A SPEECH PRESENCE PROBABILITY ESTIMATOR BASED ON FIXED PRIORS AND A HEAVY-TAILED SPEECH MODEL

Balázs Fodor

Institute for Communications Technology
Technische Universität Braunschweig
38106 Braunschweig, Germany
b.fodor@tu-braunschweig.de

Timo Gerkmann

Speech Signal Processing Group
Cluster of Excellence "Hearing4all"
Universität Oldenburg, Germany
timo.gerkmann@uni-oldenburg.de

ABSTRACT

Speech enhancement approaches are often enhanced by speech presence probability (SPP) estimation. However, SPP estimators suffer from random fluctuations of the *a posteriori* signal-to-noise ratio (SNR). While there exist proposals that overcome the random fluctuations by basing the SPP framework on smoothed observations, these approaches do not take into account the super-Gaussian nature of speech signals. Thus, in this paper we define a framework that allows for modeling the likelihoods of speech presence for smoothed observations, while at the same time assuming super-Gaussian speech coefficients. The proposed approach is shown to outperform the reference approaches in terms of the amount of noise leakage and the amount of musical noise.

1. INTRODUCTION

Many algorithms in speech signal processing require the information whether speech is present or not. Examples are the estimation of the noise power or the estimation of the clean speech coefficients. In this paper we address the estimation of the *a posteriori* speech presence probability (SPP) in each time-frequency bin of the short-time discrete Fourier transform (STFT) domain when only a noisy observation is given. It was shown that when SPP is taken into account, the performance of single channel speech enhancement algorithms can be improved [1]. To estimate the SPP, likelihood functions of speech presence and speech absence are required which are typically modeled as Gaussian distributions [1, 2].

The likelihood function of speech presence is parametrized by the *a priori* signal-to-noise ratio (SNR) which is often adapted to follow the local SNR in each time-frequency bin [1, 2]. However, it has been shown that using the local SNR leads to the conceptual disadvantage that the *a posteriori* SPP yields only the *a priori* SPP in speech absence which, by definition, is independent of the observation [3]. A typical choice for the *a priori* SPP is $P(H_1) = 0.5$. Instead of using the local SNR, in [3] a fixed SNR is employed to parametrize the speech presence model which reflects the SNR that would be expected if speech *were* present in a time-frequency bin. As a result, the *a posteriori* SPP turns out to be close to zero in speech absence without any modification or adaptation of the *a priori* SPPs. Furthermore, in [3] it was proposed to base the SPP estimates on a smoothed version of the *a posteriori* SNR to reduce estimation outliers. For this, the likelihood functions of speech presence and absence are modeled by chi-squared distributions with the shape parameter ν. It turns out that the more averaging is applied, the more ν increases.

In [4] it was argued that the speech discrete Fourier transform (DFT) coefficients are not Gaussian distributed but follow a more heavy-tailed, so-called *super-Gaussian* distribution. The resulting estimators turn out to preserve the speech better than estimators employing a Gaussian speech model. However, for SPP estimation, usually the assumption of Gaussian speech priors is maintained, resulting in a mismatch between the employed speech models for SPP estimation and the estimation of clean speech coefficients. Speech enhancement schemes under SPP for a consistently super-Gaussian speech model were introduced in [5, 6]. However, these models are only applied to non-smoothed observations and may suffer from outliers in the estimation. Therefore, the goal of this paper is to derive models for smoothed observations with an underlying super-Gaussian speech model in order to avoid outliers while still having the benefit of a super-Gaussian speech model.

The paper is structured as follows: Section 2 gives a short overview of two state-of-the-art SPP research directions. Section 3 introduces the proposed approach unifying the philosophies from Section 2. The evaluation of the proposed estimator is in Section 4. Finally, Section 5 concludes this paper.

2. REVIEW ON SPEECH PRESENCE PROBABILITY (SPP) ESTIMATION

The short-time DFT coefficients of the noisy speech signal $Y(\ell, k)$ are assumed to be an additive superposition of speech $S(\ell, k)$ and noise $N(\ell, k)$. Here, ℓ is the time frame index and

