MMSE BASED NOISE PSD TRACKING WITH LOW COMPLEXITY

Richard C. Hendriks and Richard Heusdens*

Delft University of Technology, The Netherlands
{R.C.Hendriks, R.Heusdens}@TUDelft.nl

Jesper Jensen

Oticon A/S, Denmark
jsj@oticon.dk

ABSTRACT

Most speech enhancement algorithms heavily depend on the noise power spectral density (PSD). Because this quantity is unknown in practice, estimation from the noisy data is necessary.

We present a low complexity method for noise PSD estimation. The algorithm is based on a minimum mean-squared error estimator of the noise magnitude-squared DFT coefficients. Compared to minimum statistics based noise tracking, segmental SNR and PESQ are improved for non-stationary noise sources with 1 dB and 0.25 MOS points, respectively. Compared to recently published algorithms, similar good noise tracking performance is obtained, but at a computational complexity that is in the order of a factor 40 lower.

Index Terms— Noise PSD estimation, speech enhancement

1. INTRODUCTION

An often used strategy to increase listening comfort, pleasantness and robustness of speech communication systems is to apply single-channel noise reduction. Often, these algorithms estimate the clean signal by applying a discrete Fourier transformation (DFT) to the noisy signal on a frame-by-frame basis and then estimate the noise-free DFT coefficients by applying Bayesian estimators, e.g., [1][2][3]. These algorithms depend on the noise power spectral density (PSD), which is in general unknown and must be estimated.

For rather stationary noise sources, the noise PSD can be estimated using explicit voice activity detection [4], or through more advanced methods based on minimum statistics [5] (MS). However, these methods are less suitable when the noise is fast varying and speech is continuously present at a certain frequency.

To track the PSD of non-stationary noise sources, more advanced methods could be used, e.g., the classified codebook [6] (CC) or the DFT-subspace (DFT-SS) approach [7]. However, the CC approach [6] works best for noise-types for which the algorithm is trained, and both methods might be too complex for applications with very low-complexity constraints like mobile phones, hearing aids, etc. Therefore, in [8], a high-resolution DFT (HR-DFT) approach was presented with similar performance as the method in [7], but with reduced computational complexity.

Another low complexity method for noise PSD estimation is weighted noise estimation (WNE) [9]. This method estimates the noise PSD by applying a heuristically motivated weighting function to the magnitude-squared noisy DFT coefficients. Although WNE has low complexity, its noise PSD estimates are biased by definition.

The noise PSD estimator presented in this paper has similar good performance as the methods in [7] and [8], but an even lower computational complexity. The combination of good noise tracking performance and low computational complexity is highly relevant for applications with low-complexity constraints like hearing aids. Similar to the method in [9], we use a weighting function to estimate the noise PSD. However, instead of using a heuristically motivated weighting function, we derive a weighting function that is optimal in minimum mean-squared error (MMSE) sense. Moreover, the proposed estimator is combined with an analytically derived bias-compensation that overcomes under-estimation of the noise PSD.

Recently, a similar approach was proposed by Yu [10]. However, whereas the bias compensation in [10] was motivated rather heuristically, we show in this paper that the bias compensation can be derived more rigorously based on an underlying noisy signal model.

The presented method differs from the previously presented HR-DFT [7] and DFT-SS [8] methods in the exploited speech model. Both methods assume that speech can be described with a low-rank model. The proposed does not make any assumption about the rank of the signal model, but exploits an MMSE estimator based on distributional assumptions for the speech and noise DFT coefficients.

2. NOTATION AND ASSUMPTIONS

Let $Y(k,i)$ denote the DFT coefficient for frequency bin k, and time-frame i. We assume that the noisy signal consists of speech degraded by additive noise. Due to linearity of the Fourier transform it holds that $Y(k,i) = X(k,i) + W(k,i)$, where Y, X and W are the noisy speech, clean speech and noise DFT coefficient, respectively. For W and Y we use a polar notation for mathematical convenience, i.e., $W = Ne^{j\Delta}$ and $Y = Re^{j\Theta}$, respectively. The DFT coefficients are assumed to be complex zero-mean random variables that are statistically independent across time and frequency. Further, it is assumed that X and W are statistically independent. We use uppercase letters to denote random variables and corresponding lowercase letters for their realizations. Since all expressions in this paper are per time-frame i and frequency bin-index k, we leave out these indices for notational convenience. Further, *a priori* and *a posteriori* SNR are defined as $\xi = \sigma_X^2/\sigma_W^2$ and $\zeta = |Y|^2/\sigma_W^2$, respectively, where σ_X^2 and σ_W^2 denote the variance of X and W, respectively.

