ASSESSMENT OF BROADBAND SNR ESTIMATION FOR HEARING AID APPLICATIONS

Tobias May, Borys Kowalewski, Michal Fereczkowski and Ewen N. MacDonald

Hearing Systems Group,
Technical University of Denmark,
DK-2800 Kgs. Lyngby, Denmark
{tobmay, bokowal, mfer, emcd}@elektro.dtu.dk

ABSTRACT

An accurate estimation of the broadband input signal-to-noise ratio (SNR) is a prerequisite for many hearing-aid algorithms. An extensive comparison of three SNR estimation algorithms was performed. Moreover, the influence of the duration of the analysis window on the SNR estimation performance was systematically investigated. The most accurate approach utilized an estimation of the clean speech power spectral density (PSD) and the noisy speech power across a sliding window of $1280\,\mathrm{ms}$ and achieved an total SNR estimation error below $3\,\mathrm{dB}$ across a wide variety of background noises and input SNRs.

Index Terms— Signal-to-noise ratio estimation, noise power estimation, hearing-aid algorithms

1. INTRODUCTION

Many hearing-aid algorithms require general knowledge about the acoustic environment. For example, it has been shown that the benefit of fast-acting wide dynamic range compression (WDRC) over linear amplification increases with decreasing SNR [1]. Moreover, a more recent modeling study suggests that using SNR-specific time constants might benefit speech intelligibility and quality [2]. Thus, it would be advantageous to adjust the time constants used in hearing-aid WDRC based on the *a priori* SNR, which is defined as the ratio of the speech to the noise power.

Whereas the task of estimating the SNR in individual discrete Fourier transform (DFT) bins is well studied for the application of speech enhancement [3, 4, 5, 6, 7], there has been less focus on deriving the *broadband* SNR given the noisy speech signal. The log-energy histogram of noisy speech reveals a bimodal distribution [8], reflecting both the contribution of noise and noisy speech. This observation is utilized by the SNR estimator developed by the National Institute of Standards and Technology (NIST) [9]. However, the NIST algorithm measures the peak SNR, and thus overestimates the true SNR. Another approach, termed waveform amplitude distribution analysis (WADA), examines the amplitude distribution of noisy speech [10]. Although the WADA approach was shown to be more accurate than the NIST algorithm [10], the background noise is assumed to follow a Gaussian distribution. Consequently, a violation of this assumption is likely to degrade the performance of the algorithm.

An obvious alternative is the application of speech enhancement algorithms which typically estimate the noise PSD. Given the noisy speech power and an estimation of the noise PSD, the input SNR can be estimated. Recently, several noise PSD estimators were compared in terms of their ability to estimate the broadband input SNR [11]. Among all tested noise PSD estimators, the approaches proposed by Gerkmann and Henrdiks [12] and Hendriks *et al.* [13] provided the most accurate results. However, instead of using the noise PSD directly for SNR estimation, it can also be used to estimate the clean speech PSD by employing a minimum mean-square error (MMSE)-based estimator [3, 14]. This would allow the use of the decision-directed approach, which was shown to substantially reduce the amount of speech distortions in speech enhancement applications [3]. The input SNR could subsequently be derived from the noisy speech power and the estimated clean speech PSD.

In contrast to speech enhancement applications, SNR-specific hearing-aid processing, such as WDRC strategies, do not require an estimation of the SNR for individual DFT bins. Depending on the desired temporal resolution, the SNR estimation can be integrated across time, which is likely to improve the performance. Thus, from an application point of view, it would be valuable to investigate the influence of the temporal resolution on SNR estimation accuracy. However, previous studies focused on file-based evaluation [9, 10, 11] and, thus, the impact of the window duration on SNR estimation performance has not yet been clarified.

The goal of the present study was to compare different approaches to estimate the true broadband SNR of noisy speech mixtures. Specifically, the benefit of deriving the clean speech PSD using an MMSE-based estimator was tested by comparing it to the performance of a noise PSD estimator and to the WADA algorithm. An extensive evaluation was performed using a variety of stationary and non-stationary noise types mixed with speech at a wide range of input SNRs. A particular focus was to investigate the influence of the temporal resolution on the SNR estimation performance. The accuracy of the estimated SNR was quantified by using the log-error distortion [15] measure, which allows to investigate the amount of over- and underestimation with respect to the true SNR.

