SPEECH ENHANCEMENT IN THE DFT DOMAIN USING LAPLACIAN SPEECH PRIORS

Rainer Martin and Colin Breithaupt

Institute of Communication Technology
Technical University of Braunschweig, 38106 Braunschweig, Germany
Phone: +49 531 391 2485, Fax: +49 531 391 8218, E-mail: martin@ifn.ing.tu-bs.de

ABSTRACT

In this paper we consider optimal estimators for speech enhancement in the Discrete Fourier Transform (DFT) domain. We derive an analytical solution for estimating complex DFT coefficients in the MMSE sense when the clean speech DFT coefficients are Laplacian distributed and the DFT coefficients of the noise are Gaussian or Laplacian distributed. We show that these estimators have a number of interesting properties. Compared to previously proposed estimators, which are based on Gamma speech priors, the estimators based on Laplacian speech priors have a simpler analytic form.

1. INTRODUCTION

Many 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 density. The Gaussian assumption is valid when the DFT frame size is much longer than the span of correlation of the signal under consideration [4]. In this case the central limit theorem may be invoked and Gaussianity may be assumed.

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. It was also shown in [6] that DFT coefficients of clean speech might be better modelled by a Gamma or a Laplacian distribution when the DFT frame length is in the range of 10-100 ms. The assumption of Gamma speech priors, which was used in [6], leads to estimators which use special functions like the hypergeometric function or Bessel functions. In this paper we will investigate another supergaussian density, namely the Laplacian density, as a model for the clean speech DFT coefficients. We present analytical solutions for the MMSE estimation of complex DFT coefficients with Laplacian speech priors and Gaussian or Laplacian noise priors. These estimators do have similar properties as the estimators based on Gamma densities. However, they are easier to compute and to implement.

The remainder of this paper is organized as follows: In the next Section we will briefly describe the signal models used in this work. 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 of the k-th bin 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 a Hann window of length $L = 2R = 256$.

2.1. Statistical Models

It is well known that the probability density function (pdf) of speech samples in the time domain is much better modelled by a Laplacian or a Gamma density rather than a Gaussian density [7]. We note 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 [6]. In this section we will introduce our notation and briefly review these densities.

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 and the Laplacian prior densities (real and imaginary parts) can be defined as follows, where $\sigma_s^2/2$ denotes the variance of the real and imaginary parts of the clean speech DFT coefficients.

Gaussian speech model:

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

Laplacian speech model:

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

Similar probability densities can be defined for the DFT coefficients of the noise. In order to find closed form solutions to the estimation problems we must assume that the real and the imaginary part are independent. By computing the mutual information between the real and the imaginary part, we found that the dependency between the real and the imaginary part is weak. Thus, this assumption is justified.

3. MMSE ESTIMATORS

Because of the assumed independence of the real and the imaginary parts of DFT coefficients, the MMSE estimator for the complex DFT coefficients can be split into the estimators for the real and the imaginary parts which can be treated independently,

$$E\{S \mid Y\} = E\{S_R \mid Y_R\} + jE\{S_I \mid Y_I\} \,. \tag{4}$$

Again we have dropped the time and frequency indices. Based on the above distribution 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.,

$$\widehat{S} = E\{S \mid Y\} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_n^2} Y = \frac{\xi}{1+\xi} Y \,, \tag{5}$$

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* signal-to-noise ratio (SNR).

3.2. Gaussian Noise and Laplace Speech Model

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

To facilitate the development we introduce the shorthand notations

$$\begin{aligned} L_{R+} &= \frac{\sigma_n}{\sigma_s} + \frac{Y_R}{\sigma_n} = \frac{1}{\sqrt{\xi}} + \frac{Y_R}{\sigma_n} \\ L_{R-} &= \frac{\sigma_n}{\sigma_s} - \frac{Y_R}{\sigma_n} = \frac{1}{\sqrt{\xi}} - \frac{Y_R}{\sigma_n} \\ L_{I+} &= \frac{\sigma_n}{\sigma_s} + \frac{Y_I}{\sigma_n} = \frac{1}{\sqrt{\xi}} + \frac{Y_I}{\sigma_n} \\ L_{I-} &= \frac{\sigma_n}{\sigma_s} - \frac{Y_I}{\sigma_n} = \frac{1}{\sqrt{\xi}} - \frac{Y_I}{\sigma_n} \,. \end{aligned} \tag{6}$$

