EURASIP Journal on Applied Signal Processing 2005:7, 1110–1126

Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model

Thomas Lotter

Institute of Communication Systems and Data Processing, RWTH Aachen University of Technology, RWTH Aachen, 52056 Aachen, Germany

Siemens Audiological Engineering Group, Gebbertstrasse 125, 91058 Erlangen, Germany
Email: thomas.tl.lotter@siemens.com

Peter Vary

Institute of Communication Systems and Data Processing, RWTH Aachen University of Technology, RWTH Aachen, 52056 Aachen, Germany
Email: vary@ind.rwth-aachen.de

Received 7 June 2004; Revised 17 September 2004; Recommended for Publication by Jacob Benesty

This contribution presents two spectral amplitude estimators for acoustical background noise suppression based on maximum a posteriori estimation and super-Gaussian statistical modelling of the speech DFT amplitudes. The probability density function of the speech spectral amplitude is modelled 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. Also, the statistical model can be adapted to optimally fit the distribution of the speech spectral amplitudes for a specific noise reduction system. Based on the super-Gaussian statistical model, computationally efficient maximum a posteriori speech estimators are derived, which outperform the commonly applied Ephraim-Malah algorithm.

Keywords and phrases: speech enhancement, MAP estimation, speech model.

1. INTRODUCTION

The reduction of acoustical background noise using a single microphone is an important subject to improve the quality of speech communication systems in the context of digital hearing aids, speech recognition, hands-free telephony, or teleconferencing. Although single-microphone speech enhancement has been a research topic for decades, the estimation of a clean speech signal from its noisy observation remains a challenging task, especially due to the wide variety of environmental noises.

If the disturbing noise is assumed to be truly environmental, that is, its origin is, for example, machines, cars, or several persons talking at the same time, the specific properties of speech such as *nonwhiteness, nonstationarity* and *non-Gaussianity* compared to unwanted noise allow a differentiation between speech and noise.

Nonwhiteness means that the short-time spectrum of speech is generally less flat than that of acoustic noise. This property can be exploited by separating speech and noise in the spectral domain. The concept of spectral domain noise attenuation has been introduced more than twenty years ago by Boll [1] as the subtraction of an estimated noise spectral magnitude from the noisy spectral magnitude.

这意味着语音的短时间谱通常比声学噪声的短时间谱不平坦。

不稳定性

To estimate the noise power spectral density, the second property, *nonstationarity*, is exploited by averaging DFT squared magnitudes in noise-only phases or by tracking spectral minima over time [2]. Noise reduction by spectral domain weighting has frequently been plagued by musical tones, that is, annoying fluctuations in the residual noise signal. This is especially due to the subtraction of an expectation in terms of the noise power spectral density from an instantaneous value. To overcome this problem, improved algorithms have been proposed by Ephraim and Malah [3, 4]. The clean speech spectral amplitude is estimated with respect to the minimization of a statistical error criterion. Together with a recursive estimation of the underlying speech variance, the approach results in a good speech quality without audible musical noise.

非高斯性

Recently, the third property, *non-Gaussianity*, has been included in the spectral domain noise reduction framework by Martin [5, 6]. The statistical estimation of the speech

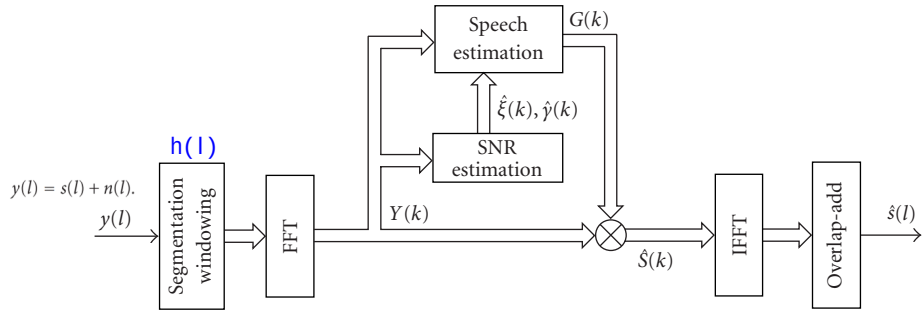


FIGURE 1: Overview of the single-channel speech enhancement system (l: time index, k: frequency index).

spectrum requires a statistical model of the undisturbed speech and noise spectral coefficients. It is well known that speech samples have a super-Gaussian distribution, which causes the speech spectral coefficients to be super-Gaussian distributed as well. By including a super-Gaussian model of speech, the mean squared error of a statistical estimator can be decreased compared to an estimation with an underlying Gaussian model. Whereas the proposed estimators by Martin with underlying Gamma or Laplace PDFs for real and imaginary parts of speech and noise DFT coefficients [5, 6] are optimal with respect to the mean squared estimation error of the estimated complex speech DFT coefficient, they are suboptimal for the estimation of the speech spectral amplitude. 相位感知不重要幅度更重要 Spectral amplitude estimation can be considered more advantageous due to the perceptual unimportance of the phase [7]. Ephraim and Malah have proposed two estimators that minimize the squared or logarithmic error of the speech spectral amplitude under a Gaussian model of the complex speech and noise DFT coefficients [3, 4].

In this contribution spectral amplitude estimators with super-Gaussian speech modelling are introduced. The probability density function of the speech spectral amplitude is approximated by a function with two parameters. With a proper choice of the parameters, for example, the probability density of the amplitude of a complex random variable (RV) with both independent Laplace and Gamma components can be approximated with high accuracy. Also, the parameters of the underlying PDF can be optimally fitted to the real distribution of the speech spectral amplitude for a specific noise reduction algorithm. Using this statistical model, computationally efficient speech estimators can be found by applying the maximum a posteriori (MAP) estimation rule. The resulting estimators, which are super-Gaussian extensions of the MAP estimators derived by Wolfe and Godsill [8], outperform the commonly applied Ephraim-Malah estimators by the more accurate statistical model.

The remainder of the paper is organized as follows. Section 2 gives an overview of the single-channel noise reduction by spectral weighting. Section 3 introduces the underlying statistical model for the speech and noise spectral amplitudes along with comparisons to experimental data. In Section 4 the statistical model is applied to derive a MAP estimator for the speech spectral amplitude and a joint MAP estimator for the speech spectral amplitude and phase. Finally, in Section 5, experimental results are presented.

2. OVERVIEW

Figure 1 shows an overview of the single-channel speech enhancement system examined in this work [9]. The noisy time signal $y(l)$ sampled at regular time intervals $l \cdot T$ is composed of clean speech $s(l)$ and additive noise $n(l)$:

$$y(l) = s(l) + n(l). \tag{1}$$

After segmentation and windowing with a function $h(l)$, for example, Hann window, the DFT coefficient of frame λ and frequency bin k is calculated with

$$Y(\lambda, k) = \sum_{l=0}^{L-1} y(\lambda Q + l) h(l) e^{-j2\pi lk/L}, \tag{2}$$

L denotes the DFT frame size. For the noise reduction system applied in this work, $L = 256$ is used at a sampling frequency of 20 kHz. For the computation of the next DFT, the window is shifted by Q samples. To decrease the disturbing effects of cyclic convolution, we apply half overlapping Hann windows with 16 zeros at the beginning and end. The effective frame size is thus only 224 samples, which corresponds to a frame size of 11.2 milliseconds and a frame shift of 5.6 milliseconds, respectively.

The noisy DFT coefficient Y consists of speech part S and noise N:

$$Y(\lambda, k) = S(\lambda, k) + N(\lambda, k), \tag{3}$$

with $S = S_{\mathrm{Re}} + jS_{\mathrm{Im}}$ and $N = N_{\mathrm{Re}} + jN_{\mathrm{Im}}$, where $S_{\mathrm{Re}} = \mathrm{Re}\{S\}$ and $S_{\mathrm{Im}} = \mathrm{Im}\{S\}$. In polar coordinates the noisy DFT coefficient of amplitude R and phase ϑ is written as ???

$$R(\lambda, k) e^{j\vartheta(\lambda,k)} = A(\lambda, k) e^{j\alpha(\lambda,k)} + B(\lambda, k) e^{j\beta(\lambda,k)}. \tag{4}$$

The speech DFT amplitude is termed as A, the noise DFT amplitude as B, and the respective phases as α, β.

The SNR estimation block calculates a priori SNR ξ and a posteriori SNR γ for each DFT bin k. The SNR calculation requires an estimate of the noise power spectral density $\sigma_N^2(\lambda, k)$. It can be estimated by averaging DFT squared magnitudes in periods of speech pauses. Assuming that noise is stationary, the measured PSD can be saved and applied as an estimate during following speech activity. This method requires a reliable voice activity detector (e.g., [10]). However, a VAD is difficult to tune and its application at low SNRs often results in clipped speech. Therefore, we apply *minimum statistics*, which tracks minima of the smoothed periodogram over a time period that greatly exceeds the speech short-time stationarity [2].

声音停顿时幅度的平方

Based on the noise estimates $\hat{\sigma}_N^2$ and the observed Fourier amplitudes R the a priori and the a posteriori SNRs are estimated by

$$\hat{\xi}(\lambda, k) = \frac{\hat{\sigma}_S^2(\lambda, k)}{\hat{\sigma}_N^2(\lambda, k)}, \qquad \hat{\gamma}(\lambda, k) = \frac{R^2(\lambda, k)}{\hat{\sigma}_N^2(\lambda, k)}. \tag{5}$$

表示语音的瞬时功率谱密度。

Here, σ_S^2 denotes the instantaneous power spectral density of the speech. Whereas the a posteriori SNRs γ can directly be computed, the a priori SNRs ξ have to be estimated. This is performed using a recursive approach proposed by Ephraim and Malah [3]:

$$\hat{\xi}(\lambda, k) = \alpha_{\text{snr}} \frac{\hat{A}^2(\lambda - 1, k)}{\hat{\sigma}_N^2(\lambda, k)} + (1 - \alpha_{\text{snr}}) F[\gamma(\hat{\lambda}, k) - 1],$$
$$F[x] = \begin{cases} x, & x > 0, \\ 0, & \text{else.} \end{cases} \tag{6}$$

An alternative estimation approach which incorporates frequency correlation is presented in [11]. It is frequently argued [12, 13] that the recursive approach is essential for a high quality of the enhanced signal. A high smoothing factor α_{snr} greatly reduces the dynamics of the instantaneous SNR in speech pauses and thus reduces musical tones. However the a priori SNR will then comprise a delayed version of the speech. Since the a priori SNR has a high impact on the noise reduction amount, it is useful to lower limit the a priori SNR according to

$$\tilde{\xi}(\lambda, k) = \begin{cases} \hat{\xi}(\lambda, k), & \hat{\xi}(\lambda, k) > \xi_{\text{thr}}, \\ \xi_{\text{thr}}, & \text{else.} \end{cases} \tag{7}$$

The task of the speech estimation block is the calculation of spectral weights G for the noisy spectral components Y, such that the estimated speech DFT coefficient $\hat{S}$ is calculated by

$$\hat{S}(\lambda, k) = G(\hat{\xi}(\lambda, k), \hat{\gamma}(\lambda, k)) \cdot Y(\lambda, k). \tag{8}$$

After IFFT and overlap-add, the enhanced time signal $\hat{s}(l)$ is obtained.