2. BROADBAND SNR ESTIMATION

For a given window size $\mathcal{W}_\xi$, the broadband *a priori* SNR $\xi^{\mathrm{dB}}[\ell]$ for time frame ℓ was defined as the ratio of the speech PSD $\sigma_{\mathrm{S}}^2[\ell, f]$ to the noise PSD $\sigma_{\mathrm{N}}^2[\ell, f]$ integrated across all frequency bins f

$$\xi^{\mathrm{dB}}[\ell] = 10\log_{10}\left(\min\left(\max\left(\frac{\sum_f \sigma_{\mathrm{S}}^2[\ell, f]}{\sum_f \sigma_{\mathrm{N}}^2[\ell, f]}, \xi_{\min}\right), \xi_{\max}\right)\right). \tag{1}$$

The dynamic range of $\xi^{\mathrm{dB}}[\ell]$ was limited by an lower $\xi_{\min}$ and an upper bound $\xi_{\max}$. Obviously both $\sigma_{\mathrm{S}}^2[\ell, f]$ and $\sigma_{\mathrm{N}}^2[\ell, f]$ are unknown and thus, $\xi^{\mathrm{dB}}[\ell]$ had to be blindly estimated from the noisy

speech mixture. Given an estimation of the SNR in the linear domain $\hat{\xi}[\ell]$, the final SNR in dB was determined by

$$\hat{\xi}^{\mathrm{dB}}[\ell] = 10\log_{10}\left(\min\left(\max\left(\hat{\xi}[\ell], \xi_{\min}\right), \xi_{\max}\right)\right). \quad (2)$$

In the following, the three tested SNR estimation approaches are described.

2.1. WADA

The WADA algorithm [10] operates in the time domain and was used to analyze the amplitude distribution of the noisy speech mixture. To obtain an SNR estimation for a specific temporal resolution, the time-domain signal was segmented into overlapping frames and the SNR $\hat{\xi}[\ell]$ was estimated for each frame separately by the WADA algorithm.

2.2. Noise PSD estimation

Given the short-time discrete Fourier transform (STFT) representation of noisy speech $X[\lambda, k]$ with λ and k indexing the time frame and the frequency bin, respectively, the approach by Hendriks *et al.* [13] was used to estimate the noise PSD denoted by $\hat{\sigma}^2_{\mathrm{N}}[\lambda, k]$. The STFT was always computed with a Hamming window $\mathcal{W}_{\mathrm{STFT}}$ of 20 ms duration and a step size $\mathcal{S}_{\mathrm{STFT}}$ of 10 ms. Subsequently the SNR $\hat{\xi}[\ell]$ was estimated by integrating the noisy speech power and the estimated noise PSD across all frequency bins and L time frames

$$\hat{\xi}[\ell] = \max\left(\frac{\sum_{m=0}^{L-1}\sum_{k}|X[\ell R + m, k]|^2}{\sum_{m=0}^{L-1}\sum_{k}\hat{\sigma}^2_{\mathrm{N}}[\ell R + m, k]} - 1, \epsilon\right), \quad (3)$$

where L and the step size R were adjusted to match the predefined window size $\mathcal{W}_{\xi}$ of the *a priori* SNR. A lower bound ϵ was employed to avoid negative values of $\hat{\xi}[\ell]$.

2.3. Speech PSD estimation

Instead of estimating the SNR directly from the estimated noise PSD $\hat{\sigma}^2_{\mathrm{N}}[\lambda, k]$, the MMSE estimator by Erkelens *et al.* [14] was used to obtain an estimation of the clean speech PSD $\hat{\sigma}^2_{\mathrm{S}}[\lambda, k]$. This estimator assumes that the distribution of clean DFT coefficients can be characterized by a generalized Gamma distribution with the two parameters γ and ν [14]. Moreover, it utilizes the decision-directed approach [3], which incorporates a non-linear smoothing, and thus, allows for a more accurate estimation of the *a priori* SNR [16]. Similar to Sect. 2.2, the noisy speech power and the speech PSD was integrated across all frequency bins and a predefined number of time frames L to match the window size $\mathcal{W}_{\xi}$ of the *a priori* SNR

$$\hat{\xi}[\ell] = \frac{1}{\max\left(\frac{\sum_{m=0}^{L-1}\sum_{k}|X[\ell R + m, k]|^2}{\sum_{m=0}^{L-1}\sum_{k}\hat{\sigma}^2_{\mathrm{S}}[\ell R + m, k]} - 1, \epsilon\right)}. \quad (4)$$

Table 1. Tested window durations $\mathcal{W}_{\xi}$ and step sizes $\mathcal{S}_{\xi}$ for the *a priori* SNR as well as the corresponding algorithm settings.