3. MMSE BASED NOISE PSD TRACKING

In order to estimate the noise PSD, we exploit an MMSE estimator of the noise magnitude-squared DFT coefficients, i.e., N^2. Let E denote the statistical expectation operator. The MMSE estimator of N^2 is then defined by the conditional expectation $E\{N^2|Y\}$. Using Bayes' rule, we can write

$$E\{N^2|Y\} = \frac{\int_0^{+\infty}\int_0^{2\pi} n^2 f_{Y|N,\Delta}(y|n,\delta) f_{N,\Delta}(n,\delta)\, d\delta\, dn}{\int_0^{+\infty}\int_0^{2\pi} f_{Y|N,\Delta}(y|n,\delta) f_{N,\Delta}(n,\delta)\, d\delta\, dn}. \quad (1)$$

*The research is supported by the Oticon foundation and the Dutch Technology Foundation STW.

Assuming that both the speech and noise DFT coefficients have a complex-Gaussian distribution, it follows that

$$f_{Y|N,\Delta}(y|n,\delta) = \frac{1}{\pi\sigma_X^2}\exp\left(\frac{2nr\cos(\delta-\theta)-r^2-n^2}{\sigma_X^2}\right) \quad (2)$$

and

$$f_{N,\Delta}(n,\delta) = \frac{n}{\pi\sigma_W^2}\exp\left(-\frac{n^2}{\sigma_W^2}\right). \quad (3)$$

Substituting Eqs. (2) and (3) into Eq. (1) and using using [11, Eqs. 8.431.5 and 6.643.2] gives

$$E\{N^2|Y\} = \left(\frac{1}{(1+\xi)^2} + \frac{\xi}{(1+\xi)\zeta}\right)|Y|^2. \quad (4)$$

Notice that a similar expression was derived in [12] for estimating the speech magnitude-squared DFT coefficients.

To be more in line with histograms of speech DFT coefficients, see e.g. [2], super-Gaussian instead of Gaussian distributions can be used to model the distribution of speech DFT coefficients. However, the distribution $f_{Y|N,\Delta}$ then becomes super-Gaussian, which complicates analytic derivations of Eq. (1). For simplicity, we therefore maintain the assumption that speech DFT coefficients are Gaussian distributed.

3.1. Biased Estimator

Computing the expectation of Eq. (4) with respect to Y, i.e., $E_Y\{\cdot\}$, it can be shown that $E_Y\{E\{N^2|Y\}\} = \sigma_W^2$, that is, the estimator in Eq. (4) is unbiased. However, in practice the true *a priori* SNR ξ is unknown and estimation of ξ might introduce a bias. Here we consider the following maximum likelihood (ML) estimator for ξ [1] based on the noise PSD from the previous time-frame, that is

$$\hat{\xi}(k,i) = \max\left(\frac{|Y(k,i)|^2}{\hat{\sigma}_W^2(k,i-1)} - 1, 0\right). \quad (5)$$

Let $E\{N^2|y;\hat{\xi}\}$ denote the estimator from Eq. (4) based on $\hat{\xi}$ in Eq. (5). Further, let B be the bias-factor defined as

$$B = \frac{\sigma_W^2}{E_Y\{E\{N^2|y;\hat{\xi}\}\}} = \frac{\sigma_W^2}{\int_R\int_\Theta E\{N^2|y;\hat{\xi}\}f_Y(y)rd\theta dr} \quad (6)$$

where the integral in Eq. (6) is expressed in polar coordinates for mathematical convenience. With the assumption that speech and noise DFT coefficients are Gaussian distributed, it follows that Y has a complex zero-mean Gaussian distribution, with variance σ_Y^2. Solving the integrals in Eq. (6) using [11, Eq. 3.381.1] it follows that

$$B^{-1}(\xi) = \left((1+\xi)\,\gamma\left(2, \frac{1}{1+\xi}\right) + e^{-\frac{1}{1+\xi}}\right), \quad (7)$$

where $\gamma(\nu,\mu)$ is the incomplete Gamma function.

