FAST NOISE PSD ESTIMATION WITH LOW COMPLEXITY

Richard C. Hendriks and Richard Heusdens

Delft University of Technology,
2628 CD Delft, The Netherlands
{R.C.Hendriks, R.Heusdens}@TUDelft.nl

Jesper Jensen and Ulrik Kjems

Oticon A/S
2765 Smørum, Denmark
{jsj, uk}@oticon.dk

ABSTRACT

Although noise PSD estimation is a crucial part of noise reduction algorithms, most noise PSD estimators have problems in tracking non-stationary noise sources. Recently, a noise PSD estimator based on DFT-subspace decompositions was proposed, which improves estimation of the PSD of such noise sources. However, as this approach is based on eigenvalue decompositions per DFT bin, it might be too computationally demanding for low-complexity applications like hearing aids. In this paper we present a method with similar noise tracking performance as the DFT-subspace approach, but with low computational costs. This method is based on computation of high resolution perodiograms, and can estimate the noise PSD when both speech and noise are present in a frequency bin. When combined with a complete noise reduction system, the proposed method can lead to an improvement for non-stationary noise sources of more than 1 dB segmental SNR and 0.3 on a PESQ scale, compared to standard noise tracking methods such as minimum statistics and the quantile based approach, while computational complexity is in the same order of magnitude.

Index Terms— speech enhancement, noise reduction, noise PSD tracking

1. INTRODUCTION

Typically, estimators that are used for noise reduction of speech signals are dependent on an estimate of the noise power spectral density (PSD), see e.g., [1][2, Chs. 3 and 4][3]. Since this quantity is unknown in advance, it has to be estimated from the noisy speech signal. Accurate estimation of this noise PSD is crucial, as an underestimate of the noise PSD leaves an unnecessary amount of residual noise in the enhanced signal, while an over-estimate leads to speech distortions and a potential loss of speech quality.

Many of the noise PSD estimation algorithms that have been proposed over the last years are based on some form of minima tracking of (smoothed) noisy speech periodograms in a certain time-interval, e.g., minimum statistics (MS) [4]. The length of this time-interval is a decisive factor for reliability of the estimated noise PSD. If the interval is too short, speech will leak into the noise PSD estimate. However, increasing the time-interval will increase tracking delay, especially, when the noise PSD is increasing in level.

A method that does not explicitly depend on minimum tracking is quantile based (QB) noise PSD estimation [5]. This method estimates the noise PSD by computing per DFT bin a temporal quantile p of noisy periodograms in a time-interval. The speed at which this method can estimate the noise PSD for non-stationary noise sources depends on the length of the time-interval. As such, QB noise PSD estimation methods are subject to a similar tradeoff as MS. Since the noise PSD estimate is based on a quantile across time and not on a minimum, QB noise PSD estimation is expected to track decreasing noise levels with larger delay than MS, while an increasing noise level can potentially be tracked faster than MS.

Recently, in [6], a DFT-subspace based method for noise tracking was proposed which improves noise PSD estimation, especially for non-stationary noise sources. Compared to other noise PSD methods the DFT-subspace approach leads to a shorter delay in tracking a time-varying noise PSD. This method exploits the tonal structure in speech and is based on the construction of correlation matrices in the DFT-domain for each time-frequency point. These correlation matrices are decomposed using an eigenvalue decomposition into a mutually orthogonal signal (+ noise) subspace and noise-only subspace. The eigenvalues that describe the energy in the noise-only subspace then allow for an update of the noise PSD, even when speech is present.

Although the DFT-subspace approach improves noise PSD tracking of non-stationary noise sources, the necessary eigenvalue decompositions might be too complex for applications with very low-complexity constraints like hearing aids.