$\mathcal{W}_{\xi}$	$\mathcal{S}_{\xi}$	$\mathcal{W}_{\mathrm{STFT}}$	$\mathcal{S}_{\mathrm{STFT}}$	L	R
20 ms	10 ms	20 ms	10 ms	2	1
40 ms	20 ms	20 ms	10 ms	4	2
80 ms	40 ms	20 ms	10 ms	8	4
160 ms	80 ms	20 ms	10 ms	16	8
320 ms	160 ms	20 ms	10 ms	32	16
640 ms	320 ms	20 ms	10 ms	64	32
1280 ms	640 ms	20 ms	10 ms	128	64
2560 ms	1280 ms	20 ms	10 ms	256	128
5120 ms	2560 ms	20 ms	10 ms	512	256

3. EVALUATION

3.1. Stimuli

Noisy speech with a sampling frequency of 16 kHz was created by corrupting clean speech with background noise from the DEMAND database [17]. The DEMAND database consists of 18 different noise types that are classified into 6 categories (domestic, office, public, transportation, street and nature), spanning across a wide variety of both stationary and non-stationary noise types. Each noisy speech mixture consisted of an initial noise-only segment of 100 ms duration followed by 5 randomly selected sentences from the TIMIT corpus [18] that were separated by 100 ms long noise-only segments.

For evaluation, 500 noisy speech mixtures with an average duration of 15.9 sec were created for each of the 18 background noises and 9 different SNRs ($-15, -10, -5, 0, 5, 10, 15, 20$ and 25 dB), leading to a set of 81000 mixtures. The accuracy of the SNR estimation $\hat{\xi}^{\mathrm{dB}}$ was evaluated across all N time frames of a given noisy speech mixture by the log-error distortion measure [15], which allowed to investigate the amount of over- and underestimation in comparison to the true SNR ξ^{dB}

$$\mathrm{LogError}_{\mathrm{Total}} = \mathrm{LogError}_{\mathrm{Over}} + \mathrm{LogError}_{\mathrm{Under}} \quad (5)$$

with

$$\mathrm{LogError}_{\mathrm{Over}} = \frac{1}{N}\sum_{\ell=1}^{N}\left|\min(0, \xi^{\mathrm{dB}}[\ell] - \hat{\xi}^{\mathrm{dB}}[\ell])\right| \quad (6)$$

$$\mathrm{LogError}_{\mathrm{Under}} = \frac{1}{N}\sum_{\ell=1}^{N}\left|\max(0, \xi^{\mathrm{dB}}[\ell] - \hat{\xi}^{\mathrm{dB}}[\ell])\right|. \quad (7)$$

3.2. Algorithm settings

All three SNR estimation methods were tested with 9 different window durations $\mathcal{W}_{\xi}$ ranging from 20 ms to 5120 ms with 50 % overlap, as listed in Tab. 1. The noise PSD and the speech PSD estimation was always carried out with a fixed window duration $\mathcal{W}_{\mathrm{STFT}}$ of 20 ms with 50 % overlap. In order to obtain the predefined temporal resolution defined by $\mathcal{W}_{\xi}$, the integration of the SNR estimation across time, according to Eq. (3) and Eq. (4), was controlled by the two parameters L and R.

