SPEECH ENHANCEMENT USING MMSE SHORT TIME SPECTRAL ESTIMATION WITH GAMMA DISTRIBUTED SPEECH PRIORS

Rainer Martin

Institute of Communication Systems and Data Processing
Aachen University of Technology, D-52056 Aachen, Germany
Phone: +49 241 802 6984, Fax: +49 241 802 2186, E-mail: martin@ind.rwth-aachen.de

ABSTRACT

In this paper we consider optimal estimators for speech enhancement in the Discrete Fourier Transform (DFT) domain. We present an analytical solution for estimating complex DFT coefficients in the MMSE sense when the clean speech DFT coefficients are Gamma distributed and the DFT coefficients of the noise are Gaussian or Laplace distributed. Compared to the state-of-the-art Wiener or MMSE short time amplitude estimators the new estimators deliver improved signal-to-noise ratios. When the noise model is a Laplacian density the enhanced speech shows less annoying random fluctuations in the residual noise than for a Gaussian density.

1. INTRODUCTION

Almost all of the known speech enhancement algorithms which operate in the Discrete Fourier Transform (DFT) domain [1, 2, 3] assume that the real and imaginary part of the clean speech DFT coefficients can be modelled by a Gaussian distribution. The Gaussian assumption is indeed true in the asymptotic case of large DFT frames when the span of correlation of the signal under consideration is much shorter than the DFT frame size [4]. For speech signals and the typical DFT frame sizes used in mobile communications this assumption is not well fulfilled. This has been recognized e.g. by Porter and Boll [5], who proposed a heuristic method to construct approximately optimal estimators from given clean speech material. As it will be shown below, the DFT coefficients of clean speech might be well modelled by a Gamma distribution. In this paper we will therefore present analytical solutions to the MMSE estimation of complex DFT coefficients with Gamma distributed speech priors. The probability density (pdf) of the DFT coefficients of the noise might be either a (complex) Gaussian or a Laplacian density.

It is well known that the MMSE estimator is a linear estimator when both the speech and the noise coefficients are complex Gaussians. Another consequence of the Gaussian model is that the spectral coefficients of the filter are real valued. The estimated clean speech coefficients then have the same short time phase as the noisy coefficients and only the magnitude needs to be manipulated. The estimator is therefore spherically invariant and easily implemented. For Gamma or Laplace distributions the MMSE estimator is highly nonlinear and complex valued. It's application, however, leads to an improved SNR in the enhanced speech.

The remainder of this paper is organized as follows: In Section 2 we briefly review the distribution of DFT coefficients of speech and noise. The results presented there will support our assumption that speech coefficients might be better modeled by Laplacian or Gamma probability densities. Section 3 presents the new MMSE estimators for our models of speech and noise. Finally, in Section 4 we will discuss experimental results.

2. STATISTICAL MODELS IN THE DFT DOMAIN

In what follows we consider a bandlimited, sampled noisy speech signal $y(i)$ which is the sum of a clean speech signal $s(i)$ and a disturbing noise $n(i)$, $y(i) = s(i) + n(i)$. i denotes the sampling time index. We further assume that $s(i)$ and $n(i)$ are statistically independent and zero mean. The noisy signal $y(i)$ is transformed into the frequency domain by applying a window $h(i)$ to a frame of L consecutive samples of $y(i)$ and by computing the DFT of size L on the windowed data. Before the next DFT computation the window is shifted by R samples. This sliding window DFT analysis results in a set of frequency domain signals which can be written as

$$Y(\lambda,k) = S(\lambda,k) + N(\lambda,k) = \sum_{\mu=0}^{L-1} y(\lambda R+\mu)h(\mu)e^{-j2\pi k\mu/L} \tag{1}$$

where λ is the subsampled time index, $\lambda \in \mathbb{Z}$, and k is the frequency bin index, $k \in \{0, 1, ..., L-1\}$, which is related to the normalized center frequency Ω_k by $\Omega_k = 2\pi k/L$. Furthermore, to facilitate our notation and to avoid additional normalization factors we assume $\sum_{\mu=0}^{L-1} h^2(\mu) = 1$. In a mobile communications application, we typically use a sampling rate of $f_s = 8000$ Hz and $L = 2R = 256$.