3. STATISTICAL MODEL

We introduce the statistical model for the speech and noise spectral amplitudes. For the sake of brevity the frame index λ and frequency index k are omitted, however the following considerations hold independently for every frequency bin k and frame λ.

Motivated by the central limit theorem, real and imaginary parts of both speech and noise DFT coefficients are very often modelled as zero-mean independent Gaussian [3, 14, 15] with equal variance. This is due to the properties of the DFT:

$$Y(\lambda, k) = \sum_{l=0}^{L-1} y(\lambda Q + l) \cos\left(\frac{2\pi k l}{L}\right) - j \sum_{l=0}^{L-1} y(\lambda Q + l) \sin\left(\frac{2\pi k l}{L}\right), \tag{9}$$

where L samples are added after multiplication with modulation terms. The central limit theorem states that the distribution of the DFT coefficients will converge towards a Gaussian PDF regardless of the PDF of the time samples $y(l)$, if successive samples are statistically independent. This also holds if the correlation in $y(l)$ is short compared to the analysis frame size [14].

For many relevant acoustic noises this assumption holds. Moreover, multiple noise sources or reverberation often reduce the noise correlation in between the analysis frame size, so that the Gaussian assumption is fulfilled. The variance of the noise DFT coefficient σ_N^2 is assumed to split equally into real and imaginary parts. Thus, the probability density function of real and imaginary parts of noise Fourier coefficients can be modelled as

$$p(N_{\text{Re}}) = \frac{1}{\sqrt{\pi}\sigma_N} \exp\left\{-\frac{N_{\text{Re}}^2}{\sigma_N^2}\right\}. \tag{10}$$

Based on (10) and the assumption of statistically independent real and imaginary parts, the PDF of the noisy spectrum Y conditioned on the speech amplitude A and phase α can be written as joint Gaussian:

$$p(Y|A, \alpha) = \frac{1}{\pi\sigma_N^2} \exp\left(-\frac{|Y - Ae^{j\alpha}|^2}{\sigma_N^2}\right). \tag{11}$$

A Rice PDF is obtained for the density of the noisy amplitude given the speech amplitude A after polar integration of (11) [15]:

$$p(R|A) = \frac{2R}{\sigma_N^2} \exp\left\{-\frac{R^2 + A^2}{\sigma_N^2}\right\} I_0\left(\frac{2AR}{\sigma_N^2}\right), \tag{12}$$

where I_0 denotes the modified Bessel function of the first kind and zeroth order.

Considering speech, the span of correlation with typical frame sizes from 10 milliseconds to 30 milliseconds cannot be neglected. The smaller the frame size, the less Gaussian

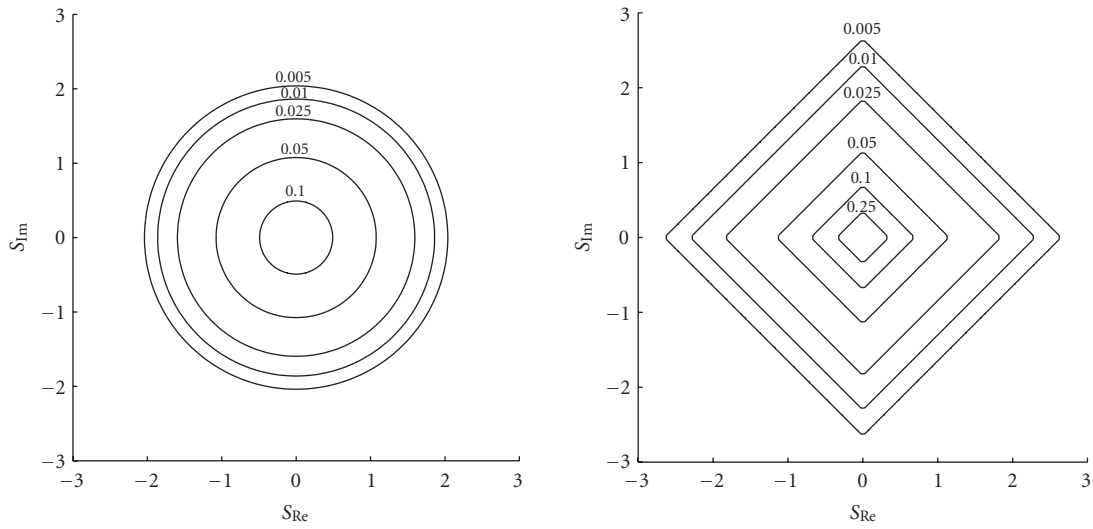


Figure 2: Contour lines of complex Gaussian model with independent Cartesian coordinates and of complex Laplace model with independent Cartesian coordinates ($\sigma_S^2 = 1$).

will the distribution of the speech real and imaginary parts of the Fourier coefficients will be. It is well known, that the PDFs of speech samples in the time domain are much better modelled by a Laplace or Gamma density [16]. In the frequency domain similar distributions can be observed. Martin [5, 6] has abandoned the Gaussian speech model according to

$$p(S_{\mathrm{Re}}) = \frac{1}{\sqrt{\pi}\sigma_S} \exp\left\{ -\frac{S_{\mathrm{Re}}^2}{\sigma_S^2} \right\}. \tag{13}$$

Instead, the Laplace probability density function

$$p(S_{\mathrm{Re}}) = \frac{1}{\sigma_S} \exp\left\{ -\frac{2\,|S_{\mathrm{Re}}|}{\sigma_S} \right\} \tag{14}$$

and Gamma PDFs for statistical independent real and imaginary parts have been proposed:

$$p(S_{\mathrm{Re}}) = \frac{\sqrt[4]{3}\,|S_{\mathrm{Re}}|^{-1/2}}{2\sqrt[4]{2}\sqrt{\pi\sigma_S}} \exp\left\{ -\frac{\sqrt{3}\,|S_{\mathrm{Re}}|}{\sqrt{2}\sigma_S} \right\}. \tag{15}$$

The same equations hold for the imaginary parts.

3.1. Modelling the spectral amplitudes

In the following a simple statistical model for the speech and noise spectral amplitudes will be presented [17], which is significantly closer to the real distribution than the commonly applied Gaussian model.

The spectral amplitudes are of special importance, because the phase of the Fourier coefficients can be considered unimportant from a perceptual point of view [7, 18]. Hence, spectral amplitude estimators are more advantageous and a statistical model for the amplitude alone is needed.

Considering noise, the Gaussian assumptions hold due to comparably low correlation in the analysis frame. Assuming statistical independence of real and imaginary parts the PDF of the noise amplitude B can easily be found as Rayleigh distributed by polar integration

$$p(B) = \int_0^{2\pi} B \cdot p(N_{\mathrm{Re}}, N_{\mathrm{Im}})\, d\beta = \frac{2B}{\sigma_N^2} \exp\left\{ -\frac{B^2}{\sigma_N^2} \right\}. \tag{16}$$

For the calculation of an appropriate PDF for A, the Gauss, Laplace, and Gamma PDFs for real and imaginary parts are taken into account. The real and imaginary parts of the Fourier coefficients can be considered statistically independent with high accuracy. Then, $p(A)$ can in general be calculated by

$$p(A) = \int_0^{2\pi} A \cdot p(A\cos\alpha) \cdot p(A\sin\alpha)\, d\alpha, \tag{17}$$

with the PDFs according to (13), (14), or (15) for $p(S_{\mathrm{Re}} = A\cos\alpha)$, $p(S_{\mathrm{Im}} = A\sin\alpha)$.

Figure 2 shows contour lines of a complex Gaussian or Laplace PDF with independent Cartesian components. Compared to the Gaussian PDF, the Laplace PDF has a higher peak, a low amplitude and decreases slower towards higher amplitudes visible by the greater distances of the contour lines compared to the complex Gaussian PDF. While the complex Gaussian PDF is rotational invariant, the Laplace amplitude depends on the phase.

Considering Gaussian components, the rotational invariance greatly facilitates the polar integration. Similar to (16) the amplitude is Rayleigh distributed:

$$p(A) = \frac{2A}{\sigma_S^2} \exp\left\{ -\frac{A^2}{\sigma_S^2} \right\}. \tag{18}$$

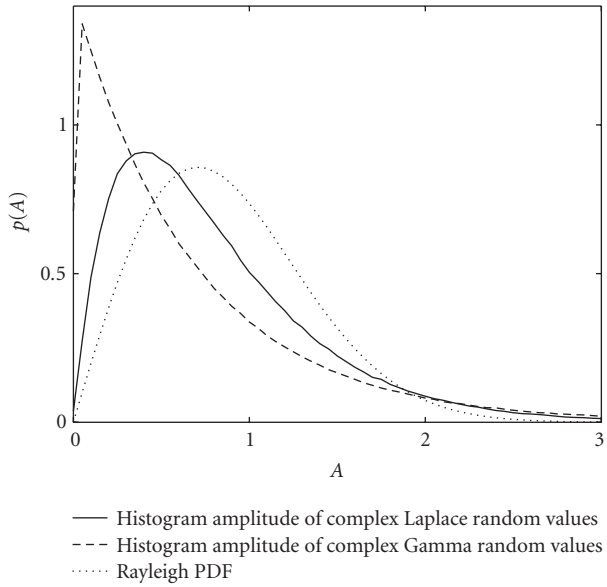


Figure 3: Measured histograms of amplitudes of complex 1.000.000 random variables with independent Cartesian Laplace (solid) or Gamma (dashed) components along with Rayleigh PDF (dotted) ($\sigma_S^2 = 1$).

The PDF of the amplitude of a complex Laplace or Gamma random variable with independent Cartesian components varies with the angle α. This makes an analytic calculation of the distribution $A = \sqrt{S_{\mathrm{Re}}^2 + S_{\mathrm{Im}}^2}$ for (14) or (15) difficult, if not impossible.

Instead of an analytic solution to (17) we are looking for a function that approximates the real PDF of the spectral amplitudes with high accuracy regardless of the underlying joint distribution of real and imaginary parts of the Fourier coefficients. However, as indication about how the function should look like the amplitude of a complex Laplace or Gamma PDF with independent components is taken into account.

Figure 3 plots histograms of the amplitude $A = \sqrt{S_{\mathrm{Re}}^2 + S_{\mathrm{Im}}^2}$ of 1.000.000 Laplace and Gamma, respectively, distributed independent random values S_{Re}, S_{Im} of variance $\sigma_S^2/2$. Whereas the Laplace-distributed random variables can easily be generated using the inverse distribution function method [19], the Gamma-distributed random values were generated according to [20]. Compared to the Rayleigh-distributed amplitude of a complex Gaussian random variable, low values are more likely, but the PDF decreases more slowly towards high values.

