Noise Tracking Using DFT Domain Subspace Decompositions

Richard C. Hendriks, Jesper Jensen, and Richard Heusdens

***Abstract*—All discrete Fourier transform (DFT) domain-based speech enhancement gain functions rely on knowledge of the noise power spectral density (PSD). Since the noise PSD is unknown in advance, estimation from the noisy speech signal is necessary. An overestimation of the noise PSD will lead to a loss in speech quality, while an underestimation will lead to an unnecessary high level of residual noise. We present a novel approach for noise tracking, which updates the noise PSD for each DFT coefficient in the presence of both speech and noise. This method is based on the eigenvalue decomposition of correlation matrices that are constructed from time series of noisy DFT coefficients. The presented method is very well capable of tracking gradually changing noise types. In comparison to state-of-the-art noise tracking algorithms the proposed method reduces the estimation error between the estimated and the true noise PSD. In combination with an enhancement system the proposed method improves the segmental SNR with several decibels for gradually changing noise types. Listening experiments show that the proposed system is preferred over the state-of-the-art noise tracking algorithm.**

***Index Terms*—Discrete Fourier transform (DFT) domain subspace decompositions, noise tracking, speech enhancement.**

I. Introduction

As a consequence of the increased use of mobile voice processors in public areas (e.g., hearing aids and cellular phones), there has been an increasing interest for these systems to work well under noisy conditions. To achieve this, single-channel speech enhancement methods can be used to reduce the noise level. Among them is the group of discrete Fourier transform (DFT)-based methods that have received significant interest recently because of their relatively low complexity and good performance. These methods estimate the clean DFT coefficients by applying either a gain function to the noisy DFT coefficients or to the magnitude of the noisy DFT coefficients. Gain functions have been derived under minimum mean square error (MMSE) and maximum *a posteriori* (MAP) criteria, where speech DFT coefficients are assumed to have a super-Gaussian density [1]–[3]. Recently, estimators based on Garch models [4] have also been proposed.

All these gain functions rely on knowledge of the noise power spectral density (PSD), which has to be estimated from the noisy speech signal. An overestimation of the noise PSD will lead to over-suppression and, as a consequence, to a potential loss of speech quality, while an underestimation will lead to an unnecessary high level of residual noise. An accurate tracking of the noise PSD is therefore essential to obtain proper quality of the enhanced speech signal. Furthermore, fast tracking is important for nonstationary noise. However, both fast and accurate noise tracking is very challenging, especially under these nonstationary noise conditions.

A conventional method for estimating the spectral noise variance is to exploit speech pauses. Here, a voice activity detector (VAD) [5], [6] is used and only in case of speech absence the noise PSD is estimated and updated. Although this is effective when the noise is stationary, it often fails when the noise statistics change during speech presence. Moreover, accurate voice activity detection under very low signal-to-noise-ratio (SNR) conditions is not trivial.

Minimum statistics (MS)-based noise trackers [7], [8] offer a more advanced alternative to VAD based methods. This method exploits the property that the minimum power level in a particular frequency bin seen across a sufficiently long time interval is due to the noise process. From this minimum, the average noise power can be estimated by applying a bias compensation. The size of the time interval should be such that there is at least one noise-only observation within the window. The minimum size of the time window is therefore dependent on the duration of speech presence in a frequency bin. If the time window is chosen too short and speech energy is constantly present in the search window, MS will track the PSD of the noisy speech instead of the noise PSD. This will lead to an overestimate of the noise level. If, on the other hand, the time window is chosen too long, changes in the noise power level are not tracked or can only be tracked with a large delay.

In this paper, we present a novel approach for noise tracking, which updates the noise PSD for each DFT coefficient even when both speech and noise are present. This method is based on the eigenvalue decomposition of correlation matrices that are constructed from time series of noisy DFT coefficients. We exploit the fact that these correlation matrices can be decomposed using an eigenvalue decomposition into two submatrices of which the columns span two mutually orthogonal vector spaces, namely a signal (+ noise) subspace and a noise-only subspace. We use the property that speech signals seen in a particular frequency bin can often be described by a low-rank model, i.e., can be expressed as a linear combination of a small number of complex exponentials [9]. In that case, the eigenvalues that describe the energy in the noise-only subspace allow for an update of the noise statistics, even when speech is constantly present. Noise types that are

Manuscript received February 28, 2007; revised December 3, 2007. This work was supported by Philips Research and the Technology Foundation STW. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yariv Ephraim.

R. C. Hendriks and R. Heusdens are with the Department of Mediamatics, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: r.c.hendriks@tudelft.nl; r.heusdens@tudelft.nl).

J. Jensen is with Oticon A/S, 2765 Smørum, Denmark (e-mail: jsj@oticon.dk).

Digital Object Identifier 10.1109/TASL.2007.914977

described by a low-rank model itself, i.e., deterministic types of noise, will be represented in the signal subspace as well and need different measures to be estimated. How to track these deterministic types of noise will be discussed as well.

Notice that a considerable amount of research has been done on the application of subspace decompositions for speech enhancement, e.g., [10]–[12]. Also, it has been proposed to estimate the noise correlation matrix using time-domain subspace decompositions; see, e.g., [13], [14]. However, the method that we propose in this paper works in the DFT domain.

The remainder of this paper is organized as follows. In Section II, we illustrate the potential of the proposed method of noise tracking. In Section III, we explain the signal model and the concept of DFT domain subspace decompositions that we use to derive the noise tracking method. In Section IV, we consider estimation of the noise PSD based on estimated noisy correlation matrices. Furthermore, in Section V, we focus on some implementational aspects of the proposed noise tracking algorithm. In Section VI, we present experimental results, and finally in Section VII concluding remarks are given.

II. Illustration of DFT Domain Subspace-Based Noise Tracking

To illustrate the potential of the proposed method of noise tracking, we compare our new method to the MS method, which is known as the state-of-the-art for noise tracking in single-microphone speech enhancement applications. To do so, we create a synthetic signal in which the speech signal is modeled by a sinusoid of approximately 190 Hz. With this simplistic, but relevant model of a speech signal, we can simulate the situation where speech energy is constantly present and demonstrate that our proposed method has great potential for tracking of the noise PSD in the presence of speech. In this example, we use frame sizes of 256 samples with 50% overlap. In the first 2 s, (125 time frames) the signal consists of white noise only. Then, after 2 s, a sinusoidal component is turned on and remains constantly present at a certain frequency bin with a global SNR of 5 dB. This sinusoid simulates the continuous presence of speech energy. Finally, 0.5 s later, at frame number $i = 157$, the noise PSD decreases by 6 dB while the sinusoid remains present. We use both the MS approach and the proposed method to estimate the noise PSD. In Fig. 1, we compare their estimated noise PSDs together with the true noise PSD obtained by recursively smoothed periodogram estimates. The dotted line denotes the true noise PSD, the dashed-dotted line the noise PSD estimated using minimum statistics, and the dashed line the noise PSD estimated with the proposed approach, all in the same frequency bin. We see that in the first approximately 156 frames both methods lead to a fairly good estimate of the true noise PSD. After 156 frames, the proposed method follows the decrease in the noise PSD even though the sinusoid is present, while the MS method, on the other hand, is not able to follow this change. Moreover, approximately 100 frames after the sinusoid is turned on, the MS approach takes the energy of the noisy sinusoid as the new minimum and wrongly updates the estimated noise PSD. To what degree these type of overestimates occur in practice heavily depends on the size of the search window. By enlarging the search window, the effect of this problem can be weakened or overcome. However, enlarging the search window will result in a larger delay and is harmful for tracking changes in the noise power.

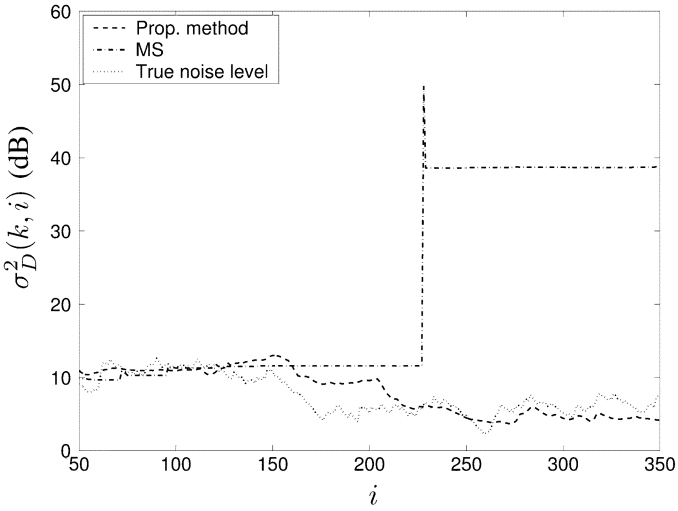


Fig. 1. Synthetic noise tracking example.

III. Signal Model and DFT Domain Subspace Decompositions

In this paper, we consider the discrete Fourier transform of speech signals as being the outcome of a random process. That is, $Y(k,i)$, $X(k,i)$, and $D(k,i)$ are complex random variables denoting the noisy speech, clean speech, and noise DFT coefficients of frame i and frequency bin k, with $k \in \{1,\ldots,K\}$, and K the total number of frequency bins. We assume the noise to be additive, i.e., $Y(k,i) = X(k,i)+D(k,i)$, zero mean and uncorrelated with the clean speech signal, i.e., $E[X(k,i)D(k,i)] = 0, \forall\ (k,i)$.