2.1. Statistical Models

It is well known that the pdf of speech samples in the time domain is much better modelled by a Laplacian or a Gamma density rather than a Gaussian density [6]. We here suggest that also in the short term DFT domain (frame size < 100 ms) the Laplace and Gamma densities are much better models for the pdf of the real and imaginary parts of speech coefficients than the commonly used Gaussian density. In this section we will briefly review these densities and provide examples of experimental data.

Let $S_R = \Re\{S(\lambda,k)\}$ and $S_I = \Im\{S(\lambda,k)\}$ denote the real and the imaginary part of a clean speech DFT coefficient, respectively. To enhance the readability of the following results we will drop both the frame index λ and the frequency index k and consider an individual speech DFT coefficient $S = S_R + jS_I$ at

a given time instant. Then the Gaussian, the Laplacian, and the Gamma prior densities (real and imaginary parts) can be defined as follows. $\sigma_s^2/2$ denotes the variance of the real and imaginary parts of the DFT coefficients. Similar probability densities can be defined for the DFT coefficients of the noise.

2.1.1. Gaussian speech model

$$p(S_R) = \frac{1}{\sqrt{\pi}\sigma_s} \exp(-\frac{S_R^2}{\sigma_s^2}) \qquad p(S_I) = \frac{1}{\sqrt{\pi}\sigma_s} \exp(-\frac{S_I^2}{\sigma_s^2}) \tag{2}$$

2.1.2. Laplacian speech model

$$p(S_R) = \frac{1}{\sigma_s} \exp(-\frac{2|S_R|}{\sigma_s}) \qquad p(S_I) = \frac{1}{\sigma_s} \exp(-\frac{2|S_I|}{\sigma_s}) \tag{3}$$

2.1.3. Two-sided Gamma speech model

$$\begin{aligned} p(S_R) &= \frac{\sqrt[4]{3}}{2\sqrt{\pi\sigma_s}\sqrt[4]{2}} |S_R|^{-\frac{1}{2}} \exp(-\frac{\sqrt{3}|S_R|}{\sqrt{2}\sigma_s}) \\ p(S_I) &= \frac{\sqrt[4]{3}}{2\sqrt{\pi\sigma_s}\sqrt[4]{2}} |S_I|^{-\frac{1}{2}} \exp(-\frac{\sqrt{3}|S_I|}{\sqrt{2}\sigma_s}) \end{aligned} \tag{4}$$

The Gamma density diverges when the argument approaches zero but, as shall be seen, provides otherwise an excellent fit to the observed data.

2.2. Experimental Data

Figure 1 plots the histogram of the real part of the DFT coefficients of high SNR ("clean") speech averaged over three male and three female speakers. Since speech can be considered to be a non-stationary random process only quasi-stationary coefficients were selected for the histogram. The coefficients represented in the histogram are taken from a narrow predefined SNR interval. For the depicted histograms only those coefficients are selected which have an SNR larger than 28 dB but smaller than 30 dB. The rationale behind this approach is that modern speech enhancement algorithms base the computation of the enhanced spectral coefficients on the (estimated) *a priori* SNR. [2, 3]. The statistics of the DFT coefficients and hence the resulting optimal estimators must be therefore characterized with respect to the (measured) *a priori* SNR. The full histogram in the top graph of Fig. 1 as well as the enlarged section in the bottom graph of Fig. 1 show that indeed the Laplace (dashed) and Gamma (solid) densities provide a much better fit to the experimental DFT data than the Gaussian (dotted) distribution. Other SNR intervals as the one stated above were tested as well with similar results.

