

SPEECH ENHANCEMENT USING A MMSE SHORT TIME SPECTRAL AMPLITUDE ESTIMATOR WITH LAPLACIAN SPEECH MODELING

*Bin Chen and Philipos C. Loizou**

Dept. of Electrical Engineering
University of Texas-Dallas, Richardson, TX 75083

ABSTRACT

This paper focuses on optimal estimators of the magnitude spectrum for speech enhancement. We present an analytical solution for estimating in the MMSE sense the magnitude spectrum when the clean speech DFT coefficients are modeled by a Laplacian distribution and the noise DFT coefficients are modeled by a Gaussian distribution. Furthermore, we derive the MMSE estimator under speech presence uncertainty and a Laplacian model. Results indicated that the Laplacian based MMSE estimator yielded less residual noise in the enhanced speech than the traditional Gaussian-based MMSE estimator.

1. INTRODUCTION

Single-channel speech enhancement algorithms based on minimum mean-square error (MMSE) estimation of the short-time spectral magnitude have received a lot of attention in the past two decades [1]. A key assumption made in the MMSE algorithms is that the real and imaginary parts of the clean DFT coefficients can be modeled by a Gaussian distribution. This Gaussian assumption, however, holds asymptotically for long duration analysis frames, in which the span of the correlation of the signal is much shorter than the DFT size. While this assumption might hold for the noise DFT coefficients, it does not hold for the speech DFT coefficients, which are typically estimated using relatively short (20-30 ms) duration windows. For that reason, several [2-5] have proposed the use of non-Gaussian distributions for modeling the real and imaginary parts of the speech DFT coefficients. In particular, the Gamma and Laplacian probability distributions can be used to model the distributions of the real and imaginary parts of the DFT coefficients.

The use of Gamma or Laplacian distributions, however, complicates the derivation of the MMSE estimate of the magnitude spectrum. This is partly because there is no analytical expression for the pdf of the magnitude of the DFT coefficients when the real and imaginary parts of the DFT

coefficients are modeled by a Laplacian (or Gamma) distribution. For that reason, alternative solutions were explored in [2-5]. For instance, in [3] the authors approximated the pdf of the magnitude of the DFT coefficients with a parametric function, and used that to derive a MAP estimator of the magnitude spectrum. In [2], they derived separately the estimators of the real and imaginary parts of the DFT coefficients assuming Gamma and Laplacian distributions for the speech DFT coefficients. The two estimators combined yielded an estimator for the signal DFT coefficients that was complex valued.

In this paper, we derive a closed-form expression for the pdf of the magnitude of the DFT coefficients, and use that to derive the MMSE estimator of the speech magnitude spectrum based on a Laplacian model for the speech DFT coefficients and a Gaussian model for the noise DFT coefficients. To further improve the amplitude estimation, we also incorporate speech presence uncertainty into the Laplacian based estimator.

The paper is organized as follows. In section II, we derive the Laplacian based MMSE estimator and in section III we derive the MMSE estimator under signal presence uncertainty. In Section IV, we evaluate the performance of the proposed estimators, and in Section V we give the summary and conclusions.

2. LAPLACIAN BASED SHORT-TIME SPECTRAL AMPLITUDE ESTIMATOR

2.1. Derivation of Amplitude Estimator

Let $y(n) = x(n) + d(n)$ be the sampled noisy speech signal consisting of the clean signal $x(n)$ and the noise signal $d(n)$. Taking the short-time Fourier transform of $y(n)$, we get:

$$Y(\omega_k) = X(\omega_k) + D(\omega_k) \quad (1)$$

for $\omega_k = 2\pi k/N$ where $k = 0, 1, 2, \dots, N-1$, and N is the frame length. The above equation can also be expected in polar form as

$$Y_k e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)} \quad (2)$$

*This work was partly supported by NIH/NIDCD. Email: loizou@utdallas.edu

According to [1], the MMSE estimator of X_k is obtained as follows:

$$\begin{aligned}\hat{X}_k &= E\{X_k|Y(\omega_k)\}, \quad k = 0, 1, 2, \dots, N-1 \quad (3) \\ &= \frac{\int_0^\infty \int_0^{2\pi} x_k p(Y(\omega_k)|x_k, \theta_{x_k}) p(x_k, \theta_{x_k}) d\theta_{x_k} dx_k}{\int_0^\infty \int_0^{2\pi} p(Y(\omega_k)|x_k, \theta_{x_k}) p(x_k, \theta_{x_k}) d\theta_{x_k} dx_k}\end{aligned}$$

