NOISE REDUCTION BY MAXIMUM A POSTERIORI SPECTRAL AMPLITUDE ESTIMATION WITH SUPERGAUSSIAN SPEECH MODELING

Thomas Lotter and Peter Vary

Institute of Communication Systems and Data Processing (ind)
Aachen University (RWTH), Templergraben 55, D-52056 Aachen, Germany
E-mail: {lotter | vary}@ind.rwth-aachen.de

ABSTRACT

This contribution presents a spectral amplitude estimator for acoustical background noise suppression based on maximum a posteriori estimation and supergaussian statistical modeling of the speech DFT coefficients. The probability density function of the speech spectral amplitude is modeled with a simple parametric function, which allows a high approximation accuracy for Laplace or Gamma distributed real and imaginary parts of the speech DFT coefficients. Based on the approximation, a computationally efficient maximum a posteriori speech estimator is derived, which outperforms the Ephraim-Malah algorithm in a single channel noise reduction framework.

1. INTRODUCTION

Most single channel speech enhancement systems rely on frequency domain weighting [1], commonly consisting of a noise power spectral density estimator and a speech spectral or spectral amplitude estimator. The speech estimator applies a statistical estimation rule based on a statistical model of the Discrete Fourier Transform (DFT) coefficients. The well known Wiener filter estimates the complex speech DFT coefficients with minimum mean square error (MMSE), whereas the Ephraim-Malah algorithm[2] is an MMSE estimator for the speech DFT amplitude. The second estimator is considered advantageous from a perceptual point of view, since the spectral phase is rather unimportant to the listener. Both estimators assume zero mean Gaussian distributions of real- and imaginary parts for Fourier coefficients of speech and noise. Whereas the Gaussian model is usually a good approximation for the noise DFT coefficients, the real- and imaginary part of the speech coefficients are better modeled with supergaussian densities [3]. Recently, MMSE complex spectrum estimators with Laplace or Gamma modeling of real- and imaginary parts have been developed [3]. These estimators have shown to provide consistently better result than the linear Wiener filter.

Instead of estimating the complex speech DFT coefficient, this contribution introduces a speech spectral amplitude estimator with an underlying supergaussian model of the speech spectral amplitude. The probability density function of the speech spectral amplitude is approximated by a function with two parameters. With a proper choice of the parameters, the probability density of the amplitude of a complex random variable (RV) with both Laplace and Gamma components can be approximated with high accuracy. Using this approximation, a computationally efficient speech estimator can be found by applying the maximum a posteriori (MAP) estimation rule.

The remainder of the paper is organized as follows: Section II introduces the underlying statistical model for the speech and noise spectral amplitude along with comparisons to experimental data. In Section III the statistical model is applied to derive a MAP estimator for the speech spectral amplitude and finally, in Section IV experimental results are presented.

2. STATISTICAL MODELING

Let $y(i) = s(i) + n(i)$ denote the sampled signal consisting of speech $s(i)$ and noise $n(i)$. After segmentation and windowing with a function $h(i)$, e.g. Hann window, the DFT coefficient of frame k and frequency bin l is calculated with:

$$Y(k,l) = \sum_{i=0}^{L-1} y(kR+i)h(i)e^{-j2\pi li/L}. \tag{1}$$

For the computation of the next DFT, the window is shifted by R samples. For the sake of brevity the index l, k is omitted is the following. Y consists of speech part S and noise N

$$Y = Re^{j\vartheta} = S + N = Ae^{j\alpha} + Be^{j\beta}, \tag{2}$$

with $S = S_{Re} + jS_{Im}$ and $N = N_{Re} + jN_{Im}$. The common Gaussian assumption for the distribution of the real- and imaginary parts of S, N is motivated by the central limit theorem [4]. The span of correlation in $n(\mu)$ is often small compared to the analysis frame size L, thus the assumption holds for the noise coefficients:

$$p(N_{Re}) = \frac{1}{\sqrt{\pi}\sigma_N} \exp\{-\frac{N_{Re}^2}{\sigma_N^2}\}. \tag{3}$$