We collect DFT coefficients per frequency bin k that originate from the time frames $i-p_1$ up to frame $i+p_2$ and form a vector $\mathbf{Y}(k,i) \in \mathbb{C}^M$ with $M = p_1 + p_2 + 1$. That is

$$\mathbf{Y}(k,i) = [Y(k,i-p_1),\ldots,Y(k,i+p_2)]^T. \tag{1}$$

Let $\mathbf{R}_Y(k,i) \in \mathbb{C}^{M\times M}$ be the noisy speech correlation matrix related to frequency bin k and time frame i defined as

$$\mathbf{R}_Y(k,i) = E[\mathbf{Y}(k,i)\mathbf{Y}^H(k,i)] \tag{2}$$

where H indicates Hermitian transposition. The construction of $\mathbf{R}_Y(k,i)$ is illustrated in Fig. 2. Similarly, we can define the speech correlation matrix $\mathbf{R}_X(k,i) \in \mathbb{C}^{M\times M}$, that is

$$\mathbf{R}_X(k,i) = E[\mathbf{X}(k,i)\mathbf{X}^H(k,i)] \tag{3}$$

and the noise correlation matrix $\mathbf{R}_D(k,i) \in \mathbb{C}^{M\times M}$, that is

$$\mathbf{R}_D(k,i) = E[\mathbf{D}(k,i)\mathbf{D}^H(k,i)]. \tag{4}$$

Using the assumption that speech and noise are additive and uncorrelated we can write the noisy speech correlation matrix $\mathbf{R}_Y(k,i)$ as

$$\mathbf{R}_Y(k,i) = \mathbf{R}_X(k,i) + \mathbf{R}_D(k,i). \tag{5}$$

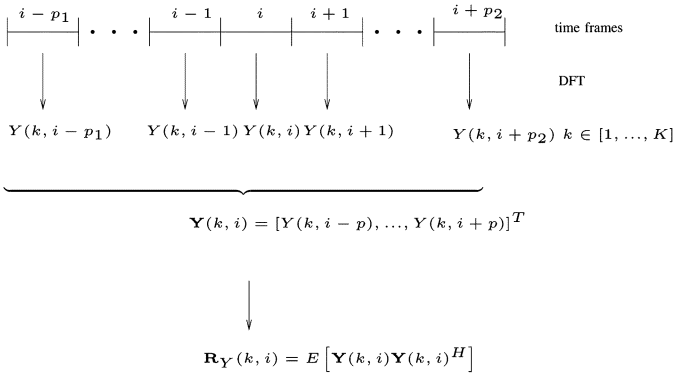


Fig. 2. Schematic overview of how correlation matrices in the DFT domain are computed.

Let us assume that $\mathbf{R}_D(k,i)=\sigma_D^2(k,i)\mathbf{I}_M$, that is, the noise DFT coefficients in $\mathbf{D}(k,i)$ are uncorrelated. This assumption is valid when frames do not overlap and the correlation time of the noise is small enough [15]. In case of overlapping frames, this assumption will be violated. This violation can be overcome by applying a prewhitening transform, as we describe in Section V.

As mentioned before, the clean speech correlation matrix $\mathbf{R}_X(k,i)$ is assumed to be of low-rank. In particular, this is true when speech sounds can be modeled by a sum of complex exponentials, e.g., voiced speech sounds [9]. Under this signal model and under assumption that the frame size is long enough, ideally each frequency bin will observe at most one complex exponential across time. The clean speech correlation matrix $\mathbf{R}_X(k,i)$ can therefore be assumed to be of low-rank. When the noise-only subspace is of full-rank and the speech signal can be described using such a low-rank signal subspace, the eigenvalues that describe the energy in the noise-only subspace allow for an update of the noise PSD, even when speech is constantly present. A validation of the low-rank assumption of $\mathbf{R}_X(k,i)$ is given in Section IV-C. Notice that for unvoiced speech sounds, in general, the speech signal is not of low rank, which means that for these type of speech sounds only few or no eigenvalues belong to the noise-only subspace. It is therefore less likely that the noise PSD can be updated during unvoiced speech sounds.

Let $\mathbf{R}_X(k,i)=\mathbf{U}\mathbf{\Lambda}_X\mathbf{U}^H$ denote the eigenvalue decomposition of the clean speech correlation matrix related to frequency bin k and time frame i. Here, $\mathbf{U}\in\mathbb{C}^{M\times M}$ is a unitary matrix and contains the eigenvectors as columns and $\mathbf{\Lambda}_X=\mathrm{diag}(\lambda_{X_1},\ldots,\lambda_{X_Q},0,\ldots,0)$, where $Q\leq M$ is the dimension of the signal subspace, a diagonal matrix with the nonnegative eigenvalues $\lambda_{X_1}\geq\lambda_{X_2}\geq\cdots\geq\lambda_{X_Q}\geq 0$ on the main diagonal. Using the assumption that $\mathbf{R}_D(k,i)$ is a scaled diagonal matrix and $X(k,i)$ and $D(k,i)$ are uncorrelated, we can write the eigenvalue decomposition of $\mathbf{R}_Y(k,i)$ as

$$\mathbf{R}_Y(k,i)=\mathbf{U}(\mathbf{\Lambda}_X(k,i)+\sigma_D^2(k,i)\mathbf{I}_M)\mathbf{U}^H \tag{6}$$

i.e., $\mathbf{R}_Y(k,i)$, $\mathbf{R}_X(k,i)$, and $\mathbf{R}_D(k,i)$ have the same eigenvectors, and the eigenvalues of $\mathbf{R}_Y(k,i)$ are simply obtained by adding the eigenvalues of $\mathbf{R}_X(k,i)$ and $\mathbf{R}_D(k,i)$.

The eigenvector matrix $\mathbf{U}$ can be partitioned as $\mathbf{U}=[\mathbf{U}_1;\mathbf{U}_2]$, where the columns of $\mathbf{U}_1\in\mathbb{C}^{M\times Q}$ form a basis for the signal subspace, and the columns of $\mathbf{U}_2\in\mathbb{C}^{M\times M-Q}$ form a basis for the noise-only subspace. Assuming that there indeed exists a low-dimensional signal subspace, i.e., $Q<M$, the eigenvalues in the noise-only subspace can be used to determine the noise PSD $\sigma_D^2(k,i)$, as the noise-only subspace eigenvalue matrix equals $\mathbf{I}_{(M-Q)}\sigma_D^2(k,i)$.

IV. Estimation of $\sigma_D^2(k,i)$

In the previous section, we considered the eigenvalue decomposition of $\mathbf{R}_Y(k,i)$ in order to estimate the noise PSD from the eigenvalues in the noise-only subspace. However, in practice, the correlation matrix $\mathbf{R}_Y(k,i)$ in (2) is unknown and estimated based on realizations. Therefore, we consider in this section estimation of $\sigma_D^2(k,i)$ based on an estimate of the correlation matrix $\mathbf{R}_Y(k,i)$.

The correlation matrix $\mathbf{R}_Y(k,i)$ can be estimated from a limited number of samples by

$$\hat{\mathbf{R}}_Y(k,i)=\frac{1}{L}\mathcal{Y}(k,i)\mathcal{Y}^H(k,i) \tag{7}$$

where $\mathcal{Y}\in\mathrm{C}^{M\times L}$ is a Hankel-structured data-matrix defined as

$$\mathcal{Y}(k,i)=\begin{pmatrix} y(k,i-n_1) & \cdots & y(k,i-n_1+L-1)\\ \vdots & & \vdots\\ y(k,i-n_1+M-1) & \cdots & y(k,i+n_2)\end{pmatrix} \tag{8}$$

where the small letters y indicate realizations of the random variable Y.

Let $\hat{\lambda}_{Y_l}(k,i)$ indicate an eigenvalue of the estimated correlation matrix $\hat{\mathbf{R}}_Y(k,i)$. Given the eigenvalue decomposition of $\hat{\mathbf{R}}_Y(k,i)$ and the dimension of the signal subspace Q, it is shown in Appendix C, that under the assumption that the vector $\mathbf{Y}(k,i)$ has a multivariate Gaussian density, a maximum-likelihood estimate of the noise PSD is given by

$$\hat{\sigma}_D^2(k,i)=\frac{1}{M-Q}\sum_{l=Q+1}^{M}\hat{\lambda}_{Y_l}(k,i). \tag{9}$$

That is, the noise PSD is estimated by taking the average of the eigenvalues in the noise-only subspace.

In order to compute (9), it is necessary to estimate the signal subspace dimension Q. Estimation of Q for noisy signals is a well-known problem for large data-records and can be performed using, e.g., Akaike information criterium (AIC) [16], [17], minimum description length (MDL) criterium [17], [18], or the Bayesian information criterium (BIC) [19]. However, when $\mathbf{R}_Y(k,i)$ is estimated based on a few data samples only, which is the case in our situation, existing model order estimators lead to inaccurate estimates of Q. Moreover, due to the inaccurate model order estimation and not always clear distinction between the noise-only and signal subspace, the noise power spectral estimate may be biased depending on whether the dimension of the signal subspace is overestimated or underestimated. To increase the accuracy of the estimated model order, we present an alternative approach for model order estimation in Section IV-A,

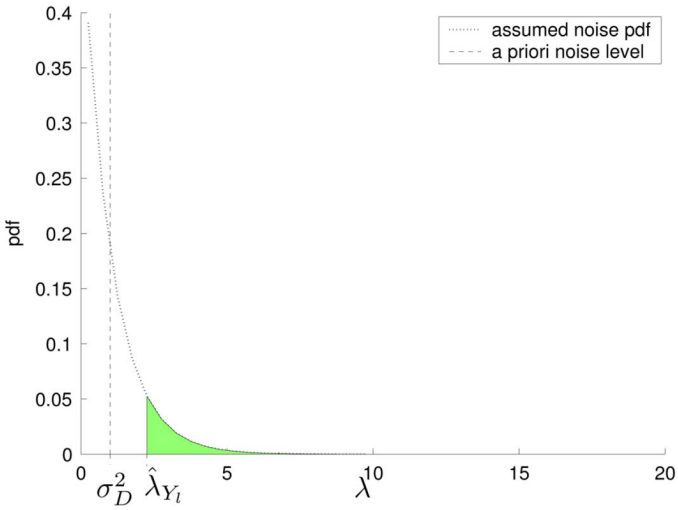


Fig. 3. Example showing how the noise-only subspace dimension is determined.