where $E[\cdot]$ denotes the expectation operator. Making the assumption that the spectral magnitudes and phases are independent and that the phases are uniformly distributed¹, we have $p(x_k, \theta_{x_k}) = \frac{1}{2\pi} p(x_k)$, where $p(x_k)$ is the density of the spectral magnitudes. Under the assumption that the real and imaginary parts of $X(\omega_k)$ are modeled by a Laplacian distribution, the density $p(x_k)$ of the spectral magnitudes is given by [6]:

$$\begin{aligned}p(x_k) &= \frac{2x_k}{\lambda_x(k)} \left[\frac{\pi}{4} I_0 \left(-\frac{\sqrt{2}}{\sqrt{\lambda_x(k)}} x_k \right) \right. \\ &\quad \left. + 2 \sum_{n=1}^{\infty} \frac{1}{n} I_n \left(-\frac{\sqrt{2}}{\sqrt{\lambda_x(k)}} x_k \right) \sin \frac{\pi n}{4} \right] \quad (4)\end{aligned}$$

where $I_n(\cdot)$ denote the modified Bessel function of n th order and $\lambda_x(k) = E\{X_k^2\}$ is the signal variance. Using a complex Gaussian distribution for $D(\omega_k)$, it is easy to show that [1]:

$$\begin{aligned}&p(Y(\omega_k)|x_k, \theta_{x_k}) \quad (5) \\ &= \frac{1}{\pi \lambda_d(k)} \exp \left[-\frac{Y_k^2 - 2x_k \operatorname{Re}\{e^{-j\theta_{x_k}} Y(\omega_k)\} + x_k^2}{\lambda_d(k)} \right]\end{aligned}$$

where $\lambda_d(k) = E\{D_k^2\}$ is the noise variance. Finally, after substituting Eq.(4) and Eq.(5) into Eq.(3) and using [7, Eq. 6.633.1] we get:

$$\hat{X}_k = \frac{A_k + B_k}{C_k + D_k} \quad (6)$$

where

$$\begin{aligned}A_k &= \frac{\left(\frac{Y_k^2}{\gamma_k}\right)^{\frac{3}{2}}}{2} \sum_{m=0}^{\infty} \frac{\Gamma(m + \frac{3}{2})}{m! \Gamma(m+1)} \left(\frac{\gamma_k}{2\xi_k^2 Y_k^2}\right)^m \\ &\quad \cdot F(-m, -m; 1; 2\xi_k^2 Y_k^2) \\ B_k &= \frac{8}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin \frac{\pi n}{4} \frac{\pi n \left(\frac{2\gamma_k}{Y_k}\right)^n \left(\frac{\gamma_k}{Y_k^2}\right)^{-\frac{n+3}{2}}}{2^{n+1} \Gamma(n+1)} \\ &\quad \cdot \sum_{m=0}^{\infty} \frac{\Gamma(m + \frac{1}{2}n + \frac{3}{2})}{m! \Gamma(m+1)} \left(\frac{\gamma_k}{2\xi_k^2 Y_k^2}\right)^m \\ &\quad \cdot F(-m, -m; n+1; 2\xi_k^2 Y_k^2)\end{aligned}$$

¹Note that these assumptions are true for the Gaussian model [1] but not for the Laplacian model. Nevertheless, for simplicity purposes we used the same assumptions.

