

Research Article

Low Complexity DFT-Domain Noise PSD Tracking Using High-Resolution Periodograms

Richard C. Hendriks,¹ Richard Heusdens,¹ Jesper Jensen (EURASIP Member),² and Ulrik Kjems²

¹Department of Mediamatics, Delft University of Technology, Mekelweg 4 2628 CD Delft, The Netherlands

²Oticon A/S, 2765 Smørum, Denmark

Correspondence should be addressed to Richard C. Hendriks, r.c.hendriks@tudelft.nl

Received 18 February 2009; Revised 16 June 2009; Accepted 26 August 2009

Recommended by Soren Jensen

Although most noise reduction algorithms are critically dependent on the noise power spectral density (PSD), most procedures for noise PSD estimation fail to obtain good estimates in nonstationary noise conditions. Recently, a DFT-subspace-based method was proposed which improves noise PSD estimation under these conditions. However, this approach is based on eigenvalue decompositions per DFT bin, and might be too computationally demanding for low-complexity applications like hearing aids. In this paper we present a noise tracking method with low complexity, but approximately similar noise tracking performance as the DFT-subspace approach. The presented method uses a periodogram with resolution that is higher than the spectral resolution used in the noise reduction algorithm itself. This increased resolution enables estimation of the noise PSD even when speech energy is present at the time-frequency point under consideration. This holds in particular for voiced type of speech sounds which can be modelled using a small number of complex exponentials.

Copyright © 2009 Richard C. Hendriks et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

The growing interest in mobile digital speech processing devices for both human-to-human and human-to-machine communication has led to an increased use of these devices in noisy conditions. In such conditions, it is desirable to apply noise reduction as a preprocessing step in order to extend the SNR range in which the performance of these applications is satisfactory.

A group of methods that is often used for noise reduction in the single-microphone setup are the so-called discrete Fourier transform (DFT) domain-based approaches. These methods work on a frame-by-frame basis where the noisy signal is divided in windowed time-frames, such that both quasistationarity constraints imposed by the input signal and delay constraints imposed by the application at hand are satisfied. Subsequently, these windowed time-frames are transformed using a DFT. From the resulting noisy speech DFT coefficients the corresponding clean speech DFT coefficients are estimated, typically by using Bayesian estimators [1] followed by an inverse DFT to the time

domain and an overlap-add procedure to synthesize the enhanced signal.

Typically, clean speech DFT estimators depend on the speech and noise power spectral density (PSD), for example, [2–5]. Since these two quantities are defined in terms of the statistical expectation operator they are unknown in practice and have to be estimated from the noisy speech signal. The speech PSD is often estimated by exploiting the so-called decision-directed approach [2]. This method is sometimes favored over maximum likelihood estimation of the speech PSD [2], because it results in a lower amount and more natural sounding residual noise [6]. Accurate noise PSD estimation is also of vital importance in order to obtain an estimated clean speech signal with good quality. Errors in the noise PSD estimate influence directly the amount of achieved noise suppression. Specifically, an overestimate of the noise PSD will typically lead to oversuppression of the noise and potentially to a loss of speech quality, while an underestimate of the noise PSD leaves an unnecessary amount of residual noise in the enhanced signal.

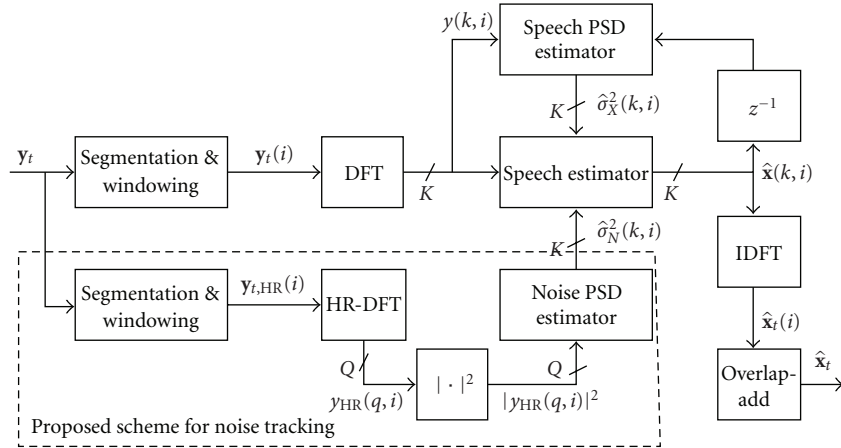


FIGURE 1: Overview of a DFT-domain-based noise reduction system with the proposed noise PSD tracking algorithm.

Under rather stationary noise conditions, the use of a voice activity detector [7, 8] (VAD) can be sufficient for estimation of the noise PSD. With a VAD the noise PSD is estimated during speech pauses. However, VAD based noise PSD estimation fails when the noise is non-stationary. An alternative is to estimate the noise PSD using algorithms based on minimum statistics [9, 10] (MS). These methods do not rely on the explicit use of a VAD, but make use of the fact that the power level of the noisy signal in a particular frequency bin seen across a sufficiently long time interval will reach the noise-power level. From the minimum value in such a time-interval the noise PSD is estimated by applying an appropriate bias compensation [11]. A crucial parameter in MS based noise PSD estimation is the length of the time-interval. If the interval is chosen too short, speech energy will leak into the noise PSD estimate, because the interval will not contain a noise-only region. However, increasing the duration of the interval will increase the tracking delay in regions where the noise PSD is increasing in level.

Another method that does not depend on a VAD is quantile-based (QB) noise PSD estimation [12]. This method relies on estimation of the noise PSD by computing per DFT bin a temporal quantile p of noisy periodograms in a certain time-interval. For the special case of a $p = 0.5$ quantile, the noise PSD is estimated by the median of the data in the time-interval. The speed at which this method can estimate the noise PSD for nonstationary 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 only on the 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. In addition, it is also more likely that QB noise PSD estimation is subject to leakage of speech into the noise PSD estimate because it exploits the quantile instead of the minimum within a time-interval.

Other recent advancements for noise PSD estimation comprise data-driven noise PSD estimation [13], improved

minima controlled recursive averaging [14], noise PSD estimation based on classified codebooks [15], and noise PSD estimation based on harmonic tunnelling [16]. The approach based on harmonic tunnelling makes explicit use of the harmonic structure in voiced speech sounds and estimates the noise PSD by exploiting the gaps between harmonics. Consequently, this method can continuously update the noise PSD under the condition that the DFT bin under consideration does not contain a speech harmonic.

Recently, in [17], a method for noise tracking was proposed which exploits the tonal structure in speech, but which can also estimate the noise PSD when speech is actually present in the DFT bin under consideration. This method, named DFT-subspace approach, 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 two submatrices of which the columns span two mutually orthogonal vector spaces, namely, a noisy signal subspace and a 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 method proposed in [17] has been shown to be effective for noise PSD estimation and can be implemented in MATLAB in real-time on a modern PC, the necessary eigenvalue decompositions might be too complex for applications with very low-complexity constraints like portable communication devices such as mobile phones and hearing aids.

A possible way to reduce the computational complexity of the algorithm in [17] is to use subspace tracking algorithms that are able to track subspaces efficiently over time, for example, [18, 19]. Although this might reduce the computational complexity of the DFT-subspace algorithm, it might also change its performance in an unpredictable way.

In this paper, we propose an alternative noise PSD tracking algorithm with approximately similar performance as the method presented in [17], but with considerably reduced computational complexity. The proposed method is outlined in Figure 1. The method makes use of the fact that often speech sounds can be modelled using a small

number of complex exponentials [20]. Notice that this holds in particular for voiced type of speech sounds, especially at lower frequencies. The noise PSD tracking method is based on noisy periodograms computed using a DFT with a frequency resolution that is typically higher than that of the DFT used in the noise reduction algorithm itself. In the following, we will use the expression HR-DFT to refer to the high-resolution DFT that is used to estimate the noise PSD. To refer to the DFT that is used to compute the noisy DFT coefficients in the noise reduction algorithm we maintain the expression DFT. For example, in the simulation experiments reported in Section 4, we use a 256-points DFT and a 1024-points HR-DFT at a sampling rate of 8 kHz. Hence, due to the difference in resolution between the DFT and the HR-DFT, every DFT bin corresponds to a sub-band of several HR-DFT bins. The high-resolution periodogram is divided in sub-bands, corresponding to the frequency bins obtained by the DFT. Analogous to the method in [17] we divide the HR-DFT bins within each sub-band to contain noisy speech and noise only. The noise-only HR-DFT bins are used to compute a maximum likelihood estimate of the noise PSD level.

The remainder of this paper is organized as follows. In Section 2 the basic notation and assumptions are introduced that will be used throughout this paper. In Section 3 the proposed noise PSD estimation method based on high-resolution periodograms is presented. Furthermore, in Section 4 experimental results will be presented followed by a discussion on the proposed noise PSD estimator in Section 5. Finally, in Section 6 concluding remarks are given.

