), \]

where the term \( 2a \) represents the number of degrees of freedom, \( \lambda_x(k) = E[X_k^2] \) is the sum of signal power associated with each \( X_k \), and \( G(\cdot) \) is the complete Gamma function. Finally, in [2], Martin modeled the speech DFT coefficients using the two-sided Gamma prior \( p(X_k) \) at frequency bin \( k \) as:

\[ p(X_R) = \frac{\sqrt{R}}{2 \sqrt{\pi} \lambda_x(k)} |X_R|^{-0.5} \exp \left( - \frac{\mu |X_R|}{\sqrt{\lambda_x(k)}} \right) \]

where \( \mu = \sqrt{2} \) and \( R \) denotes the real part of the DFT coefficient at frequency \( k \). The same pdf holds for the imaginary part. The corresponding complete complex spectrum estimator can be obtained by replacing \( X_k \) and \( Y(\omega_k) \) with \( X_R \) and \( Y_R \) respectively in (3). Note that in both speech priors discussed here, we always assume that the noise DFT coefficients are Gaussian distributed.

Now, we present the WED solutions of (3). Inserting the Chi prior pdf (4) in (3) and using [10, eqs.(6.631.1), (8.406.3)] results in the estimator,

\[ \hat{G}_k = \sqrt{\frac{R}{\gamma_k}} \left\{ \frac{\Gamma(\frac{\alpha+\beta}{2} + \frac{a}{2}) \phi(\frac{\alpha+\beta}{2} + \frac{a}{2}, 1; \nu_k)}{\Gamma(\frac{\alpha+\beta}{2}) \phi(\frac{\alpha+\beta}{2}, a; \lambda_x(k))} \right\}^{\frac{1}{2}} \]

where,

\[ \nu_k = \frac{\xi_k}{\xi_k - 1} \gamma_k, \quad \xi_k = \frac{\lambda_x(k)}{\lambda_x(k)}, \quad \lambda_x(k) = E[X_k^2], \quad \gamma_k = \frac{\nu_k}{\lambda_x(k)}. \]

and \( \phi(a, b; x) \) denotes the confluent hypergeometric function. It is assumed that the real and imaginary parts of the DFT coefficients have equal variance \( \frac{\lambda_x(k)}{2} \). The constraints in (6) are given by,

\[ a > 0, \quad \alpha \neq 0, \quad \beta > -2a, \quad \alpha + \beta > -2a. \]

This implies that both the parameters (\( \alpha, \beta \)) can take negative values which will be useful for achieving greater degrees of noise suppression. It may be noted that for \( \alpha = 1 \) and \( \alpha = 1 \), the Chi estimator in (6) becomes the estimator in [8, eq.(18)]. In [8], the constraint was \( \beta > 2 \). This constraint is relaxed in (6) if \( \alpha > 1 \) since \( \beta < -2 \) then becomes valid which helps in penalizing the spectral valleys. Since spectral valleys are regions of low SNR and noise is expected to be predominant in this region, the estimator in (6) achieves greater noise suppression. Also, if \( \alpha = 1, \beta = 0 \), and \( \alpha \) is replaced by \( 2\alpha \) in (6) it becomes the power spectrum based GMMSE estimator [5, eq.(19)]. Further, the MMSE estimator [1] is a special case of (6) when \( a = 1, \alpha = 1, \beta = 0 \).

