MMSE SPEECH ENHANCEMENT UNDER SPEECH PRESENCE UNCERTAINTY ASSUMING (GENERALIZED) GAMMA SPEECH PRIORS THROUGHOUT

Balázs Fodor and Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig
Schleinitzstr. 22, D - 38106 Braunschweig, Germany
{b.fodor, t.fingscheidt}@tu-bs.de

ABSTRACT

Several investigations showed that speech enhancement approaches can be improved by speech presence uncertainty (SPU) estimation. Although there has been a strong focus on the use of correct statistical models for spectral weighting rules for the last few decades, there is just a few publications about SPU estimation based on a speech prior consistent with the spectral weighting rule. This contribution presents a new consistent solution for MMSE speech amplitude (SA) estimation under SPU, being based on the generalized gamma distribution representing a variety of speech priors. Employing the gamma speech model which is a special case of the generalized gamma distribution, the new approach is shown to outperform both the SPU-based MMSE-SA estimator relying on a Gaussian speech prior, and the gamma MMSE-SA estimation without SPU.

1. INTRODUCTION

Speech enhancement has been a vital field of research for the last few decades. Maybe most publications on this topic are written about the improvement of spectral weighting rules, covering different optimization criteria (such as minimum mean square error (MMSE) [1–4] or maximum *a posteriori* (MAP) [3, 5]), or different statistical models for the speech and/or noise DFT coefficients (such as Gaussian [1, 6], and non-Gaussian [2–5]).

Amongst others, Martin [4] and Lotter et al. [5] showed that the speech DFT coefficients follow a distribution which has a sharper peak than the Gaussian probability density function (pdf). These distributions are called super-Gaussian (such as Laplace, gamma, etc.). It was shown in several papers [2–5] that the use of the gamma speech prior achieves better results than the Gaussian.

With the generalized gamma distribution [2], Erkelens et al. introduced a quite flexible parametric statistical model of the speech DFT coefficients which advantageously covers a wide range of typically employed speech magnitude densities.

Unfortunately, super-Gaussian pdfs do not allow for an analytical solution of the MMSE-SA error criterion [2, 3]. Therefore, in [2] an approximative analytical solution, and in [3] a numerical solution was used for the calculation of the spectral weighting rule.

Further improvement is typically obtained exploiting the *speech presence uncertainty* (SPU). In [1] Ephraim and Malah showed that under an MMSE error criterion the SPU estimation turns out to be a multiplicative pre-factor, called soft weights, for the common spectral weighting rule. Furthermore, an SPU estimator based on a Gaussian speech model was presented.

The derivation of the SPU estimation requires the pdf of the complex-valued noisy speech signal, which can be determined as a convolution of the (bivariate) speech and noise pdfs (i. e., with a complex argument). For the noise, generally the bivariate Gaussian distribution is employed. In order to model the bivariate speech pdf for purpose of the convolution, in [6] the real and imaginary parts of the speech DFT coefficients are assumed being statistically independent. In [2], however, it was shown that they are indeed uncorrelated, but not independent. Additionally, it was pointed out that the bivariate histogram of the complex-valued speech DFT coefficients is approximately rotationally symmetric.

We present in this paper a new MMSE-SA estimator under SPU assuming gamma speech priors throughout. To achieve this, we first recapitulate briefly the derivation of the MMSE-SA estimator under generalized gamma priors (see also [2]). Then, employing the parameters of a gamma distribution which is a special case of the generalized gamma pdf, the result turns out to be the spectral weighting rule proposed in [3].

In analogy to the MMSE-SA estimator above, our new SPU estimator is first derived under a *generalized* statistical model for the speech, offering a wide flexibility of choice between different speech priors. For this, the *univariate* generalized gamma distribution [2] is extended to complex variables, resulting in a *bivariate* generalized gamma distribution. Advantageously, this new pdf does not use the assumption that the real and imaginary parts of the speech DFT coefficients are statistically independent. As with the weighting rule, for the further derivation of the soft weights then the gamma distribution is employed. Both the MMSE-SA spectral weighting rule as well as the soft weights of Ephraim and Malah [1] turn out to be a special case of our new pdf-generalized approach.

Our paper is organized as follows: Section 2 gives a short review of the reference MMSE-SA estimators with Gaussian and gamma speech priors, respectively. Section 3 presents the reference and the new SPU estimator based on Gaussian and gamma speech priors, respectively, followed by the evaluation of the proposal in Section 4. Finally, Section 5 gives some concluding remarks.