Fig. 1 shows B as a function of the *a priori* SNR ξ. For high ξ, the estimator $E\{N^2|y;\hat{\xi}\}$ is unbiased while for low ξ, it is under-biased. The bias in $E\{N^2|y;\hat{\xi}\}$ can completely be compensated using the analytical expression in Eq. (7). The accuracy at which this can be done depends on how accurately ξ can be estimated.

To compute $B^{-1}(\xi)$ in Eq. (7), we will use a different estimator for ξ than for computing $E\{N^2|y;\hat{\xi}\}$ in Eq. (4). For computing $E\{N^2|y;\hat{\xi}\}$ we proposed to use the ML estimate of ξ in Eq. (5), which enables analytical derivation of the bias in $E\{N^2|y;\hat{\xi}\}$. Computing $B^{-1}(\xi)$ will be based on an estimate of ξ obtained using the decision-directed (DD) approach [1], denoted by $\hat{\xi}_{DD}$, as this will lead to accurate estimates with a smooth evolution across time.

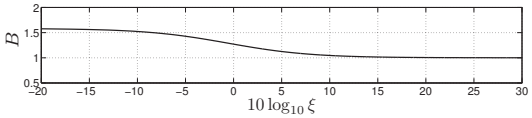

Fig. 1. Bias-factor B as a function of SNR.

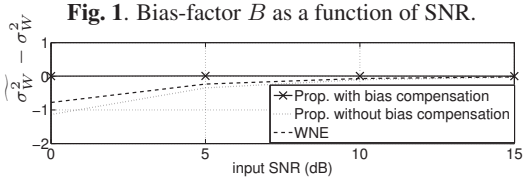

Fig. 2. Obtained bias with and without bias compensation.

To demonstrate the accuracy of the above outlined bias compensation, speech signals are degraded by white Gaussian noise at input SNRs of 0, 5 , 10 and 15 dB. Subsequently, σ_W^2 is estimated with and without bias compensation, i.e., $\widehat{\sigma_W^2} = E\{N^2|y;\hat{\xi}\}B(\hat{\xi}_{DD})$ and $\widehat{\sigma_W^2} = E\{N^2|y;\hat{\xi}\}$, and averaged across all time-frequency indices. Fig. 2 shows the difference between the estimated noise level (with and without the proposed bias compensation) and the true noise level. For comparison we also show the difference between WNE and the true noise level. From Fig. 2 we can conclude that the proposed bias compensation indeed leads to unbiased estimates.

4. ALGORITHM OVERVIEW

In this section we outline the processing steps of the proposed algorithm for a frequency-bin k and time-frame i.

1. Compute $\hat{\xi}(k,i)$ using Eq. (5).
2. Compute $E\{N^2|y;\hat{\xi}(k,i)\}$ by substituting $\hat{\xi}(k,i)$ from Eq. (5) into Eq. (4).
3. Estimate $\xi(k,i)$ using $\widehat{\sigma_W^2}(k,i-1)$ and the DD approach [1], denoted by $\hat{\xi}_{DD}$.
4. Compute $B(\hat{\xi}_{DD})$ using Eq. (7).
5. Compute $\widehat{\sigma_W^2}(k,i) = E\{N^2|y;\hat{\xi}(k,i)\}B(\hat{\xi}_{DD})$.
6. Smooth $\widehat{\sigma_W^2}(k,i)$ across time to reduce its variance, i.e., $\widehat{\sigma_W^2}(k,i) = \beta\widehat{\sigma_W^2}(k,i-1) + (1-\beta)\widehat{\sigma_W^2}(k,i)$, where we use $\beta = 0.8$ in our experimental results.