Next, we turn our attention to the WED solution of two-sided Gamma prior. In general, it can be shown that for any power \( \alpha \),

\[ \int_{-\infty}^{\infty} X_R^\theta p(Y_R|X_R) p(X_R) \, dX_R = T_1(\alpha) T_2(\alpha) \]

where,

\[ T_1(\alpha) = c \left( \frac{\lambda_x(k)}{2} \right)^{\frac{\alpha+\beta}{2}} \exp \left( - \frac{Y_R^2}{\lambda_x(k)} \right) \Gamma(\alpha + \frac{1}{2}), \]

\[ T_2(\alpha) = \exp \left( \frac{G_{R^-}^2}{2} D_{-(\alpha+\frac{1}{2})} (\sqrt{\gamma} G_{R^-}) \right) + \left(-1\right)^\alpha \exp \left( \frac{G_{R^+}^2}{2} D_{-(\alpha+\frac{1}{2})} (\sqrt{\gamma} G_{R^+}) \right), \]

\[ c = \frac{\sqrt{\pi}}{2 \Gamma(\frac{\alpha+\beta}{2}) \Gamma(\alpha+\frac{1}{2})}, \quad G_{R^-}, G_{R^+} \text{ are as defined in [2, eq.(12)], and } D_{\gamma}(z) \text{ is the parabolic cylindrical function [10, eq.(9.241.2)].} \]

Substituting (8), (9) in (3) with appropriate power terms, we get,

\[ \hat{G}_k = \sqrt{\frac{1}{\gamma_{R_k}}} \left\{ \frac{\Gamma(\alpha + \beta + \frac{1}{2}) T_2(\alpha + \beta)}{\Gamma(\beta + \frac{1}{2}) T_2(\beta)} \right\}^{\frac{1}{2}} \]

where \( \gamma_{R_k} = \frac{\sqrt{\pi} G_{R^+}^2}{2 \Gamma(\alpha+\beta+1)} \) is the \textit{a posteriori} SNR of the real or imaginary part of the DFT coefficient. The constraints in (10) are given by,

\[ \alpha 
eq 0, \quad \beta > -0.5, \quad \alpha + \beta > -0.5, \quad (\alpha, \beta) \in Z^+. \]

The integer constraint comes from the fact that for \( T_2(\alpha) \) (and hence \( \hat{G}_k \)) to be real, the term \((-1)^\alpha \) in \( T_2(\alpha) \) has to be real which happens for all integer values of \( \alpha \). Therefore, possible values of \( \alpha = 1, 2, 3, \ldots \) and \( \beta = 0, 1, 2, \ldots \). Unfortunately, \( \beta \) cannot take negative values here unlike (6). The MMSE estimator in [2, eq.(13)] is a special case of (10) when \( \alpha = \beta = 0 \).

### 2.2. Joint MAP Estimation

In this section, the joint MAP estimate [11] of speech spectral magnitude \( X_k \) and phase \( \phi_{X,k} \) is derived for the case of Chi prior.

\[ (\hat{X}_k, \hat{\phi}_{X,k}) = \underset{X_k, \phi_{X,k}}{\arg \max} p(X_k, \phi_{X,k} | Y(\omega_k)) \]

\[ = \underset{X_k, \phi_{X,k}}{\arg \max} \frac{p(Y(\omega_k)|X_k, \phi_{X,k}) p(X_k, \phi_{X,k})}{p(Y(\omega_k))} \]

The denominator \( p(Y(\omega_k)) \) can be ignored since it is only a normalization term. For a rotational invariant pdf and assuming uniform distribution of phase in \([ -\pi, +\pi ] \),

\[ p(X_k, \phi_{X,k}) = \frac{1}{2\pi} p(X_k) \]

Since the natural logarithm function is monotonically increasing, the \( \ln(\cdot) \) of (12) could be maximized. This may be represented by \( L \) and is given by,

\[ L = \ln(p(Y(\omega_k)|X_k, \phi_{X,k}) p(X_k) p(\phi_{X,k})) = \frac{Y(\omega_k) - X(\omega_k))^2}{\lambda_x k} - \frac{X_k^2}{\lambda_x(k)} + (2a - 1) \ln(X_k) + \delta \]