The fast decay of the Rayleigh PDF results from the second-order term of A in the argument of the exponential function in (18) similar to the decay of the Gauss function in (13). Similarly, the measured PDFs of the complex Laplace and Gamma amplitudes can be assumed to decay like (14) and (15) with a linear argument in the exponential function.

Apparently, the slope of the Gamma amplitude PDF differs from that of the Laplace amplitude PDF. Hence, a parameter μ is introduced, which enables to approximate both. After normalizing A by the standard deviation σ_S we thus assume

$$p(A) \sim \exp\left\{-\mu\frac{A}{\sigma_S}\right\}. \tag{19}$$

At low values of A the PDF of the Laplace and Gamma amplitudes is much higher than the Rayleigh PDF as shown in Figure 3. Considering the Rayleigh PDF according to (18), the behavior at low values is mainly due to the linear term of A, whereas the exponential term plays a minor role at small values.

Both the PDF of the Laplace amplitude and the PDF of the Gamma amplitude can be approximated by abandoning a linear term in A. Instead, A is taken to the power of a parameter ν after normalization to the standard deviation of speech, that is, $p(A) \sim (A/\sigma_S)^\nu$ in order to be able to approximate a large variety of PDFs. The smaller the parameter ν, the larger the proposed PDF at low values. The term hardly influences the behavior of the function at a high value due to the dominance of the exponential decay

$$p(A) \sim \frac{A^\nu}{\sigma_S^\nu}\exp\left\{-\mu\frac{A}{\sigma_S}\right\}. \tag{20}$$

After taking $\int_0^\infty p(A)dA = 1$ into account, the approximating function with parameters ν, μ is finally obtained using [21, equation 3.381.4]:

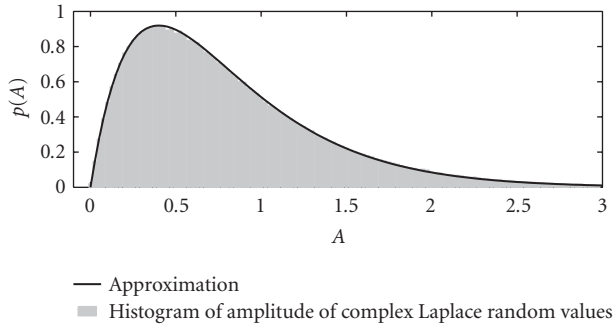
$$p(A) = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)}\frac{A^\nu}{\sigma_S^{\nu+1}}\exp\left\{-\mu\frac{A}{\sigma_S}\right\}. \tag{21}$$

Here, Γ denotes the Gamma function.

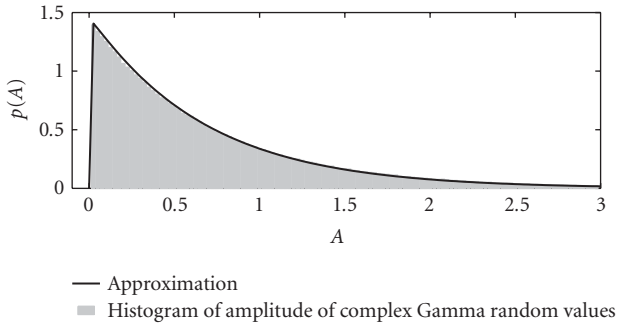
Figure 4 shows the approximation of the measured histogram of the amplitude of 1.000.000 complex Laplace or Gamma random values with independent components with $\sigma_S^2 = 1$ by (21) using different sets of parameters ν, μ. Apparently, (21) allows a very accurate approximation for both Laplace and Gamma components. To approximate the Laplace amplitude, we applied the parameter set ($\nu = 1$, $\mu = 2.5$). To approximate the Gamma amplitude we used ($\nu = 0.01$, $\mu = 1.5$). PDFs in between both or closer to the Rayleigh PDF can be approximated with different sets of parameters ν, μ.

3.1.1. Matching with experimental data

The real PDF of the speech amplitude will not be exactly like the Laplace or Gamma amplitude approximation but somewhere in between. Also, it will depend on parameters of the noise reduction system such as the analysis frame size. At a larger frame size the correlation decreases relative to the analysis frame size and thus the distribution will be less super-Gaussian. The task is therefore to find a set of parameters (ν, μ) which outperforms the above sets for Laplace or Gamma amplitude approximation for a given system.



(a)



(b)

Figure 4: Approximation of amplitudes of complex random values with Laplace and Gamma components using (21). (a) Laplace components: ($\nu = 1, \mu = 2.5$). (b) Gamma components: ($\nu = 0.01$, $\mu = 1.5$).

To measure the probability density function of the speech complex DFT coefficients S or speech DFT amplitudes A, a histogram is built using 1-hour speech from different speakers. Ideally, DFT bins, which solely contain speech of equal variance, should be taken into account.

In practice, the speech variance in a frequency bin is strongly time variant and can only be estimated in a time frame and frequency bin with a certain estimation error. Thus, we apply (6), which is commonly considered as the best performing method to estimate the speech variance in the form of the a priori SNR. Hereby, the histogram measurement process also incorporates the same method of estimating the time-varying speech variance as the noise reduction system. Data is collected for the histogram at time instances, when the frequency bin is dominated by speech. For that purpose a high and narrow a priori SNR interval is predefined, for example, 19–21 dB. The width of the interval is a tradeoff between the amount of data obtained and the demand to pick samples of same variance.

Figure 5a shows the contour lines of the measured speech DFT coefficients. The data shown has been obtained by building separate histograms for each frequency and normalizing each histogram to $\sigma_S^2 = 1$ for an averaged histogram over the frequency. Compared to the Gaussian contour lines in Figure 2, a slower decrease towards high amplitudes and faster increase towards low amplitudes is visible. Also, the observed data hardly shows any dependency on the phase as in the Laplace contour lines in Figure 2 as shown for the complex Laplace PDF in Figures 5b, 5c, 5d, 5e, 5e, 5f, and 5g which depict the histogram of phases for the six specific contour lines. Approximately, the phases can be considered as uniformly distributed. The variation visible for $A = 0.005$ is probably due to the low amount of data available here.

Figure 6a a plots the histogram of the speech amplitude, which is obtained by integration over the phase of the two-dimensional histogram along with the analytic Rayleigh PDF and the approximation according to (21) with the parameter set for Laplace and Gamma amplitude approximations, respectively. Figure 6b shows a zoom into the higher regions. Apparently, (21) provides a much better fit for the speech amplitude than the Rayleigh PDF for both Laplace and Gamma amplitude approximations. For low arguments, the Rayleigh PDF rises too slowly, while for large arguments, the density function decays too fast. The real PDF of the speech amplitude lies between the Laplace and Gamma amplitude approximations for the data measured with our system the Gamma amplitude approximation.

To find a set (ν, μ) that approximates the real PDF best, a distance measure between the analytic function and the histogram with N bins is numerically minimized. The Kullback divergence [22] can be considered optimal from an information theoretical point of view. Given two random variables of probability density $p_1(x)$ and $p_2(x)$, then $I(2:1)$ describes the mean information per observation of process 2 for discrimination in favor of process 2 and $I(1:2)$ for discrimination in favor of process 1:

$$I(1:2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx,$$
$$I(2:1) = \int p_2(x) \log \frac{p_2(x)}{p_1(x)} dx. \tag{22}$$

The sum $J(1:2) = I(1:2) + I(2:1)$ is a measure of divergence between the two processes. To differentiate between the analytical $p_A(n)$ and the histogram PDF $p_h(n)$ with N bins, the divergence can be calculated by

$$J(A:h) = \sum_{n=1}^{N} (p_h(n) - p_A(n)) \log \left(\frac{p_h(n)}{p_A(n)} \right). \tag{23}$$

Figure 7 shows the best $p(A)$ according to (21) determined by minimizing the Kullback divergence. The analytical PDF now fits even better to the observed data than the Laplace or Gamma amplitude approximation. To illustrate the improvement provided by the new model, Table 1 shows the Kullback divergences between measured data and model functions. The divergences have been normalized to that of the Rayleigh PDF, that is, the Gaussian model. When using the Laplace or Gamma amplitude approximation, the Kullback divergence is significantly lower than that for the Gaussian model. By determining an optimal parameter set, the divergence further decreases.

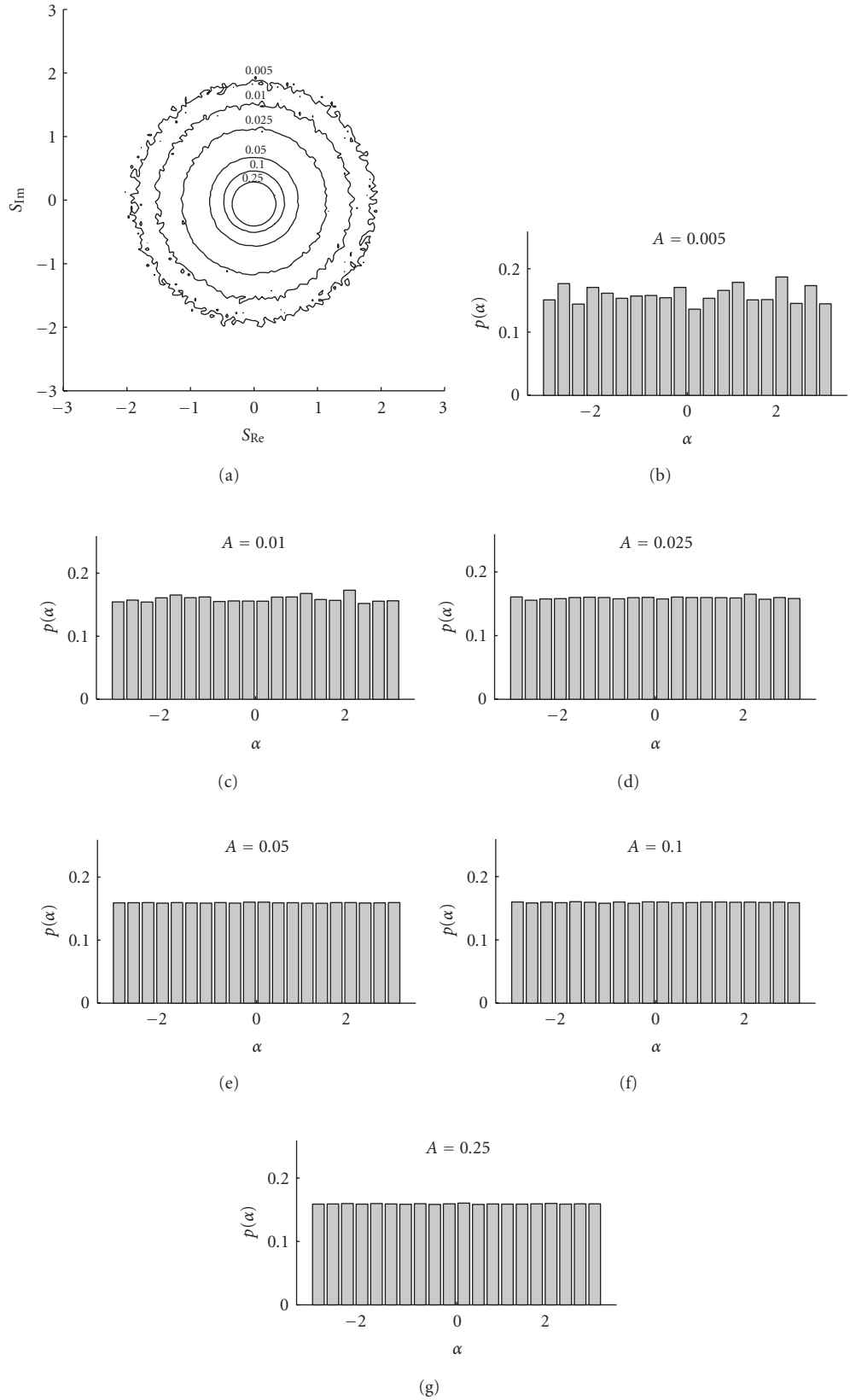


Figure 5: (a) Contour lines of measured speech DFT coefficients. ((b), (c), (d), (e), (f), (g)) Histogram of speech DFT phases for six different amplitudes.