In order to apply fast and reliable noise tracking in such low-complexity constrained applications we present in this paper an algorithm with similar performance as the DFT-subspace approach, but with considerably reduced computational complexity. As the method in [6], the proposed method also exploits the tonal structure of speech. It is based on the computation of noisy periodograms with a frequency resolution that is typically higher than the frequency resolution used in the noise reduction algorithm itself. The high resolution-periodogram is divided in sub-bands, each corresponding to a frequency bin in the noise reduction algorithm. Analogous to the method in [6], we divide the frequency bins within each sub-band to contain noisy speech and noise-only. Assuming that the true noise PSD is constant across the sub-band, the noise-only frequency bins of the high resolution periodogram are used to compute a maximum likelihood (ML) estimate of the noise PSD.

2. DFT-BASED SPEECH ESTIMATORS

Let y_n denote a sampled time-domain noisy speech signal consisting of a speech signal x_n degraded by additive noise n_n, i.e.,

$$y_n = x_n + n_n.$$

For noise reduction y_n is divided in signal-frames of length L_1 by applying a sliding window $w_1(m)$ with $m \in \{0, ...L_1 - 1\}$. For notational convenience it is assumed that the window w_1 is normalized, i.e., $\sum_{m=0}^{L_1-1} w_1^2(m) = L_1$. Let k and i be the frequency-bin

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

index and time-frame index and let $K \geq L_1$ be the DFT size. The noisy DFT coefficients $y(k,i)$ are then obtained using the DFT of the windowed time-frames, that is

$$y(k,i) = \frac{1}{\sqrt{L_1}} \sum_{m=0}^{m=L_1-1} y_n(iL_1/2+m)w_1(m) \exp\left[-2\pi kmj/K\right],$$

where $j = \sqrt{-1}$ is the imaginary unit. Similarly, let $x(k,i)$ and $n(k,i)$ be the clean speech and noise DFT coefficient at frequency bin k and time-frame i. The DFT coefficients $y(k,i)$, $x(k,i)$ and $n(k,i)$ are assumed to be realizations of the zero-mean complex-valued random variables $Y(k,i)$, $X(k,i)$ and $N(k,i)$, respectively. To estimate the clean speech DFT coefficient $x(k,i)$ a gain function $G(k,i)$ is typically applied to the noisy DFT coefficients, that is

$$\hat{x}(k,i) = G(k,i)y(k,i).$$

Although there are many ways to determine this gain function, e.g., using Bayesian principles [7], all gain functions depend on the noise PSD $\sigma_N^2(k,i) = \mathrm{E}[|N(k,i)|^2]$. Since, this quantity is generally unknown it must be estimated from the available data.

3. NOISE PSD ESTIMATION BASED ON HIGH RESOLUTION PERIODOGRAMS

Besides the signal-frames defined in Section 2 to which the actual noise reduction is applied we also define a second type of time frame that we call super-frames. The super-frames are used to estimate the noise PSD using high resolution DFTs (HR-DFTs) and have a length of L_2 samples with $L_2 > L_1$. Let $Q \geq L_2$ be the order of the HR-DFT and let w_2 be a normalized window function such that $\sum_{m=0}^{L_2-1} w_2^2(m) = L_2$. The HR-DFT coefficient of a super-frame at frequency bin q and time-frame i is given by

$$y_{HR}(q,i) = \frac{1}{\sqrt{L_2}} \sum_{m=L_1-L_2}^{m=L_1-1} y_n\left(iL_1/2+m\right) w_2(m) \exp\left[-2\pi qmj/Q\right],$$

where the subscript HR indicates that this is a coefficient of the HR-DFT of a super-frame. The HR-DFT coefficients $y_{HR}(q,i)$ are used to form a high-resolution noisy periodogram $|y_{HR}(q,i)|^2$. Hence, each frequency bin k corresponds to a band of, say P, frequency bins in the high-resolution periodogram estimate $|y_{HR}|^2$. More specifically, the kth band of the high-resolution periodogram consists of the frequency bins $q \in \{q_1, ..., q_2\}$, with $P = q_2 - q_1 + 1$ and

$$q_1 = (k-1/2)\tfrac{Q}{K} \quad \text{and} \quad q_2 = (k+1/2)\tfrac{Q}{K},$$

where it is assumed that Q and K are integer powers of 2. This high frequency resolution makes it possible to estimate the noise PSD at a frequency band k when speech is present in this frequency band as long as the clean speech signal observed in this band can be approximated using less than the P complex exponential basis function that are necessary to represent this frequency band.