On the other hand, since the span of correlation in speech cannot be neglected within an analysis frame. The PDF of the Fourier components of speech are better approximated with a Laplace

$$p(S_{Re}) = \frac{1}{\sigma_S} \exp\{-\frac{2|S_{Re}|}{\sigma_S}\} \tag{4}$$

or Gamma density[3]

$$p(S_{Re}) = \frac{\sqrt[4]{1.5}|S_{Re}|^{-\frac{1}{2}}}{2\sqrt{\pi\sigma_S}} \exp\{-\frac{\sqrt{3}|S_{Re}|}{\sqrt{2}\sigma_S}\}. \tag{5}$$

The same equations hold for the imaginary parts S_{Im}, N_{Im}.

To estimate the noise variance σ_N^2, the noise power spectral density can be tracked by averaging periodogramms in noise only phases using a voice activity detector or by tracking minima of a smoothed periodogramm over a sliding time window[5]. To estimate the variance of speech σ_S^2, the decision directed approach can be applied[2].

Assuming statistical independence of real and imaginary parts the

PDF of the noise amplitude can easily be found as Rayleigh distributed by polar integration

$$p(B) = \int_0^{2\pi} Bp(N_{Re}, N_{Im}) d\varphi \tag{6}$$

$$= \int_0^{2\pi} Bp(B\cos\varphi) \cdot p(B\sin\varphi) d\varphi = \frac{2B}{\sigma_N^2} \exp\{-\frac{B^2}{\sigma_N^2}\}. \tag{7}$$

Using the Gaussian model the PDF of the noisy amplitude R given the clean amplitude A can be derived [6] as

$$p(R|A) = \frac{2R}{\sigma_N^2} \exp\left\{-\frac{R^2 + A^2}{\sigma_N^2}\right\} I_0\left(\frac{2AR}{\sigma_N^2}\right), \tag{8}$$

where I_0 denotes the modified Bessel function of zeroth order.

An analytic calculation similar to (6) of the PDF for $A = \sqrt{S_{Re}^2 + S_{Im}^2}$ where S_{Re} and S_{Im} are distributed as given in (4) or (5) is very difficult and a solution is not known to the authors. We therefore propose to use an approximation for $p(A)$. This can be achieved by the following parametric function.

$$p(A) = \frac{\mu^{\alpha+1}}{\Gamma(\alpha+1)} \frac{A^\alpha}{\sigma_S^{\alpha+1}} \exp\{-\mu\frac{A}{\sigma_S}\}. \tag{9}$$

The parameters α, μ determine the shape of the PDF. α greatly influences the value of the PDF at small values while μ gives the slope of the decay towards higher values. Figure 1 shows the goodness of the approximation of $p(A)$, when A is composed of Laplace or Gamma components S_{Re}, S_{Im}, independent in cartesian coordinates. The histogram was taken by generating 500.000 independent RVs S_{Re} and S_{Im} according (4) and (5) respectively of variance $\sigma_S^2/2$. Note that the Laplace or Gamma RV used is only statistically independent in cartesian but not in polar coordinates. We nevertheless use the above generation of the RVs.