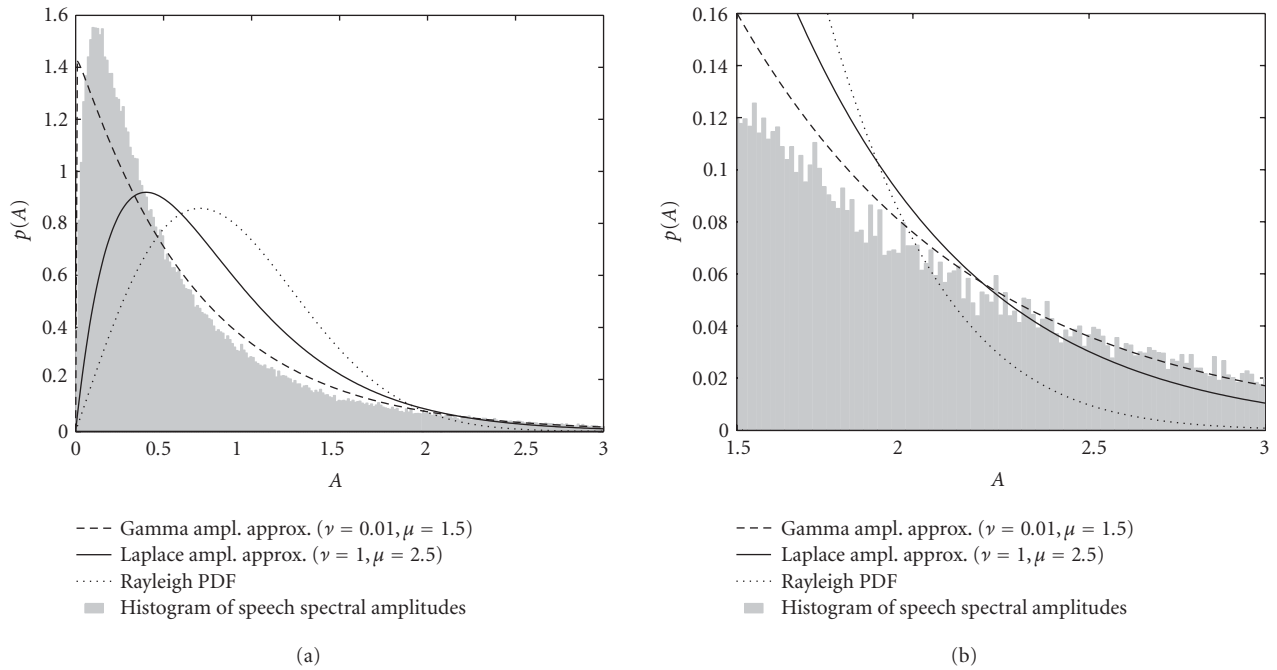


Figure 6: (a) Histogram of speech DFT amplitudes A ($\sigma_S^2 = 1$) fitted with Rayleigh PDF and Laplace/Gamma amplitude approximation (21). (b) Zoom into the area $1.5 \leq A \leq 3$.

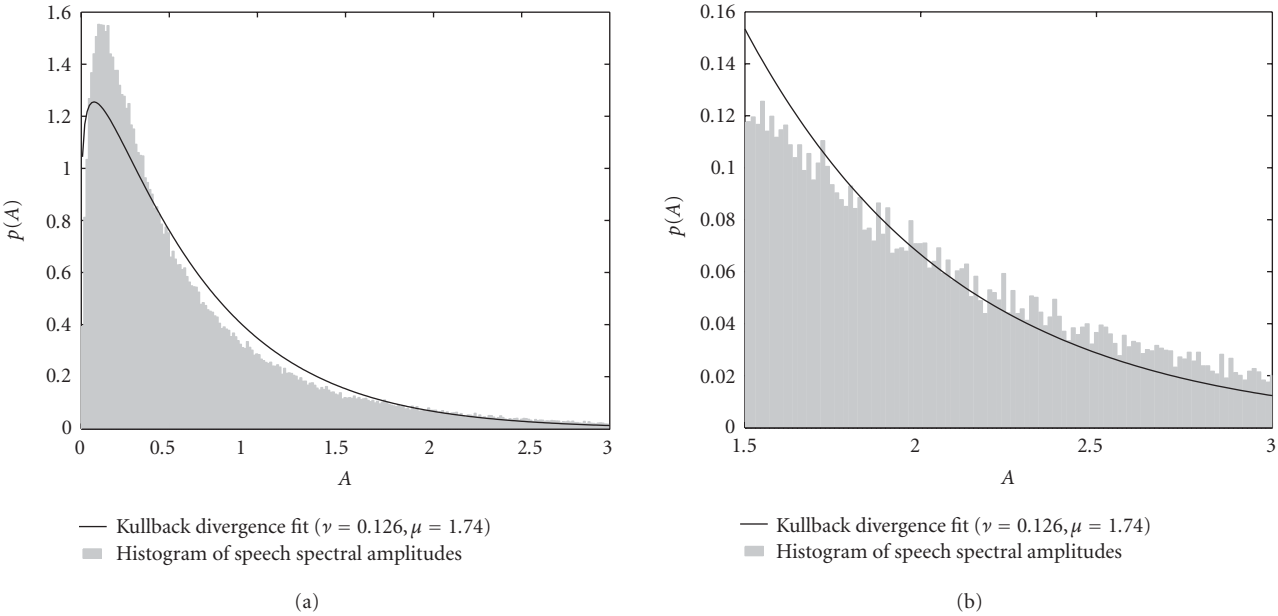


Figure 7: (a) Histogram of speech DFT amplitudes and fitted approximation by (21) according to Kullback divergence ($\sigma_S^2 = 1$). (b) Zoom into the area $1.5 \leq A \leq 3$.

3.1.2. Reverberant signal

The acoustic environment will influence the distribution of the speech spectral amplitude. Especially if the desired acoustic source is located at larger distances from the microphone, for example, in a hearing aid application, reverberation will degrade the amount of correlation in between an analysis frame and thus will lead to a less super-Gaussian distribution.

To examine the amount of influence of reverberation, the scenario depicted in Figure 8 is considered. The acoustical impulse response in a reverberant room from a source to a microphone was simulated with the image method [23], which models the reflecting walls by several image sources.

Table 1: Normalized Kullback divergence between measured speech PDF and different model functions.

$p(A)$	ν, μ	$J(A:h)/J(A:h)_{\text{Rayleigh}}$
Rayleigh (18)	—	1
Laplace amplitude approximation (21)	1, 2.5	0.35
Gamma amplitude approximation (21)	0.01, 1.5	0.05
Kullback fit (21)	0.126, 1.74	0.045

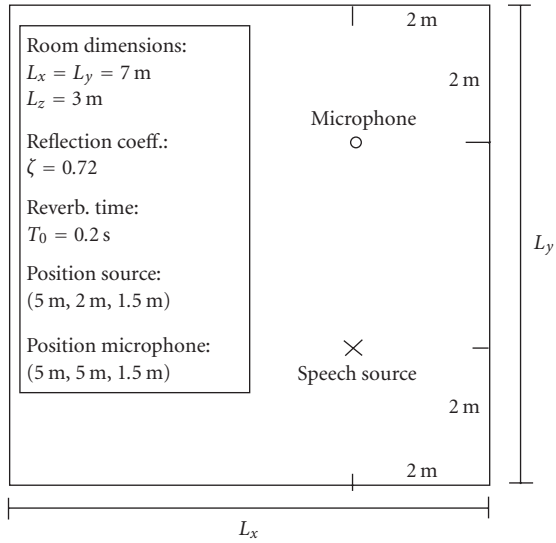


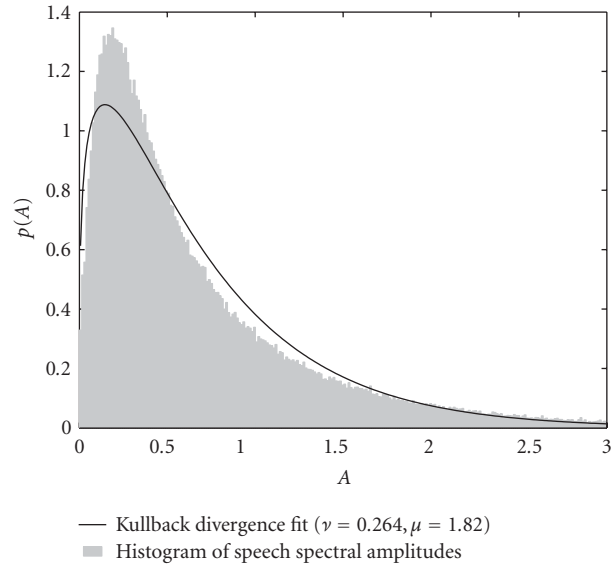
Figure 8: Simulation of impulse response between speech source and microphone in a reverberant room using the image method.