The work of T. Gerkmann was funded by the DFG Grant GE 2538/2-1

k the frequency bin index. The aim of speech enhancement is to obtain an estimate of the speech, denoted as $\widehat{S}(\ell,k)$, when only the noisy speech $Y(\ell,k)$ is observed. In the remaining of the paper, we will omit the indices ℓ and k for ease of readability. Introducing the hypotheses for speech absence H_0 and speech presence H_1 and employing minimum mean square error (MMSE) estimation for estimating the speech amplitude $A = |S|$, the MMSE short-time spectral amplitude (STSA) estimator under speech presence uncertainty is given by [7]

$$\widehat{A} = P(H_1|Y) \cdot E\{A|Y, H_1\} \tag{1}$$

with $E\{\cdot\}$ being the expectation operator. With the noise power σ_N^2, the *a posteriori* SNR is defined as $\gamma = |Y|^2/\sigma_N^2$. As the estimation is obtained in each time-frequency bin independently, it is reasonable to assume that the isolated instantaneous phase in each time-frequency point does not give information on whether speech is present or absent in the time-frequency point under consideration. Thus, the posterior SPP $P(H_1|Y)$ can also be written as a function of γ, i. e., $P(H_1|Y) = P(H_1|\gamma)$ [8].

2.1. SPP for a Smoothed Observation

In [3], it is proposed to apply smoothing to the *a posteriori* SNR to reduce random fluctuations in the *a posteriori* SPP. The effect of smoothing is denoted by a bar, e. g., $\overline{\gamma}$ represents the *a posteriori* SNR after smoothing. The *a posteriori* SPP can be obtained as

$$P(H_1|\overline{\gamma}) = \frac{\Lambda}{1+\Lambda} \tag{2}$$

with the generalized likelihood ratio (GLR)

$$\Lambda = \frac{P(H_1)}{P(H_0)} \cdot \frac{p(\overline{\gamma}|H_1)}{p(\overline{\gamma}|H_0)} \tag{3}$$

where $P(H_1)$ is the *a priori* probability of speech presence, $P(H_0)$ is the *a priori* probability of speech absence, $p(\overline{\gamma}|H_1)$ is the likelihood of speech presence, and $p(\overline{\gamma}|H_0)$ is the likelihood of speech absence.

In [3] the likelihoods of speech presence and speech absence are modeled by the chi-squared distribution with the shape parameter $\bar{\nu}$. The effect of smoothing the *a posteriori* SNR is reflected by an increase of the shape parameter $\bar{\nu}$ which can be identified by relating the first and second moment obtained from training data as [3]

$$\bar{\nu} = \frac{(E\{\overline{\gamma}\})^2}{\mathrm{var}\{\overline{\gamma}\}}. \tag{4}$$

The same shape parameter is used for the likelihoods of speech presence and absence, resulting in the GLR [3]

$$\Lambda_{[3]} = \frac{P(H_1)}{P(H_0)} \cdot \left(\frac{1}{1+\xi}\right)^{\bar{\nu}} \cdot e^{\bar{\nu}\frac{\xi}{1+\xi}\overline{\gamma}} \tag{5}$$

with the *a priori* SNR $\xi = \sigma_S^2/\sigma_N^2$. Finally, the *a posteriori* SPP is obtained by employing (2).

2.2. SPP with Super-Gaussian Speech Models

In [5, 6] the likelihood of speech presence is obtained based on a super-Gaussian model for speech, while smoothing is not taken into account. To model the super-Gaussian characteristics of speech, we model the clean speech amplitudes by a chi distribution with shape parameter μ (cf. [9]). While $\mu = 1$ reflects a Gaussian speech model, $\mu < 1$ reflects a super-Gaussian speech model. To model the noise distribution, a widely-employed Gaussian model is utilized. With these assumptions, the likelihoods result in [5]

$$p_{[5]}(\gamma|H_1) = \left(\frac{\mu}{\mu+\xi}\right)^{\mu} \cdot e^{-\gamma} \cdot {}_1F_1\!\left(\mu; 1; \frac{\gamma\cdot\xi}{\mu+\xi}\right) \tag{6}$$

and

$$p_{[5]}(\gamma|H_0) = e^{-\gamma} \tag{7}$$

with the confluent hypergeometric function ${}_1F_1(\cdot)$. Accordingly, for the GLR we obtain [5]

$$\Lambda_{[5]} = \frac{P(H_1)}{P(H_0)} \cdot \left(\frac{\mu}{\mu+\xi}\right)^{\mu} \cdot {}_1F_1\!\left(\mu; 1; \frac{\gamma\cdot\xi}{\mu+\xi}\right) \tag{8}$$

which can then be employed to obtain the *a posteriori* SPP using (2).