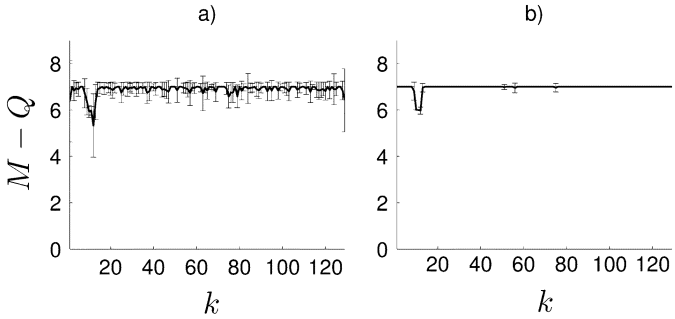


Fig. 4. (a) MDL model order estimator with *a priori* knowledge on noise variance. (b) Proposed model order estimator.

where we assume that some *a priori* knowledge of the noise level in each frequency bin is available. In order to correct for a consistent bias, we introduce a bias compensation factor for the estimation of $\hat{\sigma}_D^2(k,i)$ in Section IV-B.

A. Model Order Estimation

We consider an alternative approach, where we exploit the fact that some *a priori* information of the noise PSD is present. In this paper, we use the noise PSD estimate of the previous frame. This implicitly assumes relatively slowly varying noise, i.e., the DFT-domain noise correlation matrix $\mathbf{R}_D(k,i)$ should not change too abruptly from one frame to another. However, this does not limit the practical performance as will be shown in simulation experiments in Section VI. There it is shown that a change in the noise level of 15 dB/s can successfully be tracked. Furthermore, we assume that the eigenvalues in the noise-only subspace have an exponential distribution. Although we cannot mathematically show that the distribution is truly exponential, the choice for an exponential distribution for the noise eigenvalues shows a reasonable fit in validation experiments [20].

A noisy eigenvalue $\hat{\lambda}_{Y_l}$ is decided to belong to the signal subspace when the probability of observing an eigenvalue equal or larger than $\hat{\lambda}_{Y_l}$ is smaller than a prechosen minimum probability $P_{\min}$. We can write this as

$$\int_{\hat{\lambda}_{Y_l}}^{+\infty} f_{\Lambda_D}(\lambda_D)d\lambda_D < P_{\min} \tag{10}$$

where $f_{\Lambda_D}(\lambda_D)$ denotes the assumed pdf of the noise eigenvalues with its mean equal to the *a priori* known noise PSD, which we will take to be the noise PSD estimate of the previous frame. The decision procedure is visualized in Fig. 3. The dotted curve in Fig. 3 denotes the exponential pdf f_{Λ_D} of the noise eigenvalues belonging to the noise-only subspace. This approach can be seen within a hypothesis-based framework where H_0 and H_1 are defined as

$$\begin{aligned} &H_0\text{: } \hat{\lambda}_{Y_l} \text{ belongs to the noise-only subspace} \\ &H_1\text{: } \hat{\lambda}_{Y_l} \text{ belongs to the signal subspace.} \end{aligned} \tag{11}$$

Given a threshold λ_{th}, H_1 is decided when $\hat{\lambda}_{Y_l} > \lambda_{\text{th}}$. When $\hat{\lambda}_{Y_l} \leq \lambda_{\text{th}}$, $\hat{\lambda}_{Y_l}$ is decided to belong the noise-only subspace. The hypothesis is evaluated for all eigenvalues in increasing order until the H_0 hypothesis is rejected, which determines then the dimension of the noise and the signal subspace. The threshold λ_{th} can be expressed in terms of the false alarm probability $P_{fa} = P_{\min}$ and is given by $\lambda_{\text{th}} = -\sigma_D^2 \ln P_{fa}$ [21].

Notice that in the case of very low SNRs, the eigenvalues that fall in the noise-only and the signal subspace may be dominated by the noise and converge in their value. This is not only the case for the presented model order estimator, but holds in general for model order estimators. However, this was not observed to be a problem for the noise levels typically used in speech enhancement.

For evaluation, the proposed model order is compared to an MDL-based model order estimator. Comparing to the existing MDL-based model order estimator, [17] is not completely fair and will be in advantage of the proposed method, because it uses *a priori* knowledge on the noise variance while the MDL estimator in [17] does not. Therefore, we derived in Appendix B a modified MDL model order estimator where *a priori* knowledge on the noise variance is also taken into account.

For the comparison, a synthetic signal was constructed, consisting of a sinusoid at frequency bin number 11 in additive white noise. The sinusoid will not only have a contribution to bin $k = 11$, but to neighboring bins as well, because the period of the sinusoid is not an integer multiple of the minimum period visible with the used DFT size. The overall SNR between the sinusoid and white noise was 0 dB. For each frequency bin, we estimate a correlation matrix $\mathbf{R}_Y(k,i) \in \mathbb{C}^{7\times 7}$ and use either the proposed approach or the modified MDL method to estimate the dimension of the noise-only subspace. At those frequency bins where the sinusoid is present, a noise-only subspace dimension of 6 is expected, while at all other bins a noise-only subspace dimension of 7 is expected.

In Fig. 4, the outcome of the comparison between modified MDL derived in Appendix B and the proposed method are shown with $\mathbf{R}_Y(k,i) \in \mathbb{C}^{M\times M}$ estimated based on a data-matrix $\mathcal{Y} \in \mathbb{C}^{7\times 7}$. For each successive frequency bin, the model order is estimated. This is repeated for many frames. The average noise-only subspace dimension and the variance of noise-only subspace dimension are shown in Fig. 4. We see that the modified MDL approach leads to a larger variance in the estimated model order than the proposed approach.

We use in the following the proposed approach to estimate the model order of the noise-only subspace because of its smaller variance.

B. Bias Compensation of $\hat{\sigma}_D^2(k,i)$

When the dimension Q of the signal subspace is overestimated or underestimated, evaluating (9) can result in the introduction of a bias in the noise PSD estimate. To correct for such a bias in the estimated noise PSD as a result of consistent over or underestimates of Q, we introduce a signal subspace dimension dependent bias compensation factor $B(Q)$ and compute $\hat{\sigma}_D^2(k,i)$ as

$$\hat{\sigma}_D^2(k,i) = \frac{1}{B(Q)} \frac{1}{(M-Q)} \sum_{l=Q+1}^{M} \hat{\lambda}_{Y_l}(k,i). \tag{12}$$

The argumentation that we use to define the bias compensation factor is similar to the one introduced in [22].

The use of this bias compensation factor $B(Q)$ is based on the fact that

$$E\left[\frac{1}{(M-Q)} \sum_{l=Q+1}^{M} \hat{\lambda}_{Y_l}(k,i)\right] \tag{13}$$

is proportional to σ_D^2. We therefore write

$$\hat{\sigma}_D^2(k,i) = \frac{\sigma_D^2(k,i)}{E\left[\frac{1}{(M-Q)} \sum_{l=Q+1}^{M} \hat{\lambda}_{Y_l}(k,i)\right]} \times \frac{1}{(M-Q)} \sum_{l=Q+1}^{M} \hat{\lambda}_{Y_l}(k,i) \tag{14}$$

with

$$B(Q) = \frac{E\left[\frac{1}{(M-Q)} \sum_{l=Q+1}^{M} \hat{\lambda}_{Y_l}(k,i)\right]}{\sigma_D^2(k,i)}. \tag{15}$$

In order to compute the bias compensation factor $B(Q)$, for $Q = 0, 1, \ldots, M$, we approximate (15) by making use of a training procedure based on speech data degraded by white noise with a known variance $\sigma_D^2(k,i) = 1 \; \forall \; (k,i)$. Let $\tilde{B}(k,i)$ be defined as

$$\tilde{B}(k,i) = \frac{\frac{1}{M-Q} \sum_{l=Q+1}^{M} \hat{\lambda}_{Y_l}(k,i)}{\sigma_D^2(k,i)}. \tag{16}$$

Let $\mathcal{Q}(Q)$ be the set of time–frequency points in the training data for which the signal subspace dimension is estimated to be Q. $B(Q)$, $Q = 0, 1, \ldots, M$, is then computed by averaging $\tilde{B}(k,i)$ over the set $\mathcal{Q}(Q)$ leading to

$$B(Q) = \frac{1}{|\mathcal{Q}(Q)|} \sum_{(k,i) \in \mathcal{Q}(Q)} [\tilde{B}(k,i)] \tag{17}$$

where $|\mathcal{Q}(Q)|$ is the cardinality of the set $\mathcal{Q}(Q)$. Notice that computing the bias compensation factor in the training phase using the same signal subspace dimension estimator as when used in practice has the advantage that it can help to overcome systematic errors due to the signal subspace dimension estimator. Further, notice that $B(Q)$ can show some dependency on the SNR of the training data. This can be taken into account by computing $B(Q)$ also as a function of SNR.

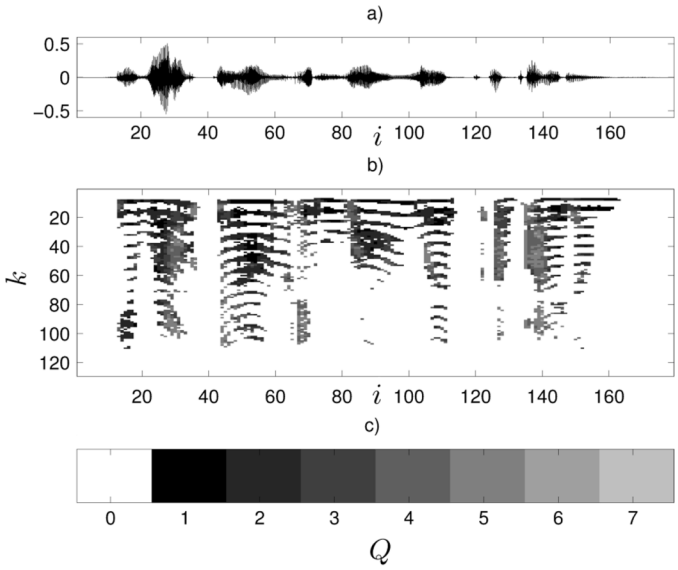


Fig. 5. (a) Clean speech signal. (b) Dimension of the signal subspace Q for each time-frequency point (k,i). Q is estimated by measuring in how many of the M eigenvalues 95% of the energy is distributed. (c) Color legend.