2. SPEECH SPECTRAL MAGNITUDE ESTIMATION

The input signal $y(n)$ of a speech enhancement system is assumed to consist of the clean speech signal $s(n)$ and the additive noise signal $n(n)$, with n being the discrete time index. After segmentation, windowing, and the discrete Fourier transform (DFT), the input signal can be rewritten as $Y(\ell,k) = S(\ell,k) + N(\ell,k)$ with ℓ being the analysis frame index, k being the frequency bin index. Using polar coordinates, the input signal can be reformulated as $R(\ell,k)e^{j\Theta(\ell,k)} = A(\ell,k)e^{j\alpha(\ell,k)} + B(\ell,k)e^{j\beta(\ell,k)}$ where R, A, B, (Θ, α, β) are the magnitudes (phases) of the short-time spectra Y, S, and N, respectively. Note that in the majority of the paper we omit frame index ℓ and frequency bin index k.

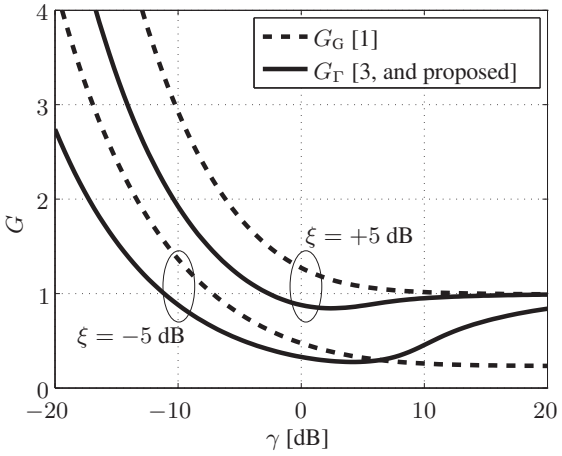

Fig. 1. Spectral amplitude (SA) estimators with Gaussian (G_{G}) and gamma (G_{Γ}) speech priors as a function of the *a posteriori* SNR γ at two different *a priori* SNRs $\xi = -5$ dB and 5 dB.

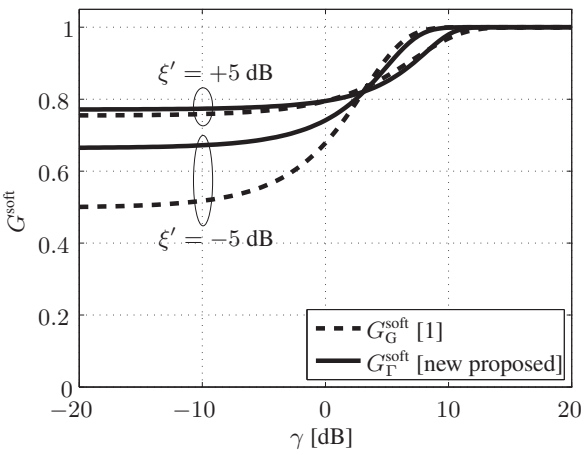

Fig. 2. Speech presence uncertainty (SPU) estimators (soft weights) with Gaussian ($G_{\mathrm{G}}^{\mathrm{soft}}$) and gamma ($G_{\Gamma}^{\mathrm{soft}}$) speech priors with $q = 0.2$ as a function of the *a posteriori* SNR γ at two different conditional *a priori* SNRs $\xi' = -5$ dB and 5 dB.

The aim of the speech enhancement algorithm is to obtain a clean speech estimate $\hat{S}$ (or its magnitude $\hat{A}$) given the noisy input signal Y. As shown in [1], the estimation of the clean speech magnitude A under the minimum mean square error (MMSE) criterion $\min_{\hat{A}} E\{(\hat{A} - A)^2 | Y\}$ leads to

$$\hat{A} = E\{A|Y\} = \frac{\int_0^{\infty}\int_0^{2\pi} A \cdot p(Y|A,\alpha) \cdot p(A,\alpha)\, d\alpha dA}{\int_0^{\infty}\int_0^{2\pi} p(Y|A,\alpha) \cdot p(A,\alpha)\, d\alpha dA}, \quad (1)$$

with $p(Y|A,\alpha)$ and $p(A,\alpha)$ being the conditional pdf of the noisy signal Y given speech, as well as the joint pdf of the clean speech magnitude A and phase α, respectively. The DFT coefficients of the additive noise N are assumed being Gaussian distributed, therefore, $p(Y|A,\alpha)$ turns out to be