The noise PSD approach by Hendriks *et al.* was configured with the default parameters reported in [13] and initialized for each noisy speech mixture by averaging the PSD across the initial noise-only segment of 100 ms. The MMSE estimator that was used to obtain the clean speech PSD was configured with the two generalized Gamma

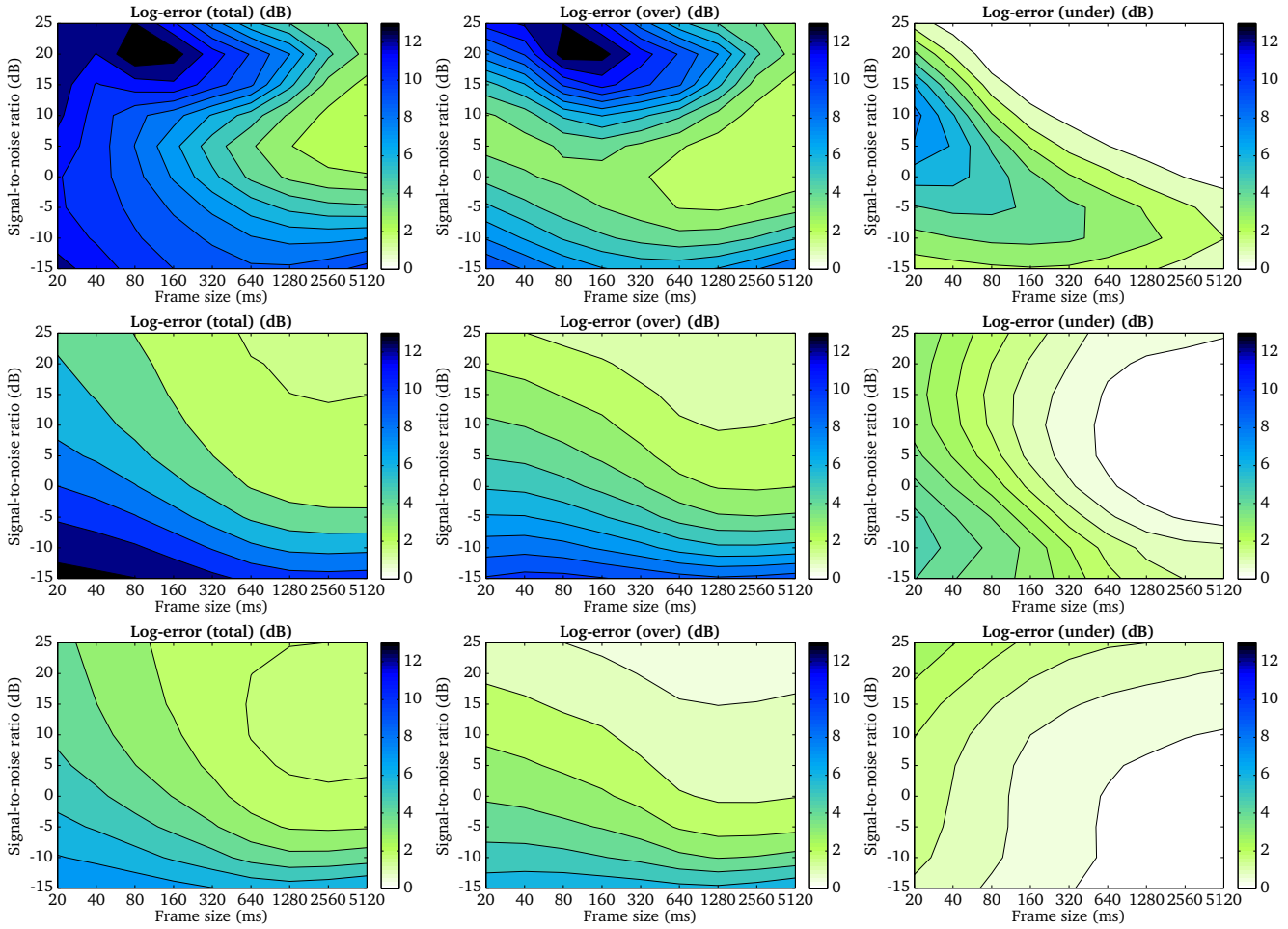

Fig. 1. SNR estimation error of the WADA algorithm (top panels), the noise PSD approach (middle panels) and the speech PSD method (bottom panels) as a function of the frame size and the input SNR averaged across all noise types. Left panels: total SNR estimation error, middle panels: SNR overestimation, right panels: SNR underestimation.

parameters $\gamma = 1$ and $\nu = 0.6$ [14]. Moreover, the smoothing factor α used by the decision-directed approach corresponded to a time constant of $0.792\,\mathrm{s}$. Finally, the lower and upper SNR bounds $\xi_{\min}$ and $\xi_{\max}$ corresponded to -30 and $30\,\mathrm{dB}$.