Another assumption which is frequently invoked in the development of spectral estimators is that the real and the imaginary part of the complex DFT coefficients are statistically independent [4]. To verify this assumption we evaluated scatter plots of the real and imaginary parts of clean speech DFT coefficients. From these plots (not shown here) we conclude that the real and imaginary part are only weakly dependent. Note, that only for the Gaussian density the components can be strictly independent in both cartesian and polar coordinates.

To conclude this Section we note that also real noise signals are not necessarily complex Gaussians in the DFT domain. For a car noise recorded at a constant speed of 90 km/h we found that also the Laplacian density could provide a reasonable fit to the experimental data.

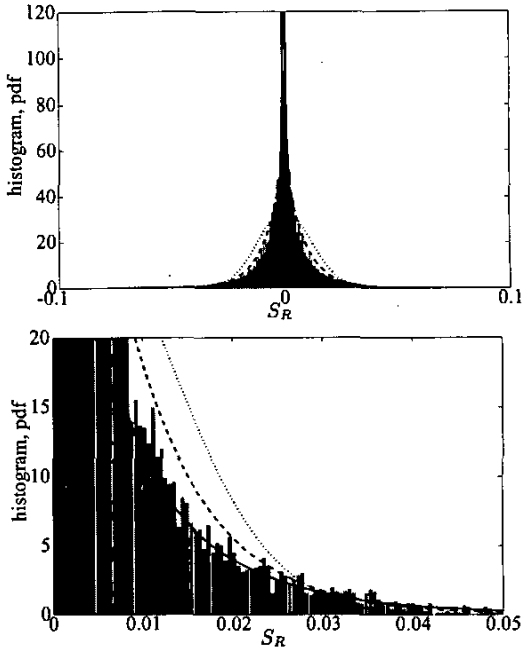

Fig. 1. Gaussian (dotted), Laplace (dashed), and Gamma (solid) density fitted to clean speech DFT coefficients. The lower graph shows an enlarged section of the top graph.

3. MMSE ESTIMATORS

If we assume independence of the real and the imaginary parts of DFT coefficients the MMSE estimator for the complex DFT coefficients can be split in the estimators for the real and the imaginary parts which can be treated independently

$$E\{S \mid Y\} = E\{S_R \mid Y_R\} + jE\{S_I \mid Y_I\}\,. \tag{5}$$

Again we have dropped the time and frequency indices. Based on the above prior models we will now develop MMSE estimators for the clean speech coefficients.

3.1. Gaussian Noise and Gaussian Speech Model

It is well known, that when both the noise and the speech coefficient pdf is a complex Gaussian the optimal estimator is linear (Wiener filter), i.e.,

$$\widehat{S} = E\{S \mid Y\} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_n^2} Y = \frac{\xi}{1+\xi} Y\,, \tag{6}$$

where σ_s^2 and σ_n^2 are the mean of $|S|^2$ and $|N|^2$, respectively. $\xi = \sigma_s^2/\sigma_n^2$ denotes the *a priori* SNR.

3.2. Gaussian Noise and Gamma Speech Model

We now derive the MMSE estimator for the complex DFT coefficients of clean speech when the speech prior is Gamma distributed and the noise is modeled by a Gaussian pdf.

For ease of notation we define

$$G_{R+} = \frac{\sqrt{3}\sigma_n}{2\sqrt{2}\sigma_s} + \frac{Y_R}{\sigma_n} = \frac{\sqrt{3}}{2\sqrt{2}\sqrt{\xi}} + \frac{Y_R}{\sigma_n} \tag{7}$$

$$G_{R-} = \frac{\sqrt{3}\sigma_n}{2\sqrt{2}\sigma_s} - \frac{Y_R}{\sigma_n} = \frac{\sqrt{3}}{2\sqrt{2}\sqrt{\xi}} - \frac{Y_R}{\sigma_n}\,. \tag{8}$$