To compute an estimate $\hat{\sigma}_N^2(k,i)$ based on the kth frequency band of $|y_{HR}|^2$ we assume that the noise level is constant across this frequency band. Further we assume that the noise DFT coefficients N have a complex Gaussian distribution, which is validated by the fact that the time-span of dependency is relative short for many noise sources [2, Chs. 3 and 4]. Let $\mathcal{M}(k,i)$ be the set of frequency bins in the kth HR-DFT frequency band that do not contain speech energy. The ML estimate of the noise PSD in frequency bin k is then given by

$$\hat{\sigma}_N^2(k,i) = \frac{1}{|\mathcal{M}(k,i)|} \sum_{q \in \mathcal{M}(k,i)} |y_{HR}(q,i)|^2, \tag{1}$$

where $|\mathcal{M}(k,i)|$ denotes the cardinality of $\mathcal{M}(k,i)$. To decrease the variance of $\hat{\sigma}_N^2(k,i)$, smoothing across time can be applied.

3.1. Determining $\mathcal{M}(k,i)$

Determination of $\mathcal{M}(k,i)$ is based on a procedure which was proposed in [6] to determine the dimension of a noise-only subspace. The procedure is based on two assumptions. First it is assumed that the squared-magnitude of the noise DFT coefficients, i.e., $|N_{HR}(q,i)|^2$, is exponentially distributed, which follows automatically from the assumption already made in Section 3 that the noise DFT coefficients are Gaussian distributed. Secondly, we assume that the noise PSD develops relatively slowly across time, which allows us to use the noise PSD estimated in the previous frame, i.e., $\hat{\sigma}_N^2(k,i-1)$, as *a priori* information when estimating the noise PSD in the current frame. This assumption does not limit the performance, since a change in noise PSD of 10 - 15 dB per second can easily be tracked.

With these assumptions, we are now in position to determine which of the frequency bins $q \in \{q_1, ..., q_2\}$ in the kth HR-DFT frequency band do not contain speech energy, by applying a Neyman-Pearson hypothesis test [8] with the following hypotheses

$$\begin{array}{ll} H_0: & |y_{HR}(q,i)|^2 \text{ consists of only noise} \\ H_1: & |y_{HR}(q,i)|^2 \text{ consists of noise and speech.} \end{array} \tag{2}$$

It can be shown that under rather general conditions, an optimal decision test compares $|y_{HR}(q,i)|^2$ to a threshold λ_{th} [8], i.e.

$$|y_{HR}(q,i)|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \lambda_{th}.$$

Using the aforementioned distributional assumption on $|N_{HR}(q,i)|^2$ we can express the threshold λ_{th} as a function of the false-alarm probability P_{fa} by $\lambda_{th} = -\sigma_N^2(k,i-1)\ln P_{fa}$ [8].

3.2. Bias Compensation

Due to spectral leakage from neighboring DFT coefficients that contain speech energy, the estimate $\hat{\sigma}_N^2(k,i)$ is generally biased high. Therefore, we introduce a bias compensation-factor, much along the same lines as in [4], dependent on the cardinality of the set $\mathcal{M}(k,i)$, i.e. $B(|\mathcal{M}(k,i)|)$. Altogether, the noise PSD is estimated by

$$\hat{\sigma}_N^2(k,i) = \frac{1}{B(|\mathcal{M}(k,i)|)|\mathcal{M}(k,i)|} \sum_{q \in \mathcal{M}(k,i)} |y_{HR}(q,i)|^2, \tag{3}$$

where $|\mathcal{M}(k,i)| \in \{1, ..., P\}$. The exact values of $B(|\mathcal{M}(k,i)|)$ are computed using an off-line training procedure, where we used more than 12 minutes of speech sentences that were degraded by white Gaussian noise with a known variance $\sigma_N^2(k,i)$. Let $\tilde{B}(k,i)$ be defined as

