NOISE REDUCTION BY JOINT MAXIMUM A POSTERIORI SPECTRAL AMPLITUDE AND PHASE ESTIMATION WITH SUPER-GAUSSIAN SPEECH MODELLING

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

For acoustical background noise reduction a computationally efficient joint MAP estimator with a super-Gaussian speech model is presented. Compared to a recently introduced MAP estimator the new joint MAP estimator allows an optimal adjustment of the underlying statistical model to the real PDF of the speech spectral amplitude. The computationally efficient estimator outperforms the Ephraim-Malah estimator and the recently proposed MAP estimator in a single microphone noise reduction framework due to the more accurate statistical model.

1. INTRODUCTION

Most single microphone speech enhancement systems rely on frequency domain weighting, commonly consisting of a noise power spectral density estimator and a speech spectrum 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[1] 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 modelled with super-Gaussian densities [2]. Recently, MMSE complex spectrum estimators with Laplace or Gamma modelling of real- and imaginary parts [2],[3] have been proposed. Moreover, a spectral amplitude estimator with a parametric super-Gaussian speech model for Laplace like distributed real- and imaginary parts has been introduced [4]. These estimators have shown to provide consistently better result than the Wiener and Ephraim-Malah estimator respectively.

In this contribution a new spectral amplitude estimator with a more general underlying statistical model than in [4] is proposed. Due to the possibility to apply more accurate models, the estimator outperforms the estimators from [4],[1].

The remainder of the paper is organized as follows: Section II gives an overview of the noise reduction system. Section III reviews the underlying statistical model for the speech spectral amplitude along with novel matching of the model to experimental data. In Section IV the statistical model is applied to derive the joint MAP estimator for the speech spectral amplitude and phase and finally, in Section V experimental results are presented.

2. NOISE REDUCTION SYSTEM

Figure 1 shows an overview of the single channel speech enhancement system examined in this work. The noisy time signal $y(l)$ is

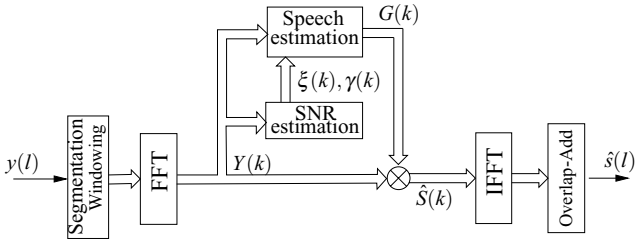

Figure 1: Single channel speech enhancement system.

composed of clean speech $s(l)$ and additive noise $n(l)$. After segmentation and windowing with $h(l)$, e.g. Hann window with zero padding, the DFT coefficient of frame λ and frequency bin k is calculated with:

$$Y(\lambda,k) = \sum_{l=0}^{L-1} y(\lambda Q + l) h(l) e^{-j2\pi lk/L}. \quad (1)$$

L denotes the DFT frame size. For the noise reduction system $L = 256$ is used at a sampling frequency of 20kHz. For the computation of the next DFT, the window is shifted by $Q = 112$ samples. The noisy DFT coefficient Y of amplitude R and phase ϑ consists of speech component S and noise N

$$Y(\lambda,k) = R(\lambda,k) e^{j\vartheta(\lambda,k)} = S(\lambda,k) + N(\lambda,k), \quad (2)$$

with $S = S_{\mathrm{Re}} + jS_{\mathrm{Im}}$ and $N = N_{\mathrm{Re}} + jN_{\mathrm{Im}}$, where $S_{\mathrm{Re}} = \mathrm{Re}\{S\}$ and $S_{\mathrm{Im}} = \mathrm{Im}\{S\}$. The speech coefficient consists of amplitude A and phase α, i.e. $S(\lambda,k) = A(\lambda,k) e^{j\alpha(\lambda,k)}$. The SNR estimation block calculates a priori SNR ξ and a posteriori SNR γ for each DFT bin k with the use of an estimate of the noise power spectral density σ_N^2, obtained by *Minimum Statistics* [5].

$$\xi(\lambda,k) = \frac{\sigma_S^2(\lambda,k)}{\sigma_N^2(\lambda,k)} \;;\; \gamma(\lambda,k) = \frac{R^2(\lambda,k)}{\sigma_N^2(\lambda,k)}. \quad (3)$$

Here, σ_S^2 denotes the estimate of the instantaneous frequency and time dependent power spectral density of the speech. For estimation of the a priori SNRs ξ we apply the well known recursive approach proposed by Ephraim and Malah [1]. The task of the speech estimation block is the calculation of spectral weights G for the noisy spectral components Y. After IFFT and overlap-add the enhanced time signal $\hat{s}(l)$ is obtained.