2. DFT-Based Speech Estimators

Let the bandlimited and sampled time-domain noisy speech signal be denoted by y_t , where the subscript t explicitly indicates that this is a time-domain signal. We assume that y_t consists of a clean speech signal x_t that is degraded by additive noise n_t , that is,

$$y_t = x_t + n_t. \quad (1)$$

The noisy signal y_t is divided in frames of length L_1 by applying a sliding window $w_1(m)$ with $m \in \{0, \dots, L_1 - 1\}$ with a window-shift M . Let k and i be the frequency-bin index and time-frame index, respectively, and let $K \geq L_1$ be the DFT order. The noisy DFT coefficients $y(k, i)$ are then given by the discrete Fourier transform of the windowed time-frames, that is,

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

where $j = \sqrt{-1}$ is the imaginary unit and where w_1 is the normalized analysis window such that $\sum_{m=0}^{L_1-1} w_1^2(m) = 1$. (This normalization is used to overcome energy differences between the DFT and HR-DFT coefficients when using different analysis windows in both transforms.) 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 . Due to linearity of the Fourier transform, it holds that

$$y(k, i) = x(k, i) + n(k, i). \quad (3)$$

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. Further, it is assumed that $X(k, i)$ and $N(k, i)$ are uncorrelated, that is,

$$E[X(k, i)N^*(k, i)] = 0 \quad \forall k, i. \quad (4)$$

In order to find an estimate of the clean speech DFT coefficient $x(k, i)$, say $\hat{x}(k, i)$, a gain function $G(k, i)$ is typically applied to the noisy DFT coefficients, that is,

$$\hat{x}(k, i) = G(k, i)y(k, i). \quad (5)$$

There exist various ways to determine this gain function, for example, based on Bayesian principles [2–5] or based on more heuristically motivated arguments, for example, spectral subtraction [21]. However, irrespective of how the gain function is derived, it holds that all gain functions are dependent on the noise PSD $\sigma_N^2(k, i) = E[|N(k, i)|^2]$. As discussed above, this quantity is generally not known with certainty, but must be estimated from the available data.

3. Noise PSD Estimation Based on High-Resolution Periodograms

In the proposed noise PSD tracking method we distinguish between two different type of time-frames. The time-frames that are used for the actual processing of the noisy signal in the noise reduction system have a length of L_1 samples and are defined in Section 2. We refer to these time-frames as signal-frames. The second type will be called super-frames and have a length of L_2 samples where generally $L_2 > L_1$. The super-frames are used to estimate the noise PSD using high-resolution DFTs (HR-DFTs). Let D be the allowed algorithmic delay in samples in addition to the delay of the signal-frame. A super-frame with index i then comprises the time samples $y_t(iM+m)$ with $m \in \{L_1-L_2+D, \dots, L_1-1+D\}$. For simplicity we assume that size and position of the super-frames with respect to the signal-frames is fixed. However, notice that size and position of the super-frames could be made adaptive with respect to the underlying noisy signal, for example, using a segmentation algorithm for noisy speech as presented in [22].

Let $Q \geq L_2$ be the order of the HR-DFT and let w_2 be a normalized window function such that $\sum_{m=0}^{L_2-1} w_2^2(m) = 1$. The HR-DFT coefficient of a super-frame at frequency bin q and time-frame i is given by

$$y_{\text{HR}}(q, i) = \sum_{m=L_1-L_2+D}^{L_1-1+D} y_t(iM + m)w_2(m) \exp\left[-\frac{2\pi qmj}{Q}\right], \quad (6)$$

where the subscript HR indicates that this is a coefficient of the HR-DFT of a super-frame. The HR-DFT coefficients

$y_{\text{HR}}(q, i)$ are used to form a high-resolution noisy periodogram $|y_{\text{HR}}(q, i)|^2$. Each DFT frequency bin k corresponds to a band of, say W , HR-DFT frequency bins in the high-resolution periodogram. More specifically, let HR-DFT-order Q and DFT-order K be related as $Q = PK$ and let the k th band of the high-resolution periodogram consist of the frequency bins $q \in \{q_1, \dots, q_2\}$, with $W = q_2 - q_1 + 1$. The bin-numbers q_1 and q_2 for which the difference between their center-frequencies equals the width of a DFT frequency bin k can then be shown as

$$\begin{aligned} q_1 &= kP - \left\lfloor \frac{1}{2}P \right\rfloor, \\ q_2 &= kP + \left\lfloor \frac{1}{2}P \right\rfloor, \end{aligned} \quad (7)$$

where $\lfloor x \rfloor$ is defined as the nearest integer $\leq x$. Because of the higher-frequency resolution in the HR-DFT, it will be possible to estimate the noise PSD at a frequency band k even when speech is actually present in this frequency band. This is possible under the condition that the clean speech signal as observed in frequency bin k can be approximated well using less than the W HR-DFT basis functions that are necessary to represent the sub-band under consideration. Notice that this holds in particular for voiced type of speech sounds.

To compute an estimate $\hat{\sigma}_N^2(k, i)$ based on the k th frequency band of $|y_{\text{HR}}(q, i)|^2$, we assume that the noise level is constant across this frequency band. This assumption can be made arbitrarily accurate by narrowing the width of the DFT frequency bins. (Notice that even when this assumption is not valid, e.g., when the noise level is not constant in a frequency-band but has a certain slope, the estimated noise PSD can still be correct as the average noise level in the k th HR-DFT frequency band might still be equal to the noise PSD level in the k th DFT bin.) Further we assume that the noise HR-DFT coefficients N_{HR} have a complex Gaussian distribution, which is validated by the fact that the time-span of dependency [23] is relative short for many noise sources [4]. Let $\mathcal{M}(k, i)$ be the set of HR-DFT frequency bins corresponding to the k th DFT frequency bin that do not contain speech energy. The maximum likelihood estimate of the noise PSD in DFT frequency bin k is then given by

$$\hat{\sigma}_N^2(k, i) = \frac{1}{|\mathcal{M}(k, i)|} \sum_{q \in \mathcal{M}(k, i)} |y_{\text{HR}}(q, i)|^2, \quad (8)$$

where $|\mathcal{M}(k, i)|$ denotes the cardinality of the set $\mathcal{M}(k, i)$. When $|\mathcal{M}(k, i)| = 0$, all HR-DFT coefficients contain speech energy, and $\hat{\sigma}_N^2(k, i)$ is not updated. To reduce the variance of the estimated values, $\hat{\sigma}_N^2(k, i)$ can be smoothed across time, for example, using exponential smoothing in combination with adaptive smoothing factors as in [10]. This will be done in the simulation experiments in Section 4.

3.1. Determining $\mathcal{M}(k, i)$. In order to evaluate (8), it is necessary to know the set $\mathcal{M}(k, i)$. To determine $\mathcal{M}(k, i)$ we make use of a procedure that is quite similar to the one that was proposed in [17] and which was used to determine the dimension of a noise-only subspace. The procedure is based

on two assumptions. As already mentioned in Section 3, the noise HR-DFT coefficients $N_{\text{HR}}(q, i)$ are assumed to be complex Gaussian distributed. Based on this assumption, it can easily be shown that the squared-magnitude of the noise HR-DFT coefficients, that is, $|N_{\text{HR}}(q, i)|^2$, is exponentially distributed. Secondly, we assume that the noise PSD develops relatively slowly across time. This assumption does not limit the practical performance, since, as it turns out, a noise PSD that changes with 10 dB per second can still be tracked. This allows us to use the noise PSD estimated in the previous frame, that is, $\hat{\sigma}_N^2(k, i - 1)$, as a priori information when estimating the noise PSD in the current frame.

With these assumptions, we are now in position to determine which of the frequency bins $q \in \{q_1, \dots, q_2\}$ in the k th HR-DFT frequency band do not contain speech energy. To do so, we apply a Neyman-Pearson hypothesis test [24] with the following H_0 and H_1 hypotheses:

$$\begin{aligned} H_0 &: |y_{\text{HR}}(q, i)|^2 \text{ consists of only noise,} \\ H_1 &: |y_{\text{HR}}(q, i)|^2 \text{ consists of noise and speech.} \end{aligned} \quad (9)$$

It can be shown that under rather general conditions, an optimal decision test compares the value $|y_{\text{HR}}(q, i)|$ to a threshold $\lambda_{\text{th}}(k, i)$ [24], that is,

$$|y_{\text{HR}}(q, i)|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \lambda_{\text{th}}(k, i). \quad (10)$$

Using the aforementioned distributional assumption on $|N_{\text{HR}}(q, i)|^2$, we can express the threshold λ_{th} as a function of the false-alarm probability P_{fa} by [24]

$$\lambda_{\text{th}}(k, i) = -\sigma_N^2(k, i) \ln P_{\text{fa}}, \quad (11)$$

where the unknown noise PSD $\sigma_N^2(k, i)$ is approximated in practice by the estimated noise PSD value $\hat{\sigma}_N^2(k, i - 1)$.