4. RESULTS

Figure 1 shows the SNR estimation performance of the WADA algorithm (top panels), the noise PSD approach (middle panels) and the speech PSD method (lower panels) as a function of the window duration and the input SNR. The total error is shown in the left panels, whereas the amount of over- and underestimation is presented in the middle and the right panels, respectively.

It can be seen that the WADA algorithm produced the largest deviations from the true SNR (top left panel), in particular for window durations shorter than $1280\,\mathrm{ms}$. When a longer window was used, the WADA algorithm was most accurate for positive input SNRs, which is in line with the results presented in [10]. The noise PSD approach substantially reduced the total SNR estimation error (middle left panel) for positive input SNRs when a window size of at least $320\,\mathrm{ms}$ was used. However, the estimation errors for negative input SNRs were still in the range of $10\,\mathrm{dB}$. Among all tested approaches, the speech PSD method achieved the lowest SNR estimation errors (bottom left panel). The total log-error was within $4\,\mathrm{dB}$ for a wide range of input SNRs and window durations. Generally, all approaches tended to overestimate the true SNR (middle panels), which is apparent when comparing the amount of overestimation with the amount of underestimation (right panels).

The total SNR estimation error averaged across all input SNRs is shown in Fig. 2 as a function of the window size. An accurate estimation of the input SNR was quite challenging when a short window duration was used, which is indicated by an average estimation error of up to $12\,\mathrm{dB}$. Moreover, it can be seen that the SNR estimation error decreased systematically for all three methods with increasing window duration. Whereas the WADA algorithm produced the lowest estimation error of $4.5\,\mathrm{dB}$ for the longest window of $5120\,\mathrm{ms}$, the performance of both the noise PSD and the speech PSD method saturated for a window duration of $1280\,\mathrm{ms}$, yielding estimation errors of $4.1\,\mathrm{dB}$ and $2.8\,\mathrm{dB}$, respectively.

The SNR estimation of the three tested methods is illustrated in Fig. 3 for a noisy speech mixture at $0\,\mathrm{dB}$ SNR. The top panel shows the time waveform, whereas the lower panels presents the *a*

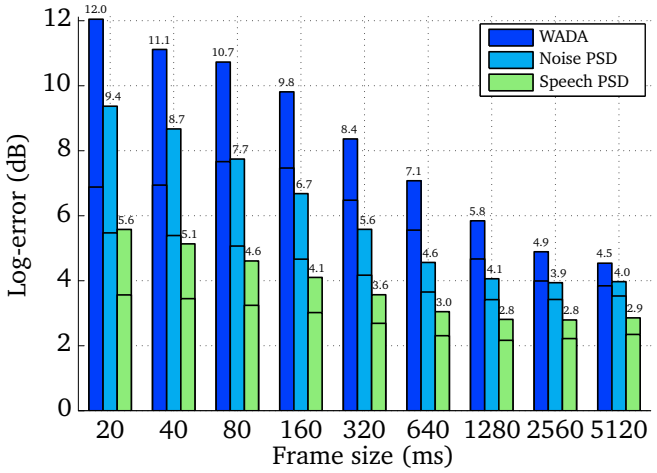

Fig. 2. Total SNR estimation error of the WADA algorithm, the noise PSD approach and the speech PSD method as a function of the frame size averaged across all SNRs and noise types. The lower part of the bars reflect the SNR overestimation, while the upper part of the bars indicate the SNR underestimation.

priori SNR along with the SNR estimation for a temporal window of 320 ms duration. It can be seen that the speech PSD estimator most closely followed the *a priori* known SNR. In contrast, the WADA approach frequently underestimated the *a priori* SNR.

5. DISCUSSION AND CONCLUSION

This study presented a systematic comparison of three algorithms to estimate the broadband input SNR of noisy speech mixtures across a wide range of noise types and input SNRs. Instead of using the estimated noise PSD directly to determine the input SNR, as presented in [11], the most accurate results were obtained by using an estimation of the clean speech PSD. It seems that the MMSE estimator proposed by Erkelens *et al.* [14] combined with the non-linear smoothing provided by the decision-directed approach allowed for an accurate estimation of the clean speech DFT coefficients, and subsequently, produced the most accurate estimation of the broadband input SNR across a wide range of acoustic conditions.