For the sake of brevity the frame index λ and frequency index k is omitted in the following.

3. STATISTICAL MODEL

Motivated by the central limit theorem, real and imaginary part of the noise DFT coefficients are very often modelled as zero mean independent Gaussian [1] with equal variance. For many relevant acoustic noises this assumption holds.
The variance of the noise DFT coefficient σ_N^2 is assumed to split equally into real and imaginary part. The PDF of the noisy spectrum Y conditioned on the speech amplitude A and phase α can then be written as joint Gaussian:

$$p(Y|A,\alpha) = \frac{1}{\pi\sigma_N^2}\exp\left(-\frac{|Y - A\,e^{j\alpha}|^2}{\sigma_N^2}\right) \tag{4}$$

The PDF of the noisy amplitude R given the speech amplitude A is Rician

$$p(R|A) = \frac{2R}{\sigma_N^2}\exp\left\{-\frac{R^2+A^2}{\sigma_N^2}\right\} I_0\left(\frac{2AR}{\sigma_N^2}\right) \tag{5}$$

I_0 denotes the modified Bessel function of zeroth order.
On the other hand the speech DFT coefficients are known to be super-Gaussian distributed. Instead of a Gaussian model Martin [2],[3] has developed spectral estimators with Laplace or Gamma model for statistical independent real and imaginary parts.
For the calculation of appropriate PDFs for the speech spectral amplitude A, the Gauss, Laplace and Gamma PDFs for real and imaginary parts are taken into account. Considering Gaussian components, the amplitude would be Rayleigh distributed. For independent Laplace or Gamma components a parametric approximation has been proposed in [4] with:

$$p(A) = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)}\frac{A^{\nu}}{\sigma_S^{\nu+1}}\exp\{-\mu\frac{A}{\sigma_S}\}. \tag{6}$$

The Gamma function is denoted as Γ. The parameters ν, μ determine the shape of the PDF. ν greatly influences the value of the PDF at small values while μ gives the slope of the decay towards higher values. It has been shown that the amplitude of a complex Laplace or Gamma random variable with independent components can be approximated with high accuracy using the parametric function and different parameter sets of (ν, μ). For the Laplace amplitude approximation ($\nu = 1, \mu = 2.5$) can be applied, while ($\nu = 0.01, \mu = 1.5$) approximates the PDF of the amplitude of a complex Gamma variable.

3.1 Matching with Experimental Data

To measure $p(A)$, DFT amplitudes were taken from a narrow speech variance interval measured with the speech enhancement system using a database of about one hour speech for DFT bins between 500Hz and 2000Hz. Figure 2 plots the histogram after normalization to $\sigma_S^2 = 1$ along with the analytic Rayleigh PDF and the approximation according to (6) with the parameter set for Laplace and Gamma amplitude approximation respectively. Apparently, (6) provides a much better fit for the speech amplitude than the Rayleigh PDF for both Laplace and Gamma amplitude approximation for both low and high arguments. The real PDF of the speech amplitude lies between the Laplace and Gamma amplitude approximation. For the data measured with our system the Gamma amplitude approximation fits the observed data better.

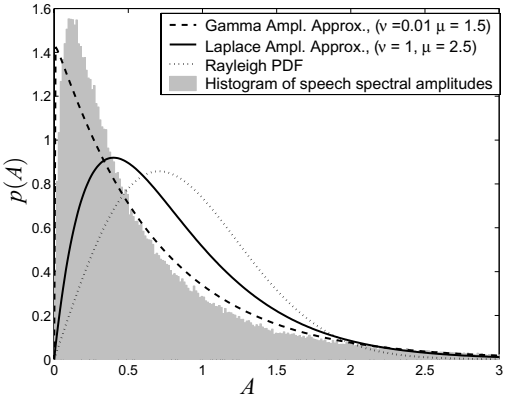

Figure 2: Histogram of speech DFT amplitudes A ($\sigma_S^2 = 1$) with Rayleigh PDF and Laplace/Gamma amplitude approximation (6).