3.2. Bias Compensation. Generally, the estimate $\hat{\sigma}_N^2(k, i)$ is biased high due to spectral leakage from neighboring DFT coefficients that contain speech energy. To overcome this bias we introduce a bias compensation-factor B , much along the same lines as in [10], that is dependent on the cardinality of the set $\mathcal{M}(k, i)$, that is, $B(|\mathcal{M}(k, i)|)$. Altogether, the noise PSD is estimated by

$$\hat{\sigma}_N^2(k, i) = \frac{1}{B(|\mathcal{M}(k, i)|)|\mathcal{M}(k, i)|} \sum_{q \in \mathcal{M}(k, i)} |y_{\text{HR}}(q, i)|^2, \quad (12)$$

where $|\mathcal{M}(k, i)| \in \{1, \dots, P\}$. The exact values of $B(|\mathcal{M}(k, i)|)$ are computed using an offline 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

$$\tilde{B}(k, i) = \frac{(1/|\mathcal{M}(k, i)|) \sum_{q \in \mathcal{M}(k, i)} |y_{\text{HR}}(q, i)|^2}{\sigma_N^2(k, i)}, \quad (13)$$

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 $|\mathcal{M}|$. The bias compensation-factor $B(|\mathcal{M}(k, i)|)$, is then computed by averaging $\tilde{B}(k, i)$ over the set $\mathcal{T}(|\mathcal{M}|)$ leading to

$$B(|\mathcal{M}|) = \frac{1}{|\mathcal{T}(|\mathcal{M}|)|} \sum_{(k,i) \in \mathcal{T}(|\mathcal{M}|)} \tilde{B}(k, i). \quad (14)$$

Although this training procedure makes use of white noise in order to compute $B(|\mathcal{M}|)$, this does not limit the applicability of the proposed noise PSD estimator as it can be used to track both white and non-white noise sources as long as the noise-level in a band can be assumed approximately constant. The training procedure is applied using only one SNR, that is, at a global SNR of 10 dB. Clearly, the bias compensation could be extended by making $B(|\mathcal{M}|)$ also a function of SNR. However, in the results presented in Section 4 we keep $B(|\mathcal{M}|)$ independent of SNR in order to keep complexity and storage requirements low.

3.3. Algorithm Overview. In this section, we give a summary of the necessary processing steps in the proposed algorithm. It is assumed that all processing steps are repeated for each time-frame index i . However, when less processing power is available the update rate could be reduced.

- (1) Compute HR-DFT of a windowed noisy super-frame using (6).
- (2) Determine the set $|\mathcal{M}(k, i)|$ for each band k using (9).
- (3) Compute $\hat{\sigma}_N^2(k, i)$ for each band k using (12).
- (4) Apply smoothing across time of the estimate noise PSD in order to reduce its variance.

Whenever $|\mathcal{M}(k, i)| = 0$, all frequency bins in the band contain speech energy in which case it is not possible to update the noise PSD in that band during time-frame i . In these situations, the estimate from the time-frame $i - 1$ is used. To overcome a complete locking of the noise PSD estimator under extreme situations when $|\mathcal{M}(k, i)| = 0$ for a very long time we adopt the safety-net proposed in [13] and compute the minimum $P_{\min}(k, i)$ of $|y(k, i)|^2$ across a long time-interval, for example, a time-interval of one second. Using $P_{\min}(k, i)$, the noise PSD is updated by

$$\hat{\sigma}'_N^2(k, i) = \max[\hat{\sigma}_N^2(k, i), P_{\min}(k, i)]. \quad (15)$$

4. Experimental Results

For performance evaluation of the proposed method for noise PSD estimation we compare its performance with three reference methods, namely, noise PSD estimation based on MS as proposed in [10], QB noise PSD estimation as proposed in [12] with quantile parameter $p = 0.5$ and a buffer length of 20 frames, and noise PSD estimation based on the DFT-subspace approach as proposed in [17]. The speech database that we used consists of more than 7 minutes of Danish speech that was read from newspapers by 17 different speakers, 9 female speakers and 8 male speakers, and does not contain long portions of silence.

These speech signals were not used for computation of the bias compensation in Section 3.2. The speech signals were degraded by a variety of noise sources at input SNRs of 0, 5, 10, and 15 dB. Both the speech and the noise signals were used at a sampling frequency of 8 kHz. All signals start with a noise-only period of 0.5 seconds. All algorithms use the first 0.1 seconds for initialization; these noise-only samples are excluded from all performance measurements. The length of the signal-frames is set to $L_1 = 256$, that is, 32 milliseconds. The length L_2 of the super-frames for the proposed method is a tradeoff between complexity constraints and stationarity requirements on the noisy speech signal on one hand, and the potential to exploit the increased frequency resolution for noise PSD estimation on the other hand. In Section 4.1.2 experiments will be performed that also reflect this tradeoff. Based on these experiments it follows that the best choice in terms of noise tracking performance for the length of the super-frames is around 70–100 milliseconds. In order to make a fair comparison possible with the DFT-subspace approach [17], we therefore chose the length L_2 such that it equals the amount of data used in [17] and use $L_2 = 640$ samples, that is, 80 milliseconds.

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, and are chosen as an integer power of 2 to facilitate an efficient implementation of the DFT using FFTs. The false-alarm probability in (11) was set to $P_{fa} = 0.001$. The estimated values of $B(|\mathcal{M}|)$ are between 1 and 3.7. Obviously, the estimated bias compensation factors $B(|\mathcal{M}|)$ depend on the chosen parameter settings, for example, super-frame length L_2 and the HR-DFT order Q . In the experimental results presented in this section we focus on real-time applications that require low algorithmic delay. Therefore, we set the allowed algorithmic delay to $D = 0$ for all methods. Further, we apply the same safety-net procedure as in (15) to the DFT-subspace approach [17] to avoid locking of the estimator.

4.1. Noise PSD Estimation Performance. Because optimal estimators used for noise reduction are always functions of the true noise variance $\sigma_N^2(k, i)$, we can evaluate the performance of noise PSD tracking algorithms by measuring directly the error between $\sigma_N^2(k, i)$ and its estimate $\hat{\sigma}_N^2(k, i)$. For this purpose we use the symmetric log-error distortion measure defined in [17] as

$$\text{LogErr} = \frac{1}{IK} \sum_{k=1}^K \sum_{i=1}^I \left| 10 \log \left[\frac{\sigma_N^2(k, i)}{\hat{\sigma}_N^2(k, i)} \right] \right| \quad (\text{dB}), \quad (16)$$

where I denotes the total number of signal-frames and $\sigma_N^2(k, i)$ denotes the ideal noise PSD that is obtained by smoothing measured noise periodograms across time using an exponential window, that is,

$$\sigma_N^2(k, i) = \alpha \sigma_N^2(k, i-1) + (1-\alpha) |n(k, i)|^2, \quad (17)$$

with a smoothing factor $\alpha = 0.9$ [10].

4.1.1. Synthetic Performance Example. To demonstrate the potential of the proposed approach, we consider a synthetic example of noise PSD estimation where the presence of speech is modelled by a sinusoid at a frequency of 937.5 Hz, that is, centered in the 31st frequency bin. This clean synthetic signal is shown in Figure 2(a). During the time instance of approximately 2 till 5 seconds, the sinusoid is continuously present in periods of 450 milliseconds, each time followed by a 150 ms period where the sinusoid is absent in order to model speech absence. Subsequently, this synthetic clean signal is degraded by white Gaussian noise. The SNR in the frequency bin under consideration is approximately 36 dB during presence of the sinusoidal component in the first 3.5 seconds. In the time span from 3.5 till 4.5 seconds the SNR decreases from 36 dB to 30 dB. For visibility the results are distributed over two subplots. Figure 2(b) shows the noise PSD estimated by the proposed method and MS, compared to the true noise PSD. Figure 2(c) shows the noise PSD estimated by the DFT-subspace approach and QB noise PSD estimation, compared to the true noise PSD.

From the comparison in Figures 2(b) and 2(c) it is clear that both the MS and the QB approach heavily overestimate the noise PSD. This is caused by the presence of the sinusoidal component, which leads to tracking of the PSD of the noisy sinusoid instead of the noise PSD. The proposed approach and the DFT-subspace approach show accurate tracking of the changing noise level. That the proposed approach is able to track the changing noise level is due to the higher frequency resolution that is exploited. This also becomes clear from Figure 2(d) where the number of HR-DFT bins is shown for the DFT bin under consideration that are classified as noise-only, that is, $|\mathcal{M}(k, i)|$. As expected, when there is no speech presence $|\mathcal{M}(k, i)|$ equals the total number of HR-DFT bins that fall within one DFT bin, that is, under the given parameter settings $|\mathcal{M}(k, i)| = 5$. When the sinusoidal component is present, $|\mathcal{M}(k, i)|$ decreases to one or two, which means that the estimated noise PSD can still be updated even though the sinusoidal component is present.