3. PROPOSED ESTIMATOR

So far we have reviewed two different estimators for the SPP. While the estimator in Section 2.1 incorporates the averaging of the observation in the statistical model, the estimator in Section 2.2 reflects the super-Gaussian characteristics of speech. The averaging has the benefit that outliers can be reduced, while the super-Gaussian speech model has the potential for a better speech preservation. The key assumption in this paper is that for moderate to large SNRs we can approximate the likelihood of speech presence for a smoothed observation $\overline{\gamma}$ by the same parameterized PDF (6), but with an increase of the shape parameter $\bar{\mu} > \mu$, where $\bar{\mu}$ is the shape parameter after averaging.

We now aim at obtaining the shape parameter $\bar{\mu}$ of the super-Gaussian model (8) to also incorporate a smoothing of the observation. For this, we now compute the first and second statistical moments of our approximate likelihood for smoothed observations (6), resulting in [10, Eq. (7.621.4)]

$$m_{\overline{\gamma}|H_1} = \int_0^{\infty} \overline{\gamma} \cdot p_{[5]}(\overline{\gamma}|H_1)\, d\overline{\gamma} = 1+\xi, \tag{9}$$

while the second central moment results in [10, Eq. (7.621.4)]

$$\sigma^2_{\overline{\gamma}|H_1} = \int_0^{\infty} \overline{\gamma}^2 \cdot p_{[5]}(\overline{\gamma}|H_1)\, d\overline{\gamma} - m^2_{\overline{\gamma}|H_1} = \frac{1}{\bar{\mu}}\xi^2 + 2\xi + 1. \tag{10}$$

Solving for the shape parameter $\bar{\mu}$ we obtain

$$\bar{\mu} = \frac{\xi^2}{\frac{\sigma^2_{\overline{\gamma}|H_1}}{m^2_{\overline{\gamma}|H_1}} \cdot (1+\xi)^2 - 2\xi - 1}. \tag{11}$$

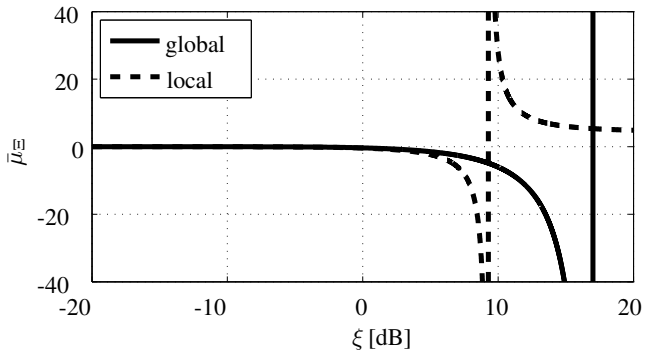

Fig. 1. Shape parameter of the likelihood of speech presence for averaged *a posteriori* SNR as a function of the *a priori* SNR

Thus, similar to (4), we can fit our super-Gaussian model also to smoothed observations by relating the first and second statistical moments. For a given smoothing process these moments can be computed using training data. In contrast to (4), the shape parameter $\bar{\mu}$ in (11) is also a function of the *a priori* SNR. As μ in (6) intrinsically models the shape of the speech distribution, large degrees of averaging are difficult to model especially in regions where the noise dominates. As a consequence, (11) yields negative—and thus invalid—values for $\bar{\mu}$ whenever

$$\frac{\sigma^2_{\overline{\gamma}|H_1}}{m^2_{\overline{\gamma}|H_1}} < \frac{2\xi+1}{(1+\xi)^2}. \tag{12}$$

However, due to the central limit theorem, the impact of a super-Gaussian model decreases with an increased averaging. Thus, for large amounts of smoothing, model (5) can be used even if the underlying speech is super-Gaussian. Therefore, in this paper we propose to fit the super-Gaussian model to moderately smoothed observations, while employing the model in (5) for large amounts of smoothing.