The intensity of the sound from an image source at the microphone array is determined by a frequency-independent reflection coefficient ζ and by the distance to the microphone. In our experiment, the reverberation time was set to $T_0 = 0.2$ seconds, which corresponds to a reflection coefficient of $\zeta = 0.72$ according to Eyring's formula

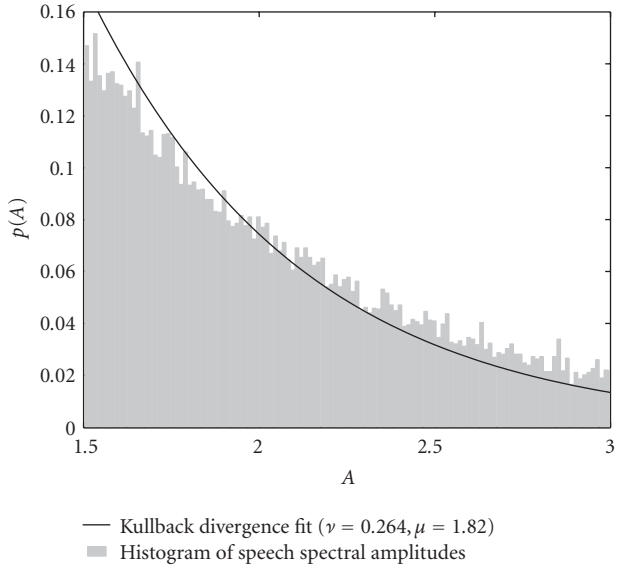
$$\zeta = \exp\left\{-13.82/\left(c\left(\frac{1}{L_x}+\frac{1}{L_y}+\frac{1}{L_z}\right)T_0\right)\right\}. \quad (24)$$

The histogram of the speech amplitude was then taken as before after convolving the database of speech with the impulse response delivered by the image method.

Figure 9 plots the histogram along with the approximation with parameters fitted according to the Kullback divergence. As expected, the speech spectral amplitude is now less super-Gaussian distributed. However the optimal parameters with respect to the Kullback divergence (i.e., $\nu = 0.264$, $\mu = 1.82$) are still much closer to the values originally obtained from the Kullback fit than to those of the Laplace amplitude approximation or even from the Rayleigh PDF. It can be concluded that accuracy of the statistical model is only slightly affected by reverberation. Whereas a slight performance gain can be expected when adapting the parameters of the statistical model during run-time, the gain might not justify the additional computational complexity of an acoustic classifier. Thus, in the following the fixed parameter set ($\nu = 0.126$, $\mu = 1.74$) is considered as optimal.



(a)



(b)

Figure 9: (a) Histogram of speech amplitudes in reverberant room and fitted approximation (21) according to Kullback divergence ($\sigma_S^2 = 1$). (b) Zoom into the area $1.5 \le A \le 3$.

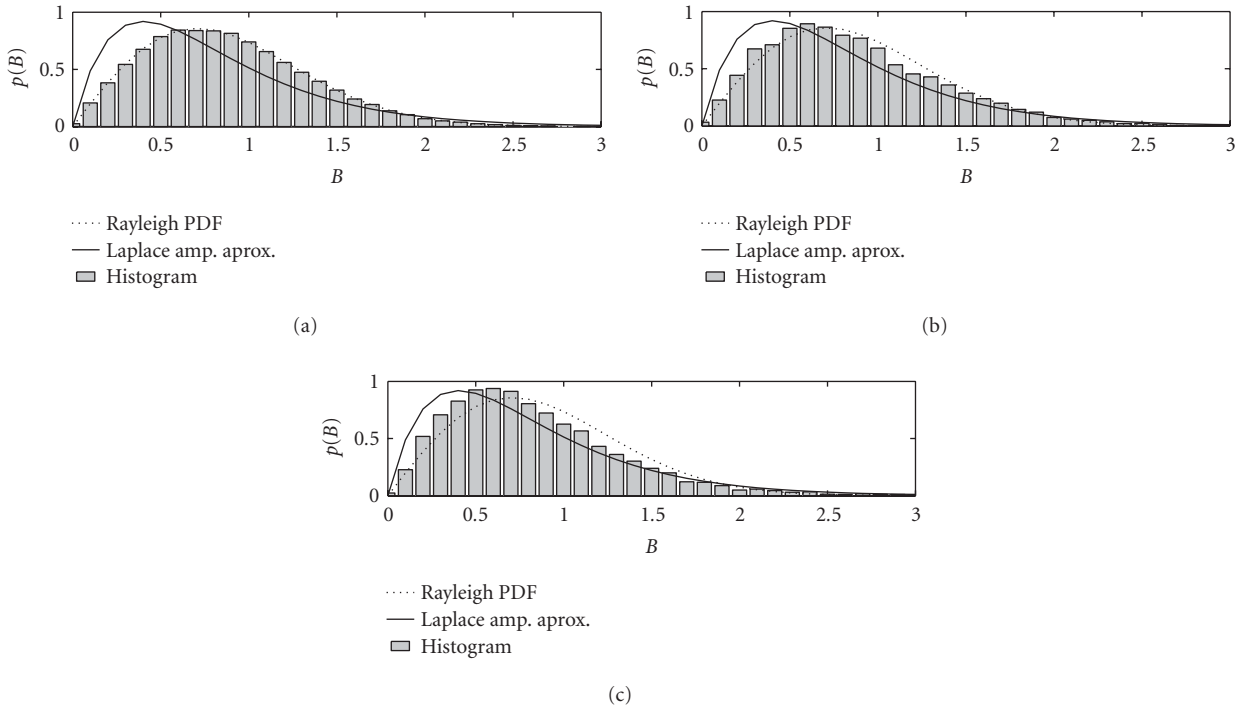


Figure 10: Histogram of noise DFT amplitudes B for (a) white uniform distributed noise, (b) fan noise, and (c) cafeteria noise ($\sigma_N^2 = 1$) fitted with Rayleigh PDF and Laplace amplitude approximation.

3.1.3. Spectral amplitude of noise

Compared to speech, the span of noise correlation in an analysis frame is much lower. Thus, the PDF of the real and imaginary parts of the noise spectral coefficients will according to the central limit theorem be closer to a Gaussian function. Martin [5, 6] has proposed spectral estimators with Laplace or Gaussian noise model (and Laplace and Gamma models for the speech coefficients). A Laplace model for noise is motivated by the observation that environmental noises are also super-Gaussian distributed to a certain degree. Figure 10 plots histograms of DFT amplitudes measured for three different noise classes. For building the histograms, the frequency- and time-dependent noise variances σ_N^2 were estimated using the same system as applied in the noise reduction algorithm, that is, minimum statistics [2]. Spectral amplitudes with corresponding estimated noise variances inside a narrow predefined interval were then collected for the histogram database. To plot the histogram together with the Rayleigh function (18) and the super-Gaussian model function (21) in Figure 10 the collected database was normalized to $\sigma_N^2 = 1$.

For the white noise, which was uniformly distributed in the time domain, a Rayleigh function perfectly models the PDF of the noise spectral amplitude. This is because there is no correlation in a time frame, resulting in Gaussian-distributed real and imaginary parts of Fourier coefficients according to the central limit theorem. For fan noise, the PDF slightly changes towards the Laplace amplitude approximation, while the effect is more visible for the cafeteria noise, which contains speech components from many speakers. The deviation for the measured histogram from the Rayleigh model is low compared to that of speech. In the following, the Gaussian assumption for the noise will therefore be kept.

4. SPEECH ESTIMATORS

The task of the speech estimator lies in calculating an estimate for the speech spectral amplitude $\hat{A} = G \cdot R$ given the observed noisy coefficient Y or the noisy amplitude R and the variances of speech σ_S^2 and noise σ_N^2. With probability one, the estimate will not be identical to the real value, therefore a cost function $C(A, \hat{A})$ is introduced [24], which assigns a value to each combination of undisturbed and estimated speech spectral amplitudes. The Bayesian estimators aim at minimizing the expectation of the cost according to

$$E\{C(A,\hat{A})\} = \int_{-\infty}^{\infty}\int_{0}^{\infty} C(A,\hat{A})\, p(A,Y)\, dA\, dY. \tag{25}$$

For $C(A,\hat{A}) = (A-\hat{A})^2$ the Ephraim-Malah or conditional expectation estimator [3] is obtained:

$$G = \frac{\sqrt{\nu}}{\gamma} \cdot \Gamma(1.5) F_1(-0.5, 1, -\nu), \quad \nu = \gamma \frac{\xi}{1+\xi}, \tag{26}$$

where the confluent hypergeometric series F_1 can be calculated with

$$F_1(-0.5, 1, -\nu) = e^{-\nu/2}\left[(1+\nu) I_0\left(\frac{\nu}{2}\right) + \nu I_1\left(\frac{\nu}{2}\right)\right], \tag{27}$$

where I_0, I_1 denote the modified Bessel function of zeroth and first order. The cost function $C(A, \hat{A}) = \log A - \log \hat{A}$ leads to the logarithmic Ephraim-Malah estimator [4]. Alternatively the β-order MMSE estimator [25] allows an estimation in between both rules.

By choosing a uniform cost function according to

$$C = \begin{cases} 0, & |S - \hat{S}| < \epsilon, \\ 1, & \text{else.} \end{cases} \tag{28}$$

MAP estimators can be obtained, which are in general computationally more efficient.

Wolfe and Godsill [8, 26] introduced alternatives to the Ephraim-Malah spectral amplitude estimator based on the maximum a posteriori estimation rule. The spectral weights obtained by the MAP estimators are similar to those of the Ephraim and Malah estimator, thus a quality improvement cannot be expected. However, straightforward implementations without the use of computational expensive Bessel or exponential function are possible.

In the following, we introduce two speech spectral amplitude estimators, which keep the computational simplicity of the Wolfe and Godsill estimators but also achieve a quality gain by applying the super-Gaussian speech model according to (21) and a Gaussian model for noise.

First, a MAP estimator for the speech spectral amplitude is derived. Secondly, a joint MAP estimator for the amplitude and phase is introduced. Both estimators are extensions of the MAP estimators proposed by [8].

4.1. MAP spectral amplitude estimator

A computationally efficient MAP solution following

$$\hat{A} = \arg\max_A p(A|R) = \arg\max_A \frac{p(R|A)p(A)}{p(R)} \tag{29}$$

similar to [26], where Gaussian-distributed S_{Re}, S_{Im} are assumed, can be found. Now, the super-Gaussian function (21) is used to model the PDF of the speech spectral amplitude $p(A)$. The Gaussian assumption of noise allows to apply (12) for $p(R|A)$. 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

$$I_0(x) \approx \frac{1}{\sqrt{2\pi x}} e^x. \tag{30}$$

Figure 11 shows that the approximation is reasonable for larger arguments and becomes erroneous for low arguments.

After insertion of (30) and (21) in (12) we get

$$p(R|A)p(A) \sim A^{\nu-1/2} \exp\left\{ -\frac{A^2}{\sigma_N^2} - A\left(\frac{\mu}{\sigma_S} - \frac{2R}{\sigma_N^2}\right)\right\}. \tag{31}$$

Note that the approximation of the Bessel function has introduced a negative exponent for $\nu > 0.5$.