4.1.2. Super-Frame Size L_2 . In this section, we investigate the relation between the length of the super-frames L_2 and noise tracking performance. To do so, we degraded the speech signals in the database by two different noise sources, namely, white noise and non-stationary white noise. The non-stationary white noise consists of white noise that is modulated by the following function:

$$f(m) = 1 + 0.5 \sin\left(\frac{2\pi m f_{\text{mod}}}{f_s}\right), \quad (18)$$

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, that is, a maximum change of the noise PSD of approximately 10 dB per second. An example of such a modulated white noise sequence can be seen in Figure 6. Subsequently, the proposed noise tracking algorithm is applied with several super-frame sizes

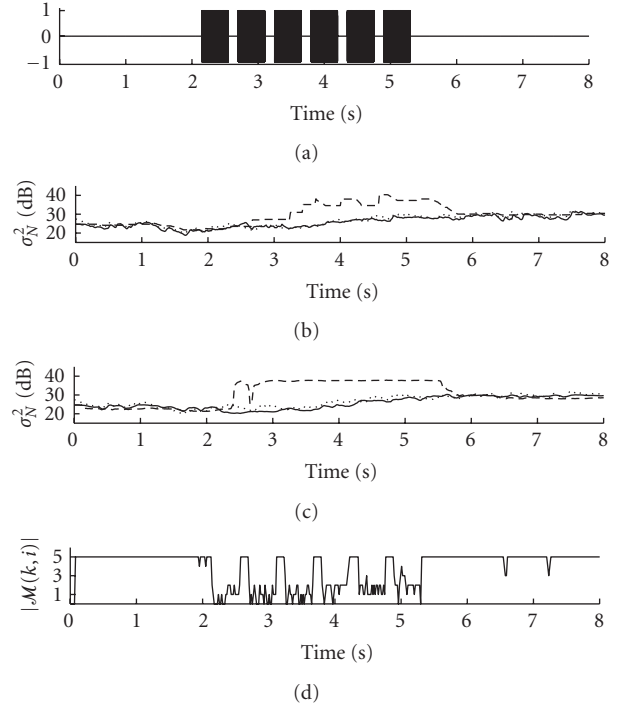


FIGURE 2: Synthetic noise tracking example. (a) Clean synthetic signal. (b) Comparison between true noise PSD (dotted line), proposed approach (solid line), and MS (dashed line) for DFT bin centered around 937.5 Hz. (c) Comparison between true noise PSD (dotted line), DFT-subspace approach (solid line), and QB approach (dashed line) for DFT bin centered around 937.5 Hz. (d) Cardinality of the set $\mathcal{M}(k, i)$ for the frequency bin centered around 937.5 Hz.

L_2 . The outcome of this experiment is shown in Figure 3. As expected, the optimal length L_2 is dependent on noise type and noise level as the optimal L_2 -value is a tradeoff between stationarity requirements on the noisy speech signal on one hand and the potential to exploit the increased frequency resolution for noise PSD estimation on the other hand. This tradeoff results in the bowl-shaped performance curves in Figure 3. With increasing super-frame size the LogErr distortion decreases due to increased frequency resolution. However, the noisy data within the super-frame is likely to become non-stationary for a super-frame size that becomes too large. In that case, more of the W HR-DFT basis functions are necessary to model the clean speech signal as observed in the sub-band under consideration and cannot be used to estimate the noise PSD. Therefore, eventually, the LogErr distortion will increase again. In general, the optimal super-frame size is around 70–100 milliseconds. For the experiments in the remaining sections of this paper, we will use a super-frame size of 80 milliseconds, that is, $L_2 = 640$, such that it equals the amount of data used by the DFT-subspace approach in [17].

Using a super-frame size that is too short will lead to a worse frequency resolution of the HR-DFT coefficients. To demonstrate the effect of having a poor frequency resolution, we consider in Figure 4 a similar synthetic example as in

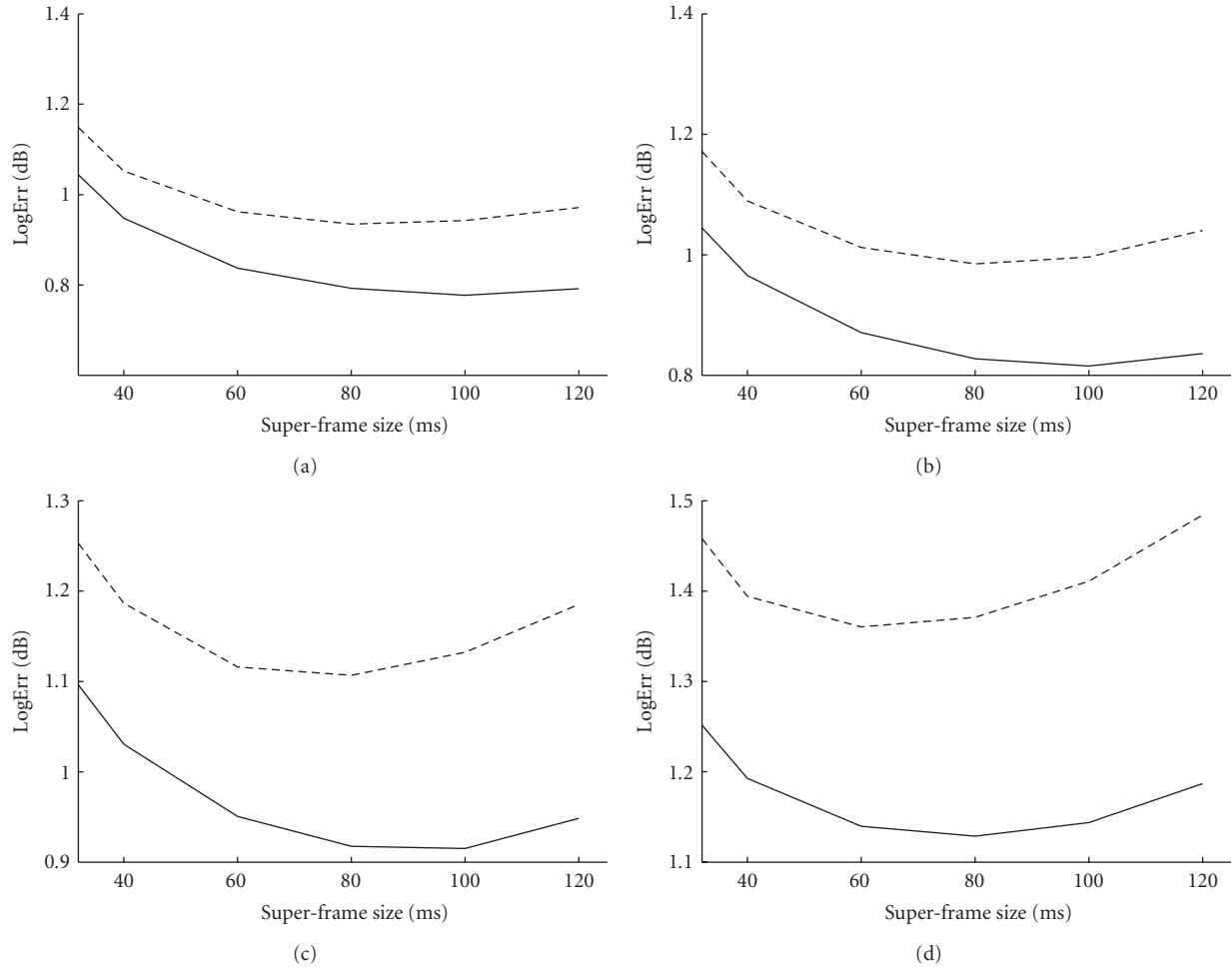


FIGURE 3: Noise tracking performance in terms of LogErr (dB) as a function of the length of the super-frames for stationary Gaussian white noise (solid line) and nonstationary Gaussian white noise (dashed line) at an input SNR of (a) 0 dB (b) 5 dB (c) 10 dB (d) 15 dB.

Figure 2, but then with a super-frame size of only $L_2 = 320$ samples (40 milliseconds). Let us first consider the time span from 0 up till 3.5 seconds. Similar as for the synthetic example in Figure 2, the number of noise-only HR-DFT bins $|\mathcal{M}(k, i)|$ equals the total number of HR-DFT bins that fall within one DFT bin when the sinusoidal component is absent. However, in contrast to the example in Figure 2, the cardinality of the set $\mathcal{M}(k, i)$ is zero when the sinusoidal component is present. This is due to the lower resolution that is obtained for the HR-DFT and means that the noise PSD cannot be updated when the sinusoidal component is present. When the noise level increases after 3.5 seconds, the noise tracking algorithm can hardly distinguish the noise-only HR-DFT bins from the speech-plus-noise HR-DFT bins due to the poor frequency resolution. In this particular situation, too many HR-DFT bins are classified as being noise-only resulting in an overestimated noise PSD. The behavior to wrongly classify HR-DFT bins as being noise-only is influenced by the false alarm probability P_{fa} in (11). By increasing the false alarm probability, the Neyman-Pearson hypothesis test in (9) will become more conservative with respect to updating the noise PSD. The hypothesis test will

classify more HR-DFT bins as consisting of speech-plus-noise and will not use these to update the noise PSD. Setting P_{fa} , for example, to $P_{fa} = 0.005$ instead of $P_{fa} = 0.001$, in combination with a super-frame size of only $L_2 = 320$ samples, we obtain the example in Figure 5.