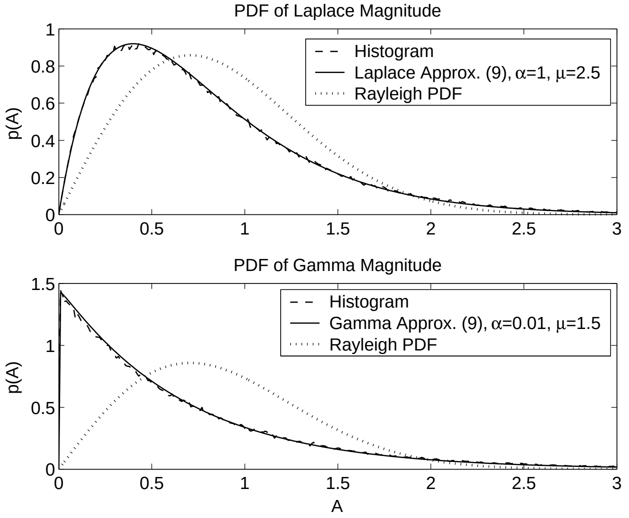

Figure 1: Approximation of $p(A)$ ($\sigma_S^2 = 1$). Laplace components: ($\alpha = 1$, $\mu = 2.5$). Gamma components: ($\alpha = 0.01$, $\mu = 1.5$)

With a proper choice of the parameters ($\alpha = 1$, $\mu = 2.5$ for Laplace approximation, $\alpha = 0.01$, $\mu = 1.5$ for Gamma approximation) the amplitude of a complex RV, that is composed out of two statistically independent Laplace or Gamma components can be approximated with high accuracy. Compared to the Rayleigh distributed amplitude of two Gaussian, low values are more likely and the PDF decreases more slowly towards high values.

2.1. Experimental Data

The real PDFs of the speech and noise amplitude depend on parameters of the noise reduction system. The analysis frame size will influence the distribution of the speech coefficients. At a larger frame size the correlation decreases relative to the analysis frames and thus the distribution will be less supergaussian.

In our experiments we have used a system with half overlapping frames of duration 10ms. To measure the PDF of the speech amplitudes, DFT coefficients were taken from a narrow a priori SNR interval between 19 and 21 dB using a database of about one hour speech. After normalization to variance $\sigma_S^2 = 1$, values from several DFT bin were used. Figure 2 plots the histogram along with the analytic PDFs. Apparently, (9) provides a much better fit for the speech amplitude than the Rayleigh PDF. The real PDF of the speech amplitude lies somewhere between the Laplace and Gamma approximation.

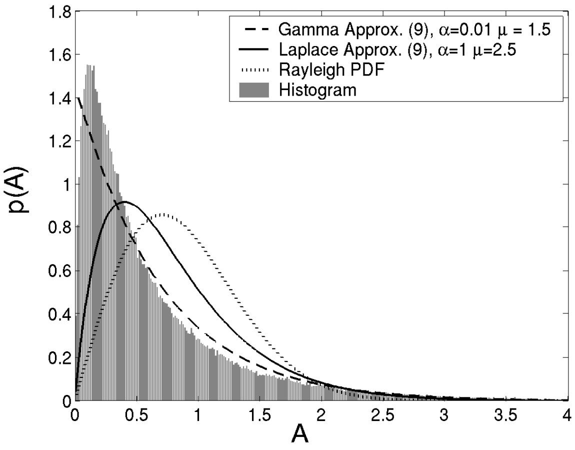

Figure 2: Histogram of speech DFT amplitudes A ($\sigma_S^2 = 1$) fitted with Rayleigh PDF and Laplace/Gamma amplitude approximation

Figure 3 plots the histogram of noise DFT amplitudes measured by a database of noise recorded inside a cafeteria, a crowded market place and a lecture room. Here, the DFT coefficients were taken from an a priori SNR interval of -10 to -12 dB. It can be seen that the Rayleigh distribution is a very good approximation for $p(B)$.

3. DERIVATION OF ESTIMATOR

The given statistical model can be utilized to estimate A. The resulting estimator will be formulated in terms of the a priori ξ and the a posteriori SNR γ

$$\xi = \frac{\sigma_S^2}{\sigma_N^2} \;\; ; \;\; \gamma = \frac{R^2}{\sigma_N^2}. \tag{10}$$

Whereas the a posteriori SNR can directly be computed, the a priori SNR must be estimated by i.e. using the decision-directed approach [2].