3.1. *A Posteriori* SNR Averaging Framework

The averaging of γ can be obtained for instance in the cepstral domain [11] or directly in the time-frequency domain [3]. In this work, we adopt the time-frequency averaging from [3]. There, it is proposed to use two averaging windows of different sizes, namely the local and the global averaging window. The resulting *a posteriori* SPPs are finally combined by multiplication to obtain the final SPP estimator

$$\widehat{P(H_1|\gamma)} = P(H_1|\overline{\gamma}_{\text{local}}) \cdot P(H_1|\overline{\gamma}_{\text{global}}). \tag{13}$$

Please note that different parameters, e. g., $\bar{\mu}_\Xi$, $\bar{\nu}_\Xi$, are used for the locally and globally averaged *a posteriori* SNR. Here, Ξ stands for either "local" or "global". Averaging is obtained in the vicinity of the current time-frequency bin as [3]

$$\overline{\gamma}_\Xi(\ell,k) = \frac{1}{|\mathbb{K}_\Xi| \cdot |\mathbb{L}_\Xi|} \cdot \sum_{\substack{\lambda_\Xi \in \mathbb{L}_\Xi \\ \kappa_\Xi \in \mathbb{K}_\Xi}} \gamma(\lambda_\Xi, \kappa_\Xi) \tag{14}$$

with $\mathbb{L}_\Xi$, $\mathbb{K}_\Xi$ being a set of frames and a set of frequency bins within the averaging window, respectively. Each averaging window consists of the current frame ℓ and the previous $\Delta\ell_\Xi$ frames, hence, the width of each averaging window is $|\mathbb{L}_\Xi| = \Delta\ell_\Xi + 1$. The height of each averaging window is $|\mathbb{K}_\Xi| = 2\Delta k_\Xi + 1$, i. e., besides the current frequency bin k, Δk_Ξ frequency bins below it and Δk_Ξ frequency bins above it are employed for averaging. Therefore, each averaging window is of the size $|\mathbb{K}_\Xi| \cdot |\mathbb{L}_\Xi|$.

Ξ	Δk_Ξ	$\Delta\ell_\Xi$	$\lvert\mathbb{K}_\Xi\rvert\lvert\mathbb{L}_\Xi\rvert$	$\bar{\mu}_\Xi$	$\bar{\nu}_\Xi$	ζ_Ξ
local	1	2	9	71.4	—	9.6 dB
global	8	2	51	—	25.7	2.9 dB

Table 1. Parameters of the averaging framework

To fit the super-Gaussian model to the smoothed data, $\bar{\mu}_\Xi$ in (8) was calculated as follows: First, we generated Y samples artificially by superimposing complex-valued Gaussian-distributed random samples (representing the noise) as well as complex-valued random samples with chi-distributed amplitude and uniformly distributed phase (representing speech) at a given SNR ξ. The γ samples were then obtained by taking the magnitude square of Y and normalizing by a given noise variance. Then, the moving average of the artificial γ values was calculated by using a local and a global averaging window of the lengths $|\mathbb{K}_\Xi| \cdot |\mathbb{L}_\Xi|$ from Table 1. Finally, $\bar{\mu}_\Xi = f_\Xi(\xi)$ was estimated employing (11) within a local and a global window for different *a priori* SNRs ξ. The result is depicted in Figure 1.

As in [3] we argue that the *a priori* SNR ξ in (8) and (5) is a parameter of our speech presence model. As such it should not reflect the true local SNR, but an SNR that can be expected if speech is present in a time-frequency bin. Therefore, we employ a fixed *a priori* SNR ζ_Ξ in each averaging window which can be found by minimizing the probability of misdetection of speech presence for a given range of true local SNRs [3]. This optimization using the new likelihoods (6) and (7) for smoothed observations yields $\zeta_{\text{local}} = 9.6\,\text{dB}$ for local averaging and $\zeta_{\text{global}} = 6.6\,\text{dB}$ for global averaging. The corresponding shape parameters can be obtained by applying the resulting fixed *a priori* SNRs to (11), resulting in $\bar{\mu}_{\text{local}} = 71.4$ and a negative $\bar{\mu}_{\text{global}}$. Thus, the super-Gaussian model can be fitted to the training data for the moderate local averaging, but not for the strong global averaging. As a consequence, for the global window we propose to use the model (5), while for the local averaging we employ the model in (8), resulting in the GLRs

$$\Lambda_{\text{local}} = \frac{P(H_1)}{P(H_0)} \cdot \left(\frac{\bar{\mu}_{\text{local}}}{\bar{\mu}_{\text{local}} + \zeta_{\text{local}}}\right)^{\bar{\mu}_{\text{local}}} \cdot {}_1F_1\left(\bar{\mu}_{\text{local}}; 1; \frac{\overline{\gamma} \cdot \zeta_{\text{local}}}{\bar{\mu}_{\text{local}} + \zeta_{\text{local}}}\right), \tag{15}$$

$$\Lambda_{\text{global}} = \frac{P(H_1)}{P(H_0)} \cdot \left(\frac{1}{1+\zeta_{\text{global}}}\right)^{\bar{\nu}_{\text{global}}} \cdot e^{\bar{\nu}_{\text{global}} \frac{\zeta_{\text{global}}}{1+\zeta_{\text{global}}} \overline{\gamma}} \tag{16}$$