C. Dimension of $\mathbf{R}_Y(k,i)$

A requirement for the noise-only subspace to exist is that the signal subspace is not of full rank. For many speech sounds, it holds that they can be modeled using a (limited) number of basis functions. Consider, for example, the voiced speech sounds that can be modeled using a sum of complex exponentials. In that case, a particular frequency bin containing a harmonic will only observe a small number of complex exponentials and results in a low-dimensional signal subspace. The dimension of the correlation matrix can then be chosen such that the noise-only subspace has sufficiently high dimension to make an accurate estimate of the noise variance. To show that the dimension of the signal subspace is usually relatively low, we estimated the model order of clean speech signals. To do so, we estimated for each DFT coefficient in the time–frequency plane the correlation matrix $\mathbf{R}_X \in \mathcal{C}^{M \times M}$ with $M = 7$. For each estimated correlation matrix, we defined the model order as the number of eigenvalues needed to contain at least 95% of the energy. In Fig. 5, we illustrate this experiment. The clean speech signal is shown in Fig. 5(a). The sentence that is used originates from the Noizeus database [23] and reads the text "He wrote down a long list of items." For each time–frequency point, the estimated model order Q is indicated in Fig. 5(b) using colors from the legend in Fig. 5(c). The white color in the legend indicates speech absence, i.e., $Q = 0$. Time–frequency points are classified as speech absence when their energy is 40 dB below the DFT coefficient with maximum energy. We see that in general the dimension of the signal subspace Q is relatively low, especially at the harmonic tracks. Further, we see that $M = 7$ is a sufficient dimension for the correlation matrix, since the model order of 5 is hardly exceeded.

V. Implementational Aspects

In this section, we focus on some implementational aspects and present a summary of the proposed algorithm.

A. Prewhitening

In Section III, the assumption was made that $\mathbf{R}_D(k,i) = \sigma_D^2(k,i)\mathbf{I}_M$. Although this assumption holds as long as the DFT coefficients in $\mathbf{D}(k,i)$ are computed from time frames that are not overlapping and/or when the correlation time of the noise is small enough [15], this assumption becomes less valid when an overlap is introduced. In this section, we show how the interframe correlation is affected by the window overlap and indicate how a prewhitening matrix can be obtained such that the aforementioned assumption is fulfilled.

Let $D_t(m)$ denote a time domain sample considered as a random variable, let $\overline{D_t(m)}$ indicate complex conjugation of $D_t(m)$, and let P denote the frame shift. Let $R_D(k,i;p)$ denote the correlation between a noise DFT coefficient $D(k,i+p)$ and $D(k,i)$ with frame lag p. The correlation $R_D(k,i;p)$ can then be written as shown by (18)–(21) at the bottom of the page. We conclude that the correlation $R_D(k,i;p)$ consists of two components: a term $\tilde{R}_D(k,i;p)$ and a term $R_C(k,i;p)$. $R_C(k,i;p)$ contains all the cross-terms and is dependent on the cross-correlation between the time samples. In general, it holds that $R_C(k,i;p)$ decreases for increasing P. Also, the shorter the correlation time in the noise, the smaller $R_C(k,i;p)$ becomes. For $R_D(k,i;p)$ with $p > 0$, it follows from (21) that even if the time domain process $D_t(\cdot)$ is completely uncorrelated $R_D(k,i;p) \neq 0$, unless $P > K-1$, which means no overlap between consecutive frames.

Using simulations with white noise training data, we can estimate the first term $\tilde{R}_D(k,i;p)$ for a given overlap. The second term $R_C(k,i;p)$ is signal dependent and is therefore in general unknown.

We can write $\tilde{R}_D(k,i;p)$ in Toeplitz matrix form similar as (2), that is

$$\tilde{\mathbf{R}}_D(k,i) = b\begin{pmatrix} \tilde{R}_D(k,i;0) & \tilde{R}_D(k,i;1) & \cdots & \tilde{R}_D(k,i;p) \\ \tilde{R}_D(k,i;-1) & \tilde{R}_D(k,i;0) & & \\ \vdots & & \ddots & \\ \tilde{R}_D(k,i;-p) & \cdots & & \tilde{R}_D(k,i;0) \end{pmatrix}. \tag{22}$$

Let the relative error between the two correlation matrices $\mathbf{R}_D(k,i)$ and $\tilde{\mathbf{R}}_D(k,i)$ be defined as

$$\mathrm{Err}_{\mathrm{rel}}(\mathbf{R}_D(k,i), \tilde{\mathbf{R}}_D(k,i)) = \frac{\|\mathbf{R}_D(k,i) - \tilde{\mathbf{R}}_D(k,i)\|_{\mathrm{F}}^2}{\|\mathbf{R}_D(k,i)\|_{\mathrm{F}}^2} \tag{23}$$

with $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm [24]. In a simulation environment, we can then compute the error that would have been made between $\mathbf{R}_D(k,i)$ and $\tilde{\mathbf{R}}_D(k,i)$ by neglecting the second correlation term $R_C(k,i;p)$.

To investigate the influence of neglecting the second correlation term $R_C(k,i;p)$, we conducted an experiment where $K = 256$ and $P = 32$, i.e., the overlap between time frames was 87.5%. Then we computed for three different nonwhite noise sources that originate from the Noisex-92-database [25], i.e., babble noise, factory noise 1 and factory noise 2, the true correlation matrix $\mathbf{R}_D(k,i)$, and computed $\tilde{\mathbf{R}}_D(k,i)$ based on white noise. Factory noise 1 and 2 are two rather different noise types; factory noise 2 has more low-frequency spectral components, while factory noise 1 contains more high-frequency spectral components and has a somewhat broader spectrum. The relative error $\mathrm{Err}_{\mathrm{rel}}(\mathbf{R}_{\mathrm{D}}(\mathrm{k,i}), \tilde{\mathbf{R}}_{\mathrm{D}}(\mathrm{k,i}))$ that is made by replacing $\mathbf{R}_D(k,i)$ by $\tilde{\mathbf{R}}_D(k,i)$ based on white noise and averaged over all frequency bins is shown in Table I. We see that the relative error is always lower than $12 \cdot 10^{-3}$. This indicates that neglecting the the cross-terms leads to a relatively small error for these type of noise sources and that $\mathbf{R}_D(k,i)$ is mainly

$$R_D(k,i;p) = E[D(k,i+p)\overline{D(k,i)}] \tag{18}$$

$$= E\left[\left(\sum_{m=0}^{K-1} D_t(m+(i+p)P)e^{-j2\pi km/K}\right)\overline{\sum_{n=0}^{K-1} D_t(n+iP)e^{-j2\pi kn/K}}\right] \tag{19}$$

$$= e^{j2\pi kpP/K} E\left[\sum_{m=pP}^{K-1+pP} D_t(m+iP)e^{-j2\pi km/K}\sum_{n=0}^{K-1}\overline{D_t(n+iP)}e^{j2\pi kn/K}\right] \tag{20}$$

$$= \underbrace{e^{j2\pi kpP/K}\sum_{m=pP}^{K-1} E\left[|D_t(m+iP)|^2\right]}_{\tilde{R}_D(k,i;p)} + \underbrace{\sum_{\substack{m=pP \\ m\neq n+pP}}^{K-1+pP}\sum_{n=0}^{K-1} e^{j2\pi k(pP+n-m)/K}E[D_t(m+iP)\overline{D_t(n+iP)}]}_{R_C(k,i;p)}. \tag{21}$$

TABLE I
RELATIVE ERROR FOR THREE NONWHITE NOISE SOURCES

noise type	Babble	Factory 1	Factory 2
$\mathrm{Err}_{\mathrm{rel}}(\mathbf{R}(\mathrm{k,i}), \tilde{\mathbf{R}}(\mathrm{k,i}))$	$12 \cdot 10^{-3}$	$5.5 \cdot 10^{-3}$	$8.2 \cdot 10^{-3}$

determined by $\tilde{\mathbf{R}}_D(k,i)$. In the experimental results presented in Section VI, we will therefore neglect the correlation term $R_C(k,i;p)$ and use a correlation matrix $\tilde{\mathbf{R}}_D(k,i)$ trained on white noise to whiten possibly colored noise in $\mathbf{Y}(k,i)$.

Let $\mathbf{R}^{(1/2)}$ denote the principle square root of a matrix $\mathbf{R}$ [26]. The whitening of a vector $\mathbf{Y}(k,i)$ can then be written as

$$\mathbf{Y}_{\mathrm{pre}}(k,i) = \tilde{\mathbf{R}}_D^{-\frac{1}{2}}(k,i)\mathbf{Y}(k,i). \tag{24}$$

$\mathbf{Y}_{\mathrm{pre}}(k,i)$ is then used in (2). We denote the noise PSD when estimated in the whitened domain by $\hat{\sigma}^2_{D,\mathrm{pre}}(k,i)$. Notice, that if $\sigma^2_D(k,i)$ is estimated in the whitened domain, we have to correct with a scaling factor $(\mathrm{tr}[\tilde{\mathbf{R}}_D(k)])/(M)$, with $\mathrm{tr}[\cdot]$ the trace operator [26], to obtain the noise PSD estimate in the nonwhitened domain.

For some highly correlated noise types, i.e., with long correlation time, the aforementioned assumption of neglecting the correlation term $R_C(k,i;p)$ might be less valid. In that case, (24) is not sufficient to whiten the noise process. A possible solution is to estimate the whitening transform matrix $\mathbf{R}_D(k,i)$ online during speech absence using a VAD. A somewhat more advanced method would be to exploit signal subspace dimension estimator and update the estimated correlation matrix when the estimated noise-only subspace is full rank, i.e., $Q = 0$. However, the experimental results that are presented in Section VI are obtained using (24).