To overcome a complete locking of the algorithm in the unlikely situation that the noise level would make an abrupt step from one sample to another, we adopt the safety-net proposed in [13] and compute the minimum $P_{\min}(k,i)$ of $|y(k,i)|^2$ across a time-interval of 0.8 seconds. Using $P_{\min}(k,i)$, the noise PSD is updated by $\widehat{\sigma_W^2}(k,i) = \max\left[\widehat{\sigma_W^2}(k,i), P_{\min}(k,i)\right]$.

5. EVALUATION

The proposed algorithm is compared to five reference methods, namely, MS [5], DFT-SS [7], HR-DFT [8], WNE [9] and the method by Yu [10]. Evaluations are performed using a data-base of more than 7 minutes of Danish speech spoken by 9 female and 8 male speakers. The signals were degraded by noise sources at input SNRs of 0, 5, 10, and 15 dB. The noise sources are circle saw noise, passing train noise, passing car noise, and white noise modulated by the function,

$$f(m) = 1 + 0.5\sin(2\pi\, m f_{mod}/f_s), \quad (8)$$

where m is the sample index, f_s the sampling frequency, and f_{mod} the modulation frequency, which increases linearly in 25 seconds from 0 Hz to 0.5 Hz, i.e. a maximum change of the noise PSD of approximately 10 dB per second. Fig. 3 shows an example of a realization of this noise source. All signals are sampled at a frequency of $f_s = 8$ kHz and start with a noise-only period of 0.5 seconds. All algorithms use the first 0.1 seconds for initialization, which is therefore excluded in performance measurements. The time-frames have a length of $K = 256$ samples with 50% overlap, and are windowed using a square-root-Hann window.

5.1. Performance Measures

The noise tracking performance is evaluated using the symmetric log-error distortion measure [7]

$$\text{LogErr} = \frac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I}\left|10\log_{10}\left[\frac{\sigma_W^2(k,i)}{\widehat{\sigma_W^2}(k,i)}\right]\right| \quad \text{[dB]}, \qquad (9)$$

where I is the number of signal-frames and $\sigma_W^2(k,i)$ is the ideal noise PSD, which is obtained by smoothing noise periodograms across time using an exponential window, i.e.

$$\sigma_W^2(k,i) = 0.9\sigma_W^2(k,i-1) + 0.1|w(k,i)|^2. \qquad (10)$$

To evaluate speech enhancement performance, all methods are combined with a single-channel DFT-based noise reduction system. This algorithm uses the DD approach [1] for *a priori* SNR estimation. To estimate the clean speech DFT coefficients we use the magnitude-DFT MMSE estimator presented in [3], which assumes that speech magnitude-DFT coefficients are generalized Gamma distributed with parameters $\gamma = 1$ and $\nu = 0.6$. Speech enhancement performance is evaluated using PESQ [14] and segmental SNR defined as

$$\text{SNR}_{\text{seg}} = \frac{1}{I}\sum_{i=0}^{I-1}\mathcal{T}\left\{10\log_{10}\frac{\|x_t(i)\|^2}{\|x_t(i)-\hat{x}_t(i)\|^2}\right\} \quad \text{[dB]},$$

where $x_t(i)$ and $\hat{x}_t(i)$ denote time-frame i of the clean speech signal x_t and the enhanced speech signal $\hat{x}_t$, respectively, and $\mathcal{T}(x) = \min\{\max(x,-10),35\}$ constrains the estimated SNR per frame to the range of -10 dB to 35 dB.

5.2. Performance Evaluation

Fig. 3 shows an example of noise PSD tracking for a frequency bin centered at 900 Hz. In this example we compare noise PSD estimation using MS, HR-DFT and WNE, for a female speech signal degraded by modulated white noise at an overall SNR of 5 dB. Together with the estimated noise PSDs we also show the ideal noise PSD $\sigma_W^2(k,i)$ obtained using Eq. (10). Fig. 3(a) shows the noisy signal. Fig. 3(b) shows the noise PSD estimated by the proposed method, MS and the true noise PSD and subplot Fig. 3(c) shows the noise PSD estimated by the HR-DFT approach, WNE and the true noise PSD. It is clear that WNE underestimates the noise PSD, independent of how slow the true noise PSD changes. For a rather slowly changing noise PSD, e.g., in the time-span from 0 - 7 seconds, the proposed method, MS and the HR-DFT approach lead to similar estimates of the noise PSD. When the noise level shows faster variations, MS is not able to follow this. The proposed and the HR-DFT method are still able to track these fast changes rather accurately.