$$p(Y|A,\alpha) = \frac{1}{\pi\sigma_N^2} e^{-\frac{|Y - Ae^{j\alpha}|^2}{\sigma_N^2}}, \quad (2)$$

with σ_N^2 being the variance of the additive noise process N. If the magnitude of the speech DFT coefficients is modeled by the generalized gamma pdf assuming independence between speech magnitude A and phase α [2], then the joint pdf $p(A,\alpha)$ turns out to be

$$p(A,\alpha) = p(\alpha) \cdot p(A) = \frac{1}{2\pi} \cdot \frac{\eta\beta^{\nu}}{\Gamma(\nu)} A^{\eta\nu-1} e^{-\beta A^{\eta}}, \quad (3)$$

with η, β, ν being the parameters of the generalized gamma pdf, and $\Gamma(\cdot)$ being the gamma function. Substituting (3) and (2) as well as doing some manipulations by means of [7, (8.431.5)], (1) turns out to be [2]

$$\hat{A} = \frac{\int_0^{\infty} A^{v} \cdot e^{-\frac{1}{\sigma_N^2}A^2 - \beta A^{\eta}} \cdot I_0(2\frac{R}{\sigma_N^2}A) dA}{\int_0^{\infty} A^{v-1} \cdot e^{-\frac{1}{\sigma_N^2}A^2 - \beta A^{\eta}} \cdot I_0(2\frac{R}{\sigma_N^2}A) dA}, \quad (4)$$

with $I_0(\cdot)$ being the modified Bessel function of the first kind and zeroth order. Employing the Gaussian parameters $\eta = 2$, $\beta = 1/\sigma_S^2$ with σ_S^2 being the speech spectral variance and $\nu = 1$, (4) leads to the well-known MMSE-SA weighting rule G_{G} with $\hat{A} = R \cdot G_{\mathrm{G}}$ (index G denotes the Gaussian speech prior) [1]

$$G_{\mathrm{G}} = \Gamma(1.5)\frac{\sqrt{v}}{\gamma} M(-0.5; 1; -v), \quad (5)$$

with $v = \gamma\xi/(1+\xi)$, and $\gamma = R^2/\sigma_N^2$, $\xi = \sigma_S^2/\sigma_N^2$, $M(\cdot)$ being the *a posteriori* signal-to-noise ratio (SNR), the *a priori* SNR, as well as the confluent hypergeometric function, respectively. This weighting rule G_{G} is plotted as dashed lines in Figure 1.

Now we want to derive the MMSE-SA weighting rule under gamma speech prior (applying $\eta = 1$ and $\beta = \sqrt{\nu(\nu+1)}/\sigma_S$ according to [2]). It was shown in [2, 3] that for $\eta = 1$ (4) cannot be solved analytically. Therefore, a numerical solution was sought as it was done in [3]. In order to make the solution $\hat{A}$ dependent on ξ and γ, the variable of integration A is substituted by Rg. Thus, (4) becomes

$$\hat{A} = R \cdot \frac{\int_0^{\infty} g^{v} \cdot e^{-\gamma g^2 - \sqrt{\nu(\nu+1)}\sqrt{\frac{\gamma}{\xi}}g} \cdot I_0(2\gamma g) dg}{\int_0^{\infty} g^{v-1} \cdot e^{-\gamma g^2 - \sqrt{\nu(\nu+1)}\sqrt{\frac{\gamma}{\xi}}g} \cdot I_0(2\gamma g) dg}. \quad (6)$$

Since $\hat{A} = R \cdot G$, (6) can be rewritten as

$$G_{\Gamma} = \frac{\Psi(0)}{\Psi(1)}, \quad (7)$$

$$\text{with } \Psi(c) = \int_0^{\infty} g^{v-c} \cdot e^{-\gamma g^2 - \sqrt{\nu(\nu+1)}\sqrt{\frac{\gamma}{\xi}}g} \cdot I_0(2\gamma g) dg.$$

Different to the proposal of Andrianakis and White [3] which is based on *performance measurements of the estimator*, we used the shaping parameter of the gamma pdf $\nu = 1.126$ suggested by Lotter and Vary [5] which is based on a *matched statistical model*. G_{Γ} was then computed by means of the adaptive Gauss-Kronrod quadrature [8], which can be seen in Figure 1 as solid lines.