For the Laplacian speech prior we obtain the optimal MMSE estimator of the real part [8, Theorem 3.462,1]

$$\begin{aligned} E\{S_R \mid Y_R\} &= \frac{1}{\sqrt{\pi}\sigma_n\sigma_s p(Y_R)} \\ &\cdot \int_{-\infty}^{\infty} S_R \exp\left(-\frac{(Y_R - S_R)^2}{\sigma_n^2}\right) \exp\left(-\frac{2|S_R|}{\sigma_s}\right) dS_R \\ &= \frac{\sigma_n \exp(\sigma_n^2/\sigma_s^2)}{2\sigma_s p(Y_R)} \Big\{ L_{R+} \exp(2\frac{Y_R}{\sigma_s}) \mathrm{erfc}(L_{R+}) \\ &\qquad - L_{R-} \exp(-2\frac{Y_R}{\sigma_s}) \mathrm{erfc}(L_{R-}) \Big\} \end{aligned} \tag{7}$$

with [8, Theorem 3.322,2]

$$\begin{aligned} p(Y_R) &= \frac{1}{\sqrt{\pi}\sigma_n\sigma_s} \\ &\cdot \int_{-\infty}^{\infty} \exp\left(-\frac{(Y_R - S_R)^2}{\sigma_n^2}\right) \exp\left(-\frac{2|S_R|}{\sigma_s}\right) dS_R \\ &= \frac{\exp(\sigma_n^2/\sigma_s^2)}{2\sigma_s} \\ &\cdot \Big\{ \exp(2\frac{Y_R}{\sigma_s}) \mathrm{erfc}(L_{R+}) + \exp(-2\frac{Y_R}{\sigma_s}) \mathrm{erfc}(L_{R-}) \Big\} \end{aligned} \tag{8}$$

where $\mathrm{erfc}(z)$ denotes the complementary error function [8, Theorem 8.250]. The optimal estimator for the imaginary part is derived in the same fashion.

$$\begin{aligned} E\{S_R \mid Y_R\} &= \frac{\sigma_n \left[L_{R+} \exp(2Y_R/\sigma_s) \mathrm{erfc}(L_{R+}) - L_{R-} \exp(-2Y_R/\sigma_s) \mathrm{erfc}(L_{R-}) \right]}{\exp(2Y_R/\sigma_s) \mathrm{erfc}(L_{R+}) + \exp(-2Y_R/\sigma_s) \mathrm{erfc}(L_{R-})} \\ &= \frac{\sigma_n \left[L_{R+} \exp(L_{R+}^2) \mathrm{erfc}(L_{R+}) - L_{R-} \exp(L_{R-}^2) \mathrm{erfc}(L_{R-}) \right]}{\exp(L_{R+}^2) \mathrm{erfc}(L_{R+}) + \exp(L_{R-}^2) \mathrm{erfc}(L_{R-})} \end{aligned} \tag{9}$$

$$\begin{aligned} E\{S_I \mid Y_I\} &= \frac{\sigma_n \left[L_{I+} \exp(2Y_I/\sigma_s) \mathrm{erfc}(L_{I+}) - L_{I-} \exp(-2Y_I/\sigma_s) \mathrm{erfc}(L_{I-}) \right]}{\exp(2Y_I/\sigma_s) \mathrm{erfc}(L_{I+}) + \exp(-2Y_I/\sigma_s) \mathrm{erfc}(L_{I-})} \\ &= \frac{\sigma_n \left[L_{I+} \exp(L_{I+}^2) \mathrm{erfc}(L_{I+}) - L_{I-} \exp(L_{I-}^2) \mathrm{erfc}(L_{I-}) \right]}{\exp(L_{I+}^2) \mathrm{erfc}(L_{I+}) + \exp(L_{I-}^2) \mathrm{erfc}(L_{I-})} \end{aligned} \tag{10}$$

The optimal estimator for the complex speech coefficient is therefore given by $E\{S \mid Y\} = E\{S_R \mid Y_R\} + jE\{S_I \mid Y_I\}$ with $E\{S_R \mid Y_R\}$ and $E\{S_I \mid Y_I\}$ given in (9) and (10), respectively. We note that both $E\{S_R \mid Y_R\}$ and $E\{S_I \mid Y_I\}$ are odd symmetric functions of Y_R and Y_I, respectively. The function $\mathrm{erfcx}(x) = \exp(x^2)\mathrm{erfc}(x)$ is known as the scaled complementary error function and is available, e.g., in MATLAB$^{\mathrm{TM}}$. Figure 1 plots the resulting estimate for $0 \leq Y_R \leq 5$, $\sigma_s^2 + \sigma_n^2 = 2$, and three different *a priori* SNR values. For high *a priori* SNR values the estimate is almost identical to the estimate delivered by the Wiener filter. Very little signal distortion occurs. For low SNR values the new estimator is highly non-linear. An interesting feature of the new estimator is that it provides significantly less attenuation to the noisy input coefficient than the Wiener filter when the input coefficient is several times larger than its standard deviation. Given the heavy tailed speech prior, it is very likely that speech is present in this case. For small input values the new estimator delivers more attenuation than the Wiener filter. These two characteristics which were also observed for the estimators developed in [6], both contribute to the improved SNR of the output coefficients with respect to the linear estimator. It is interesting to see that these properties follow nicely from the assumed statistical model.