B. Algorithm Summary

In order to apply the proposed algorithm, the following steps should be taken.

Step 1) Compute $\hat{\mathbf{R}}_Y(k,i)$ using (7) and (8). The DFT coefficients necessary to form data-matrix $\mathcal{Y}$ in (8) are computed using an FFT of frames with a predefined overlap. The choice for this overlap is a tradeoff between variance reduction of $\hat{\mathbf{R}}_Y(k,i)$ and stationarity of the data in the data-matrix.

Step 2) Apply prewhitening using (24) to remove the correlation in the noise introduced in Step 1.

Step 3) Compute the eigenvalue decomposition of the prewhitened correlation matrix.

Step 4) Estimate the noise PSD $\sigma^2_{D,\mathrm{pre}}(k,i)$ using (12).

Step 5) Correct for scaling due to the prewhitening in Step 2

$$\hat{\sigma}^2_D(k,i) = \frac{\mathrm{tr}[\tilde{\mathbf{R}}_D(k,i)]}{M}\hat{\sigma}^2_{D,\mathrm{pre}}(k,i). \tag{25}$$

VI. EXPERIMENTAL RESULTS

For performance evaluation, we compare the proposed method with the minimum statistics-based noise tracking algorithm implemented as described in [8] and with the situation where the noise PSD is computed using an ideal VAD, i.e., during silence intervals preceding speech activity. The speech and noise signals originate from the Noizeus [23] database. This database was extended with stationary computer generated white Gaussian noise, babble noise from the Noisex-92-database [25], noise originating from a passing train, and nonstationary white Gaussian noise, respectively. Noisy signals are constructed synthetically at input SNRs of 0, 5, 10, and 15 dB. For the nonstationary white Gaussian noise, the initial noise level is 0, 5, 10, and 15 dB, respectively, and then gradually increases in 1 s by 15 dB where it stays at that level for 2 s after which it decreases again by 15 dB in 1 s. All signals are filtered at telephone bandwidth and sampled at 8 kHz. The noisy time domain signals are divided in frames of 256 samples with 50% overlap. For both analysis and synthesis a square root Hann window is used. The DFT coefficients that are used to form the data-matrix $\mathcal{Y}$ originate from time frames taken with an overlap of 87.5%. The dimensions of $\mathcal{Y}$ were chosen as $M = L = 7$ and $n_1 = n_2 = 6$. The estimated noise PSDs $\hat{\sigma}^2_D(k,i)$ are smoothed using an exponential smoother with adaptive smoothing factors [8].

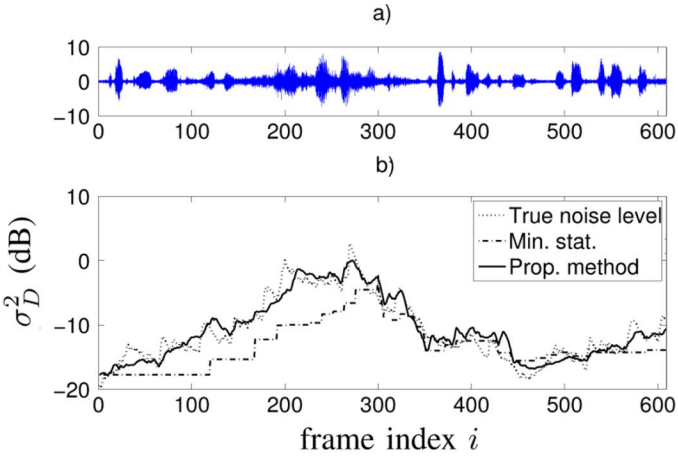


Fig. 6. (a) Noisy speech signal degraded by nonstationary train noise at an overall input SNR of 5 dB. (b) Comparison between proposed method and minimum statistics. The estimated noise levels are shown for bin $k = 20$.

A. Performance Evaluation

To illustrate the noise tracking performance of the proposed approach within a typical example of noisy speech, we concatenated four speech signals and degraded this by noise originating from a passing train at 5-dB global SNR. In Fig. 6, the estimated noise PSDs are shown for the proposed approach and the MS approach together with the true noise variance for a single frequency bin $k = 20$. This bin index corresponds to a frequency band centered around 625 Hz. We see that the proposed approach follows the increase in the noise level much better than the minimum statistics approach. This is due to the fact that the proposed approach can track changes in the noise level during speech presence. The MS approach on the other hand is limited in its update rate due to its search window and the fact that it can not track the noise when speech is continuously present in a bin. This results for MS in the delayed tracking of a rising noise level in Fig. 6. When the noise level decreases, we see that both methods track approximately equally well. This difference in behavior of minimum statistics towards increasing

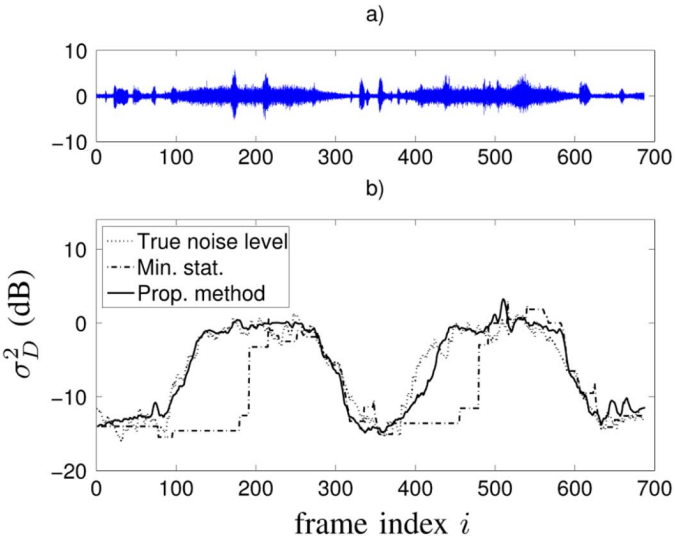


Fig. 7. (a) Noisy speech signal degraded by nonstationary white noise. (b) Comparison between proposed method and minimum statistics. The estimated noise levels are shown for bin $k = 20$.

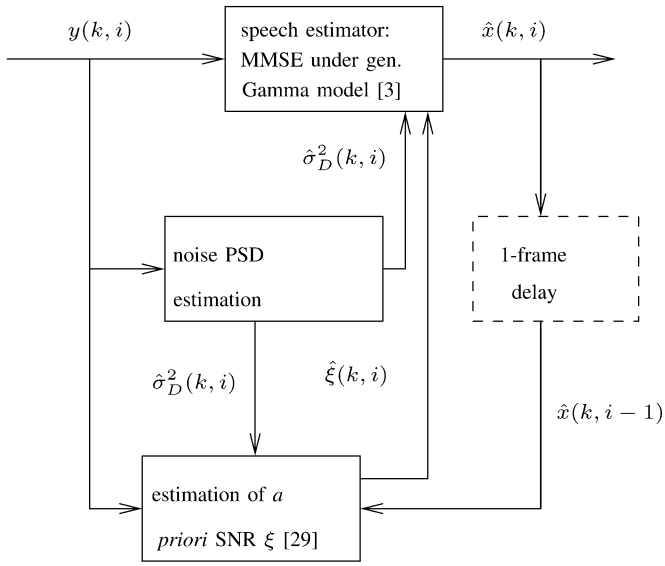


Fig. 8. Blockdiagram of DFT-domain-based enhancement algorithm.

and decreasing noise levels is due to the fact that MS tries to find the minimum. For a decreasing noise level, the minimum will in general be found among the most recent samples in the search window resulting in a much smaller additional delay for decreasing noise levels than for increasing noise levels.

In Fig. 7, another example is shown where the same speech signal is degraded by the nonstationary white noise described above. The initial part of the speech signal is degraded at an SNR of 10 dB. We again see that the proposed approach tracks the increase in noise level much faster than the MS approach.

1) Objective Performance Evaluation: For further objective performance evaluation, we use the segmental relative estimation error defined in [27] as

$$\mathrm{Err}_{\mathrm{seg}} = \frac{1}{I}\sum_{i=1}^{I}\frac{\sum_k \left[\hat{\sigma}_D^2(k,i) - \sigma_D^2(k,i)\right]^2}{\sum_k \sigma_D^4(k,i)} \tag{26}$$

where I is the total number of frames in the signal and where $\sigma_D^2(k,i)$ is the ideal noise PSD measured using noise periodograms smoothed over time using an exponential window, i.e.,

$$\sigma_D^2(k,i) = \alpha\sigma_D^2(k,i-1) + (1-\alpha)|D(k,i)|^2 \tag{27}$$

with a smoothing factor $\alpha = 0.9$ [8]. The measure $\mathrm{Err}_{\mathrm{seg}}$ is nonsymmetric and is more sensitive to overestimates than to underestimates. Therefore, we propose a symmetric segmental logarithmic estimation error, defined as

$$\mathrm{LOG{-}Err}_{\mathrm{seg}} = \frac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I}\left|10\log\left[\frac{\sigma_D^2(k,i)}{\hat{\sigma}_D^2(k,i)}\right]\right|. \tag{28}$$

In order to evaluate the influence of the proposed noise tracking algorithm on speech enhancement performance, we use the estimated noise PSDs within a DFT domain-based speech enhancement algorithm. In Fig. 8, a blockscheme of the used DFT-domain enhancement algorithm is shown. This algorithm works on a frame-by-frame basis, where per frame the clean speech DFT coefficients are estimated. As an estimator, we use the MMSE amplitude estimator under the generalized Gamma model as presented in [28], [3] with $\gamma = 2$ and $\nu = 0.1$. The maximum suppression was limited to 0.1 for perceptual reasons. For *a priori* SNR estimation, we use the decision-directed (DD) approach [29] where a smoothing factor $\alpha = 0.98$ was used as proposed in [29]. For a performance comparison, we use segmental SNR, i.e.,

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{I}\sum_{i}^{I} 10\log_{10}\frac{\sum_k |x(k,i)|^2}{\sum_k |x(k,i) - \hat{x}(k,i)|^2} \tag{29}$$

where $x(k,i)$ is a realization of a clean speech DFT and $\hat{x}(k,i)$ is its clean speech DFT estimate, respectively. Notice that the performance measured using $\mathrm{SNR}_{\mathrm{seg}}$ is unlike $\mathrm{LOG{-}Err}_{\mathrm{seg}}$ and $\mathrm{Err}_{\mathrm{seg}}$ not only influenced by the noise tracking algorithm, but also by the chosen gain function and *a priori* SNR estimator.