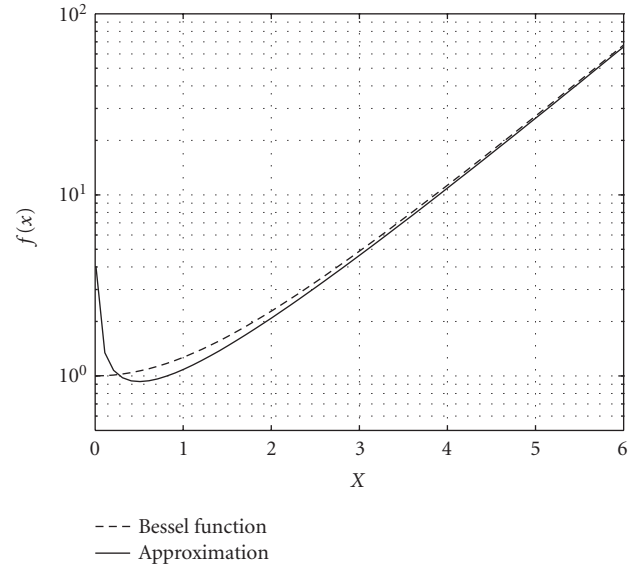


Figure 11: Modified Bessel function of zeroth-order $f(x) = I_0(x)$ and approximation (30), $f(x) = (1/\sqrt{2\pi x})e^x$.

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:

$$\frac{d\log[p(R|A)p(A)]}{dA} = \left(\nu - \frac{1}{2}\right)\frac{1}{A} - \frac{2A}{\sigma_N^2} - \frac{\mu}{\sigma_S} + \frac{2R}{\sigma_N^2} \stackrel{!}{=} 0. \tag{32}$$

After multiplication with A, one reasonable solution $\hat{A} = GR$ to the quadratic equation is found, because the second solution delivers spectral amplitudes $A < 0$ at least for $\nu > 0.5$. The second derivative at $\hat{A}$ is negative, thus a local maximum is guaranteed:

$$G = u + \sqrt{u^2 + \frac{\nu - 1/2}{2\gamma}}, \quad u = \frac{1}{2} - \frac{\mu}{4\sqrt{\gamma\xi}}. \tag{33}$$

Whereas the MAP spectral amplitude estimator is very useful 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 (30). 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 (33).

Figure 12 shows the dependency of the weights on the a posteriori SNR γ for two a priori SNRs ξ for the parameter set (ν, μ), that approximates the amplitude of a complex Laplace PDF. Most of the time, the weights of the super-Gaussian 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 a posteriori SNRs the Ephraim-Malah weights converge towards the Wiener weights, that is, $\xi/(1 + \xi)$. The weights of the super-Gaussian MAP estimator however increase due to the slower

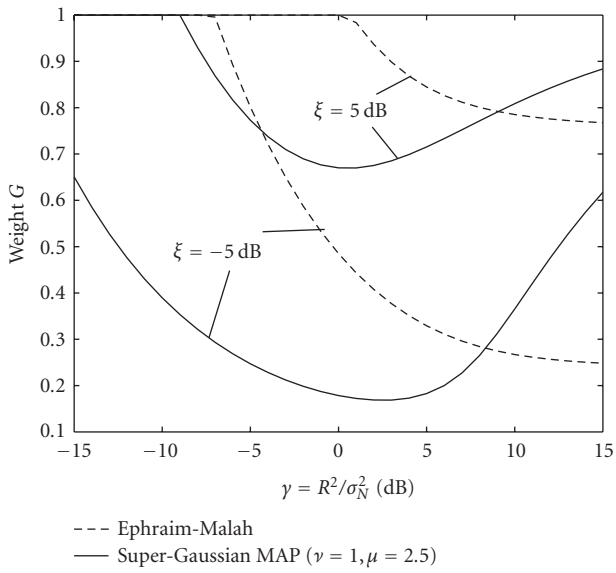


Figure 12: Weights of the super-Gaussian MAP estimator with Laplace amplitude approximation ($\nu = 1$, $\mu = 2.5$) compared to the Ephraim-Malah weighting rule depending on the a posteriori SNR γ for two a priori SNRs $\xi = -5$ dB and $\xi = 5$ dB.

decay of the model function towards larger values. Higher observed spectral amplitudes R will result in a higher spectral output compared to the Wiener filter or Ephraim-Malah estimator. This effect is due to the underlying more accurate statistical model of the spectral amplitude of speech, in which high amplitudes are considered more likely than in the Rayleigh model. Consequently, high observed noisy amplitude will be judged to contain more speech components by the super-Gaussian MAP estimator.

4.2. Joint MAP amplitude and phase estimator

To overcome the inability of the proposed MAP estimator with approximation of the Bessel function to cope with an underlying Gamma model or the model that minimizes the Kullback divergence towards the measured data, we introduce a joint MAP estimator of the amplitude and phase. 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, that is, $p(A, \alpha|Y)$:

$$\hat{A} = \arg\max_A p(A, \alpha|Y) = \arg\max_A \frac{p(Y|A,\alpha)p(A,\alpha)}{p(Y)},$$
$$\hat{\alpha} = \arg\max_\alpha p(A, \alpha|Y) = \arg\max_\alpha \frac{p(Y|A,\alpha)p(A,\alpha)}{p(Y)}. \tag{34}$$

If the problem is formulated this way, the Bessel function and its erroneous approximation are avoided. $p(Y|A, \alpha)$ is given by (11) using the Gaussian assumption of noise. Up to now we have only dealt with the probability of the speech amplitude, that is, $p(A)$, while the joint PDF of the amplitude and phase $p(A, \alpha)$ is now required. For a rotational invariant PDF,

$$p(A, \alpha) = \frac{1}{2\pi} p(A). \tag{35}$$

Formulas (34) can be solved similar to the MAP estimator. Again, the natural logarithm greatly facilitates the optimization process. After insertion of (11) and (21) we get

$$\log\left(p(Y|A,\alpha)p(A,\alpha)\right) = \log\left(\frac{\mu^{\nu+1}}{2\pi^2\sigma_N^2\sigma_S^{\nu+1}\Gamma(\nu+1)}\right) - \frac{\left|Y - Ae^{j\alpha}\right|^2}{\sigma_N^2} + \nu\log A - \mu\frac{A}{\sigma_S}. \tag{36}$$

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. Differentiating with respect to α yields

$$\frac{\delta}{\delta\alpha}\log\left(p(Y|A,\alpha)p(A,\alpha)\right) = -\frac{\left(Y^* - Ae^{-j\alpha}\right)\left(-jAe^{j\alpha}\right) + \left(Y - Ae^{j\alpha}\right)\left(jAe^{-j\alpha}\right)}{\sigma_N^2}. \tag{37}$$

Setting to zero and substituting $Y = Re^{j\vartheta}$ yields

$$\hat{\alpha} = \vartheta. \tag{38}$$

The candidate for the joint MAP phase estimate is simply the noisy phase. Differentiating with respect to the speech amplitude gives

$$\frac{\delta}{\delta A}\log\left(p(Y|A,\alpha)p(A,\alpha)\right) = \frac{\left(Y^* - Ae^{-j\alpha}\right)e^{j\alpha} + \left(Y - Ae^{j\alpha}\right)e^{-j\alpha}}{\sigma_N^2} + \frac{\nu}{A} - \frac{\mu}{\sigma_S}. \tag{39}$$

Setting to zero and replacing $\alpha = \vartheta$, the following quadratic equation is obtained:

$$A^2 + A\left(\frac{\mu\sigma_N^2}{2\sigma_S} - R\right) - \frac{\nu}{2}\sigma_N^2 \stackrel{!}{=} 0. \tag{40}$$

Solving the equation leads to an estimation rule similar to that of the super-Gaussian MAP estimator:

$$G = u + \sqrt{u^2 + \frac{\nu}{2\gamma}}, \quad u = \frac{1}{2} - \frac{\mu}{4\sqrt{\gamma\xi}}. \tag{41}$$

Again, checking the second derivatives guarantees that the extremum found by (41) is a local maximum. Figures 13 and 14 plot the weights of the joint MAP estimator in dependence on the a posteriori SNR for two different a priori SNRs and different set of parameters (ν, μ), that is, Laplace and Gamma amplitude approximations as well as Kullback divergence matching.

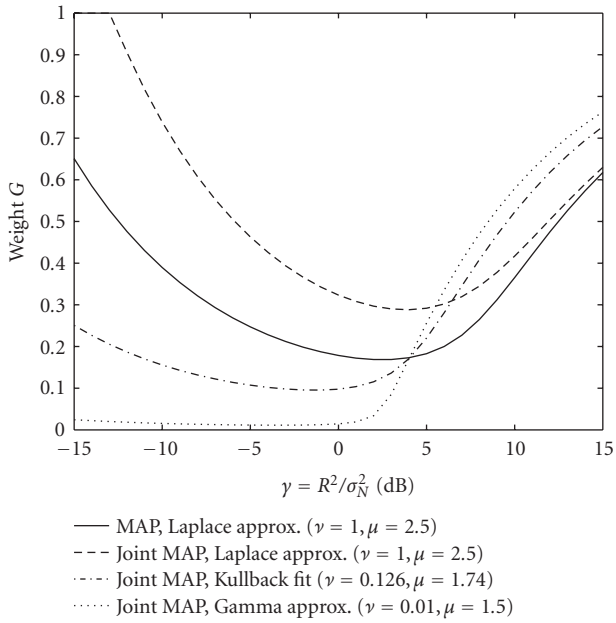


Figure 13: Weights of the joint MAP estimator as a function of the a posteriori SNR γ with different parameter sets, that is, Laplace and Gamma amplitude approximations as well as Kullback divergence matching, compared to the MAP estimator with Laplace approximation model for $\xi = -5$ dB.

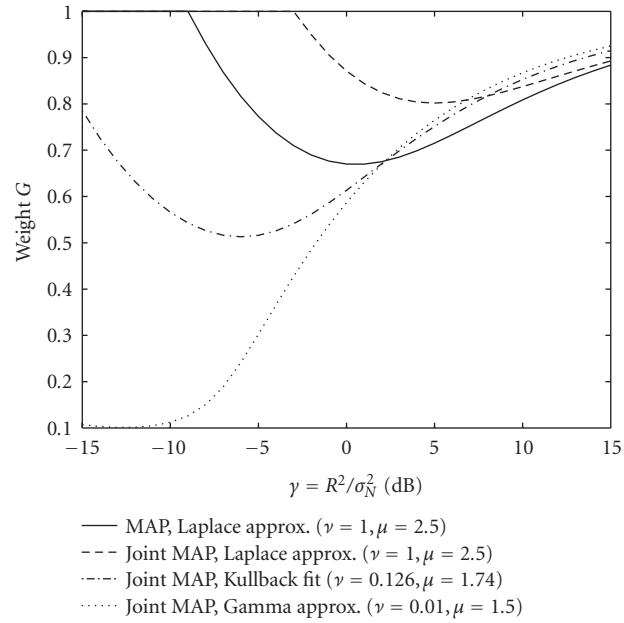


Figure 14: Weights of the joint MAP estimator as a function of the a posteriori SNR γ with different parameter sets, that is, Laplace and Gamma amplitude approximations as well as Kullback divergence matching, compared to the MAP estimator with Laplace approximation model for $\xi = 5$ dB.