(14)
where $\delta$ is a constant term. After differentiating $L$ with respect to $\phi X, k$ and setting the derivative to zero, we get the optimal phase estimate as the noisy phase, i.e., $\hat{\phi} X, k = \phi Y, k$. Similarly, differentiating $L$ with respect to $X_k$ and setting the derivative to zero yields the quadratic,

$$X_k^2 - \left( \frac{Y_k \xi_k}{1 + \xi_k} \right) X_k - \frac{(2a - 1) \lambda_3(k) \xi_k}{1 + \xi_k} = 0. \quad (15)$$

The root of the quadratic is $\hat{X}_k$ and letting $\hat{X}_k = \hat{G}_k Y_k$ we get,

$$\hat{G}_k = \frac{\xi_k + \sqrt{\xi_k^2 + 4 \left( a - \frac{1}{2} \right) (1 + \xi_k) \left( \frac{\xi_k}{\lambda_3} \right)}}{2(1 + \xi_k)}. \quad (16)$$

The constraints of (16) are given by,

$$a > 0, \quad a \geq 1 - \frac{\nu_k}{4}. \quad (17)$$

The second constraint should be met to get real values from the square root term in (16). It may be noted that the joint MAP estimate of the Chi prior magnitude in (16) differs from the joint MAP estimate of Wolfe and Godsill [11, eq.(29)] by a term $4 \left( a - \frac{1}{2} \right)$ inside the square root instead of 2. Setting $a = 1$ in (16), yields the estimate of [11].

3. Experimental Results

A set of 5 (3 female, 2 male) phonetically balanced test utterances (downsampled to 8kHz) from the TIMIT test corpus was used for objective quality evaluations. The corpus was 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 5dB. The quality of enhanced speech was assessed using objective speech quality measures such as the segmental SNR (SegSNR) and the DFT distortion introduced in [12, eq.(27)]. In all the experiments, the $a \text{ priori}$ SNR $\xi_k$ was evaluated using the decision-directed approach [1] using a value of 0.98 for the smoothing constant. The authors used the code in [13] for running the baseline MMSE enhancement routines.

In Fig.1, the SegSNR performance of the MMSE WED Chi estimator solution (6) is illustrated for FLN and LCR noise types using different values of $a, \alpha, \beta$. With $a, \alpha$ fixed and $\beta$ varying, performance improved using lower values of $\beta$. This is consistent across all noise levels and types. Note that for the family of curves representing $a = 1, \beta$ appears truncated since these cannot be extended to $\beta \leq -2.0$ due to the constraint in (7). For the same reason, the curve representing $a = 3, \alpha = -0.9$ cannot be extended to $\beta = -5.5$. Since, low values of $\beta$ exhibited better performance than high $\beta$ values, we can focus our attention to the regions of the curves representing low and valid $\beta$ values in subsequent discussions. Next, with $a, \beta$ fixed and $\alpha$ varying, SegSNR performance was found to be dependent on the SNR level. Lower values of $\alpha$ produced better SegSNR improvement at lower SNRs (-5dB) whereas at higher SNRs (+5dB) higher values of $\alpha$ are preferred. For example, in the region $\beta \in [-4.5, -5.5]$, $\alpha_{min} = -0.9$ had the best performance in FLN -5dB but it gradually degraded with 0dB and 5dB. In the 5dB subplot, $\alpha_{max} = 2$ is the best choice. A similar behavior was observed in the case of LCR noise. Additionally, it becomes clear that negative values are good choices of both $\alpha, \beta$ at low SNRs. Finally, $\alpha, \beta$ fixed and $\alpha$ varying, the nett effect observed was a shift in the SegSNR curves. That is, the constraint (7) on $\beta$ values is imposed much earlier for lower values of Chi parameter $\alpha$ than with those achieved using higher $\alpha$ values.