To find a set (ν, μ), that approximates the real PDF best the Kullback divergence according to (7) between the analytic function and the histogram with N bins is minimized.

$$J(A:h) = \sum_{n=1}^{N}(p_h(n) - p_A(n))\log\left(\frac{p_h(n)}{p_A(n)}\right). \tag{7}$$

Figure 3 shows the best $p(A)$ according to (6) determined by minimizing the Kullback divergence. For our system the parameter set

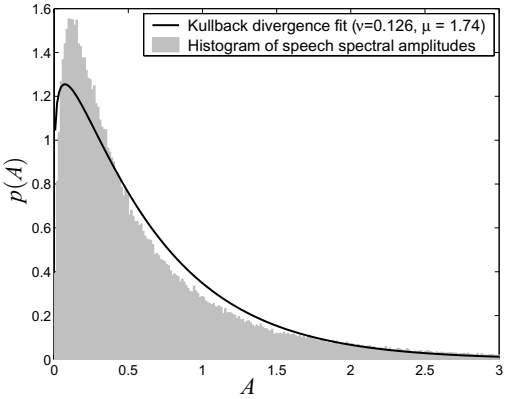

Figure 3: Histogram of speech DFT amplitudes and fitted approximation by (6) according to Kullback divergence ($\sigma_S^2 = 1$).

($\nu = 0.126, \mu = 1.74$) fits best to the observed data.

4. SPEECH ESTIMATORS

In the following, the MAP estimator proposed in [4] is briefly reviewed. Secondly, a joint MAP estimator for the amplitude and phase is introduced, which is a super-Gaussian extension of the joint MAP estimators proposed by [6].

4.1 MAP Spectral Amplitude Estimator

A computationally efficient MAP solution following

$$\hat{A} = \arg\max_A p(A|R) = \arg\max_A \frac{p(R|A)p(A)}{p(R)} \tag{8}$$

similar to [6], where Gaussian distributed S_{Re}, S_{Im} are assumed was found in [4]. The super-Gaussian function (6) is used to model the PDF of the speech spectral amplitude $p(A)$. The Gaussian assumption of noise allows to apply (5) for $p(R|A)$. A closed form solution was found after considering the modified Bessel function I_0 asymptotically with $I_0(x) \approx \frac{1}{\sqrt{2\pi x}}\, e^x$:

$$G_{\mathrm{MAP}} = u + \sqrt{u^2 + \frac{\nu - \frac{1}{2}}{2\gamma}} \quad \text{with} \quad u = \frac{1}{2} - \frac{\mu}{4\sqrt{\gamma\xi}} \tag{9}$$

Whereas the efficient MAP spectral amplitude estimator outperforms the Ephraim-Malah estimator for an estimation with an underlying Laplace model of the DFT coefficients, it cannot be applied using a Gamma model or the optimal parameter set. This is due to the inaccuracy introduced by the approximation of the Bessel function. For $\nu < 0.5$, the approximated a posteriori density $p(A|R)$ has a pole at $A = 0$, which will misplace the maximum found by (9).

4.2 Joint MAP Amplitude and Phase Estimator

To overcome the inability of the MAP estimator from [4] to cope with an underlying Gamma model or the model, that minimizes the Kullback divergence towards the measured data, a joint MAP estimator similar to [6] is introduced. Instead of maximizing the a posteriori probability $p(A|R)$, we now jointly maximize the probability of amplitude and phase conditioned on the observed complex coefficient, i.e. $p(A, \alpha|Y)$.

$$\hat{A} = \arg\max_A p(A,\alpha|Y) = \arg\max_A \frac{p(Y|A,\alpha)p(A,\alpha)}{p(Y)} \tag{10}$$

$$\hat{\alpha} = \arg\max_\alpha p(A,\alpha|Y) = \arg\max_\alpha \frac{p(Y|A,\alpha)p(A,\alpha)}{p(Y)} \tag{11}$$

If the problem is formulated this way, the Bessel function and its erroneous approximation are avoided. It can be shown by measurements that the PDF of amplitude and phase is rotationally invariant, thus we can write: $p(A,\alpha) = \frac{1}{2\pi} p(A)$. (10) and (11) can be solved similar to the MAP estimator. Taking the natural logarithm greatly facilitates the optimization process. After insertion of (4) and (6) we get

$$\log(p(Y|A,\alpha)p(A,\alpha)) \sim -\frac{|Y - Ae^{j\alpha}|^2}{\sigma_N^2} + \nu \log A - \mu \frac{A}{\sigma_S} \tag{12}$$

The partial derivatives of $\log(p(Y|A,\alpha)p(A,\alpha))$ with respect to the phase α and amplitude A need to be zero. Taking the partial derivative w.r.t. α results in $\hat{\alpha} = \vartheta$, i.e. the best MAP estimate for the clean phase is the noisy phase. Solving the derivative w.r.t. the amplitude A then leads to an estimation rule similar to that of the super-Gaussian MAP estimator.

$$G_{\mathrm{JMAP}} = u + \sqrt{u^2 + \frac{\nu}{2\gamma}} \quad \text{with} \quad u = \frac{1}{2} - \frac{\mu}{4\sqrt{\gamma\xi}} \tag{13}$$

Figure 4 compares the weights of the joint MAP estimator with optimal parameter set with those of the MAP estimator with Laplace amplitude model and those of the Ephraim-Malah estimator in dependence of the a posteriori SNR for two different a priori SNRs. Most of the time, the weights of the super-Gaussian estimators 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.