3. SPEECH PRESENCE UNCERTAINTY

As published in [1], the MMSE-SA estimation under SPU $\hat{A}' = G' \cdot R$ turns out to be the multiplicative relationship of the common spectral weighting rule G and the soft weights G^{soft}:

$$G'(\gamma, \xi') = G(\gamma, \xi') \cdot G^{\text{soft}}(\gamma, \xi'), \tag{8}$$

employing $\xi' = E\{A^2|H_1\}/\sigma_N^2$ being the conditional *a priori* SNR supposing speech presence. Hypotheses H_0 and H_1 denote speech absence and presence, respectively. The soft weights can be calculated by means of the so-called likelihood ratio Λ

$$G^{\text{soft}} = \frac{\Lambda}{1+\Lambda}, \tag{9}$$

where Λ is defined as

$$\Lambda = \frac{P(H_1)}{P(H_0)} \cdot \frac{p(Y|H_1)}{p(Y|H_0)}, \tag{10}$$

with $P(H_0) = q$, $P(H_1) = (1-q)$, $p(Y|H_0)$, and $p(Y|H_1)$ being the speech absence and presence probability, as well as the pdf of the noisy signal Y assuming speech absence and presence, respectively. In speech absence H_0, assuming a Gaussian noise model, $p(Y|H_0)$ turns out to be

$$p(Y|H_0) = \frac{1}{\pi\sigma_N^2} e^{-\frac{|Y|^2}{\sigma_N^2}}. \tag{11}$$

In speech presence H_1, the distribution of the noisy signal as a sum of the clean speech and noise signal $p(Y|H_1) = p(Y \equiv S + N)$ is needed. Assuming the speech S and the additive noise N being statistically independent random processes, $p(Y|H_1)$ turns out to be

$$p(Y|H_1) = p(Y \equiv S) * p(Y \equiv N), \tag{12}$$

where $*$ denotes the convolution operation. Please note the identity $p(Y \equiv N) = p(Y|H_0)$.

Employing again a generalized pdf to allow for a free choice of different speech priors, and assuming that the pdf of the (complex-valued) speech DFT coefficients is zero-mean and rotationally symmetric, the noisy speech signal Y in presence of speech and in absence of noise can be described by a new *bivariate generalized gamma distribution*:

$$p(Y \equiv S) = \frac{1}{2\pi} \cdot \frac{\eta\beta^\nu}{\Gamma(\nu)} \cdot |Y|^{\eta\nu-2} \cdot e^{-\beta|Y|^\eta}, \tag{13}$$

with $Y \in \mathbb{C}$. The relation between this bivariate (13) and the univariate (3) generalized gamma distribution is shown in the Appendix.

After substituting (13) and (11) into (12), and using polar coordinates to solve the complex convolution integral, applying again a variable substitution, the likelihood ratio Λ in (9) for our new soft weights is (index gΓ denotes the generalized gamma speech model)

$$\Lambda_{\text{g}\Gamma} = \frac{(1-q)}{q} \cdot \frac{\eta\beta^\nu}{\Gamma(\nu)} \int_0^\infty x^{\eta\nu-1} e^{-\beta R^\eta x^\eta} e^{-\gamma x^2} I_0(2\gamma x) dx, \tag{14}$$

with $x \in \mathbb{R}$. Substituting the Gaussian parameters $\eta = 2$, $\beta = 1/\sigma_S^2$ and $\nu = 1$, (14) exactly reduces to Ephraim-Malah's likelihood ratio [1] (index G denotes the Gaussian speech model)

$$\Lambda_{\text{G}} = \frac{(1-q)}{q} \frac{e^v}{1+\xi'}, \tag{15}$$

with $v = \gamma\xi'/(1+\xi')$. The resulting soft weights $G_{\text{G}}^{\text{soft}}$ for $q = 0.2$ are plotted in Figure 2 as dashed lines.

Applying the gamma parameters $\eta = 1$, and $\beta = \sqrt{\nu(\nu+1)}/\sigma_S$, (14) turns out to be analytically not solvable [2]. Therefore, again, the adaptive Gauss-Kronrod quadrature [8] was used employing the shaping parameter $\nu = 1.126$. The resulting soft weights G_Γ^{soft} for $q = 0.2$ are plotted in Figure 2 as solid lines.