$$\begin{aligned}C_k &= \frac{Y_k^2}{2\gamma_k} \sum_{m=0}^{\infty} \frac{1}{m!} \left(\frac{\gamma_k}{2\xi_k^2 Y_k^2}\right)^m \\ &\quad \cdot F(-m, -m; 1; 2\xi_k^2 Y_k^2)\end{aligned}$$

$$\begin{aligned}D_k &= \frac{8}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin \frac{\pi n}{4} \frac{\pi n \left(\frac{2\gamma_k}{Y_k}\right)^n \left(\frac{\gamma_k}{Y_k^2}\right)^{-\frac{n}{2}-1}}{2^{n+1} \Gamma(n+1)} \\ &\quad \cdot \sum_{m=0}^{\infty} \frac{\Gamma(m + \frac{1}{2}n + 1)}{m! \Gamma(m+1)} \left(\frac{\gamma_k}{2\xi_k^2 Y_k^2}\right)^m \\ &\quad \cdot F(-m, -m; n+1; 2\xi_k^2 Y_k^2)\end{aligned}$$

where $\xi_k = \lambda_x(k)/\lambda_d(k)$ and $\gamma_k = Y_k^2/\lambda_d(k)$ are the *a priori* and *a posteriori* signal-to-noise (SNR) ratios respectively, $\Gamma(\cdot)$ is the gamma function and $F(a, b, c; x)$ is the Gaussian hypergeometric function [7, Eq. 9.100]. Equation (6) gives the MMSE estimator of the spectral magnitudes based on the assumption that the real and imaginary parts of $X(\omega_k)$ are modeled by a Laplacian distribution. We will be referring to this estimator as the LapMMSE estimator.

3. DERIVATION OF AMPLITUDE ESTIMATOR UNDER SPEECH PRESENCE UNCERTAINTY

In this section we derive the MMSE magnitude estimator under the assumed Laplacian model and uncertainty of speech presence. This is motivated by the fact that speech might not be present at all times and at all frequencies. We therefore consider a two-state model for speech events, i.e., that either speech is present at a particular frequency bin (hypothesis H_1) or that is not (hypothesis H_0). Intuitively, this amounts to multiplying the estimator by a term that provides an estimate of the probability that speech is present at a particular frequency bin. Following [1], this new estimator is given by:

$$\hat{X}_k = E(X_k|Y(\omega_k), H_1^k) P(H_1^k|Y(\omega_k)) \quad (7)$$

where H_1^k denotes the hypothesis that speech is present in frequency bin k , and $P(H_1^k|Y(\omega_k))$ denotes the conditional probability that speech is present in frequency bin k given the noisy speech (complex) spectrum $Y(\omega_k)$. The conditional probability $P(H_1^k|Y(\omega_k))$ can be computed using Bayes' rule [1]:

$$P(H_1^k|Y(\omega_k)) = \frac{\Lambda(Y(\omega_k), q_k)}{1 + \Lambda(Y(\omega_k), q_k)} \quad (8)$$

where $\Lambda(Y(\omega_k), q_k)$ is the generalized likelihood ratio defined by:

$$\Lambda(Y(\omega_k), q_k) = \frac{1 - q_k p(Y(\omega_k)|H_1^k)}{q_k p(Y(\omega_k)|H_0^k)} \quad (9)$$

where $q_k = P(H_0^k)$ denotes the *a priori* probability of speech absence for frequency bin k .

Under hypothesis H_0 , $Y(\omega_k) = D(\omega_k)$, and assuming that the noise DFT coefficients are modeled by a Gaussian distribution with zero mean and variance $\lambda_d(k)$, it follows that $p(Y(\omega_k)|H_0^k)$ will also have a Gaussian distribution with the same variance, i.e.,

$$p(Y(\omega_k)|H_0^k) = \frac{1}{\pi \lambda_d(k)} \exp\left(-\frac{Y_k^2}{\lambda_d(k)}\right) \quad (10)$$

Under hypothesis H_1 , $Y(\omega_k) = X(\omega_k) + D(\omega_k)$, and $p(Y(\omega_k)|H_1^k)$ will have the form [6]:

$$p_{Y(\omega_k)}(y) = p(z_r, z_i) = p_{Z_r(k)}(z_r) p_{Z_i(k)}(z_i) \quad (11)$$

where $Z_r(k) = \text{Re}\{Y(\omega_k)\}$, $Z_i(k) = \text{Im}\{Y(\omega_k)\}$, and $p_{Z_r(k)}(z_r)$ and $p_{Z_i(k)}(z_i)$ are given by [6]:

$$p_{Z_r(k)}(z_r) = \frac{\sqrt{\gamma_k} \exp\left(\frac{1}{2\xi_k}\right)}{2\sqrt{2\xi_k Y_k}} \left[\exp\left(-\frac{\sqrt{\gamma_k} z_r}{Y_k \sqrt{\xi_k}}\right) + \exp\left(\frac{\sqrt{\gamma_k} z_r}{\sqrt{\xi_k} Y_k}\right) + \exp\left(-\frac{\sqrt{\gamma_k} z_r}{\sqrt{\xi_k} Y_k}\right) \text{erf}\left(\frac{\sqrt{\gamma_k} z_r}{\sqrt{\xi_k} Y_k} - \frac{1}{\sqrt{\xi_k}}\right) - \exp\left(\frac{\sqrt{\gamma_k} z_r}{\sqrt{\xi_k} Y_k}\right) \text{erf}\left(\frac{\sqrt{\gamma_k} z_r}{\sqrt{\xi_k} Y_k} + \frac{1}{\sqrt{\xi_k}}\right) \right] \quad (12)$$

where $\text{erf}(\cdot)$ is the error function. The $p_{Z_r(k)}(z_r)$ density was obtained by performing the convolution of the speech Laplacian density with the noise Gaussian density. After substituting Eq.(8), (10) and (11) into Eq.(7) we get the final estimator (given by Eq.7) that incorporates speech-presence uncertainty.

4. IMPLEMENTATION AND PERFORMANCE EVALUATION

4.1. Implementation

As shown in Eq. (6), the derived LapMMSE estimator was highly nonlinear as it involved infinite summations. We initially truncated the infinite summations to a large number of terms, however, simulations indicated that such an approximation led to numerical instability issues. For that reason, we chose to use numerical integration techniques [8] to evaluate the integrals in (3).

The proposed estimator was applied to 20-ms duration frames of speech using a Hamming window, with 50% overlap between frames. The “decision-directed” approach [1] was used in the proposed estimators to compute the *a priori* SNR ξ_k , with $\alpha = 0.98$. The enhanced signal was combined using the overlap and add approach. The *a priori* probability of speech absence, q_k , was set to $q_k = 0.3$ in (9).

4.2. Performance Evaluation

Twenty sentences from the TIMIT database were used for the objective evaluation of the proposed LapMMSE estimator, 10 produced by female speakers and 10 produced by male speakers. The TIMIT sentences were downsampled to 8 kHz. Speech-shaped noise constructed from the long-term spectrum of the TIMIT sentences as well as F-16 cockpit noise were added to the clean speech files at 0, 5 and 10 dB SNR. An estimate of the noise spectrum was obtained from the initial 100-ms segment of each sentence. The noise spectrum estimate was not updated in subsequent frames.

Objective measures were used to evaluate the performance of the proposed estimators implemented with and without speech presence uncertainty (SPU) and denoted as LapMMSE-SPU and LapMMSE respectively. For comparative purposes we evaluated the performance of the traditional MMSE estimator [1] with and without incorporating speech presence uncertainty indicated as MMSE-SPU and MMSE respectively. Table 1 lists the (absolute) segmental SNR values averaged across the 20 sentences tested. As can be seen, higher segmental SNR values were obtained consistently by the proposed LapMMSE estimators. Particularly large improvements were noted at higher SNR levels, 5 and 10 dB.

Informal listening tests indicated that speech enhanced by the LapMMSE estimators had less residual noise. This was confirmed by visual inspection of spectrograms of the enhanced speech signals. Figure 1 shows the spectrograms of the TIMIT sentence “The kid has no manners, boys” enhanced by the LapMMSE-SPU and MMSE-SPU estimators. The sentence was originally embedded in +5 dB S/N speech shaped noise. Clearly, the sentence enhanced by the LapMMSE-SPU estimator had less residual noise with no compromise in speech distortion.

Estimator	Speech-Shaped			F-16 noise		
	0dB	5dB	10dB	0dB	5dB	10dB
MMSE	0.763	1.96	2.979	1.414	2.285	3.048
LapMMSE	1.149	4.647	7.182	1.819	5.122	7.528
MMSE-SPU	0.859	2.027	3.067	1.495	2.362	3.125
LapMMSE-SPU	1.867	5.113	7.792	2.65	5.657	8.198

Table 1. Comparative performance, in terms of segmental SNR, of the Gaussian-based MMSE and Laplacian-based MMSE estimators.