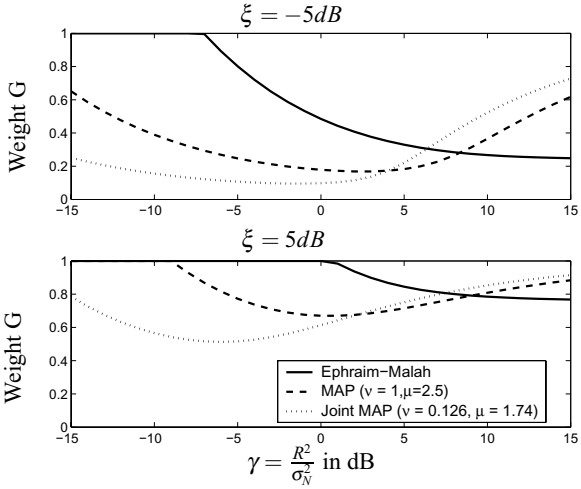

Figure 4: Weights of the joint MAP estimator with Kullback divergence parameters (dotted), MAP estimator with Laplace amplitude parameters (dashed) and Ephraim-Malah (solid) estimator as a function of the a posteriori SNR γ for $\xi = -5dB$ (upper plot) and $\xi = 5dB$ (lower plot).

At high a posteriori SNRs the Ephraim-Malah weights converge towards the Wiener weights, i.e. $\xi/(1+\xi)$. The weights of the super-Gaussian MAP estimators however increases due to the slower decay of the model function towards larger values. The behavior is more extreme for the joint MAP estimator because the underlying speech PDF is farer away from the Rayleigh PDF.

5. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed super-Gaussian spectral amplitude estimators in comparison to the state-of-the-art Ephraim-Malah spectral amplitude estimator by instrumental measurements. To measure the quality of the filter, speech s and noise n are superposed with a given SNR. The noisy signal y is processed with the noise reduction algorithm. Afterwards the desired and the interfering signal are separately processed with the resulting filter coefficients. Hence, the system enables separate tracking of speech quality and noise reduction amount by comparing outputs to inputs of the fixed filters.

The parameters $(\nu,\ \mu)$ determine the underlying statistical model of the speech amplitude. For the super-Gaussian MAP estimator we favor $(\nu = 1,\ \mu = 2.5)$, which approximate the amplitude of a complex RV with independent Laplace components. In general, the super-Gaussian MAP estimator [4] cannot be applied for $\nu < 0.5$. The super-Gaussian joint MAP estimator however can be applied to every non-negative set of parameters $(\nu,\ \mu)$. Here, we favor the parameters, that were determined by minimizing the Kullback divergence towards the measured data, i.e. $(\nu = 0.126,\ \mu = 1.74)$. For the reason of comparability the weights of the super-Gaussian estimators are scaled by a constant factor greater one so that approximately the same speech quality is reached for all estimators. The amount of noise reduction achieved then allows a comparison between the estimators. In all versions we include the soft weight given by Ephraim and Malah [1] with tracking of speech presence uncertainties [7].

5.1 Performance in White noise

The results for white noise and the three different estimators, i.e. Ephraim-Malah, MAP with $(\nu = 1, \mu = 2.5)$ and joint MAP with

($\nu = 0.126, \mu = 1.74$) are shown in Figure 5. Figure 6 plots the performance of the estimators for speech with fan noise and Figure 7 for cafeteria noise. All estimators deliver approximately the same

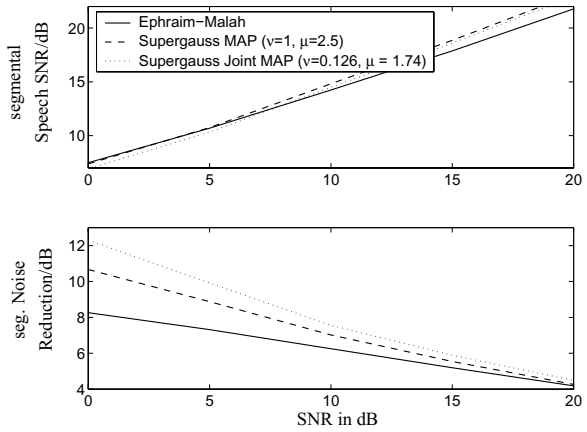

Figure 5: Speech quality and noise reduction amount of statistical filter with Ephraim-Malah estimator (solid) with super-Gaussian MAP estimator (dashed) and super-Gaussian joint MAP estimator (dotted) for speech corrupted with white noise.