$$\tilde{B}(k,i) \quad = \quad \frac{\frac{1}{|\mathcal{M}(k,i)|}\sum_{q \in \mathcal{M}(k,i)} |y_{HR}(q,i)|^2}{\sigma_N^2(k,i)}, \tag{4}$$

and let $\mathcal{T}(|\mathcal{M}|)$ be the set of time-frequency points in the training data for which the number of noise-only bins in a frequency band is estimated to be $|\mathcal{M}|$. $B(|\mathcal{M}(k,i)|)$, is then computed by averaging $\tilde{B}(k,i)$ over the set $\mathcal{T}(|\mathcal{M}|)$ leading to

$$B(|\mathcal{M}(k,i)|) = \frac{1}{|\mathcal{T}(|\mathcal{M}|)|} \sum_{(k,i) \in \mathcal{T}(|\mathcal{M}|)} \tilde{B}(k,i), \tag{5}$$

where $|\mathcal{T}(|\mathcal{M}|)|$ is the cardinality of the set $\mathcal{T}(|\mathcal{M}|)$.

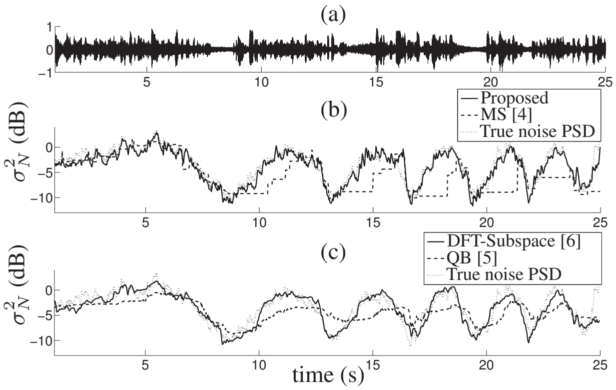

Fig. 1. (a) Speech degraded by modulated white noise at 5 dB SNR. (b)-(c) Comparison between true noise PSD and proposed and reference noise PSD estimators for DFT bin centered around 0.9 kHz.

4. EXPERIMENTAL RESULTS AND COMPLEXITY

To evaluate the performance of the proposed approach for noise PSD estimation we compare its performance with three reference methods, namely, MS [4], QB noise PSD estimation with quantile parameter $p = 0.5$ and a buffer length of 20 frames [5], and noise PSD estimation based on the DFT-subspace approach [6]. The speech data base that we used consists of more than 7 minutes of Danish speech spoken by 9 female speakers and 8 male speakers, and does not contain long portions of silence. The speech signals were degraded by several noise sources at input SNRs of 0, 5, 10, and 15 dB. As noise sources we use white noise, passing train noise, passing car noise, and white noise that is modulated by the following function,

$$f(m) = 1 + 0.5\sin(2\pi\, m f_{mod}/f_s), \qquad (6)$$

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. All signals are used at a sampling frequency of 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 from all performance measurements. The proposed method has signal-frames and super-frames of length $L_1 = 256$ and $L_2 = 640$ samples, respectively. For a fair comparison with the DFT-subspace approach [6], the length L_2 is chosen such that it equals the amount of data used in [6]. The signal-frames have an overlap of 50 % and are windowed using a square-root-Hann window. The super-frames are windowed using a Hann window. The order of the DFT and the HR-DFT are $K = 256$ and $Q = 1024$, respectively. Obviously, the estimated values of $B(|\mathcal{M}(k,i)|)$ depend on frame-length L_2 and HR-DFT order Q. In our implementation, the estimated values of $B(|\mathcal{M}(k,i)|)$ are between 1 and 3.7. For the hypothesis test in Eq. (2) we use $P_{fa} = 10^{-3}$.

4.1. Performance Evaluation

Fig. 1 shows an example of noise PSD estimation at a frequency bin centered around 900 Hz. The speech signal under consideration originates from a female speaker and is degraded by the aforementioned modulated white noise at an overall SNR of 5 dB. Together with the estimated noise PSDs we also show the ideal noise PSD $\sigma_N^2(k,i)$ obtained by smoothing noise periodograms across time using an exponential window, i.e.

$$\sigma_N^2(k,i) = 0.9\sigma_N^2(k,i-1) + 0.1|n(k,i)|^2. \qquad (7)$$

For visibility the results are distributed over two subplots. Subplot (b) shows the noise PSD estimated by the proposed method, MS and the true noise PSD and subplot (c) shows the noise PSD estimated by the DFT-subspace approach, QB noise PSD estimation and the true noise PSD.