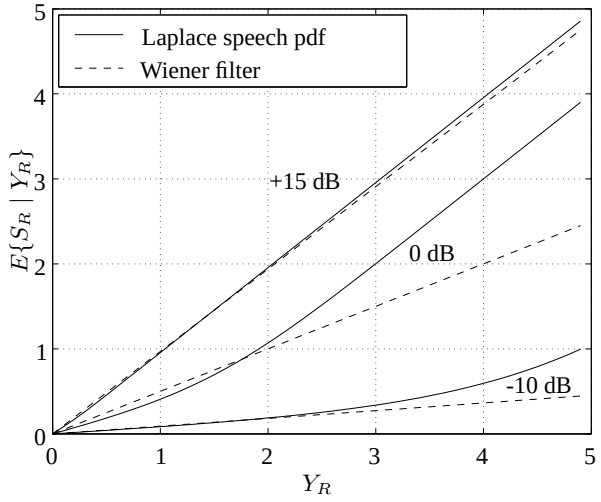

Fig. 1. $E\{S_R \mid Y_R\}$ for the Laplace speech model and a Gaussian noise model (solid), and for three *a priori* SNR values $10\log(\sigma_s^2/\sigma_n^2) = 15, 0, -10$ dB. $\sigma_s^2 + \sigma_n^2 = 2$. The Wiener filter solution is indicated with dashed lines.

3.3. Laplacian Noise and Speech Models

When both the real and the imaginary components of the noise and the speech coefficients can be modelled by a Laplacian pdf, the optimal estimator of the real part is given by ($\sigma_n \neq \sigma_s$)

$$E\{S_R \mid Y_R\} = \frac{1}{\sigma_n \sigma_s p(Y_R)} \int_{-\infty}^{\infty} S_R \exp\left(-\frac{2|Y_R - S_R|}{\sigma_n}\right) \cdot \exp\left(-\frac{2|S_R|}{\sigma_s}\right) dS_R \tag{11}$$

which evaluates to

$$E\{S_R \mid Y_R\} = \frac{sign(Y_R)}{p(Y_R)} \left\{ \frac{\sigma_s^2 \sigma_n^2}{(\sigma_n - \sigma_s)^2(\sigma_n + \sigma_s)^2} \left(\exp\left(\frac{-2|Y_R|}{\sigma_n}\right) - \exp\left(\frac{-2|Y_R|}{\sigma_s}\right)\right) - |Y_R| \exp\left(\frac{-2|Y_R|}{\sigma_s}\right) \frac{\sigma_s}{(\sigma_n - \sigma_s)(\sigma_n + \sigma_s)} \right\} \tag{12}$$

with

$$\begin{aligned} p(Y_R) &= \frac{1}{\sigma_n \sigma_s} \int_{-\infty}^{\infty} \exp\left(-\frac{2|Y_R - S_R|}{\sigma_n}\right) \exp\left(-\frac{2|S_R|}{\sigma_s}\right) dS_R \\ &= \exp\left(\frac{-2|Y_R|}{\sigma_n}\right) \frac{\sigma_n}{(\sigma_n - \sigma_s)(\sigma_n + \sigma_s)} \\ &\quad - \exp\left(\frac{-2|Y_R|}{\sigma_s}\right) \frac{\sigma_s}{(\sigma_n - \sigma_s)(\sigma_n + \sigma_s)} . \end{aligned} \tag{13}$$

For $\sigma_n \neq \sigma_s$ we obtain

$$E\{S_R \mid Y_R\} = \frac{sign(Y_R)}{\exp\left(\frac{-2|Y_R|}{\sigma_n}\right)\sigma_n - \exp\left(\frac{-2|Y_R|}{\sigma_s}\right)\sigma_s} \cdot \frac{\sigma_s^2 \sigma_n^2}{\sigma_n^2 - \sigma_s^2} \left(\exp\left(\frac{-2|Y_R|}{\sigma_n}\right) - \exp\left(\frac{-2|Y_R|}{\sigma_s}\right)\right) - |Y_R| \exp\left(\frac{-2|Y_R|}{\sigma_s}\right) \sigma_s \tag{14}$$

and for $\sigma_n = \sigma_s$ we have

$$E\{S_R \mid Y_R\} = \frac{Y_R}{2} \tag{15}$$

which is identical to the Wiener solution. Analogous relations can be derived for the imaginary part. The attenuation characteristics of this estimator is shown in Figure 2. For an SNR greater than 0 dB the characteristics is similar to the characteristics of the Gaussian noise case as shown in Figure 1. For an SNR smaller than 0 dB the estimator delivers an almost constant output value which is almost independent from the actual input value. This feature of the attenuation characteristics has a profound influence on the naturalness of the perceived residual noise. In the low *a priori* SNR case, fluctuations in the input coefficients will have little effect on the estimated DFT coefficients. Therefore, as listening tests show, the appearance of "musical noise" is less pronounced than with other estimators.