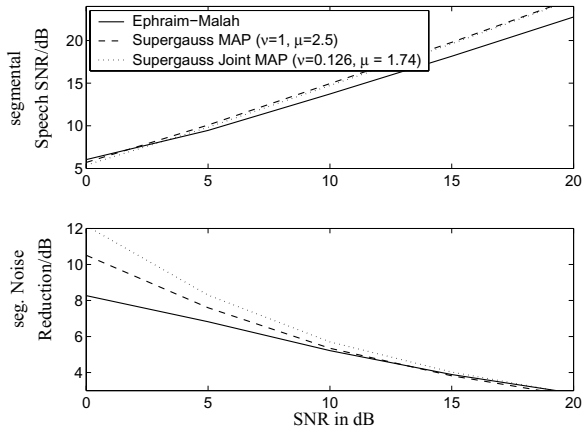

Figure 6: Speech quality and noise reduction amount of statistical filter with Ephraim-Malah estimator (solid) with super-Gaussian MAP estimator (dashed) and super-Gaussian joint MAP estimator (dotted) for speech corrupted with fan noise.

speech quality due to multiplication of the MAP estimates with a constant factor. The super-Gaussian MAP estimator achieves a significantly higher noise attenuation than the Ephraim-Malah estimator. By applying the super-Gaussian joint MAP estimator with parameters optimally adjusted to the measured data, the noise reduction amount can be increased further without decreasing the speech quality. The performance gain is slightly lower for the fan noise.

6. CONCLUSION

We have derived a computationally efficient joint MAP estimator with a super-Gaussian model for the speech spectral amplitude and phase which outperforms the Ephraim-Malah estimator and a recently proposed MAP estimator. The joint MAP estimator delivers a computationally efficient calculation rule for real spectral weights,

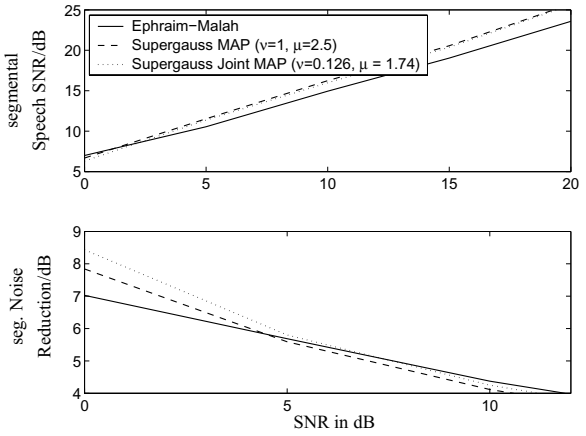

Figure 7: Speech quality and noise reduction amount of statistical filter with Ephraim-Malah estimator (solid) with super-Gaussian MAP estimator (dashed) and super-Gaussian joint MAP estimator (dotted) for speech corrupted with cafeteria noise.

i.e. the noisy phase is not modified. Compared to the recently introduced MAP estimator the new joint MAP estimator allows an optimal adjustment of the underlying statistical model to PDFs of the speech spectral amplitude with Gamma like distributed real and imaginary parts and thus improves the overall quality of the noise reduction system.

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 Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[2] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed priors," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Orlando, USA, May 2002.

[3] R. Martin and C. Breithaupt, "Speech Enhancement in the DFT Domain using Laplacian Speech Priors," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, September 2003, pp. 87–90.

[4] T. Lotter and P. Vary, "Noise Reduction by Maximum a Posteriori Spectral Amplitude Estimation with Supergaussian Speech Modelling," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, September 2003, pp. 83–86.

[5] 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.

[6] P.J. Wolfe and S.J. Godsill, "Efficient Alternatives to the Ephraim-Malah Suppression Rule for Audio Signal Enhancement," *EURSIP Journal on Applied Signal Processing, Special Issue: Digital Audio for Multimedia Communications*, pp. 1043–1051, 2003.

[7] D. Malah, R.V. Cox, and A.J. Accardi, "Tracking speech presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Phoenix, USA, May 1999.