The optimal estimator for the real part of S_R in the MMSE sense is then given by the conditional expectation

$$E\{S_R \mid Y_R\} = \frac{\sqrt[4]{1.5}}{2\pi\sigma_n\sqrt{\sigma_s}p(Y_R)} \int_{-\infty}^{\infty} S_R|S_R|^{-0.5} \cdot \exp\left(-\frac{Y_R^2}{\sigma_n^2} + \frac{2Y_RS_R}{\sigma_n^2} - \frac{S_R^2}{\sigma_n^2} - \frac{\sqrt{3}|S_R|}{\sqrt{2}\sigma_s}\right) dS_R \tag{9}$$

$$= \frac{\sigma_n}{2\sqrt{2}Z_{GR}} \{ \exp(G_{R-}^2/2) D_{-1.5}(\sqrt{2}G_{R-}) - \exp(G_{R+}^2/2) D_{-1.5}(\sqrt{2}G_{R+}) \} \tag{10}$$

where Z_{GR} is given by

$$Z_{GR} = \exp(G_{R-}^2/2) D_{-0.5}(\sqrt{2}G_{R-}) + \exp(G_{R+}^2/2) D_{-0.5}(\sqrt{2}G_{R+})\,. \tag{11}$$

$D_p(z)$ denotes a parabolic cylinder function [7, Theorem 9.240]. We note that the denominator of $E\{S_R \mid Y_R\}$ is an even function of Y_R while the numerator is an odd function of Y_R. The same estimator can be used for the imaginary part if Y_R is substituted by Y_I in (7)-(11). Figure 2 plots the estimated coefficients (solid) and the coefficients estimated by a Wiener filter (dashed) for input coefficients $0 \le Y_R \le 5$ and $\sigma_y^2 = \sigma_s^2 + \sigma_n^2 = 2$. We find that for high *a priori* SNR conditions both filters are more or less transparent. However, for low SNR conditions the new estimator is highly nonlinear and distinctly different from the Wiener filter. If the disturbed input coefficient is smaller than the variance σ_y^2 the new estimator achieves a higher attenuation than the Wiener filter. If the disturbed input coefficient is larger than the variance σ_y^2 (speech is predominant) the new estimator delivers a significantly larger output than the Wiener filter. It can be therefore assumed that the new estimator results in less speech distortions than the linear estimator.

3.3. Laplacian Noise and Gamma Speech Model

In this Section we derive the MMSE estimator when the noise pdf is a Laplacian and the speech coefficients can be modelled by a Gamma pdf. We define

$$G_+ = \frac{\sqrt{1.5}\sigma_n + 2\sigma_s}{2\sigma_s\sigma_n} = \frac{\sqrt{1.5}/\sqrt{\xi} + 2}{2\sigma_n} \tag{12}$$

$$G_- = \frac{\sqrt{1.5}\sigma_n - 2\sigma_s}{2\sigma_s\sigma_n} = \frac{\sqrt{1.5}/\sqrt{\xi} - 2}{2\sigma_n}\,. \tag{13}$$

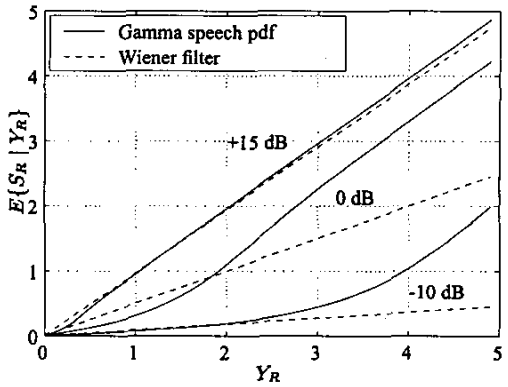

Fig. 2. $E\{S_R \mid Y_R\}$ for the Gamma speech model and a Gaussian noise model (solid), and for three *a priori* SNR values $10\log(\sigma_s^2/\sigma_n^2) = 15, 0, -10$ dB. $\sigma_s^2 + \sigma_n^2 = 2$. The Wiener filter solution is indicated with dashed lines.