Fig. 4 shows a more detailed evaluation of noise tracking performance in terms of the symmetric log-error distortion measure. Figs. 5 and 6 demonstrate the impact on speech enhancement performance

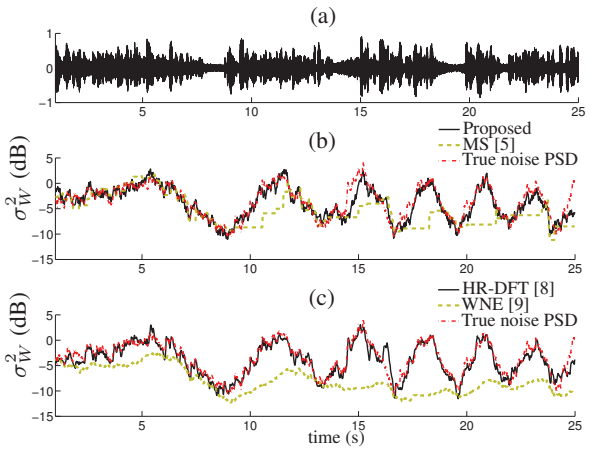

Fig. 3. (a) Speech degraded by modulated white noise at 5 dB SNR. (b)-(c) Comparison between true noise PSD and noise PSD estimators for DFT bin centered around 0.9 kHz.

Table 1. Normalized processing-time.

	DFT-SS	HR-DFT	MS	WNE	Prop.	Yu
time	42	2.67	7.8	1.2	1.0	0.3

as measured by segmental SNR and PESQ, respectively, of the various noise tracking algorithms for all aforementioned noise sources.

The figures show that WNE is inferior compared to the other methods, in terms of all three performance measures. The general trend is that MS and the method presented by Yu [10] have worse noise tracking and speech enhancement performance than DFT-SS, HR-DFT and the proposed approach. In general, the performance of the proposed approach and the DFT-SS and HR-DFT approach is more or less equal. Speech enhancement improvements of the proposed approach over MS for non-stationary noise sources, e.g., passing car noise and modulated white noise, are in the order of 0.25 in terms of PESQ and 1 dB in terms of segmental SNR.

5.3. Complexity

The computational complexity in terms of processing time of matlab implementations of all six algorithms is given in Table 1. Notice that the numbers given in Table 1 are rough estimates that are meant as an indication. In general they depend on implementational details. For the proposed method, the numbers in Table 1 reflect all processing steps outlined in Sec. 4 including the safety-net adopted from [13].

The evaluation in the preceding section reveals that the performance of the proposed approach is similar to the previously presented DFT-SS and HR-DFT method. However, as we see from Table 1, computational complexity of the proposed approach is much lower compared to the DFT-SS approach and also somewhat lower than that of the HR-DFT approach. The lower computational complexity is mainly determined by the fact that no additional spectral transforms are needed as is the case for the HR-DFT and DFT-SS approach. From Table 1 we see that compared to MS, the proposed approach has a complexity that is about a factor 7 lower. Compared to WNE, the computational complexity is in the same order of magnitude, while both MS and WNE have a worse noise tracking performance than the proposed approach. Compared to the method from [10], performance is improved, while computational complexity is slightly higher. This slightly higher computational complexity is mainly due to the safety-net that is proposed in Sec. 4.