The example in Figure 5 is comparable with the situation in Figure 4. However, due to the higher false alarm probability, the Neyman-Pearson hypothesis test classifies all HR-DFT coefficients as being speech-plus-noise when the sinusoidal component is present also after the time instance of 3.5 seconds. This results in an empty set $\mathcal{M}(k, i)$, and, consequently, the noise PSD is only updated when the sinusoidal component is clearly absent.

4.1.3. Natural Performance Examples. To further illustrate the performance of the proposed method in comparison to the three reference methods with natural speech we consider an example where a speech signal obtained from a female speaker is degraded by non-stationary white noise described by (18) at an SNR of 5 dB. In Figure 6 examples of noise PSD estimation at the frequency bin centered around 0.9 kHz (left

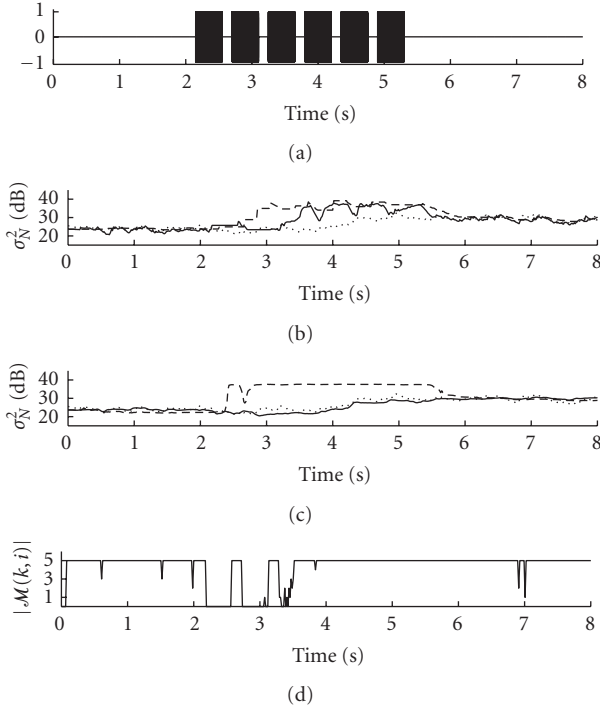


FIGURE 4: Synthetic noise tracking example with super-frame size of 40 milliseconds. (a) Clean synthetic signal. (b) Comparison between true noise PSD (dotted line), proposed approach (solid line), and MS (dashed line) for DFT bin centered around 937.5 Hz. (c) Comparison between true noise PSD (dotted line), DFT-subspace approach (solid line), and QB approach (dashed line) for DFT bin centered around 937.5 Hz. (d) Cardinality of the set $\mathcal{M}(k, i)$ for the frequency bin centered around 937.5 Hz.

column) and 2.0 kHz (right column) are shown. Together with the estimated noise PSDs we also show the ideal noise PSD $\sigma_N^2(k, i)$ that is obtained using (17). For visibility the results are shown per frequency bin and distributed over two subplots. Subplot (c) and (d) show the noise PSD estimated by the proposed method, MS and the true noise PSD at a DFT bin centered around 0.9 kHz and 2.0 kHz, respectively. Subplots (e) and (f) show the noise PSD estimated by the DFT-subspace approach, QB noise PSD estimation and the true noise PSD at a DFT bin centered around 0.9 kHz and 2.0 kHz, respectively.

From Figure 6, we see that for a low modulation frequency the noise tracking performance is approximately similar and close to the true noise PSD for all four noise PSD tracking methods. However, as the modulation frequency increases over time we see that MS is not able to track the changes when the noise PSD increases. The QB noise PSD estimator is slightly better in following the increasing noise levels, however, compared to MS, it has more problems in tracking the noise PSD for decreasing noise levels. The DFT-subspace and the proposed noise PSD tracking method on the other hand keep track of the changing noise PSD and obtain estimates that are fairly close to the true noise PSD.

In Figure 7 we show a second example at frequency bins centered around 0.9 kHz (left column) and 2.0 kHz

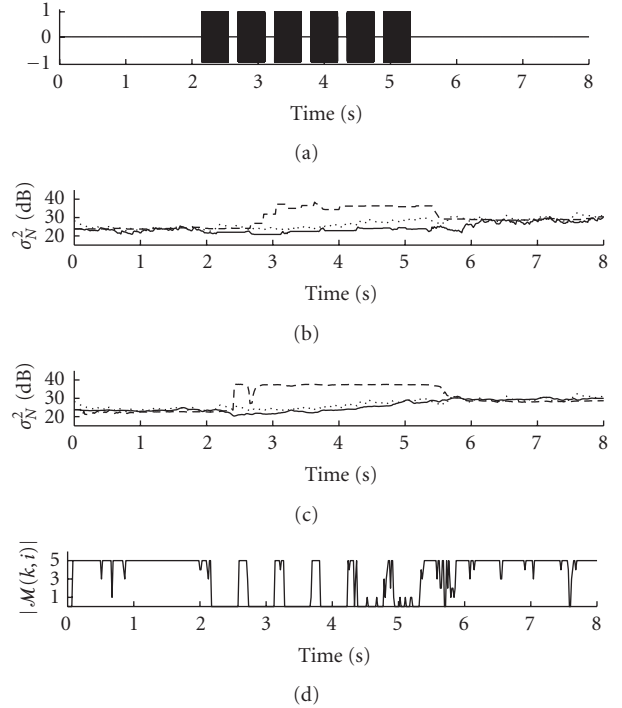


FIGURE 5: Synthetic noise tracking example with super-frame size of 40 ms and $P_{fa} = 0.005$. (a) Clean synthetic signal. (b) Comparison between true noise PSD (dotted line), proposed approach (solid line), and MS (dashed line) for DFT bin centered around 937.5 Hz. (c) Comparison between true noise PSD (dotted line), DFT-subspace approach (solid line), and QB approach (dashed line) for DFT bin centered around 937.5 Hz. (d) Cardinality of the set $\mathcal{M}(k, i)$ for the frequency bin centered around 937.5 Hz.

(right column). In this example the same speech signal is degraded with noise originating from passing cars at an overall SNR of 10 dB. We see that all four methods have similar performance when the noise is stationary, that is, in the time-interval from 10 till 15 seconds. When the noise level changes rather fast both the proposed and DFT-subspace-based noise PSD tracker show almost immediate tracking of the changing noise PSD, while both the QB approach and MS are unable to track these fast increasing noise levels. Similar to the previous example, QB noise PSD estimation has the tendency to estimate increasing noise levels with slightly less delay than MS. However, decreasing noise levels are generally overestimated. As overestimates generally lead to oversuppression and a potential loss in speech quality this is an undesired effect.

4.1.4. Evaluation of Noise Tracking Performance. For a more comprehensive study of noise tracking performance, we degraded the speech signals in our database by a wide variety of noise sources. Some of these noise sources are rather stationary, some rather nonstationary, and some are a mixture between stationary and non-stationary elements. The individual noise sources can be described as follows: as completely stationary noise sources we use computer generated pink noise and white noise. Party noise consists

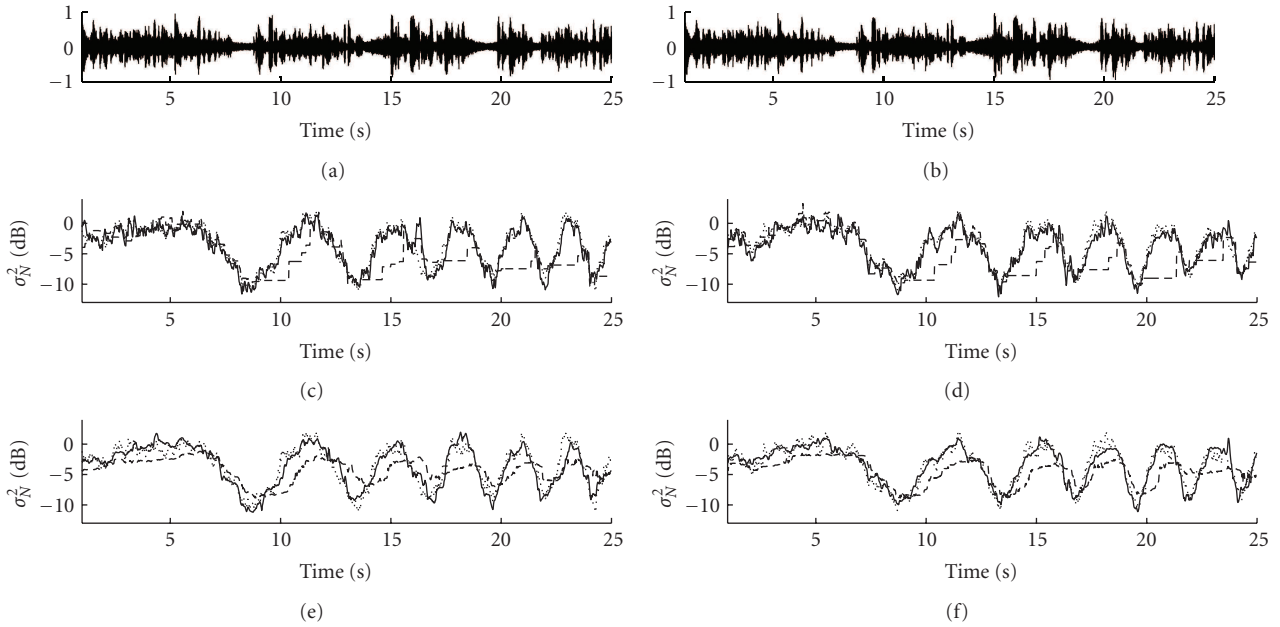


FIGURE 6: Comparison between estimated noise PSD and the true noise PSD. (a)-(b) Speech signal degraded by modulated white noise at an overall SNR of 5 dB. (c)-(d) Comparison between true noise PSD (dotted line), proposed approach (solid line), and MS (dashed line) for DFT bin centered around (c) 0.9 kHz and (d) 2.0 kHz. (e)-(f) Comparison between true noise PSD (dotted line), DFT-subspace approach (solid line), and QB approach (dashed line) for DFT bin centered around (e) 0.9 kHz and (f) 2.0 kHz.