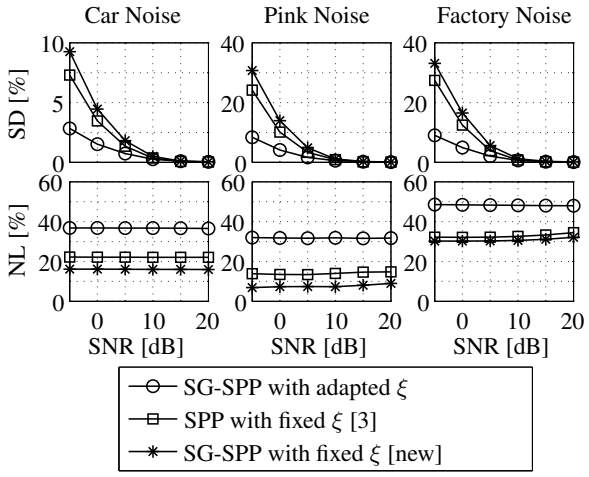

Fig. 2. Evaluation results of the reference and the proposed SPP estimators w. r. t. NL and SD [3]

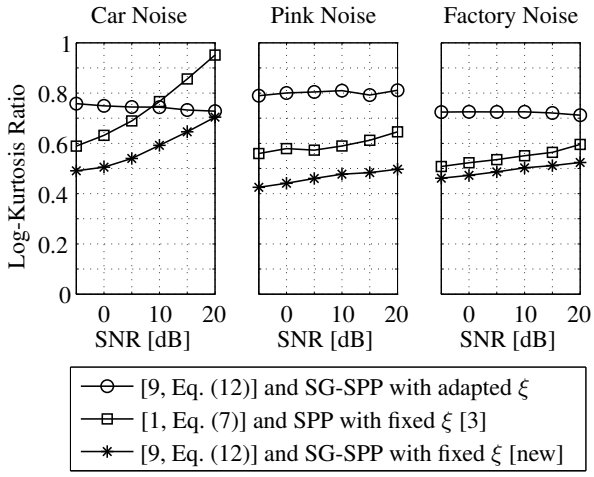

Fig. 3. Evaluation results of speech enhancement approaches based on an MMSE-STSA weighting rule and an SPP estimator w. r. t. amount of musical noise based on the log-kurtosis ratio (the lower the value the less the amount of musical noise) [16]

with the shape parameters $\bar{\mu}_{\text{local}}$ and $\bar{\nu}_{\text{global}}$ and the fixed *a priori* SNRs ζ_Ξ from Table 1. We employed the same averaging parameters $\Delta k_\Xi, \Delta \ell_\Xi, \bar{\nu}_{\text{global}}$, and ζ_{global} as in [3]. From the likelihood ratios (15), (16), the local and global *a posteriori* SPPs are obtained by using (2) and the final SPP estimator is obtained by combining the local and global SPPs via (13).

4. EVALUATION

In order to evaluate the proposed approach, the following simulations were carried out: 96 speech signals were taken from the NTT Multi-Lingual Speech Database [12]. As noise we employed car noise and factory noise signals from the NTT Ambient Noise Database [13], as well as pink noise which was generated by filtering a white noise signal by a filter with a $1/f$ frequency response. All database signals were downsampled to 8 kHz sampling rate. The desired input SNR was adjusted between -5 dB and 20 dB in 5 dB steps by scaling the speech and noise components separately according to the ITU-T Recommendation P.56 [14]. The noisy speech signal was processed based on frames with a length of $L = 256$, a frame shift of 50 %, and square-root Hann windows for spectral analysis and synthesis.

Each frequency bin in every frame was processed as follows: The noise power $\sigma_N^2(\ell, k)$ was estimated using [15], the *a priori* SNR $\xi(\ell, k)$ was obtained by the decision-directed method with a smoothing factor of 0.98 [1].

Overall, the following three approaches were evaluated: The first reference approach is based on SPP estimation employing a super-Gaussian model without averaging from Section 2.2, combined with a super-Gaussian amplitude estimator [9, Eq. (12)]. Both estimators are consistent regarding the speech PDF assumptions, utilizing the same shape parameter value $\mu = 0.5$. Furthermore, here the *a priori* SNR is not fixed in the SPP estimator but adapted using the decision-directed approach. This method is referred to as "SG-SPP with adapted ξ".