For $Y_R \ge 0$ we obtain [7, Theorem 3.381]

$$E\{S_R \mid Y_R\} = \frac{\sqrt[4]{1.5}}{2\sigma_n\sqrt{\pi\sigma_s}p(Y_R)} \int_{-\infty}^{\infty} S_R|S_R|^{-0.5} \exp\left(-\frac{2|Y_R - S_R|}{\sigma_n}\right) \exp\left(-\frac{\sqrt{3}|S_R|}{\sqrt{2}\sigma_s}\right) dS_R \tag{14}$$

$$= \frac{\sqrt[4]{1.5}}{2\sigma_n\sqrt{\pi\sigma_s}p(Y_R)} \left\{ \frac{2}{3} \exp\left(\frac{-2Y_R}{\sigma_n}\right) Y_R^{3/2} \Phi(\frac{3}{2}, \frac{5}{2}; -2G_-Y_R) + \exp\left(-\frac{\sqrt{1.5}}{\sigma_s}Y_R\right) (2G_+)^{-3/2} \Psi(-\frac{1}{2}, -\frac{1}{2}; 2G_+Y_R) - \exp\left(\frac{-2Y_R}{\sigma_n}\right) (2G_+)^{-3/2} \Gamma\left(\frac{3}{2}\right) \right\} \tag{15}$$

with

$$p(Y_R) = \frac{\sqrt[4]{1.5}}{2\sigma_n\sqrt{\pi\sigma_s}} \int_{-\infty}^{\infty} |S_R|^{-0.5} \exp\left(-\frac{2|Y_R - S_R|}{\sigma_n}\right) \cdot \exp\left(-\frac{\sqrt{3}|S_R|}{\sqrt{2}\sigma_s}\right) dS_R \tag{16}$$

$$= \frac{\sqrt[4]{1.5}}{2\sigma_n\sqrt{\pi\sigma_s}} \left\{ 2\exp\left(\frac{-2Y_R}{\sigma_n}\right) \sqrt{Y_R}\Phi(\frac{1}{2}, \frac{3}{2}; -2G_-Y_R) + \exp\left(-\frac{\sqrt{1.5}}{\sigma_s}Y_R\right) \frac{1}{\sqrt{2G_+}} \Psi(\frac{1}{2}, \frac{1}{2}; 2G_+Y_R) + \exp\left(\frac{-2Y_R}{\sigma_n}\right) \sqrt{\frac{\pi}{2G_+}} \right\} \tag{17}$$

where $\Phi(\alpha, \gamma; z) = {}_1F_1(\alpha; \gamma; z)$ denotes a confluent hypergeometric function and $\Psi(\alpha, \gamma; z)$ another confluent hypergeometric

function which is defined as [7, Theorem 9.210]

$$\Psi(\alpha,\gamma;z) = \frac{\Gamma(1-\gamma)}{\Gamma(\alpha-\gamma+1)}\Phi(\alpha,\gamma;z) + \frac{\Gamma(\gamma-1)}{\Gamma(\alpha)} z^{1-\gamma}\Phi(\alpha-\gamma+1,2-\gamma;z)\,. \quad (18)$$

$\Gamma(z)$ is the complete Gamma function [7, Theorem 8.310]. For $Y_R < 0$ we have $E\{S_R \mid Y_R\} = -E\{S_R \mid |Y_R|\}$.

Figure 3 plots the estimation characteristics for the Gamma speech model in combination with the Laplacian noise model. For high *a priori* SNR conditions we find a behaviour similar to the case of the Gaussian noise model, however, below a certain *a priori* SNR threshold the estimator delivers an almost constant enhanced DFT coefficient regardless of what the magnitude of the input coefficient is. As a result, this estimator delivers enhanced speech with a very low level of unnatural fluctuations in the residual noise.