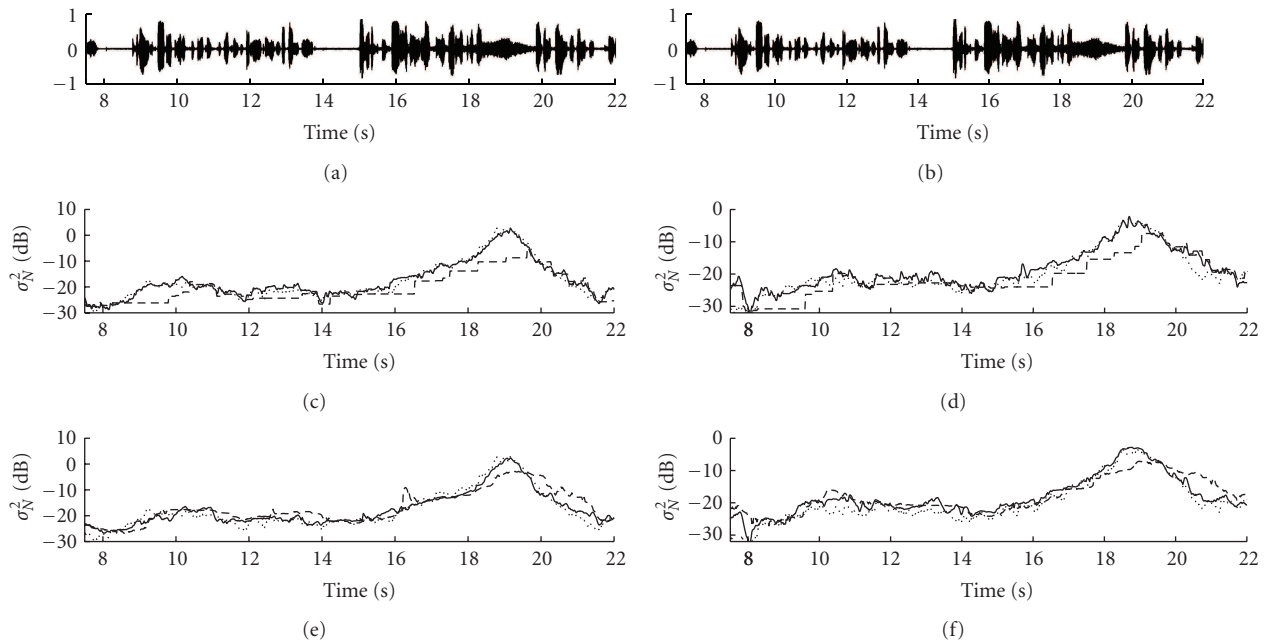


FIGURE 7: Comparison between estimated noise PSD and the true noise PSD. (a)-(b) Speech signal degraded by noise originating from passing cars at an overall SNR of 10 dB. (c)-(d) Comparison between true noise PSD (dotted line), proposed approach (solid line), and MS (dashed line) for DFT bin centered around (c) 0.9 kHz and (d) 2.0 kHz. (e)-(f) Comparison between true noise PSD (dotted line), DFT-subspace approach (solid line), and QB approach (dashed line) for DFT bin centered around (e) 0.9 kHz and (f) 2.0 kHz.

of many background speakers. Although this noise source consists of a large amount of speakers being nonstationary noise-sources individually, the sum of all these noise-sources can be perceived as being rather stationary. Noise originating from a circle saw and waves at the beach are both locally

non-stationary, but also contain long stretches of rather stationary noise. Noise originating from a passing train and passing cars both consist of gradually changing noise sources and some shorter stretches of rather stationary background noise. Modulated white and modulated pink noise are

TABLE 1: Required processing-time normalized by the processing-time of the proposed approach.

Method	DFT-sub. [17]	Prop.	MS [10]	QB [12]
Proc. time	13.5	1.0	2.4	0.3

computer generated noise sources that are modulated using the function in (18).

The performance of MS, the QB approach, the DFT-subspace approach, and the proposed approach is shown in Table 2 in terms of the LogErr distortion measure. From the results in Table 2 we see that in general the performance of the proposed approach is better than MS and the QB approach, and close to the DFT-subspace approach. Especially for gradually changing noise sources, such as passing cars and modulated noise, the proposed approach improves over MS, and the QB approach.

An exception on this are the results for pink noise. For pink noise the noise level across a sub-band is not completely constant. This means that the assumption on which (8) is based is not completely valid. A similar argument holds for the DFT-subspace approach, where it is assumed that the eigenvalues in the noise-only DFT-subspace have a flat spectrum. The assumptions that underly MS are completely valid and therefore MS has a slightly better performance for this noise source.

4.2. Influence of Noise PSD Estimator on Noise Reduction Performance. Although it is reasonable to evaluate the performance of a noise PSD tracking method directly on the estimated noise PSD as in the previous paragraph, it is also of interest to investigate the impact in a noise reduction framework. We, therefore, combined the proposed and the three reference noise PSD estimators within a single-microphone DFT-based noise reduction system, as indicated in Figure 1. In this noise reduction system, we estimate the speech PSD using the decision-directed approach [2]. For the speech estimator we use a magnitude MMSE estimator derived under the generalized-Gamma distribution with distribution parameters $\gamma = 1$ and $\nu = 0.6$ [5]. For performance evaluation we measure PESQ [25] available from [26] and segmental SNR defined as [27]

$$\text{SNR}_{\text{seg}} = \frac{1}{I} \sum_{i=0}^{I-1} \mathcal{F} \left\{ 10 \log_{10} \frac{\|x_t(i)\|^2}{\|x_t(i) - \hat{x}_t(i)\|^2} \right\}, \quad (19)$$

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, I is the number of frames, and $\mathcal{F}(x) = \min\{\max(x, -10), 35\}$ constrains the estimated SNR per frame to the range between -10 dB and 35 dB [27]. The results in terms of SNR_{seg} and PESQ are given in Tables 3 and 4, respectively. These results are in line with the performance directly measured on the estimated noise PSDs, except for the QB approach. The QB approach generally has worse performance in terms of both PESQ and segmental SNR in comparison to the proposed and other reference methods. This can be explained by the

fact that it quite regularly leads to overestimates of the noise PSD.

The general tendency is that the proposed noise PSD estimator improves on MS for the more nonstationary noise sources and shows performance close to the DFT-subspace based. For rather stationary noise sources, MS, the DFT-subspace approach, and the proposed approach lead to quite similar performance. Notice that the performance measured in such a noise reduction system is only partly determined by the noise PSD estimator. Other aspects that determine the performance are estimation of the speech PSD and the speech estimator. Although all speech estimators are dependent on the true noise PSD, different estimators might react differently on over- or underestimates of the noise PSD.

5. Discussion

5.1. Signal Model and Complexity. From Sections 4.1 and 4.2, we see that the performance of the proposed method is quite similar to the recently presented DFT-subspace based method [17]. The latter approach is based on a Karhunen-Loève transform (KLT) of a sequence of complex DFT coefficients observed in the same frequency bin across time. This implies the use of a KLT for each DFT bin, while the proposed method is based on one single HR-DFT per super-frame; the DFT-subspace approach and the proposed method are based on different signal models. Specifically, the proposed method assumes that the speech signal can be represented by a sum of undamped complex exponentials of which the frequencies are constrained to be at the center of a HR-DFT bin. The DFT-subspace approach applies a KLT, that is, a signal-adaptive transform, to a sequence of DFT coefficients. This does not require that the sequence of DFT coefficient consist of undamped complex exponentials, but allows the use of damped complex exponentials with unrestricted frequencies as well. In theory, the DFT-subspace approach should therefore have better access to the underlying noise level. However, this is at the cost of a much higher complexity, which cannot always be justified for applications where only few computational resources are available.