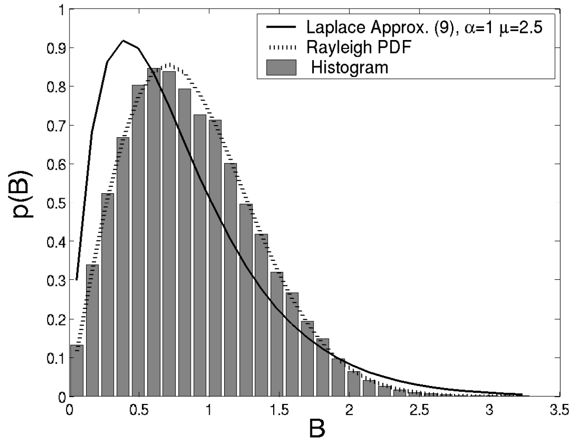

Figure 3: Histogram of noise DFT amplitudes B ($\sigma_N^2 = 1$) fitted with Rayleigh-PDF and Laplace amplitude approximation

An MMSE estimator is given by

$$\hat{A} = E\{A|R\} = \int_0^\infty A \cdot p(A|R) dA \qquad (11)$$

Although the proposed approximation for $p(A)$ in (9) is achieved by a rather simple function, the resulting integrals in (11) are even more complicated than those calculated for the derivation of the Ephraim-Malah estimator due to the additional linear term in the exponential function. However, a computationally efficient MAP solution following

$$\hat{A} = \arg\max_A p(A|R) = \arg\max_A \frac{p(R|A)p(A)}{p(R)} \qquad (12)$$

similar to [7], where Gaussian distributed S_{Re}, S_{Im} are assumed, can be found. Now, (9) is used to model the PDF of the speech spectral amplitude $p(A)$. We need to maximize only $p(R|A) \cdot p(A)$, since $p(R)$ is independent of A. A closed form solution can be found if the modified Bessel function I_0 is considered asymptotically with $I_0(x) \approx \frac{1}{\sqrt{2\pi x}} e^x$. After insertion in (8) we get

$$p(R|A)p(A) \sim A^{\alpha-\frac{1}{2}} \exp\{-\frac{A^2}{\sigma_N^2} - A(\frac{\mu}{\sigma_S} - \frac{2R}{\sigma_N^2})\}. \qquad (13)$$

Instead of differentiating $p(R|A)p(A)$, the maximization can be performed better after applying the natural logarithm, because the product of the polynomial and exponential converts into a sum.

$$\frac{d\log[p(R|A)p(A)]}{dA} = (\alpha - \frac{1}{2})\frac{1}{A} - \frac{2A}{\sigma_N^2} - \frac{\mu}{\sigma_S} + \frac{2R}{\sigma_N^2} = 0 \qquad (14)$$

After multiplication with A, one reasonable solution to the quadratic equation is found, because the second solution delivers spectral amplitudes $A < 0$ at least for $\alpha > 0.5$

$$\hat{A} \quad = \quad R\left(u + \sqrt{u^2 + \frac{\alpha - \frac{1}{2}}{2\gamma}}\right) \qquad (15)$$
$$\text{with } u = \frac{1}{2} - \frac{\mu}{4\sqrt{\gamma\xi}}$$

Compared to the spectral estimators with supergaussian speech modeling, the estimation rule is much simpler and the parameters of the underlying statistical model α, μ can be adapted during runtime without using tables. Also, the weights $G = \hat{A}/R$ are real valued, which is desirable to avoid musical tones.
Figure 4 shows the dependency of the weights on the a posteriori SNR γ for two a priori SNRs ξ.