4. EVALUATION

In order to show the merit of the proposed SPU-based estimator, we performed the following simulations: A total of 96 speech signals (spoken by four male and four female speakers) was taken from the NTT Multi-Lingual Speech Database [9], and downsampled to 8 kHz sampling rate. Car noise signals were taken from the NTT Ambient Noise Database [10]. The active speech level was set to -26 dB$_{\text{ov}}$, the noise signal level was adjusted to the desired input SNR, according to ITU-T Recommendation P.56 [11], followed by superposition of both signals. At a sampling frequency of 8 kHz, the segmentation of the noisy speech signal $y(n)$ was done by a Hann window, the analysis frame length was $T = 256$ samples, the analysis frame shift took 128 samples.

As noise power estimation, we applied the minimum statistics (MS) algorithm [12]. In order to estimate the *a priori* SNR ξ, the widely employed decision-directed estimator [1] was used. Following [1], the value of the smoothing factor β was 0.98, and 0.99, for experiments without and with SPU estimation, respectively. Then, the spectral weighting rules introduced in Section 2 were utilized as a table lookup with both γ and ξ varying from $-20...+20$ dB in 0.4 dB steps. The soft weights were also calculated numerically with $q = 0.2$ according to [1] and implemented also as table lookup, in the same fashion as the spectral weighting rules. Please note that employing (8), the conditional *a priori* SNR assuming speech presence has to be computed as $\xi' = \xi/(1-q)$.

We evaluated the performance of the proposed approach w. r. t. the speech component quality and the amount of noise suppression. Given a noisy speech signal $y(n)$ we employed the respective clean speech $\tilde{s}(n)$ and noise component $\tilde{n}(n)$ of the enhanced signal $\hat{s}(n) = \tilde{s}(n) + \tilde{n}(n)$. Through the clean speech signal $s(n)$ and its processed replica $\tilde{s}(n)$, the speech preservation performance was represented by the segmental speech to speech distortion ratio (SSDR) [13]:

$$\text{SSDR}_{\text{seg}} = \frac{1}{N_\Phi} \sum_{\ell \in \Phi} \text{SSDR}(\ell) \tag{16}$$

$$\text{SSDR}(\ell) = \text{li}\left\{ 10\log_{10} \frac{\sum_{\tau=1}^T s^2(\tau + \ell T)}{\sum_{\tau=1}^T e^2(\tau + \ell T)} \right\} \tag{17}$$

where $e(n) = s(n) - \tilde{s}(n)$, Φ is the set of frames belonging to speech activity, $N_\Phi = |\Phi|$ is the number of frames with speech activity, and the operator li$\{\cdot\}$ limits SSDR(ℓ) to [-10,30] dB.

Meanwhile, we assessed the noise attenuation performance by computing the segmental noise attenuation measure based on the noise signal $n(n)$ and the processed noise component $\tilde{n}(n)$ [13]:

$$\text{NA}_{\text{seg}} = 10\log_{10} \frac{1}{L} \sum_{\ell=1}^L \frac{\sum_{\tau=1}^T n^2(\tau + \ell T)}{\sum_{\tau=1}^T \tilde{n}^2(\tau + \ell T)} \tag{18}$$

where L is the total number of frames.

In Figure 3 the reference MMSE-SA weighting rule with (Gaussian assumption) and without (Gaussian and gamma assumption) SPU estimation are shown along with the proposed MMSE-SA estimation under SPU (gamma assumption throughout) for different input SNR values. The optimum in Figure 3 resides in the right top corner,

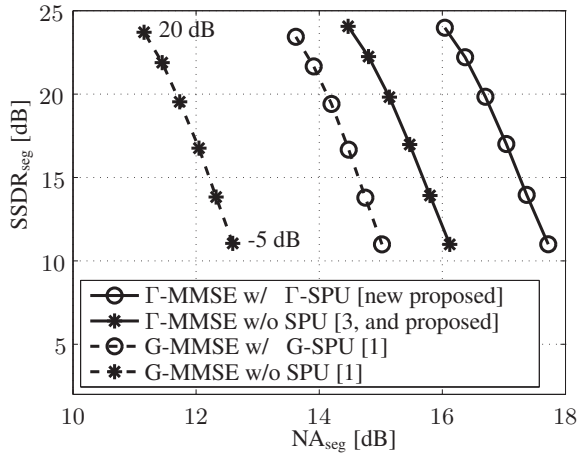

Fig. 3. Simulation results reflected by the measures NA_{seg} and SSDR_{seg} for input SNR ratios from -5 dB until +20 dB in 5 dB steps. G, Γ, and SPU denote the Gaussian and the gamma speech prior, as well as speech presence uncertainty estimation, respectively.