We compare the computational complexity of the proposed method and the DFT-subspace approach in terms of necessary operations per time-frame and in terms of processing-time. The computational complexity of the proposed method is mainly determined by the HR-DFT of order Q that needs to be computed. Based on the Cooley-Tukey algorithm [28] this leads to a complexity that is in the order of $Q \log_2 Q \approx 1.0 \cdot 10^4$ operations per time-frame. The DFT-subspace approach requires the singular values of a matrix with dimensions $L \times M$ at each frequency bin, where we used the same settings as in [17], that is, $L = M = 7$. The computational complexity for obtaining singular values only is in the order of $2.67L^3$ operations [29, 30]. This means that per time-frame the computational complexity of the DFT-subspace approach is in the order of $(K/2 + 1)2.67L^3 \approx 1.2 \cdot 10^5$ operations. Hence, for the specific parameter settings as used in the experimental results presented in this section, the proposed approach has a complexity reduction in the

TABLE 2: Performance in terms of LogErr (dB).

noise source	input SNR (dB)	MS [10]	DFT-Sub. [17]	prop. method	QB [12]
pink noise	0	1.0	1.1	1.3	1.3
	5	1.1	1.2	1.3	1.2
	10	1.2	1.3	1.4	1.3
	15	1.3	1.6	1.7	1.7
white noise	0	1.1	0.8	0.8	1.5
	5	1.2	0.9	0.8	1.5
	10	1.3	1.0	0.9	1.6
	15	1.4	1.2	1.1	2.0
party noise	0	2.2	1.7	1.7	2.0
	5	2.2	1.8	1.8	1.9
	10	2.3	2.1	2.0	2.3
	15	2.4	2.6	2.4	3.3
waves at the beach	0	2.0	1.3	1.4	2.0
	5	2.0	1.5	1.5	2.1
	10	2.1	1.7	1.6	2.3
	15	2.2	2.2	2.0	3.0
circle saw	0	3.0	2.3	2.3	3.1
	5	3.2	2.4	2.3	3.4
	10	3.4	2.6	2.5	3.9
	15	3.7	3.0	2.9	4.6
passing train	0	3.7	2.0	2.0	2.9
	5	3.6	2.3	2.2	3.2
	10	3.5	2.8	2.5	3.8
	15	3.7	3.5	3.2	5.0
passing cars	0	3.9	2.2	2.1	3.5
	5	3.9	2.5	2.5	4.1
	10	4.1	3.1	3.1	5.3
	15	4.6	3.9	3.9	7.1
modulated white noise	0	2.7	1.0	0.9	2.4
	5	2.8	1.0	1.0	2.5
	10	2.8	1.2	1.1	2.7
	15	2.8	1.4	1.4	3.0
modulated pink noise	0	2.7	1.2	1.4	2.3
	5	2.7	1.3	1.5	2.3
	10	2.7	1.5	1.6	2.4
	15	2.7	1.8	1.9	2.8

order of 11.5 in comparison to the DFT-subspace approach. Notice that there do exist other subspace tracking algorithms than the ones in [29, 30] that can reduce the complexity in a predictable way, for example, [18, 19, 31], but might change the performance of the DFT-subspace approach in a rather unpredictable way.

In Table 1 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. Next to the DFT-subspace approach and the proposed approach, we also show the processing-time for the MS and QB approach. The proposed and MS approach have a processing-time that is in the same

order of magnitude, while the quantile based 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 1 should only be considered as a rough estimate since they will in general depend on implementation details.

5.2. Unvoiced Speech Sounds. The assumption under which the proposed method is able to estimate the noise level in

TABLE 3: Performance in terms of SNR_{seg} (dB).

Noise source	Input SNR (dB)	MS [10]	DFT-Sub. [17]	Prop. method	QB [12]
Pink noise	0	0.8	1.4	1.1	0.7
	5	3.8	4.2	4.0	3.2
	10	7.0	7.2	7.0	5.6
	15	10.1	10.2	10.1	7.7
White noise	0	2.2	3.0	2.6	1.6
	5	5.2	5.6	5.3	3.9
	10	8.0	8.3	8.1	5.9
	15	10.8	11.1	11.0	7.8
Party noise	0	-0.4	0.3	0.0	-0.5
	5	2.5	3.1	3.0	2.2
	10	5.7	6.2	6.1	4.8
	15	9.1	9.4	9.4	7.2
waves at the beach	0	0.5	1.2	1.0	0.3
	5	3.4	4.0	3.9	2.9
	10	6.6	7.0	7.0	5.4
	15	9.9	10.1	10.2	7.7
Circle saw	0	0.9	1.0	1.5	0.8
	5	3.7	4.0	4.5	3.1
	10	6.8	7.1	7.5	5.3
	15	9.9	10.3	10.6	7.3
Passing train	0	0.8	1.6	1.4	0.8
	5	3.8	4.3	4.4	3.3
	10	7.2	7.4	7.6	5.8
	15	10.6	10.8	10.9	8.0
Passing cars	0	5.6	6.3	6.9	4.4
	5	8.8	9.4	9.9	6.5
	10	12.0	12.5	12.9	8.4
	15	15.0	15.6	15.9	9.9
Modulated white noise	0	1.3	3.1	2.9	1.2
	5	4.2	5.8	5.6	3.6
	10	7.2	8.6	8.4	5.7
	15	10.3	11.4	11.3	7.7
Modulated pink noise	0	0.4	1.7	1.4	0.5
	5	3.4	4.5	4.3	3.1
	10	6.6	7.4	7.3	5.5
	15	9.9	10.5	10.5	7.6

the k th frequency band is that the speech signal as observed in this band can be represented by less than the W complex exponential basis functions that are necessary to completely represent the noisy sub-band signal under consideration. It is well known that this is possible for voiced speech sounds which can be modelled using a small number of complex exponentials [20]. For unvoiced speech sounds however, this assumption will generally not be valid. Therefore, it is interesting to investigate the behavior of the proposed method during these speech sounds. To illustrate this situation we focus on a speech sentence saying “*since this story hap*”, which contains some clearly pronounced /s/ sounds. To give a clear example we use in this particular situation a speech signal at a sampling frequency of 20 kHz, since these

unvoiced sounds are especially dominantly present at higher frequencies. Ideally, to prevent leakage of speech energy in the noise PSD estimate, the noise PSD should not be updated in this situation. The clean speech time-domain signal is shown in Figure 8(a); three noise bursts representing the /s/ sounds are clearly visible. This signal is degraded by street noise at an SNR of 10 dB and processed using the proposed noise PSD estimator. The PSD of both the clean speech signal and the noise at the time interval 11.85 till 11.88 seconds are shown in Figure 8(c), where it is clearly visible that the speech signal is dominant at higher frequencies. In Figure 8(b) we show in the time-frequency plane for each frequency band the estimated number of noise-only bins in a band. We can see that during the unvoiced speech sounds the cardinality

TABLE 4: Performance in terms of PESQ.

Noise source	Input SNR (dB)	MS [10]	DFT-Sub. [17]	Prop. method	QB [12]
Pink noise	0	1.91	1.98	1.95	1.91
	5	2.33	2.38	2.36	2.31
	10	2.67	2.67	2.67	2.64
	15	2.96	2.92	2.94	2.89
White noise	0	1.86	1.96	1.91	1.82
	5	2.26	2.33	2.29	2.19
	10	2.57	2.61	2.60	2.51
	15	2.86	2.86	2.86	2.77
Party noise	0	1.62	1.63	1.66	1.63
	5	2.02	2.05	2.07	2.02
	10	2.40	2.43	2.44	2.39
	15	2.74	2.75	2.78	2.72
waves at the beach	0	1.75	1.86	1.85	1.74
	5	2.14	2.26	2.26	2.13
	10	2.49	2.55	2.57	2.47
	15	2.81	2.82	2.85	2.77
Circle saw	0	1.52	1.51	1.60	1.54
	5	1.88	1.90	1.97	1.89
	10	2.23	2.25	2.32	2.24
	15	2.57	2.59	2.65	2.56
Passing train	0	1.87	1.96	1.97	1.89
	5	2.26	2.34	2.36	2.28
	10	2.62	2.65	2.69	2.61
	15	2.93	2.91	2.96	2.88
passing cars	0	2.09	2.39	2.40	2.09
	5	2.40	2.67	2.70	2.41
	10	2.72	2.92	2.95	2.68
	15	3.00	3.14	3.15	2.91
modulated white noise	0	1.59	1.97	1.92	1.64
	5	1.98	2.33	2.29	2.02
	10	2.34	2.60	2.60	2.37
	15	2.68	2.86	2.86	2.67
modulated pink noise	0	1.75	2.00	1.97	1.79
	5	2.14	2.38	2.37	2.18
	10	2.50	2.66	2.67	2.52
	15	2.83	2.91	2.93	2.81

of the set $\mathcal{M}(k, i)$, that is, the number of noise-only bins in a band, is determined to be $|\mathcal{M}(k, i)| = 0$. Consequently, the noise PSD is not updated at these time-frequency points.