The second reference approach is the SPP estimator from Section 2.1 combined with the speech spectral amplitude estimator [1, Eq. (7)], both being consistently based on a Gaussian speech model. Further, the SPP estimator is based on a smoothed *a posteriori* SNR and employs fixed *a priori* SNRs to model the likelihoods of speech presence. Therefore, this method is denoted as "SPP with fixed ξ [3]".

The proposed approach, denoted as "SG-SPP with fixed ξ [new]", combines the proposed approach of Sections 3 and 3.1 and combines them with the super-Gaussian MMSE estimator [9, Eq. (12)] with $\mu = 0.5$. Thus, here we assume a super-Gaussian speech model consistently for MMSE-STSA estimation and SPP estimation.

The performance of the SPP estimators was assessed w. r. t. the missed-hit rate and false-alarm rate using the measures speech distortion (SD) and noise leakage (NL), respectively, as proposed in [3]. The measure SD indicates the percentage of the speech energy that the corresponding SPP estimator neglects, while the measure NL indicates in percent how much energy from the noise-only bins remains unattenuated. Therefore, the lower the NL value, the lower the residual noise level and the lower the SD value, the lower the speech distortion.

Outliers in the processed noise may be perceived as annoying musical noise. While in [3] the amount of processing outliers was assessed by the heavy-tailedness of processed noise histograms, in this work we employ the so-called weighted log-kurtosis ratio (LKR) [16]. The idea is to compare the kurtosis of the *noise component* of the noisy speech signal before and after processing in speech pauses, resulting in the LKR [16, 17]. Large values of the LKR indicate a large amount of processing outliers that may be perceived as annoying musical noise. Please note that while the measures SD and NL are applied to merely the SPP estimators, the LKR takes also the performance of the spectral weighting

rules into account.

The results are summarized in Figures 2 and 3. As can be seen in Figure 2, the proposed SPP estimator outperforms the reference approaches by achieving the lowest NL level, followed by the estimators "SPP with fixed ξ [3]" and "SG-SPP with adapted ξ" for all input SNR levels. However, the reversed order is true for the SD levels which reflects a typical trade-off in speech enhancement: The amount of speech distortion is inversely proportional to the amount of residual noise. However, the proposed approach outperforms the reference approaches w. r. t. the amount of musical noise, as can be seen in Figure 3. The proposed approach achieves the best results (the lowest LKRs), followed by the estimators "SPP with fixed ξ [3]" and "SG-SPP with adapted ξ". Informal listening tests confirmed these results.

5. CONCLUSIONS

In speech presence probability (SPP) estimation several improvements have been proposed in the past years. Among those are the smoothing of the observation to reduce outliers or the incorporation of a super-Gaussian speech model with the potential of a better speech preservation. In this work, we combine those previous approaches, resulting in an estimator that both incorporates a smoothed observation and a super-Gaussian speech model. The resulting estimator is shown to outperform the reference approaches in terms of noise leakage and the amount of musical noise.

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

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

[3] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

[4] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Orlando, FL, USA, May 2002, vol. 1, pp. I–253—I–256.

[5] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011.

[6] B. Fodor and T. Fingscheidt, "MMSE speech enhancement under speech presence uncertainty assuming (generalized) gamma speech priors throughout," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 4033–4036.

[7] R. McAulay and M. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[8] Richard C. Hendriks, Timo Gerkmann, and Jesper Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-art*, Morgan & Claypool, Feb. 2013.

[9] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

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

[11] T. Gerkmann, M. Krawczyk, and R. Martin, "Speech presence probability estimation based on temporal cepstrum smoothing," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 4254–4257.

[12] NTT, "Multi-lingual speech database for telephonometry," NTT Advanced Technology Corporation, 1994.

[13] NTT, "Ambient noise database for telephonometry," NTT Advanced Technology Corporation, 1996.

[14] ITU, "Objective measurement of active speech level," International Telecommunication Union, Telecommunication Standardization Sector (ITU-T) Recommendation P.56, Mar. 1993.

[15] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[16] H. Yu and T. Fingscheidt, "Instrumental musical tones measurement of arbitrary noise reduction systems," in *Proc. of 38th German Annual Conf. on Acoust. (DAGA)*, Darmstadt, Germany, Mar. 2012, pp. 255–256.

[17] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Seattle, WA, USA, Sept. 2008.