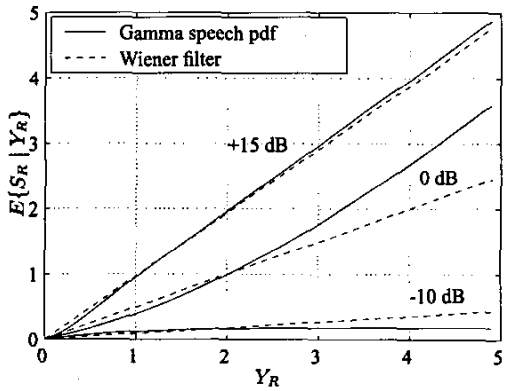

Fig. 3. $E\{S_R \mid Y_R\}$ for the Gamma speech model and a Laplacian noise model (solid), and for three *a priori* SNR values $10\log(\sigma_s^2/\sigma_n^2) = 15, 0, -10$ dB. $\sigma_s^2 + \sigma_n^2 = 2$. The Wiener filter solution is indicated with dashed lines.

4. EXPERIMENTAL RESULTS

The proposed estimators are implemented in MATLAB and embedded into a standard DFT based speech enhancement program with $L = 2R = 256$ and a Hann window for spectral analysis. The apriori SNR is estimated using the "decision directed" approach of [3]. We evaluate the newly derived estimators on a speech data base with 6 different speakers and 3 minutes of speech. Computer generated stationary Gaussian noise as well as prerecorded car noise is added at several SNR levels. When the computer generated Gaussian noise is used its variance is assumed to be perfectly known. To determine the variance of the slightly non-stationary car noise a Minimum Statistics noise estimator is employed [8, 9]. The results are presented in terms of the segmental SNR before and after the processing. Speech pauses are excluded from the computation of the segmental SNR. Table 1 shows the results of processing the noisy speech with either the Wiener filter (case Gaussian/Gaussian), the MMSE estimator with a Gaussian noise pdf and a Gamma speech pdf (case Gaussian/Gamma), or the Laplacian noise model and the Gamma speech model (case Laplace/Gamma). The application of the Gaussian/Gamma estimator results in a consistent improvement of the measured segmental SNR. The improvement using the Laplace/Gamma estimator is somewhat smaller but, as listening tests confirm, significantly less "musical noise" is audible.

noise/speech model	Gaussian noise: SNR			car noise: SNR		
	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB
Gaussian/Gaussian	7.31	14.30	22.04	6.65	13.97	20.90
Gaussian/Gamma	7.73	14.61	22.30	6.81	14.11	20.94
Laplace/Gamma	7.32	14.47	22.26	6.35	13.98	21.00

Table 1. Segmental SNR in dB for our speech and noise models before (0, 10, 20 dB) and after enhancement.

5. CONCLUSIONS

In this contribution we have derived two new estimators for speech enhancement in the DFT domain. Experimental results show that these estimators provide consistently better results than the well known linear estimator (Wiener filter).

6. ACKNOWLEDGEMENT

The author is grateful to Colin Breithaupt for performing the simulations in Table 1.

7. REFERENCES

[1] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, 1979.

[2] R.J. McAulay and M.L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, December 1980.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[4] D.R. Brillinger, *Time Series: Data Analysis and Theory*, Holden-Day, 1981.

[5] J.E. Porter and S.F. Boll, "Optimal Estimators for Spectral Restoration of Noisy Speech," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1984, pp. 18A.2.1–18A.2.4.

[6] H. Brehm and W. Stammler, "Description and Generation of Spherically Invariant Speech-Model Signals," *Signal Processing, Elsevier*, vol. 12, pp. 119–141, 1987.

[7] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, 5th edition, 1994.

[8] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. Euro. Signal Processing Conf. (EUSIPCO), 1994*, pp. 1182–1185.

[9] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.