In Tables II–IV, we show performance evaluations for several noise types averaged over speech signals originating from the Noizeus database. We compare noise tracking using VAD, MS, and the proposed approach. We see that in general for all three objective measures, the performance is increased when using the proposed approach. Especially for noise sources that are characterized by a gradual change in the noise power (passing train and nonstationary white Gaussian noise), we see that the proposed approach outperforms MS and VAD. This is mainly due to the fact that a continuous update of the noise PSD allows for a faster update of changes in the noise power.

2) Subjective Performance Evaluation: For subjective evaluation, an OAB listening test was performed with eight participants, the authors not included. Here, O is the original clean speech signal, and A and B are two noisy signals that are enhanced using the scheme in Fig. 8 with two different methods for noise tracking. Method A uses the proposed noise tracking method, and method B uses the minimum statistics approach. The listeners were presented first the original signal followed by the two different enhanced signals A and B played in random order. The

TABLE II
PERFORMANCE IN TERMS OF $\mathrm{Err}_{\mathrm{seg}}$

noise source	input SNR (dB)	VAD	MS	prop. method
train	0	1.6	0.31	0.17
	5	0.77	0.35	0.20
	10	0.75	0.35	0.22
	15	1.2	0.42	0.29
street	0	26.7	0.41	0.35
	5	18.3	0.43	0.31
	10	17.4	0.48	0.34
	15	18.6	0.94	0.53
white	0	0.19	0.13	0.074
	5	0.20	0.13	0.093
	10	0.18	0.15	0.15
	15	0.19	0.45	0.24
babble	0	0.47	0.48	0.34
	5	0.47	0.47	0.37
	10	0.47	0.47	0.40
	15	0.47	0.50	0.43
passing train	0	0.52	0.33	0.16
	5	0.52	0.38	0.18
	10	0.52	0.45	0.28
	15	0.52	1.2	0.37
non-stationary WGN	0	0.66	0.33	0.068
	5	0.66	0.35	0.075
	10	0.66	0.38	0.096
	15	0.66	0.42	0.13

TABLE III
PERFORMANCE IN TERMS OF $\mathrm{LOG-Err}_{\mathrm{seg}}$

noise source	input SNR (dB)	VAD	MS	prop. method
train	0	3.3	2.6	1.9
	5	3.2	2.9	1.9
	10	3.0	2.6	2.0
	15	3.1	2.7	2.1
street	0	4.1	2.3	2.0
	5	4.0	2.8	2.0
	10	4.4	3.1	2.2
	15	3.6	2.7	2.5
white	0	1.6	1.3	1.0
	5	1.6	1.3	1.1
	10	1.6	1.4	1.2
	15	1.6	1.5	1.5
babble	0	4.4	3.9	2.1
	5	4.4	3.7	2.3
	10	4.4	3.5	2.6
	15	4.4	3.4	3.0
passing train	0	7.7	3.7	1.6
	5	7.7	3.6	1.9
	10	7.7	3.5	2.3
	15	7.7	3.6	3.0
non-stationary WGN	0	8.6	4.0	0.94
	5	8.6	4.1	1.0
	10	8.6	4.1	1.1
	15	8.6	4.0	1.4

participants had to indicate their preference for excerpt A or B. Each series was repeated four times, with each time a randomized order of the signals A and B. In this listening test, we used four different types of additive noise at two different SNRs, namely, white noise, street noise, noise originating from a passing train, and nonstationary white noise at SNRs of 5 and 15 dB. For each noise type and noise power level, we presented the listeners two female sentences and two male sentences. The average preference for method A under each test condition is shown in Table V. Under all test conditions, the proposed method for noise tracking was preferred over the minimum statistics approach.

TABLE IV
PERFORMANCE IN TERMS OF $\mathrm{SNR}_{\mathrm{seg}}$ (DB)

noise source	input SNR (dB)	VAD	MS	prop. method
train	0	-3.3	-4.1	-2.3
	5	-0.29	-0.48	1.1
	10	3.9	3.6	4.7
	15	7.4	7.4	8.0
street	0	-4.1	-4.1	-3.5
	5	-1.1	-0.85	0.34
	10	2.9	3.2	4.1
	15	6.8	7.0	7.1
white	0	-0.65	-0.22	0.42
	5	2.9	3.1	3.7
	10	6.3	6.4	6.9
	15	9.8	9.8	10.0
babble	0	-8.4	-8.8	-6.8
	5	-4.1	-4.2	-2.8
	10	0.26	0.25	1.2
	15	4.6	4.7	5.1
passing train	0	-6.2	-4.3	-1.7
	5	-1.8	-0.037	1.8
	10	2.6	4.2	4.9
	15	7.0	8.3	8.4
non-stationary WGN	0	-19.2	-14.5	-9.4
	5	-14.6	-10.5	-5.8
	10	-10.0	-6.5	-2.5
	15	-5.5	-2.5	0.60

TABLE V
LISTENING TEST RESULTS

noise source	input SNR	mean score for method A
white noise	5 dB	82.0 %
	15 dB	85.9 %
street noise	5 dB	77.3 %
	15 dB	67.2 %
passing train	5 dB	77.3 %
	15 dB	92.2 %
Non-stat. white noise	5 dB	96.1 %
	15 dB	89.1 %

B. Deterministic Noise

Deterministic noise components can in principle not be tracked with the proposed method, since they will appear in the signal subspace and not in the noise-only subspace. The noise is thus implicitly assumed to be stochastic. This is not only a property of the proposed method. Minimum statistics [8] implicitly assumes the noise to be stochastic as well. More specifically, the bias-compensation that is applied within minimum statistics is based on the assumption that the noise is stochastic. However, it is applied to deterministic components as well. A consequence of this is that after bias-compensation the deterministic noise components are in general slightly overestimated. However, in practice, minimum statistics is less sensitive than the proposed method when violating this assumption.

When deterministic noise components are present, they are often mixed with stochastic noise components. Therefore, it is not obvious how to estimate them. One way to estimate the deterministic noise components as well is to make use of the fact that for stochastic noise the minimum of the last T minimum statistics-based noise PSD estimates $\hat{\sigma}^2_{D,\min}$ is always smaller or equal than the current noise PSD estimate made by the proposed noise tracker (12), i.e.,

$$\min\left[\hat{\sigma}^2_{D,\min}(k,i-T+1),\ldots,\ \hat{\sigma}^2_{D,\min}(k,i)\right] \leq \hat{\sigma}^2_D(k,i). \tag{30}$$

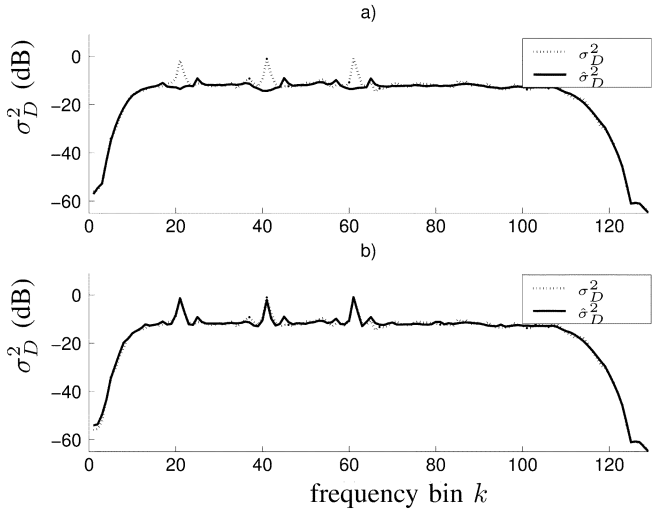


Fig. 9. (a) Noise tracking performed with DFT domain subspace decompositions only. (b) Noise tracking performed with DFT domain subspace decompositions combined with a tracker for deterministic components.

TABLE VI
PERFORMANCE IN TERMS OF $\mathrm{LOG-Err_{seg}}$ TO COMPARE THE INFLUENCE OF A DETERMINISTIC NOISE TRACKER

noise source	input SNR (dB)	MS	prop. noise tracker	prop. noise tracker combined with (31)
white noise with sinusoids	0	1.2	2.1	1.1
	5	1.2	1.7	1.1
	10	1.2	1.5	1.1
	15	1.2	1.3	1.2
Destroyer operations room	0	1.8	1.7	1.5
	5	1.9	1.9	1.7
	10	2.0	2.1	1.9
	15	2.1	2.5	2.3

Whenever this minimum is larger than $\hat{\sigma}_D^2(k,i)$, it is due to the fact that deterministic noise components are present. In that case, we can estimate the deterministic part of $\sigma_D^2(k,i)$ by

$$\hat{\sigma}_{D,\mathrm{det}}^2(k,i) = \left[\hat{\sigma}_{D,\mathrm{min}}^2(k,i) - \hat{\sigma}_D^2(k,i)\right] B_{\mathrm{min}}^{-1}(k,i) \quad (31)$$

where B_{min} is the bias-compensation as used in the minimum statistics method and which is used here to correct for the wrongly applied bias compensation on the deterministic component. The total estimate of $\sigma_D^2(k,i)$ is then given by adding $\hat{\sigma}_{D,\mathrm{det}}^2(k,i)$ and the estimate obtained by the proposed method in (12).