5. SUMMARY AND CONCLUSIONS

An MMSE estimator was derived for the speech magnitude spectrum based on a Laplacian model for the speech DFT

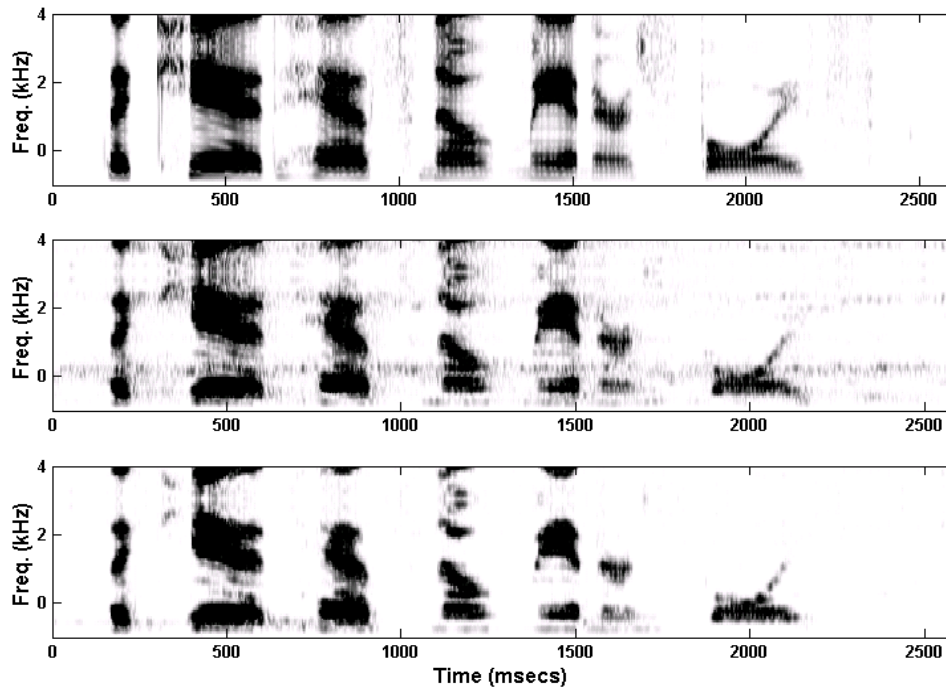


Fig. 1 Spectrogram of a TIMIT sentence in quiet (upper panel), speech enhanced by the Gaussian-based MMSE estimator (middle panel) and by the Laplacian-based MMSE estimator (bottom panel).

coefficients and a Gaussian model for the noise DFT coefficients. An estimator was also derived under speech presence uncertainty and a Laplacian model assumption. Results, in terms of objective measures, indicated that the proposed MMSE estimator yielded better performance than the traditional MMSE estimator based on a Gaussian model [1].

6. ACKNOWLEDGMENTS

The authors would like to thank Prof. Ali Hooshyar for all his help with the numerical integration techniques.

7. REFERENCES

- [1] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol.32, pp. 1109-1121, Dec.1984
- [2] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Priors," *IEEE Proc. Intern. Conf. on Acoustics, Speech and Signal Processing*, pp. 504-512, May 2002.
- [3] T. Lotter and P. Vary, "Noise Reduction by Maximum a Posteriori Spectral Amplitude Estimation With Super-gaussian Speech Modeling," *Intern. Workshop. Acoust. Echo Noise Control*, Kyoto, Japan, September 2003.
- [4] C. Breithaupt and R. Martin, "MMSE Estimation of Magnitude-Squared DFT Coefficients with SuperGaussian Priors," *IEEE Proc. Intern. Conf. on Acoustics, Speech and Signal Processing*, vol. I, pp. 896-899, April 2003.
- [5] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," *IEEE Proc. Intern. Conf. on Acoustics, Speech and Signal Processing*, pp. 18A.2.1-18A.2.4, 1984.
- [6] B. Chen and P. Loizou, "Speech Enhancement Using a MMSE Short Time Spectral Amplitude Estimator with Laplacian Speech Modeling," submitted to *IEEE Trans. Speech, Audio Proc.*
- [7] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, 6th ed., 2000.
- [8] Y. Kwon and H. Bang, *The Finite Element Method Using Matlab*, 2nd ed, CRC Press: New York, p. 35, 2000.