From Fig. 1 we see that for a low modulation frequency all four noise PSD tracking methods are close to the true noise PSD. However, as the modulation frequency increases over time we see that MS is not able to track the increasing noise PSD. The QB noise PSD estimator is slightly better in following increasing noise levels, however, it has more problems in tracking the noise PSD for decreasing noise levels. The DFT-subspace and the proposed noise PSD tracking method keep track of the changing noise PSD and obtain estimates that are fairly close to the true noise PSD.

For a further performance evaluation we measure both noise tracking performance as well as the speech enhancement performance. Noise tracking performance is measured using the symmetric log-error distortion measure [6]

$$\text{LogErr} = \frac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I}\left|10\log_{10}\left[\frac{\sigma_N^2(k,i)}{\hat{\sigma}_N^2(k,i)}\right]\right| \quad [dB], \qquad (8)$$

where I denotes the total number of signal-frames and $\sigma_N^2(k,i)$ denotes the ideal noise PSD.

To measure speech enhancement performance we applied the noise PSD estimators within a single-microphone DFT-based noise reduction system. In this system the speech PSD is estimated using the decision-directed approach [1]. As speech estimator we use a magnitude DFT MMSE estimator derived under the assumption that speech DFT coefficients have a generalized-Gamma distribution with parameters $\gamma = 1$ and $\nu = 0.6$ [3]. The speech enhancement performance of this system is evaluated using PESQ [9] and segmental SNR defined as [10]

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

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

The results in terms of of the LogErr distortion measure, PESQ and SNR_{seg} are given in Table 1. In terms of noise tracking performance, i.e., LogErr, the performance of the proposed approach is very close to the DFT-subspace approach and better than MS and the QB approach. Although this holds for all noise sources used in the experiments, it is especially noticeable for more non-stationary noise sources. In terms of speech enhancement performance, i.e., SNR_{seg} and PESQ, the results are in line with the noise tracking performance in terms of LogErr. Depending on noise type and SNR, improvements of 1 to 2 dB in terms of SNR_{seg} and 0.3 to 0.4 in terms of PESQ can be observed over MS and QB noise PSD estimation.

4.2. Computational Complexity

The computational complexity of the proposed method is mainly determined by the HR-DFT of order Q, which has a complexity in the order of $Q\log_2 Q \approx 1.0\cdot 10^4$ operations per time-frame [11]. The DFT-subspace approach requires the singular values of a matrix with dimensions $L\times M$ at each frequency bin. With the settings in [6], i.e., $L = M = 7$, the computational complexity to obtain the singular values is in the order of $2.67L^3$ operations [12], leading to a computational complexity that is in the order of $(K/2+1)2.67L^3 \approx 1.2\cdot 10^5$

Table 1. Performance in terms of LogErr (dB), $\mathrm{SNR}_{\mathrm{seg}}$ (dB) and PESQ.