For comparison the weights of the MAP estimator with Laplace amplitude approximation are also plotted. The weights of the joint MAP estimator with Laplace approximation model are always higher than that of the MAP amplitude estimator. Using the Gamma amplitude approximation or the Kullback fit, the weighting rule delivers significantly lower values at low observed SNRs. Moreover, the weights rise faster towards higher a posteriori SNRs compared to the Laplace estimation. This behavior is directly due to the different underlying statistical models of the speech amplitude by using different parameters (ν, μ) in (21). Low observed a posteriori SNRs compared to the ratio of variances in the form of the a priori SNR will highlight the effect of the statistical model at low values of A, while the behavior at high a posteriori SNRs will be influenced by the values of the PDF towards high speech spectral amplitudes. Since the Gamma amplitude approximation model assumes the highest values of the speech spectral amplitude PDF at low amplitudes and also shows the slowest decay towards high amplitude, its resulting weight rule deviates most from the Ephraim-Malah rule both at low and high a posteriori SNRs.

Comparison of computational burden

Table 2 lists the computational burden of the proposed estimators compared to other existing rules in the form of basic operations, and the evaluation of functions. A differentiation has been made between common functions like square root or exponential function, which are in digital signal processors OFTEN realized by dedicated memory tables and other more exotic functions, which are hardly considered for real-time implementations.

Among the estimators that apply a Gaussian model of speech and noise, the Wiener filter requires by far the fewest computations. The Ephraim-Malah spectral amplitude estimator needs to evaluate a square root, an exponential function, and also two Bessel functions. The MAP estimators derived by Wolfe can be realized at significantly less computations.

Considering the spectral estimators with super-Gaussian speech model, Martin's Laplace-Gauss estimator requires some divisions and a special function to be evaluated four times, especially because the estimation rule has to be executed independently for both real and imaginary parts. The proposed super-Gaussian estimators consume one square root operation more than the efficient Wolfe estimator. In a real-time implementation, the special functions for the Ephraim-Malah or Martin estimator will be realized as lookup tables. Such a table can be spared when using the proposed estimators.

5. EXPERIMENTAL RESULTS

While in informal listening tests, the super-Gaussian estimators seem to deliver a higher noise reduction at a similar speech quality compared to the Ephraim-Malah estimator, we also evaluate the performance by instrumental measurements.

Table 2: Computations required for different estimation rules (for each frequency bin).

Estimation rule	Add	Multiply	Divide	Function	Special functions
Wiener rule	1	—	1	—	—
Ephraim-Malah MMSE	3	8	2	Sqrt. (1x), exp (1x)	Bessel-fct. (2x)
Martin Laplace-Gauss	10	4	7	Sqrt. (2x)	Scaled compl. error-fct. (4x)
Wolfe MAP	4	3	2	Sqrt. (1x)	—
Super-Gaussian MAP	3–4	3	2	Sqrt. (2x)	—

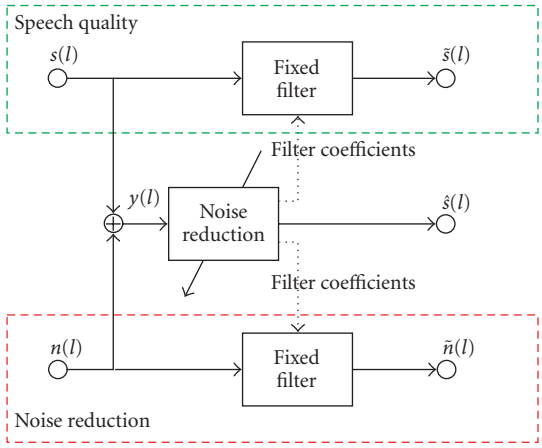


Figure 15: Instrumental performance evaluation of the noise reduction system.

The Ephraim-Malah MMSE estimator was taken as a reference, because it is considered as the best-performing speech spectral amplitude estimator. The MAP estimator derived by Wolfe results in approximately the same spectral weight, which can be calculated with much less computations. A detailed discussion about the difference in spectral weights and performance between the MAP estimators and the Ephraim-Malah rule can be found in [8]. The behavior of the proposed super-Gaussian MAP estimators with respect to the Ephraim-Malah reference is similar to the performance gain obtained by Martin's complex spectrum estimators [5, 6] with Laplace and Gamma speech model with respect to the Wiener reference. Some additional performance gain can be expected when the parameters of the super-Gaussian model function are optimally adjusted to the real distribution. Also , the resulting estimation rule is much more simple for the proposed super-Gaussian spectral amplitude estimators. Compared to approaches that model the DFT coefficient vector with Gaussian mixture models [27], the proposed estimators require less training in advance.

The noise reduction filter was applied to a speech signal with additive noise at different SNRs. To measure the quality of the filter, the system described in [28, 29] depicted in Figure 15 was applied to judge the performance of a noise reduction algorithm. The desired signal s and the interfering undesired signal n are superposed with a given SNR. The noisy signal $y(l)$ 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. Using the master-slave system depicted in Figure 15 the speech quality is tracked using the segmental signal-to-noise ratio, that is,

$$\text{segmental speech SNR/dB} = \frac{1}{P}\sum_{p=1}^{P}\left(10\cdot\log_{10}\left(\frac{\sum_{i=1}^{I} s^2(i+pI)}{\sum_{i=1}^{I}\left(s(i+pI)-\tilde{s}(i+pI)\right)^2}\right)\right). \quad (42)$$

Here M is the length of the signal, I denotes the length of the segment and P the number of segments, such that $P\cdot I = M$. On the other hand, the noise reduction amount is measured in terms of segmental noise power attenuation as

$$\text{segmental noise reduction/dB} = \frac{1}{P}\sum_{p=1}^{P}\left(10\cdot\log_{10}\left(\frac{\sum_{i=1}^{I} n^2(i+pI)}{\sum_{i=1}^{I}\tilde{n}^2(i+pI)}\right)\right). \quad (43)$$

To highlight the noise reduction during speech we only take segments p with global speech activity into account. The global activity is detected in advance by applying a VAD on the clean speech signal. The parameters (ν, μ) 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. If the parameters are adjusted for Gamma-distributed components or in order to minimize the Kullback divergence, the enhanced signal is greatly disturbed. This is due to the approximation of the Bessel function, which generates an uncompensated pole at $A = 0$ for $\nu < 0.5$. In general, the proposed super-Gaussian MAP estimator cannot be applied for $\nu < 0.5$.

The super-Gaussian joint MAP estimator however can be applied to every reasonable set of parameters (ν, μ). Here, we favor the parameters that were determined by minimizing the Kullback divergence towards the measured data, that is, $(\nu = 0.126, \mu = 1.74)$.

The amount of noise reduction using (33) with $(\nu = 1, \mu = 2.5)$ or (41) with $(\nu = 0.126, \mu = 1.74)$ is significantly higher than that for the Ephraim-Malah algorithm. The more super-Gaussian the statistical model for the speech spectral amplitude, the higher the noise reduction. Consequently, a lower speech quality will be reached. Comparing speech quality and noise reduction of the super-Gaussian

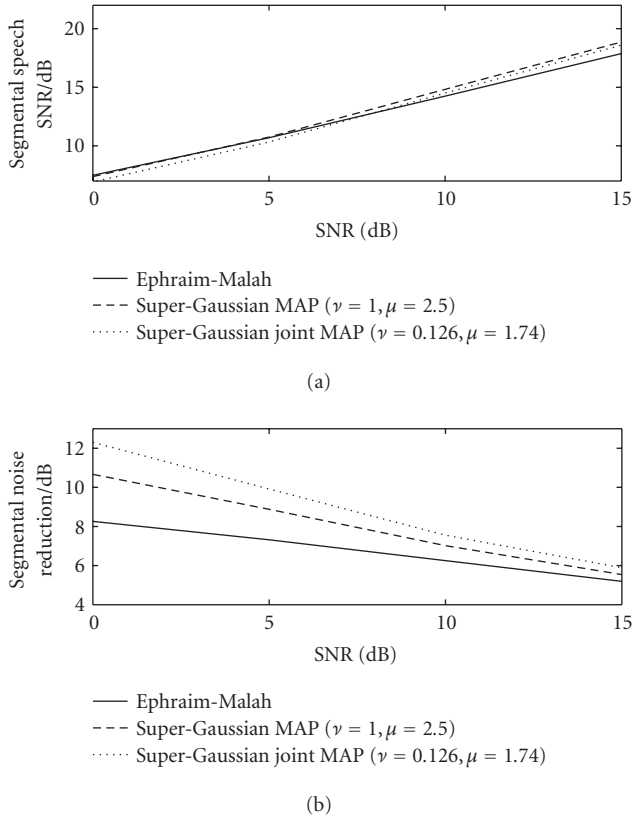


Figure 16: Speech quality and noise reduction amount of statistical filter with Ephraim-Malah estimator (solid), super-Gaussian MAP estimator (dashed), and super-Gaussian joint MAP estimator (dotted) for speech corrupted with white noise.

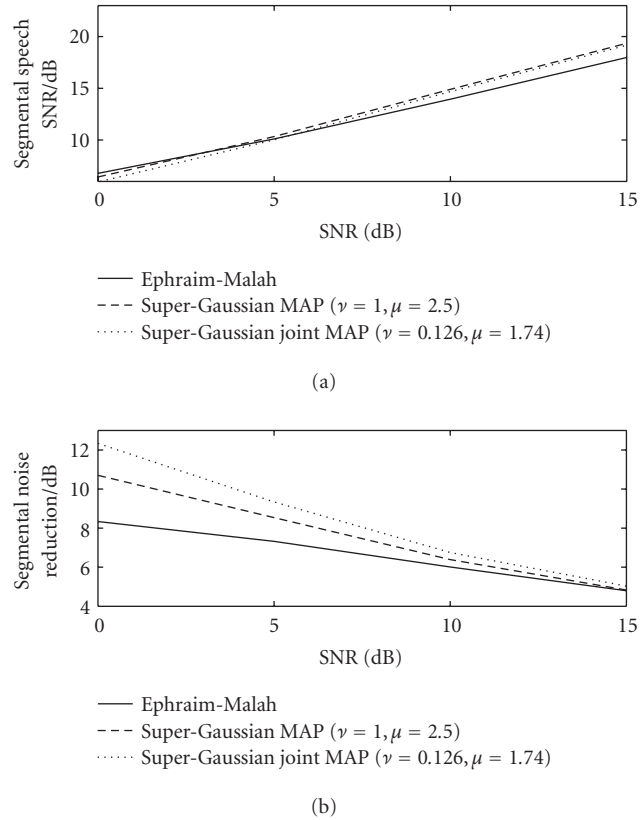


Figure 17: Speech quality and noise reduction amount of statistical filter with Ephraim-Malah estimator (solid), super-Gaussian MAP estimator (dashed), and super-Gaussian joint MAP estimator (dotted) for reverberant speech corrupted with white noise.

estimators to the Ephraim-Malah estimator would thus be of limited value. For comparability the weights of the super-Gaussian estimators are scaled by a constant factor greater than 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 [3] with tracking speech absence probabilities [30].