The SNR estimation performance increased with increasing window duration for all tested approaches. Whereas the WADA algorithm produced the lowest error of 4.5 dB with the longest analysis window of 5120 ms, the performance of both the noise PSD and the speech PSD approach saturated around an error of 4.1 dB and 2.8 dB for a window of 1280 ms duration. In general, all tested estimation algorithms showed the tendency to overestimate the true SNR. This bias could potentially be reduced by a non-linear mapping function, e.g. as presented in [11]. However, such a mapping function may only be suitable if the speech material used to create the mapping is known *a priori* and does not change.

A recent study analyzed broadband SNRs in realistic environments recorded by hearing aid users [19]. The underlying SNR was estimated using a manual noise tracking procedure. The broadband SNR estimator based on the speech PSD presented here could be used to replace the manual procedure. This would enable a fully automated analysis of a large number of recordings. Moreover, future work will incorporate the presented SNR estimator into a hearing-aid WDRC system in order to select SNR-specific time constants and to evaluate the effects of such scheme on speech intelligibility, perceived quality and listening effort.

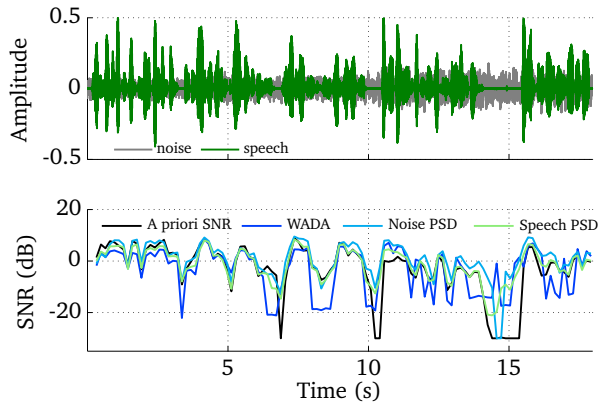

Fig. 3. Illustration of the SNR estimation for a temporal window of 320 ms duration. The top panel shows a sequence of 5 TIMIT sentences mixed with subway noise at 0 dB SNR. The bottom panel presents the *a priori* known SNR and the estimation using the WADA algorithm, the noise PSD approach and the speech PSD method.

6. ACKNOWLEDGEMENTS

This work was supported by the EU FET grant TWO!EARS, ICT-618075 and by the Centre for Applied Hearing Research (CAHR).

7. REFERENCES

[1] E. W. Yund and K. M. Buckles, "Enhanced speech perception at low signal-to-noise ratios with multichannel compression hearing aids," *J. Acoust. Soc. Amer.*, vol. 97, no. 2, pp. 1224–1240, 1995.

[2] J. Kates, "Understanding compression: Modeling the effects of dynamic-range compression in hearing aids," *Int. J. Audiol.*, vol. 49, pp. 395–409, 2010.

[3] 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, 1984.

[4] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[5] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*, 1996, pp. 629–632.

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

[7] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 3, pp. 184–192, 2003.

[8] 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, 1980.

[9] "NIST Speech quality assurance (SPQA) package," Software is available at http://www.itl.nist.gov/iad/mig/tools/.

[10] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech*, 2008, pp. 2598–2601.

[11] M. B. J. Eaton and P. A. Naylor, "A comparison of non-intrusive SNR estimation algorithms and the use of mapping functions," in *Proc. EUSIPCO*, 2013, pp. 3–7.

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

[13] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE-based noise PSD tracking with low complexity," in *Proc. ICASSP*, 2010, pp. 4266–4269.

[14] 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, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

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

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

[17] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. ICA*, 2013.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic-phonetic continuous speech corpus CD-ROM," *National Inst. Standards and Technol. (NIST)*, 1993.

[19] K. Smeds, F. Wolters, and M. Rung, "Estimation of signal-to-noise ratios in realistic sound scenarios," *J. Am. Acad. Audiol.*, vol. 26, no. 2, pp. 183–196, 2015.