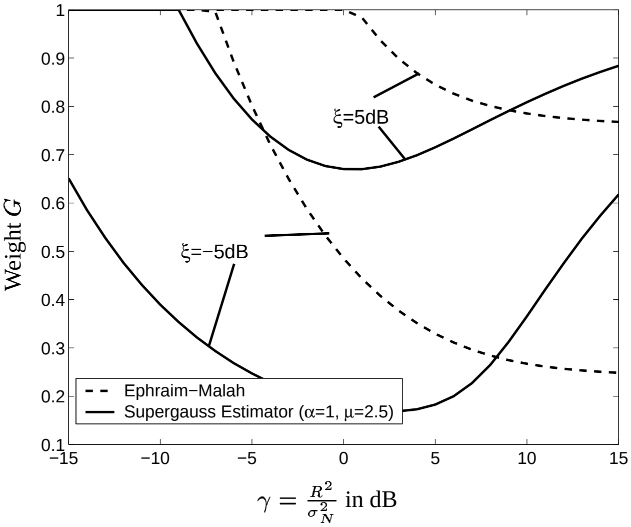

Figure 4: Weights of the Supergauss estimator compared to the Ephraim-Malah weighting rule depending on the instantaneous SNR γ for two a priori SNRs ξ

Most of the time, the weights of the supergauss estimator are smaller than those of the Ephraim-Malah algorithm due to the larger value of $p(A)$ at low amplitudes compared to the Rayleigh PDF. At high instantaneous SNRs the weights increase due to the slower decay of the approximation function towards larger values.

4. PERFORMANCE RESULTS

We evaluate the performance of the proposed estimator within a single channel noise reduction system, where the noise variances σ_N^2 were estimated by means of *Minimum Statistics*[5]. The system operates at a sampling frequency of $f_s = 20kHz$ and a DFT length of $L = 256$.
The parameters α, μ determine the underlying statistical model of the speech amplitude. From the audio impression we favor $\alpha = 1$, $\mu = 2.5$, which approximate the amplitude of a complex RV with independent Laplace components. If the parameters are adjusted for Gamma distributed components the enhanced signal contains very little residual noise but suffers from speech distortion and musical tones. This is due to the approximation of the Bessel function, which generates an uncompensated pole at $A = 0$ for $\alpha < 0.5$.
The amount of noise reduction using (15) with $\alpha = 1$, $\mu = 2.5$ is significantly better than for the Ephraim-Malah algorithm. For an instrumental comparison to the Ephraim-Malah estimator, $\hat{A}$ was multiplicated with a constant factor greater one.
The noise reduction filter was applied to speech signals with added noise for different SNRs. The resulting filter was then utilized to process speech and noise separately [8]. Instead of only considering the segmental SNR improvement obtained by the

noise reduction algorithm, this methods allows separate tracking of speech quality and noise reduction amount. The speech quality of the noise-reduced signal was measured by the segmental speech SNR between original and processed speech. On the other hand, the amount of noise reduction was measured by dividing the input noise power by the output noise power.

The results for three different noises, i.e. white noise, ventilator noise and cafeteria noise, are shown in Figure 5, 6, 7. The noise reduction amount is highest for white noise, because it is easy to track by the noise estimation algorithm due to its stationarity.

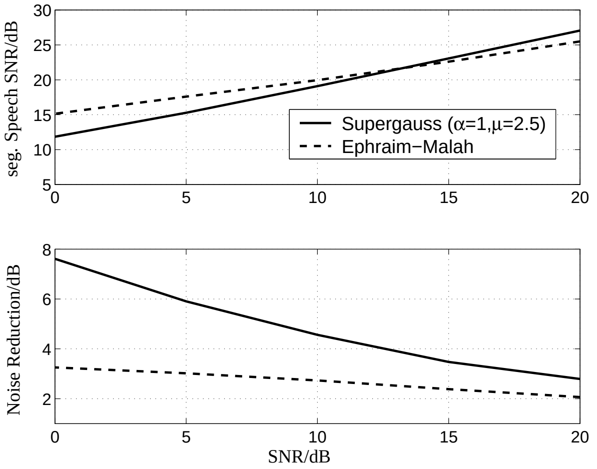

Figure 5: Comparison: Ephraim-Malah – Supergauss estimator for speech corrupted with white noise