5.3. Noise PSD Estimation in High SNR Situation. Although accurate noise PSD estimation is important for applying noise reduction on noisy speech signals, it is also relevant to investigate the situation when very little noise is present. Clearly, the higher the SNR, the lower the noise-to-signal ratio (NSR) and consequently a worse noise PSD estimate is to be expected. Obviously, for very high SNRs the noise PSD will be overestimated due to leakage of speech energy into the noise PSD estimate. However, the question is whether

the level of the estimated noise PSD is low enough to not influence the amount of suppression applied to the speech signal afterwards by the noise suppression system. To investigate this situation, an experiment is performed with a speech signal degraded by white noise at an SNR of 60 dB. Subsequently, the proposed noise PSD estimator and the reference noise PSD estimators are applied to this signal. The a priori SNR, defined as $\xi(k, i) = \sigma_X^2(k, i)/\sigma_N^2(k, i)$, is estimated using the decision-directed approach [2] after which it is used to compute the value of the gain function used in Section 4. Figure 9(a) shows the original clean speech signal. Figure 9(b) shows the estimated *a priori* SNRs in a frequency bin centered around 1.25 kHz. This is compared with the a priori SNR computed using knowledge of the

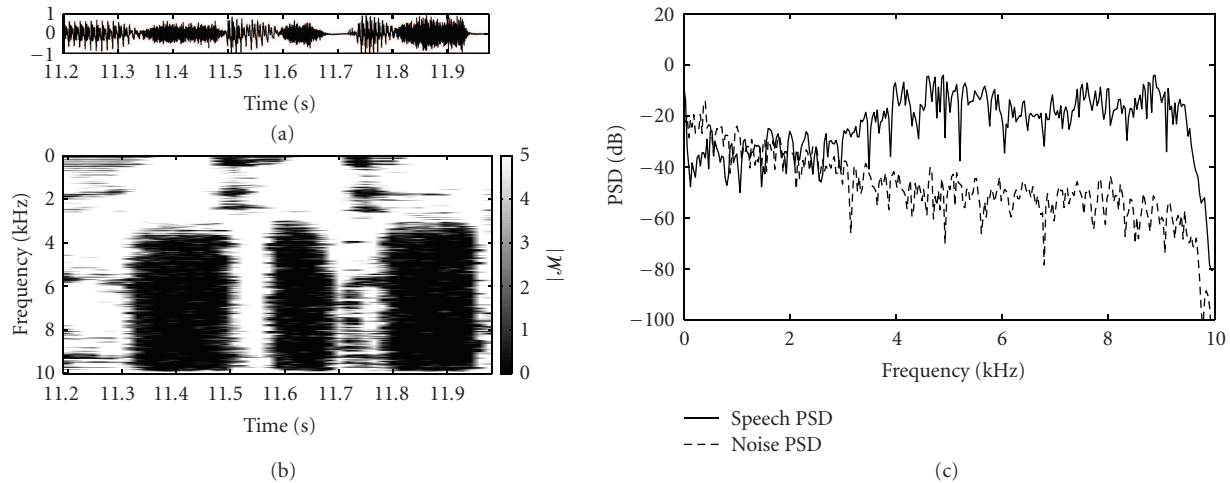


FIGURE 8: (a) Clean speech signal with the text ..since this story hap... (b) The number of noise-only bins per band. (c) PSD of clean speech and noise signal during the time interval 11.85–11.88 seconds.

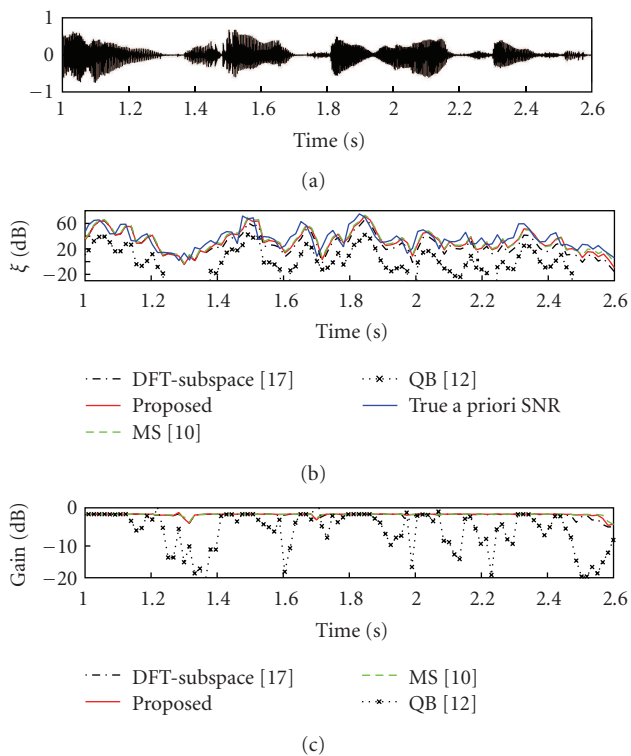


FIGURE 9: (a) Clean speech signal. (b) Comparison at frequency bin centered around at 1.25 kHz between the estimated a priori SNR and the true a priori for speech degraded by white noise at an SNR of 60 dB. (c) The amount of suppression that is applied using estimated a priori SNR for speech degraded by white noise at an SNR of 60 dB.

clean speech signal and the noise signal, which we refer to as the true a priori SNR. Clearly, all noise PSD estimators lead to a somewhat underestimated a priori SNR due to an overestimation of the noise PSD for this very high SNR.

Further, we see that all algorithms have a one frame delay with respect to the true a priori SNR, which is due to estimation of the a priori SNR using the decision-directed approach [6]. To verify whether the estimated a priori SNR is high enough not to apply any unwanted suppression we computed the value of the gain function used in Section 4.2. The resulting amount of suppression is shown in Figure 9(c). We see that for all noise PSD estimators, except for QB, the amount of suppression is generally 0 dB. The QB noise PSD estimator applies too much suppression due to leakage of speech into the noise PSD estimate.

6. Concluding Remarks

In general, noise reduction methods for speech enhancement rely on knowledge of the noise PSD. Because this quantity is defined in terms of expected values and is generally unknown, estimation from the noisy signal is necessary. In this paper, we presented a method which aims at accurate noise PSD estimation under both stationary and non-stationary noise conditions with low complexity. The proposed method makes use of the fact that speech sounds can often be modelled using a small number of complex exponentials. This is exploited by computing periodograms using an DFT with a higher order than the DFT as used in the noise reduction algorithm itself. Experiments demonstrated that the presented method leads to approximately similar noise tracking performance as the recently proposed DFT-subspace approach. However, this is at the cost of 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.

Acknowledgments

The research is supported by the Oticon foundation and the Dutch Technology Foundation STW. The authors would like to thank the anonymous reviewers whose constructive remarks helped to improve the presentation of this work.

References

- [1] H. L. van Trees, *Detection, Estimation and Modulation Theory*, vol. 1, John Wiley & Sons, New York, NY, USA, 1968.
- [2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [3] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, 2005.
- [4] 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.
- [5] 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 Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.
- [6] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.
- [7] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.
- [8] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006.
- [9] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of the European Signal Processing Conference (EUSIPCO '94)*, pp. 1182–1185, 1994.
- [10] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.
- [11] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215–1229, 2006.
- [12] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00)*, vol. 3, pp. 1875–1878, Istanbul, Turkey, June 2000.
- [13] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1112–1123, 2008.
- [14] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.
- [15] S. Srinivasan, *Knowledge-based speech enhancement*, Ph.D. thesis, Royal Institute of Technology (KTH), 2005.
- [16] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunneling: tracking non-stationary noises during speech," in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech '01)*, pp. 437–440, Aalborg, Denmark, September 2001.
- [17] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 3, pp. 541–553, 2008.
- [18] B. Yang, "Projection approximation subspace tracking," *IEEE Transactions on Signal Processing*, vol. 43, no. 1, pp. 95–107, 1995.
- [19] Y. Miao and Y. Hua, "Fast subspace tracking and neural network learning by a novel information criterion," *IEEE Transactions on Signal Processing*, vol. 46, no. 7, pp. 1967–1979, 1998.
- [20] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.
- [21] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.
- [22] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation for improved speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2064–2074, 2006.
- [23] D. R. Brillinger, *Time Series: Data Analysis and Theory*, SIAM, Philadelphia, Pa, USA, 2001.
- [24] S. K. Kay, *Fundamentals of Statistical Signal Processing*, vol. 2, Prentice-Hall, Upper Saddle River, NJ, USA, 1998.
- [25] ITU, "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. ITU-T P.862, 2000.
- [26] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, Fla, USA, 2007.
- [27] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, Piscataway, NJ, USA, 2000.
- [28] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of Fourier series," *Mathematics of Computation*, vol. 19, pp. 297–301, 1965.
- [29] G. H. Golub and C. F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Md, USA, 3rd edition, 1996.
- [30] E. Z. Anderson, *LAPACK Users' Guide*, SIAM, Philadelphia, Pa, USA, 3rd edition, 1999.
- [31] R. Badeau, B. David, and G. Richard, "Fast approximated power iteration subspace tracking," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2931–2941, 2005.