which means a significant noise reduction and a good quality of the (naturally noise-free, but potentially distorted) speech component, simultaneously. It can generally be said that at higher input SNRs a better speech preservation can be observed, reflected by higher SSDR values. Meanwhile, at very low input SNRs greater noise suppression can be obtained. Please note, that for ease of comparison a noise overestimation factor was used and adjusted in such a way that all four approaches achieve the same clean speech component quality at -5 dB input SNR. It turns out that in consequence the compared approaches only differ w. r. t. noise attenuation.
It can clearly be seen that the amount of noise suppression of Ephraim and Malah's MMSE-SA weighting rule can be improved of approximately 2.5 dB by using SPU estimation based on the Gaussian speech model. The spectral weighting rule based on a gamma speech prior performs roughly 3.5 dB better than the Gaussian-based one. The new MMSE-SA estimator under SPU based on a gamma speech prior achieves another 1.5 dB more noise suppression, clearly outperforming all other approaches. The merit of the proposal was supported by informative listening tests.

5. CONCLUSION

This paper presents an MMSE speech spectral amplitude estimation with speech presence uncertainty (SPU) estimation. Both the spectral weighting rule and the SPU estimator are optimal w. r. t. the MMSE criterion and are based consistently on gamma speech priors. It is shown that extending the gamma MMSE-SA weighting rule by soft weights based on gamma speech priors, the noise suppression performance can significantly be improved, without any quality degradation of the speech component.

6. APPENDIX

Taking the bivariate generalized gamma distribution from (13) with the complex-valued variable $Y = Y_{\text{Re}} + jY_{\text{Im}}$, we are interested in the pdf of the distribution of $R = |Y| = \sqrt{Y_{\text{Re}}^2 + Y_{\text{Im}}^2}$. According to [14], the (cumulative) distribution function of R can be calculated by integration of (13) w. r. t. Y on a circle with a radius of $|Y|$. Moreover, the first derivative of the (cumulative) distribution function yields then the probability density function of the magnitude $R = |Y|$. Using polar integration with the integration variable $r = |Y|$ as well as $dY = rd\theta dr$ we get

$$p(R) = \frac{\partial}{\partial R} \int_0^R \int_0^{2\pi} \frac{1}{2\pi} \frac{\eta \beta^\nu}{\Gamma(\nu)} r^{\eta\nu - 2} e^{-\beta r^\eta} r d\theta dr. \qquad (19)$$

Solving the integral w. r. t. θ and using that $\frac{\partial}{\partial \rho} \int_0^\rho g(\zeta) d\zeta = g(\rho)$, (19) turns out to be the *univariate* generalized gamma pdf $p(A)$ as known from [2] and (3).

7. REFERENCES

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

[2] Erkelens, J. S.; Hendriks, R. C.; Heusdens, R.; Jensen, J., "Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients with Generalized Gamma Priors," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1741 – 1752, Aug. 2007.

[3] Andrianakis, I.; White, P.R., "MMSE Speech Spectral Amplitude Estimators With Chi and Gamma Speech Priors," in *Proc. of ICASSP 2006*, Toulouse, France, May 2006, pp. III–1068–III–1071.

[4] Martin, R., "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. of ICASSP 2002*, Orlando, FL, USA, May 2002, vol. 1, pp. I–253—I–256.

[5] Lotter, T.; Vary, P., "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 1110–1126, 2005.

[6] Bin, C.; Loizou, P. C., "A Laplacian-based MMSE estimator for speech enhancement," *Speech Communication*, vol. 49, pp. 134–143, Feb. 2007.

[7] Gradshteyn, I. S.; Ryzhik, I. M., *Table of Integral, Series, and Products*, Academic Press, 4th edition, 1965.

[8] Shampine, L. F., "Vectorized Adaptive Quadrature in MATLAB," *Journal of Computational and Applied Mathematics*, vol. 211, pp. 131–140, Jan. 2008.

[9] "Multi-Lingual Speech Database for Telephonometry," NTT Advanced Technology Corporation (NTT-AT), 1994.

[10] "Ambient Noise Database for Telephonometry," NTT Advanced Technology Corporation (NTT-AT), 1996.

[11] "Objective Measurement of Active Speech Level," ITU-T P.56, Mar. 1993.

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

[13] Fingscheidt, T.; Suhadi, S.; Stan, S., "Environment-Optimized Speech Enhancement," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 4, pp. 825–834, 2008.

[14] Papoulis, A.; Pillai, U., *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th edition, 2002.