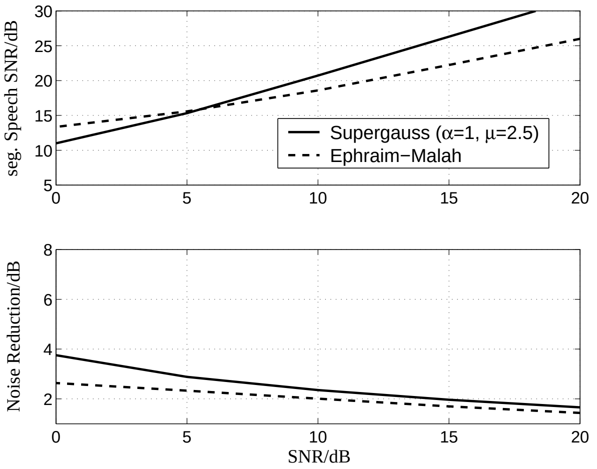

Figure 6: Comparison: Ephraim-Malah – Supergauss estimator for speech corrupted with ventilator noise

The proposed supergaussian estimator delivers a better noise reduction amount than the Ephraim-Malah algorithm at approximately the same speech quality for all three noise classes.

5. SUMMARY

We have derived a computationally efficient MAP estimator for the speech spectral amplitude. The estimator uses a Gaussian model for the noise coefficients, and a supergaussian model for the speech coefficients. The underlying supergaussian model can be adjusted to the demands of the specific noise reduction system. For a parameter setting that approximates the amplitude of a complex random variable with independent Laplace components, we obtained a consistent performance gain compared to the Ephraim-Malah estimator.

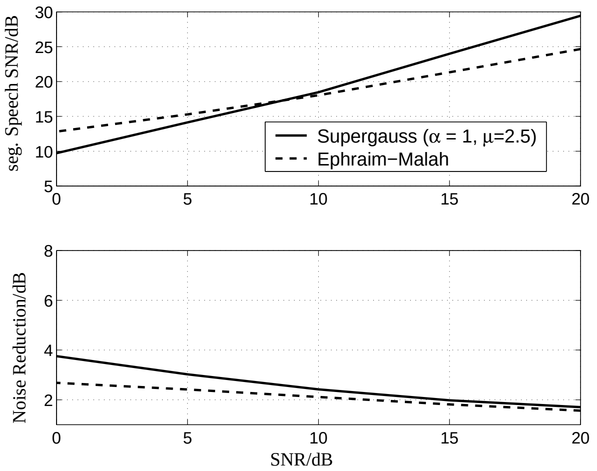

Figure 7: Comparison: Ephraim-Malah – Supergauss estimator for speech corrupted with cafeteria noise

REFERENCES

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

[2] Y. Ephraim and D. Malah, “Speech enhancement uing a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[3] R. Martin, “Speech enhancement using MMSE short time spectral estimation with gamma distributed priors,” *Proc. International Conference on Acoustics, Speech and Signal Processing*, pp. 504–512, May 2002.

[4] D. Brillinger, *Time series, data analysis and theory*. McGraw-Hill, 1981.

[5] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Speech and Audio Processing*, vol. 9, July 2001.

[6] R. McAulay and M. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoustics, Speech and Signal Processing*, pp. 137–145, April 1980.

[7] P. Wolfe and S. Godsill, “Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement,” *Proc. 11th IEEE Workshop on Statistical Signal Processing*, pp. 496–499, August 2001.

[8] S. Gustafsson, R. Martin, and P. Vary, “On the optimization of speech enhancement systems using instrumental measures,” in *Proc. Workshop on Quality Assessment in Speech, Audio, and Image Communication*, (Darmstadt, (in Germany)), pp. 36–40, Mar. 1996.