Next we explore the performance of the WED solution (10) of the two-sided Gamma prior. Since (10) is optimized based on the DFT coefficients, we evaluate its performance using $D_{DFT}$ as the root mean square of the DFT distortion [12] normalized to 100. The plot in Fig. 2 depicts the DFT distortion performance with fixed $\beta = 0$ and $a = 1, 2, ..., 5$. In both noise types, the estimator with $a = 1$ was the best performer. This is the special case estimator derived in [2]. Intuitively, we would expect lowering $a$ or $\beta$ to achieve better performance than the special case estimator. However, the drawback here is that $a \geq 1, \beta \geq 0$ following the constraint mentioned in (11). The second best estimator for both noise types and at all SNR levels was achieved with $a = 1, \beta = 2$ (not given in the plot) with the DFT distortion lower by about 30-40 counts than the $a = 2, \beta = 0$ estimator in the plot. We also evaluated the performance of the estimators resulting from the remaining higher values of $\alpha, \beta$ (not shown in the plot) and the second best estimator $\alpha = 1, \beta = 2$. Finally, we confirmed that the trend in performance of (10) using SegSNR is similar to that of $D_{DFT}$.

The SegSNR performance of the joint MAP Chi estimator (16) is evaluated for different values of the Chi parameter $\alpha$. The performance is fairly even from $a = 0.001 - 0.1$ and degrades rapidly for $a > 1$. The best performance was observed in the range $a = 0.25 - 0.5$ for most cases with the exception of LCR -5dB whose best case was at $a = 0.07$. The improvement over the special case estimator of [11] is in the range of 1 to 2dB.
A summary of the objective evaluations across the three proposed algorithms and baseline algorithms in FLN noise is tabulated in Table 1. The parameters used in each of the algorithm is indicated by \( \{a, \alpha, \beta\} \) in the leftmost column of the table. A “-” indicates that the corresponding parameter is not applicable and a “+” indicates that the parameter was varied, and not automatically adapted, across SNR levels to achieve the scores tabulated for that algorithm. Hence, the baseline methods have the limitation of fixed parameters. For the case of FLN noise, best SegSNR performance was obtained by the proposed joint MAP Chi estimator with \( \alpha = 0.25 - 0.5 \). In MMSE WED Chi estimator, we fixed the Chi parameter \( \alpha \) at 3 and used the preferred range of parameters that was discussed earlier with reference to Fig. 1. In general, higher values of \( (\alpha, \beta) \) were used as we move across lower to higher SNR levels. In the second part of Table 1, a comparison of the best DFT estimator is made against noisy speech. The best proposed DFT estimator is also the special case estimator \([2]\). In Table 2 for LCR noise, a similar behavior in the performance was observed with the joint MAP Chi estimator yielding the best results. The best Chi parameter at -5dB was 0.07 and for higher SNRs it was in the range 0.25-0.5.

### 4. Conclusion

Bayesian short time spectral magnitude and DFT coefficient estimators were presented using super-Gaussian priors. The first Table 1: Segmental SNR and DFT distortion evaluations of the baseline and proposed estimators in FLN noise.

<table>
<thead>
<tr>
<th>Global SNR (FLN noise)</th>
<th>Baseline MMSE(1, 1, 0)[1]</th>
<th>Baseline Joint MAP(1, -, -)[11]</th>
<th>MMSE WED Chi(3, *, *)</th>
<th>Joint MAP Chi(*, -, -)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td>-5.95</td>
<td>-1.19</td>
<td>0.33</td>
<td>0.38</td>
</tr>
<tr>
<td>Baseline</td>
<td>-3.71</td>
<td>1.47</td>
<td>1.84</td>
<td>2.42</td>
</tr>
<tr>
<td>0dB</td>
<td>-0.65</td>
<td>3.98</td>
<td>4.55</td>
<td>5.03</td>
</tr>
<tr>
<td>5dB</td>
<td>3.84</td>
<td>3.84</td>
<td>3.84</td>
<td>3.84</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MMSE DFT Gamma(-, 1, 0)[2]</th>
<th>304.17</th>
<th>361.7</th>
<th>265.22</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFT</td>
<td>35.15</td>
<td>43.10</td>
<td>47.21</td>
</tr>
</tbody>
</table>

5. References