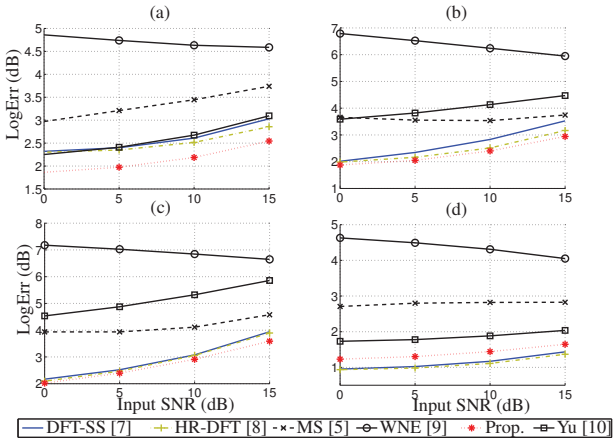

Fig. 4. LogErr (dB) (a) circle saw noise, (b) passing train noise, (c) passing car noise, (d) modulated white noise.

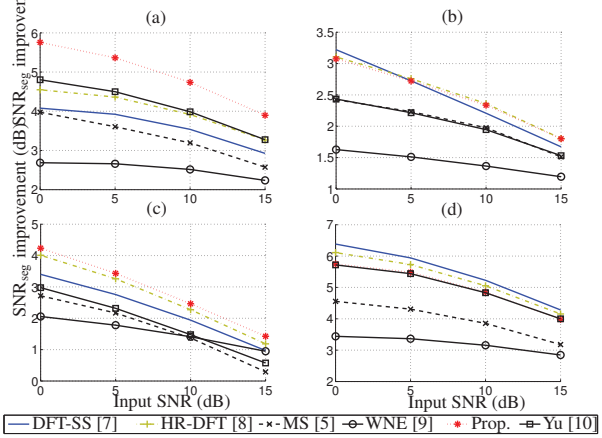

Fig. 5. $\mathrm{SNR}_{\mathrm{seg}}$ improvement (dB) (a) circle saw noise, (b) passing train noise, (c) passing car noise, (d) modulated white noise.

6. CONCLUSIONS

We proposed a low-complexity MMSE estimator of the noise power spectral density. In comparison to reference methods like minimum statistics and weighted noise estimation, both noise tracking and speech enhancement performance is improved. Compared to previously presented DFT-subspace and high resolution DFT based noise PSD estimation, the proposed method has similar performance, but achieves this at a much lower computational complexity.

7. 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 Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[2] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[3] 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 and Language Processing*, vol. 15, no. 6, pp. 1741 – 1752, August 2007.

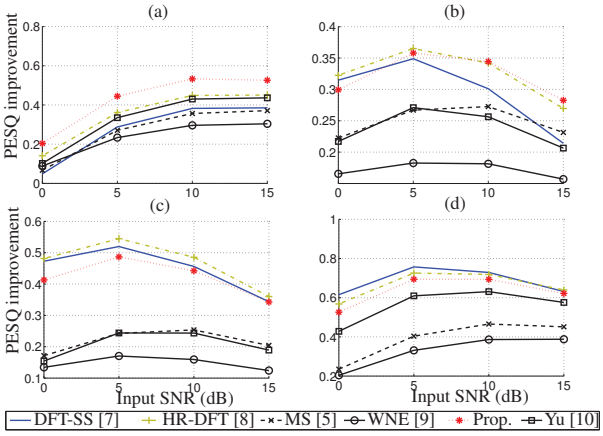

Fig. 6. PESQ improvement (a) circle saw noise, (b) passing train noise, (c) passing car noise, (d) modulated white noise.

[4] 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, January 1999.

[5] 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, July 2001.

[6] S. Srinivasan, *Knowledge-Based Speech Enhancement*, Ph.D. thesis, KTH, 2005.

[7] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio Speech and Language Processing*, vol. 16, no. 3, pp. 541–553, March 2008.

[8] R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "Fast noise psd estimation with low complexity," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2009, pp. 3881–3884.

[9] M. Kato, A. Sugiyama, and M. Serizawa, "Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA," in *IWAENC*, 2001.

[10] R. Yu, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2009, pp. 4421–4424.

[11] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series and Products*, New York: Academic, 6th ed. edition, 2000.

[12] 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. 10, pp. 1043–1051, 2003.

[13] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio Speech and Language Processing*, vol. 16, no. 6, pp. 1112–1123, Ausgust 2008.

[14] ITU-T P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assesment of narrowband telephone networks and speech codecs," Tech. Rep., 2000.