noise source	input SNR (dB)	LogErr (dB) MS [4]	LogErr DFT-Sub. [6]	LogErr prop. meth.	LogErr QB [5]	$\mathrm{SNR}_{\mathrm{seg}}$ (dB) MS [4]	$\mathrm{SNR}_{\mathrm{seg}}$ DFT-Sub. [6]	$\mathrm{SNR}_{\mathrm{seg}}$ prop. meth.	$\mathrm{SNR}_{\mathrm{seg}}$ QB [5]	PESQ MS [4]	PESQ DFT-Sub. [6]	PESQ prop. meth.	PESQ QB [5]
white noise	0	1.1	0.8	0.8	1.5	2.2	3.0	2.6	1.6	1.86	1.96	1.91	1.82
	5	1.2	0.9	0.8	1.5	5.2	5.6	5.3	3.9	2.26	2.33	2.29	2.19
	10	1.3	1.0	0.9	1.6	8.0	8.3	8.1	5.9	2.57	2.61	2.60	2.51
	15	1.4	1.2	1.1	2.0	10.8	11.1	11.0	7.8	2.86	2.86	2.86	2.77
passing train	0	3.7	2.0	2.0	2.9	0.8	1.5	1.4	0.8	1.87	1.96	1.97	1.89
	5	3.6	2.3	2.2	3.2	3.8	4.3	4.4	3.3	2.26	2.34	2.36	2.28
	10	3.5	2.8	2.5	3.8	7.2	7.4	7.6	5.8	2.62	2.65	2.69	2.61
	15	3.7	3.5	3.2	5.0	10.6	10.8	10.9	8.0	2.93	2.91	2.96	2.88
passing cars	0	3.9	2.2	2.1	3.5	5.6	6.3	6.9	4.4	2.09	2.39	2.40	2.09
	5	3.9	2.5	2.5	4.1	8.8	9.4	9.9	6.5	2.40	2.67	2.70	2.41
	10	4.1	3.1	3.1	5.3	12.0	12.5	12.9	8.4	2.72	2.92	2.95	2.68
	15	4.6	3.9	3.9	7.1	15.0	15.6	15.9	9.9	3.00	3.14	3.15	2.91
modulated white noise	0	2.7	1.0	0.9	2.4	1.3	3.1	2.9	1.2	1.59	1.97	1.92	1.64
	5	2.8	1.0	1.0	2.5	4.2	5.8	5.6	3.6	1.98	2.33	2.29	2.03
	10	2.8	1.2	1.1	2.7	7.2	8.6	8.4	5.7	2.34	2.60	2.60	2.37
	15	2.8	1.4	1.4	3.0	10.3	11.4	11.3	7.7	2.68	2.86	2.86	2.67

Table 2. Normalized processing-time.

method	DFT-sub. [6]	Prop.	MS [4]	QB [5]
Proc. time	13.5	1.0	2.4	0.3

operations per time-frame. Hence, the proposed approach reduces complexity with approximately a factor 11.5.

In Table 2 the computational complexity is reflected in terms of processing-time of Matlab implementations of the noise PSD tracking methods, normalized by the processing-time of the proposed approach. The proposed and MS approach have a processing-time that is in the same order of magnitude, while the QB approach is a bit faster. In comparison to the DFT-subspace approach, the proposed approach has a processing-time which is a factor 13.5 smaller. This reduction in terms of processing-time is in the same order of magnitude as the aforementioned reduction in terms of required operations per time-frame. Notice, that the processing times as given in Table 2 should only be considered as a rough estimate since they will in general depend on implementation details.

5. CONCLUDING REMARKS

In this paper we presented a method for fast noise PSD estimation with low complexity. The method is based on computation of periodograms using a DFT with a higher order than the DFT usually used in the noise reduction algorithm itself. Experiments show that the presented method has similar noise tracking performance as the recently proposed DFT-subspace approach. However, with a computational complexity that is more than a factor 10 lower.

In comparison to other noise PSD estimators, like minimum statistics and quantile based noise PSD estimation, the proposed approach improves noise PSD tracking performance and speech enhancement performance while computational complexity is in the same order of magnitude.

6. 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] J. Benesty, S. Makino, and J. Chen, editors, *Speech Enhancement*, New York: Springer, 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.

[4] 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.

[5] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, June 2000, pp. 1875–1878.

[6] 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.

[7] H. L. van Trees, *Detection, Estimation and Modulation Theory*, vol. 1, John Wiley and Sons, 1968.

[8] S. K. Kay, *Fundamentals of Statistical signal processing*, vol. 2, Prentice Hall, Upper Saddle River, NJ, 1998.

[9] 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.

[10] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, Piscataway, NJ, 2000.

[11] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of Fourier series," *Math. Comput.*, vol. 19, pp. 297–301, 1965.

[12] E. Z. Anderson et al., *LAPACK users's guide*, SIAM, Philadelphia, 3rd edition edition, 1999.