4. EXPERIMENTAL RESULTS

The proposed estimators are implemented in MATLAB and embedded into a standard DFT based speech enhancement program with $L = 2R = 256$. The *a priori* SNR is estimated using the "decision directed" approach of [2]. 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

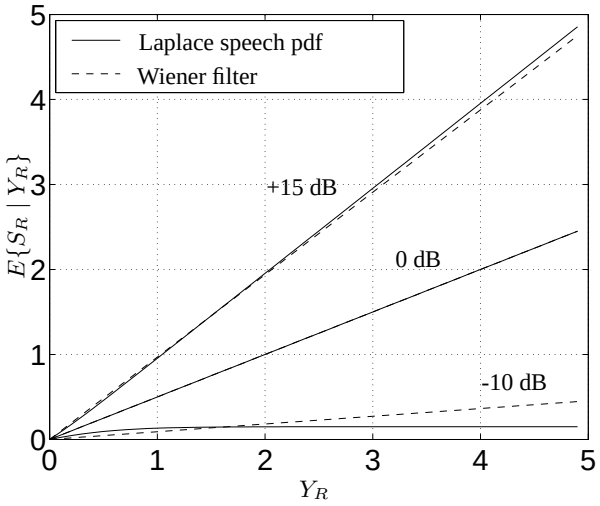

Fig. 2. $E\{S_R \mid Y_R\}$ for the Laplace speech model and Laplacian noise model (solid), and for three *a priori* SNR values $10\log(\sigma_s^2/\sigma_n^2) = 15, 0, -10$ dB. $\sigma_s^2 + \sigma_n^2 = 2$. The Wiener filter solution is indicated with dashed lines.

of the slightly non-stationary car noise a Minimum Statistics noise estimator is employed [9, 10]. 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 Laplace speech pdf (case Gaussian/Laplace), or the Laplacian noise model and the Laplacian speech model (case Laplace/Laplace). The application of the Gaussian/Laplace estimator results in a consistent improvement of the measured segmental SNR. Using the Laplace/Laplace estimator improvements are only obtained for the high *a priori* SNR conditions, however, as listening tests confirm, no "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	7.33	14.09	21.73	6.65	13.97	20.90
Gaussian/Laplace	7.68	14.47	22.13	6.86	14.22	21.19
Laplace/Laplace	6.91	13.98	21.78	6.14	13.84	20.95

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), either in the sense of an improved segmental SNR (case Gaussian/Laplace) or in the sense of better quality of the residual noise (case Laplace/Laplace). We found that the estimator based on Laplacian speech and noise priors is especially attractive: It has an simple analytical form which requires only the evaluation of exponential functions and does not produce "musical noise".

Finally, we note that the proposed model densities may be also used to derive "soft-decision" weighting functions [1] $\frac{\Lambda(\lambda,k)}{1+\Lambda(\lambda,k)}$ and $\frac{1}{1+\Lambda(\lambda,k)}$ where $\Lambda(\lambda,k)$ denotes the generalized likelihood ratio,

$$\Lambda(\lambda,k) = \frac{(1-q(\lambda,k))}{q(\lambda,k)} \frac{p(Y(\lambda,k) \mid H^{(1)}(\lambda,k))}{p(Y(\lambda,k) \mid H^{(0)}(\lambda,k))} . \qquad (16)$$

$q(\lambda,k)$ is the *a priori* probability of speech absence, and $H^{(0)}(\lambda,k)$ and $H^{(1)}(\lambda,k)$ are the hypotheses of speech absence and presence, respectively. Results on estimators which incorporate these soft-decision estimators will be included in a future paper [11].

6. REFERENCES

[1] R. McAulay and M. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, pp. 137–145, December 1980.

[2] 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, pp. 1109–1121, December 1984.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, pp. 443–445, April 1985.

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

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

[6] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. I, pp. 253–256, 2002.

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

[8] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*. Academic Press, 5th ed., 1994.

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

[10] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 504–512, July 2001.

[11] R. Martin, "Speech Enhancement Based on Minimum Mean Square Error Estimation and Supergaussian Priors," *IEEE Trans. Speech and Audio Processing*, 2003 (accepted).