In Fig. 9(a), a comparison is shown where a speech signal was degraded by white noise (filtered at telephone bandwidth) at an SNR of 5 dB. As deterministic noise, a signal consisting of a sum of three harmonically related sinusoids with fundamental frequency of 656 Hz was added at an SNR of 10 dB with respect to the original clean speech signal. We see in Fig. 9(a) that with the DFT domain subspace noise tracking approach, it is not possible to estimate the sinusoidal noise components. In Fig. 9(b), we combine DFT domain subspace noise tracking method with (31) and see that we also determine the deterministic noise components. In Table VI, we show a comparison in terms of $\mathrm{LOG-Err_{seg}}$ for speech signals degraded by the above described noise. Here, the SNR between the stochastic noise and the speech signal is 5 dB, and the SNR between the deterministic noise and the speech signal is 0, 5, 10, and 15 dB, respectively. Moreover, we also show a comparison for the natural noise source *Destroyer operations room background noise* that originates from the Noisex-92 database [25]. This is a noise source containing both stochastic and some deterministic components. The comparison is made between minimum statistics, the proposed DFT domain subspace noise tracking approach, and the DFT domain subspace noise tracking method combined with (31). The obtained distortion for these partly deterministic noise types is decreased by combining the proposed noise tracker with (31). Notice that the experimental results in Section VI-A are based on the use of the DFT domain subspace noise tracker without the use of a deterministic noise tracker.

VII. CONCLUDING REMARKS

In this paper, we presented a novel approach for noise tracking. The method is based on construction of correlation matrices in the DFT domain per time-frequency point. Each correlation matrix can be decomposed into a signal subspace and a noise-only subspace. When the signal subspace is not full rank, the noise-only subspace can be used to estimate the noise PSD. The advantage of this approach is that the noise PSD can be updated for a DFT coefficient where both speech and noise are present. Comparisons showed that the presented method decreases the error between the true noise and the estimated noise spectrum. Further, enhancement performance is improved, especially for speech signals degraded by noise types that change gradually in power. Deterministic noise sources appear in the signal subspace and cannot be estimated by observing the noise-only subspace. However, these noise components can be tracked by observing T last minimum statistics-based noise PSD estimates.

The improved noise tracking performance of the proposed DFT subspace domain noise tracker over minimum statistics comes with an increase in the computational complexity. Although the dimensions of the correlation matrices are rather small, most of the computation time is spent on eigenvalue decompositions of the noisy correlation matrices. However, the MATLAB implementation of the proposed algorithm runs approximately two times real time on a PC with a Pentium 4 processor.

APPENDIX A
DERIVATION OF MDL-BASED MODEL ORDER ESTIMATOR WITHOUT *A PRIORI* KNOWLEDGE ON THE NOISE LEVEL

For completeness, the most important steps in deriving the standard MDL model order estimator as derived in [17] (assuming no knowledge of the noise variance) are given here.

The MDL criterion is defined as [17]

$$\mathrm{MDL} = -\log(f(\mathbf{y}_1,\ldots,\mathbf{y}_N \,|\, \Theta)) + \frac{1}{2} z \log N \quad (32)$$

where $\mathbf{y}_1,\ldots,\mathbf{y}_N$ are N i.i.d. zero mean M-dimensional multivariate Gaussian observation vectors, Θ a parameter vector of the model under consideration, and z the degree of freedom. Let Θ^Q be the parameter vector of the assumed model, i.e., $\Theta^Q = [a_1,\ldots a_Q, \sigma_D^2, C_1^H,\ldots,C_Q^H]$, where a_l, with $l \in \{1,\ldots,Q\}$ are the eigenvalues in the signal subspace, σ_D^2 is the noise variance, and C_l, with $l \in \{1,\ldots,Q\}$, are

the eigenvectors in the signal subspace. The joint probability density $f(\mathbf{y}_1,\ldots,\mathbf{y}_N|\Theta^Q)$ can then be written as

$$f(\mathbf{y}_1,\ldots,\mathbf{y}_N|\Theta^Q)=\prod_{i=1}^{N}\frac{1}{\pi^M\det\mathbf{R}^{(Q)}}\exp\left[-\mathbf{y}_i^H\mathbf{R}^{(Q)^{-1}}\mathbf{y}_i\right]. \tag{33}$$

The log likelihood of (33) is then given by

$$L(\Theta^{(Q)})=-N\log[\det\mathbf{R}^{(Q)}]-N\mathrm{tr}\left[\mathbf{R}^{(Q)^{-1}}\hat{\mathbf{R}}\right] \tag{34}$$

where $\hat{\mathbf{R}}$ is the estimate of the correlation matrix

$$\hat{\mathbf{R}}=U\begin{pmatrix}\Lambda_Q & 0\\ 0 & \Lambda_{M-Q}\end{pmatrix}U^H.$$

$\mathbf{R}^{(Q)}$ is now substituted with ML estimates

$$\mathbf{R}^{(Q)}=U\begin{pmatrix}\Lambda_Q & 0\\ 0 & \hat{\sigma}_D^2\mathbf{I}_{M-Q}\end{pmatrix}U^H \tag{35}$$

where $\Lambda_Q \in \mathbb{R}^{Q\times Q}$ is a diagonal matrix with the estimated eigenvalues $\hat{\lambda}_l$ with $l\in\{1,\ldots,Q\}$ of the assumed Q-dimensional signal subspace on the main diagonal. Further, U is a ML estimate of the eigenvector matrix and $\hat{\sigma}_D^2 = (1)/(M-Q)\sum_{l=Q+1}^{M}\hat{\lambda}_l$ is the ML estimate of the noise under the assumed Q-dimensional signal subspace. That U, Λ_Q, and $\hat{\sigma}_D^2=(1)/(M-Q)\sum_{l=Q+1}^{M}\hat{\lambda}_l$ are ML estimates of the eigenvector matrix, the signal subspace eigenvalues, and noise-only subspace eigenvalues will be shown in Appendix C.

Using the relation

$$\det\hat{\mathbf{R}}=\left(\prod_{l=1}^{Q}\hat{\lambda}_l\right)\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right) \tag{36}$$

$$\Leftrightarrow\left(\prod_{l=1}^{Q}\hat{\lambda}_l^{-1}\right)=\frac{\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right)}{\det\hat{\mathbf{R}}} \tag{37}$$

it can be shown that $L(\Theta^{(Q)})$ can be written as

$$L(\Theta^{(Q)})\equiv N\log\left[\frac{\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right)}{\left(\frac{1}{M-Q}\sum_{l=Q+1}^{M}\hat{\lambda}_l\right)^{(M-Q)}}\right]. \tag{38}$$

Equation (38) agrees with the result in [17].

Appendix B
MDL Model Order Estimator With *A Priori* Knowledge on σ_D^2

When *a priori* information on the noise level is present $\mathbf{R}^{(Q)}$ in (33) is substituted with

$$\mathbf{R}^{(Q)}=U\begin{pmatrix}\Lambda_Q & 0\\ 0 & \sigma_D^2\mathbf{I}_{M-Q}\end{pmatrix}U^H \tag{39}$$

$L(\Theta^{(Q)})$ then becomes

$$L(\Theta^{(Q)})=\underbrace{-N\log\left[\det\begin{pmatrix}\Lambda_Q & 0\\ 0 & \sigma_D^2\mathbf{I}_{M-Q}\end{pmatrix}\right]}_{A}-\underbrace{N\mathrm{tr}\left[\begin{pmatrix}\Lambda_Q^{-1} & 0\\ 0 & \sigma_D^{-2}\mathbf{I}_{M-Q}\end{pmatrix}\begin{pmatrix}\Lambda_Q & 0\\ 0 & \Lambda_{M-Q}\end{pmatrix}\right]}_{B}. \tag{40}$$

Part A

$$A=-N\log\left[\det\begin{pmatrix}\Lambda_Q & 0\\ 0 & {\sigma_D}^2\mathbf{I}_{M-Q}\end{pmatrix}\right] \tag{41}$$

$$=N\log\left[\frac{\left(\prod_{l=1}^{Q}\lambda_l^{-1}\right)}{(\sigma_D^2)^{(M-Q)}}\right] \tag{42}$$

using the relation:

$$\det\hat{\mathbf{R}}=\left(\prod_{l=1}^{Q}\hat{\lambda}_l\right)\left(\prod_{l=Q+1}^{M}\lambda_l\right)$$

$$\Leftrightarrow\left(\prod_{l=1}^{Q}\hat{\lambda}_l^{-1}\right)=\frac{\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right)}{\det\hat{\mathbf{R}}}$$

we obtain

$$A=N\log\left[\frac{\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right)}{\frac{\det\hat{\mathbf{R}}}{({\sigma_D}^2)^{(M-Q)}}}\right] \tag{43}$$

$$=N\log\left[\frac{\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right)}{(\sigma_D^2)^{(M-Q)}}\right]-N\log[\det\hat{\mathbf{R}}]. \tag{44}$$

Part B

$$B=N\mathrm{tr}\left[\begin{pmatrix}\Lambda_Q^{-1} & 0\\ 0 & \sigma_D^{-2}\mathbf{I}_{M-Q}\end{pmatrix}\begin{pmatrix}\Lambda_Q & 0\\ 0 & \Lambda_{M-Q}\end{pmatrix}\right] \tag{45}$$

$$=N\mathrm{tr}\left[\begin{pmatrix}\mathbf{I}_Q & 0\\ 0 & \sigma_D^{-2}\Lambda_{M-Q}\end{pmatrix}\right] \tag{46}$$

$$=N\left(Q+\sigma_D^{-2}\sum_{l=Q+1}^{M}\hat{\lambda}_l\right) \tag{47}$$

$$=N\left(Q+\sigma_D^{-2}(M-Q)\hat{\sigma}_D^2\right). \tag{48}$$

Combining Parts A and B gives

$$L(\Theta^{(Q)})=N\log\left[\frac{\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right)}{(\sigma_D^2)^{(M-Q)}}\right]-N\log\left[\det\hat{\mathbf{R}}\right]-N\left(Q+(M-Q)\frac{\hat{\sigma}^2}{\sigma_D^2}\right) \tag{49}$$

$$\equiv N\log\left[\frac{\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right)}{(\sigma_D^2)^{(M-Q)}}\right] - N\left(Q+(M-Q)\frac{\hat{\sigma}_D^2}{\sigma_D^2}\right) \quad (50)$$

$$= N\log\left[\frac{\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right)}{(\sigma_D^2)^{(M-Q)}}\right] - N\left(Q+\frac{\sum_{l=Q+1}^{M}\hat{\lambda}_l}{\sigma_D^2}\right) \quad (51)$$

where we left out the constant $N\log[\det\hat{\mathbf{R}}]$, since that does not influence the maximum of $L(\Theta^{(Q)})$.