In the following, different experiments are documented. First, the system is applied to the speech disturbed by white noise at different SNRs and the performance when using the Ephraim-Malah estimator, the super-Gaussian MAP estimator with Laplace amplitude approximation, and the super-Gaussian joint MAP estimator with optimal parameters is compared. The experiment is then extended to reverberant speech with additive white noise. Thirdly, the experiments are conducted with fan noise and finally, the performance of the estimators is compared with the speech disturbed by cafeteria noise.

5.1. Performance in white noise

The results for white noise and the three different estimators, that is, Ephraim-Malah, MAP with ($\nu = 1$, $\mu = 2.5$), and joint MAP with ($\nu = 0.126$, $\mu = 1.74$) are shown in Figure 16. 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.

Generally, the single-microphone noise reduction system is comparably robust against reverberation. However, reverberation will degrade its performance, especially because it is harder for the noise estimation algorithm to differentiate between noise and weak reverberating parts of the speech. While this will degrade the performance of all estimators, the proposed super-Gaussian estimators are also affected by the change of distribution of the speech DFT coefficients as shown in Figure 9. To examine the performance of the proposed estimators, the acoustic scenario depicted in Figure 8 was simulated using the image method. The clean speech was filtered with the impulse response delivered by the image method and was processed by the noise reduction algorithm after adding white noise at different SNRs.

Figure 17 plots the performance in terms of instrumental speech quality and noise reduction. The reverberation hardly affects the performance gain provided by the super-Gaussian estimators. Still a significant advantage compared to the Ephraim-Malah estimator can be expected. Also, the

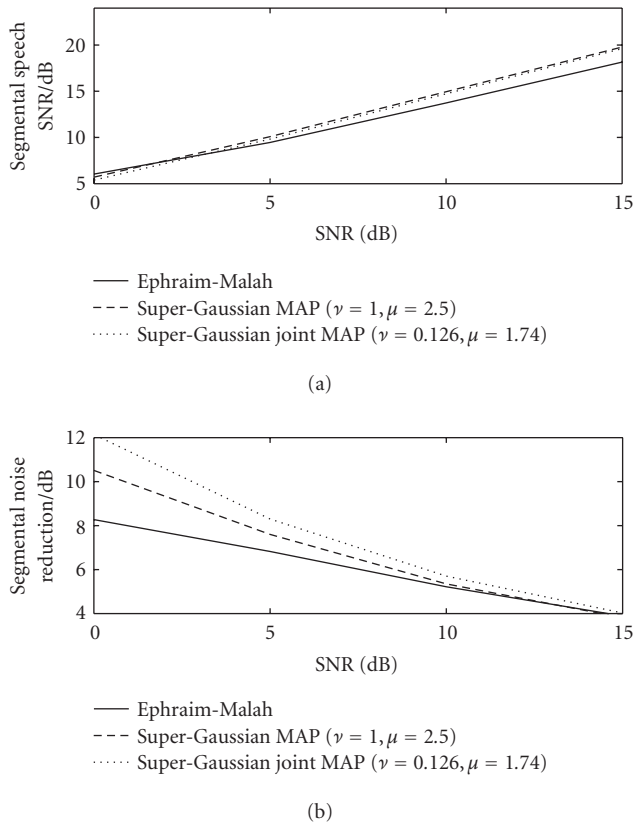


FIGURE 18: Speech quality and noise reduction amount of statistical filter with Ephraim-Malah estimator (solid), super-Gaussian MAP estimator (dashed), and super-Gaussian joint MAP estimator (dotted) for speech corrupted with fan noise.

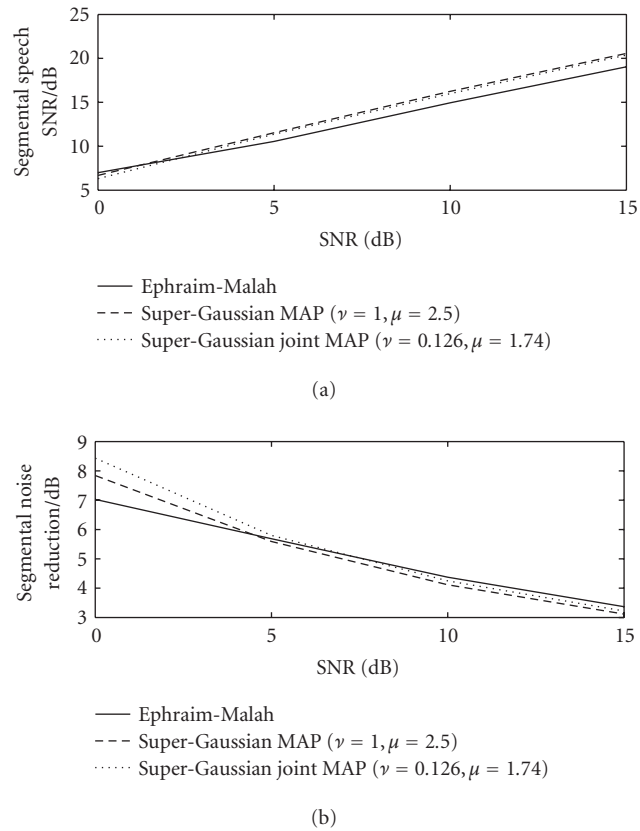


FIGURE 19: Speech quality and noise reduction amount of statistical filter with Ephraim-Malah estimator (solid), super-Gaussian MAP estimator (dashed), and super-Gaussian joint MAP estimator (dotted) for speech corrupted with cafeteria noise.

joint MAP estimator with optimal parameters for anechoic conditions outperforms the MAP estimator with Laplace approximation. This is because the anechoic approximation is still closer to the real PDF than the Laplace amplitude approximation as depicted in Figure 9.

5.2. Performance in realistic noise

Figure 18 plots the performance of the estimators for speech with fan noise and Figure 19 shows the performance for speech disturbed by cafeteria noise.

The noise reduction amount is lower for white noise, because the nonstationary cafeteria and fan noise are harder to track by the noise estimation algorithm.

The proposed super-Gaussian estimators still outperform the Ephraim-Malah algorithm although the performance gain is lower for the white noise. Again, the joint MAP estimator with optimal parameters performs best.

6. CONCLUSION

We have derived a computationally efficient MAP estimator for the speech spectral amplitude and a joint MAP estimator for the speech spectral amplitude and phase. Both estimators apply a Gaussian model for the noise coefficients, and a super-Gaussian model for the speech DFT coefficients. The underlying super-Gaussian model can be adjusted to the demands of the specific noise reduction system. While the MAP estimator allows an estimation with respect to a Laplace amplitude model for the speech DFT magnitude, the joint MAP estimator also allows an optimal adjustment of the underlying statistical model to the real PDF of the speech spectral amplitude for a specific noise reduction system.

The proposed super-Gaussian spectral amplitude estimators significantly improve the quality of the enhanced signal. The performance gain comes for free, it is obtained by applying a more accurate statistical model. Also, the weighting rules do not require the use of tables for special complicated functions compared to the state-of-the art speech spectral amplitude estimator derived by Ephraim-Malah or the super-Gaussian speech spectral estimators derived by Martin.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[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, 1984.

[4] 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, no. 2, pp. 443–445, 1985.

[5] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, vol. 1, pp. 253–256, Orlando, Fla, USA, May 2002.

[6] 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 (IWAENC '03)*, pp. 87–90, Kyoto, Japan, September 2003.

[7] P. Vary, "Noise suppression by spectral magnitude estimation—mechanisms and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387–400, 1985.

[8] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 10, pp. 1043–1051, 2003, special issue: Digital Audio for Multimedia Communications.

[9] T. Lotter, *Single and multimicrophone speech enhancement for hearing aids*, Ph.D. thesis, Aachen University (RWTH), Aachen, Germany, 2004.

[10] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[11] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001, Elsevier.

[12] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[13] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '96)*, vol. 2, pp. 629–632, Atlanta, Ga, USA, May 1996.

[14] D. R. Brillinger, *Time Series: Data Analysis and Theory*, McGraw-Hill, New York, NY, USA, 1981.

[15] 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, 1980.

[16] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Processing*, vol. 12, no. 2, pp. 119–141, 1987, Elsevier.

[17] T. Lotter and P. Vary, "Noise reduction by maximum a posteriori spectral amplitude estimation with supergaussian speech modeling," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC '03)*, pp. 83–86, Kyoto, Japan, September 2003.

[18] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[19] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, NY, USA, 1991.

[20] N. D. Wallace, "Computer generation of gamma random variates with non-integral shape parameters," *Communications of the ACM*, vol. 17, no. 12, pp. 691–695, 1974.

[21] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, San Diego, Calif, USA, 1994.

[22] S. Kullback, *Information Theory and Statistics*, Dover Publication, New York, NY, USA, 1968.

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[24] J. L. Melsa and D. L. Cohn, *Decision and Estimation Theory*, McGraw-Hill, New York, NY, USA, 1978.

[25] C. You, S. Koo, and S. Rahardja, "Adaptive β-order MMSE estimation for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '03)*, vol. 1, pp. 900–903, Hong Kong, China, April 2003.

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

[27] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.

[28] 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*, pp. 36–40, Darmstadt, Germany, March 1996.

[29] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 39–60, Springer-Verlag, New York, NY, USA, 2001.

[30] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '99)*, vol. 2, pp. 789–792, Phoenix, Ariz, USA, March 1999.

Thomas Lotter received the Dipl.-Ing. degree in electrical engineering in 2000 from the Aachen University of Technology, RWTH Aachen. He received the Ph.D. degree from the RWTH Aachen in 2004 after working at the Institute of Communication Systems and Data Processing in the area of single- and multimicrophone speech enhancement. In 2004, he joined Siemens Audiological Engineering Group, Erlangen, Germany with focus on wireless hearing aid applications. His main research interests include speech enhancement, signal processing for wireless systems, wireless standards, and audio coding.

Peter Vary received the Dipl.-Ing. degree in electrical engineering in 1972 from the University of Darmstadt, Darmstadt, Germany. In 1978, he received the Ph.D. degree from the University of Erlangen-Nuremberg, Germany. In 1980, he joined Philips Communication Industries (PKI), Nuremberg, where he became Head of the Digital Signal Processing Group. Since 1988, he has been a Professor at Aachen University of Technology, Aachen, Germany, and Head of the Institute of Communication Systems and Data Processing. His main research interests are in speech coding, channel coding, error concealment, adaptive filtering for acoustic echo cancellation and noise reduction, and concepts of mobile radio transmission.