Appendix C
ML Estimates for MDL and Modified MDL Estimator

In this Appendix, we derive maximum-likelihood estimates for the noise variance σ_D^2, the eigenvectors C_l, and the eigenvalues a_l, for $l\in\{1,\ldots,Q\}$.

The ML estimate $\hat{\sigma}_D^2 = (1)/(M-Q)\sum_{l=Q+1}^{M}\hat{\lambda}_l$ can be derived by maximization of (34) with respect to $\hat{\sigma}_D^2$, that is

$$\max_{\hat{\sigma}_D^2} L(\Theta^{(Q)}) \quad (52)$$

$$\frac{dL(\Theta^{(Q)})}{d\hat{\sigma}_D^2} = -N\frac{(M-Q)}{\hat{\sigma}_D^2} + N\left(\frac{1}{\hat{\sigma}_D^2}\right)^2\sum_{l=Q+1}^{M}\hat{\lambda}_l = 0 \quad (53)$$

which leads when solving for $\hat{\sigma}_D^2$ to $\hat{\sigma}_D^2 = (1)/(M-Q)\sum_{l=Q+1}^{M}\hat{\lambda}_l$.

ML estimates of the eigenvectors and signal subspace eigenvalues of $\mathbf{R}^{(Q)}$ can be derived by considering the EV decomposition of $\mathbf{R}^{(Q)}$

$$\mathbf{R}^{(Q)} = \mathbf{C}\begin{pmatrix} A_Q & 0 \\ 0 & A_{M-Q}\end{pmatrix}\mathbf{C}^H. \quad (54)$$

Since we use *a priori* information on the noise level, we can write $A_{M-Q} = \sigma^2\mathbf{I}_{M-Q}$. To find ML estimates of the eigenvectors $\mathbf{C}$ we consider the log-likelihood of (33), i.e.,

$$L(\Theta^{(Q)}) = -N\log\left[\det\mathbf{R}^{(Q)}\right] - N\mathrm{tr}\left[\mathbf{R}^{(Q)^{-1}}\hat{\mathbf{R}}\right] \quad (55)$$

$$= -N\log\left[\prod_{l=1}^{M}a_l\right] - N\mathrm{tr}[\mathbf{C}\mathbf{A}^{-1}\mathbf{C}^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H] \quad (56)$$

$$= -N\log\left[\prod_{l=1}^{M}a_l\right] - N\mathrm{tr}\left[\mathbf{A}^{-1}\underbrace{\mathbf{C}^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{C}}_{\mathbf{G}}\right]. \quad (57)$$

Let $\mathbf{G} = \mathbf{C}^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{C}$. The matrix $\mathbf{C}^H\mathbf{U}$ is now an orthogonal matrix and $\mathbf{C}^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{C}$ the eigenvalue decomposition of $\mathbf{G}$. Now we can write

$$\mathrm{tr}[\mathbf{A}^{-1}\mathbf{G}] = \mathrm{tr}\left[\begin{pmatrix} a_1^{-1} & 0 & \cdots & 0 \\ 0 & a_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_M^{-1}\end{pmatrix} \times \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1M} \\ g_{21} & g_{22} & & \\ \vdots & & \ddots & \\ g_{M1} & & & g_{1MM}\end{pmatrix}\right] \quad (58)$$

$$= \sum_{l=1}^{M}a_l^{-1}g_{ll} \quad (59)$$

and

$$L(\Theta^{(Q)}) = -N\sum_{l=1}^{M}\log[a_l] - N\sum_{l=1}^{M}a_l^{-1}g_{ll}. \quad (60)$$

ML estimates of a_l are then found as

$$\frac{\partial L}{\partial a_l} = -Na_l^{-1} + Na_l^{-2}g_{ll} \quad (61)$$

which gives $a_l = g_{ll}$.

Inserting this in L leads to

$$L = -N\log\left[\prod_{l=1}^{M}g_{ll}\right] - NM. \quad (62)$$

To maximize L, we need to minimize $\prod_{l=1}^{M}g_{ll}$. To find this minimum, we use *Hadamards* inequality

$$\det\mathbf{G} \le \prod_{l=1}^{M}g_{ll} \quad (63)$$

with equality if and only if $\mathbf{G}$ is diagonal and $\mathbf{G}$ should be positive definite. We know that $\mathbf{G} = \mathbf{C}^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{C}$. The orthogonal matrix $\mathbf{C}$ does not influence the determinant of G. Therefore, we can choose $\mathbf{C} = \mathbf{U}$ such that *Hadamards* inequality leads to equality.

Let us now use the fact that $\mathbf{U}$ are ML estimates of $\mathbf{C}$. ML estimates of the eigenvalues of $\mathbf{R}^{(Q)}$ can then be computed by taking partial derivatives of (55), i.e.,

$$\frac{\partial L(\Theta^{(Q)}}{\partial a_j} = -N\frac{1}{a_j} + Na_j^{-2}\hat{\lambda}_j = 0 \quad (64)$$

so that

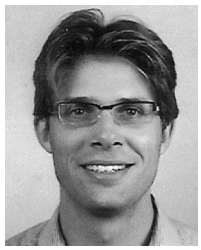
$$a_j = \hat{\lambda}_j. \quad (67)$$

References

[1] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

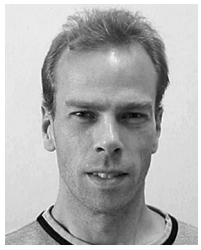
[2] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Applied Signal Process.*, vol. 7, pp. 1110–1126, May 2005.

[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 Lang. Process.*, vol. 6, no. 6, pp. 1741–1752, Aug. 2007.
[4] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Process.*, vol. 86, no. 4, pp. 698–709, 2006.
[5] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[6] J. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.
[7] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. Eur. Signal Process. Conf.*, 1994, pp. 1182–1185.
[8] 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, Jul. 2001.
[9] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.
[10] Y. Ephraim and H. L. van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.
[11] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted with colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.
[12] K. Hermus, P. Wambacq, and H. van Hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP J. Applied Signal Process.*, pp. 1–15, 2007.
[13] V. Buhnjun and M. Brookes, "Narrowband noise estimation in the subspace domain," in *Proc. Int. Symp. Intell. Multimedia, Video, Speech Process.*, 2004, pp. 1–4.
[14] V. Buhnjun, M. Brookes, and J. Y. C. Wen, "Eigendomain-based noise estimation with the minimum statistics approach," in *Int. Workshop Acoust. Echo Noise Control (IWAENC)*, 2006, pp. 1–4.
[15] D. R. Brillinger*, Time Series: Data Analysis and Theory*. Philadelphia, PA: SIAM, 2001.
[16] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.
[17] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 387–392, Apr. 1985.
[18] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.
[19] P. Stoica and Y. Selén, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.
[20] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," Tech. Rep. ICT-2007-03, 2007.
[21] S. K. Kay*, Fundamentals of Statistical Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1998, vol. 2.
[22] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Process.*, vol. 86, no. 6, pp. 1215–1229, Jun. 2006.
[23] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006, vol. 1, pp. 153–156.
[24] C. W. Therrien*, Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
[25] A. Varga and H. J. M. Steeneken, "Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–253, 1993.
[26] R. A. Horn and C. R. Johnson*, Matrix Analysis*. : Cambridge Univ. Press, 1999.
[27] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 446–475, Sep. 2003.
[28] R. C. Hendriks, J. S. Erkelens, J. Jensen, and R. Heusdens, "Minimum mean-square error amplitude estimators for speech enhancement under the generalized gamma distribution," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Sep. 2006, pp. 1–4.
[29] 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. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.



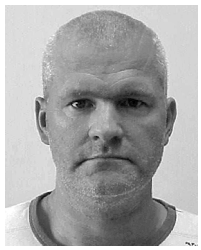
Richard C. Hendriks received the B.Sc. and M.Sc. degrees in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2001 and 2003, respectively. He is currently pursuing the Ph.D. degree at Delft University of Technology.

In September 2003, he joined the Department of Mediamatics, Delft University of Technology. From September 2005 to December 2005, he was a Visiting Researcher at the Institute of Communication Acoustics, Ruhr-University Bochum, Bochum, Germany. His main research interests are digital speech and audio processing, including acoustical noise reduction and speech enhancement.



Jesper Jensen received the M.Sc. and Ph.D. degrees in electrical engineering from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively.

From 1996 to 2001, he was with the Center for PersonKommunikation (CPK), Aalborg University, as a Researcher, Ph.D. student, and Assistant Research Professor. In 1999, he was a Visiting Researcher at the Center for Spoken Language Research, University of Colorado, Boulder. From 2000 to 2007, he was a Postdoctoral Researcher and Assistant Professor at the Delft University of Technology, Delft, The Netherlands. He is currently with Oticon, Smørum, Denmark. His main research interests are digital speech and audio signal processing, including coding, synthesis, and enhancement.



Richard Heusdens received the M.Sc. and Ph.D. degrees from the Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively.

Since 2002, he has been an Associate Professor in the Department of Mediamatics, Delft University of Technology. In the spring of 1992, he joined the Digital Signal Processing Group, Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997, he joined the Circuits and Systems Group, Delft University of Technology, where he was a Postdoctoral Researcher. In 2000, he moved to the Information and Communication Theory (ICT) Group, where he became an Assistant Professor responsible for the audio and speech processing activities within the ICT group. He is involved in research projects that cover subjects such as audio and speech coding, speech enhancement, and digital watermarking of audio.