Noise Correlation Matrix Estimation for Multi-Microphone Speech Enhancement

Richard C. Hendriks and Timo Gerkmann, *Member, IEEE*

***Abstract*—For multi-channel noise reduction algorithms like the minimum variance distortionless response (MVDR) beamformer, or the multi-channel Wiener filter, an estimate of the noise correlation matrix is needed. For its estimation, it is often proposed in the literature to use a voice activity detector (VAD). However, using a VAD the estimated matrix can only be updated in speech absence. As a result, during speech presence the noise correlation matrix estimate does not follow changing noise fields with an appropriate accuracy. This effect is further increased, as in nonstationary noise voice activity detection is a rather difficult task, and false-alarms are likely to occur. In this paper, we present and analyze an algorithm that estimates the noise correlation matrix without using a VAD. This algorithm is based on measuring the correlation of the noisy input and a noise reference which can be obtained, e.g., by steering a null towards the target source. When applied in combination with an MVDR beamformer, it is shown that the proposed noise correlation matrix estimate results in a more accurate beamformer response, a larger signal-to-noise ratio improvement and a larger instrumentally predicted speech intelligibility when compared to competing algorithms such as the generalized sidelobe canceler, a VAD-based MVDR beamformer, and an MVDR based on the noisy correlation matrix.**

***Index Terms*—Multi-microphone, noise correlation matrix, noise reduction, speech enhancement.**

I. Introduction

THE demand for speech processing applications like cell phones, hearing aids, and speech recognition systems to work anywhere and at anytime, makes them also more vulnerable for disturbances like environmental noise. To reduce degradations in terms of speech intelligibility and speech quality, single- and multi-microphone noise reduction algorithms are often employed. Although single-channel noise reduction algorithms are generally able to increase speech quality, see, e.g., [1], improvements in terms of intelligibility are reported only rarely. On the other hand, due to the possibility to perform spatial filtering, multi-microphone noise reduction algorithms have a better ability to increase speech quality as well as intelligibility of speech in noise [2]. Most often, single- and multi-channel noise reduction algorithms are implemented in the temporal-spectral domain, e.g., by computing a discrete Fourier transform (DFT). Single-channel noise reduction algorithms then estimate the clean speech DFT coefficients by applying a gain to the noisy DFT coefficients, e.g., [3]–[6], while multi-channel NR algorithms estimate the clean speech DFT coefficients by taking a linear combination of several noisy DFT coefficients from multiple microphones and form a so-called beamformer, e.g., [7]–[9].

One of the important parameters on which all single-channel estimators depend, is the noise power spectral density (PSD). This quantity is an unknown expected value and its estimation is particularly difficult in nonstationary noise fields which are common in many daily life situations. During the last decade more research focused on the estimation of the noise PSD for non-stationary noise sources. Important contributions on this topic are the methods based on so-called minimum statistics [10], [11]. Due to its underlying principle, the minimum statistics approach results in only little speech leakage into the noise PSD estimate. However, the worst case delay of tracking an increasing noise level at a particular frequency bin can be rather long compared to the speed at which certain noise sources tend to change. This motivated the further development of noise PSD estimators that can estimate the noise PSD of nonstationary noise sources with a shorter tracking delay, see, e.g., [12]–[15].

For multi-microphone noise reduction the noise PSD per microphone is extended with the noise cross power spectral densities between microphones, altogether known as the noise cross-correlation matrix. Using the spatial information on the noise field that is contained in the noise correlation matrix it is possible to adaptively steer a beamformer in the direction of interest and to cancel or reduce the effect of noise sources in other directions. This is done by multi-channel noise reduction methods like the minimum variance distortionless response (MVDR) beamformer [16] and the multi-channel Wiener filter [9]. The noise correlation matrix directly determines the spatial filtering that is applied by these algorithms. Therefore, for these types of algorithms to be able to optimally adapt to the surrounding noise field, knowing the noise correlation matrix with high accuracy is of high importance. Wrong estimates of the noise correlation matrix can either lead to the situation that disturbances from certain angles are not optimally suppressed, or, worse, that noise coming from certain angles is amplified.

Similar as for the noise PSD, the noise correlation matrix is an unknown expected value that needs to be estimated from the noisy microphone signals. Although noise PSD estimation received significant interest, estimation of the noise correlation

Manuscript received December 10, 2010; revised April 19, 2011; accepted June 08, 2011. Date of publication June 16, 2011; date of current version November 09, 2011. The work was supported by the Dutch Technology Foundation STW and the European Community's Seventh Framework Program under Grant Agreement PIAP-GA-2008-214699. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Seltzer.

R. C. Hendriks is with the Signal and Information Processing Lab, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: r.c.hendriks@tudelft.nl).

T. Gerkmann is with the Sound and Image Processing Lab, KTH–Royal Institute of Technology, SE-100 44 Stockholm, Sweden, (e-mail: gerkmann@kth.se).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2159711

matrix appear to have been less explored. Where estimation of the noise PSD is challenging because noise sources can be nonstationary across time, estimation of the noise correlation matrix can be even more challenging since the noise field can also be spatially nonstationary.

An approach that underlies many noise correlation matrix estimation methods is the usage of an energy-based voice activity detector (VAD), see, e.g., [17]–[20]. Most of these methods rely on a single-channel noise PSD estimator per microphone signal and use this to determine speech presence. When speech is not present, the off-diagonal terms of the noise correlation matrix can then be updated by means of, e.g., exponential smoothing. For noise sources that are stationary across time and space, VAD-based estimation of the noise correlation matrix can be sufficiently accurate. However, many noise sources encountered in reality are nonstationary across time and space, e.g., consider a passing car or passing train. The fact that a VAD does not allow to update the correlation matrix estimate during speech presence, might result for such temporally and/or spatially nonstationary noise fields in a wrongly estimated noise correlation matrix. As a consequence of this, the shape of the beamformer response is adapted towards the wrong direction.

Other more recently developed methods that do not rely on a VAD are based on additional assumptions on the type of noise field. In [21], a method was proposed that relies on the assumption that the noise field is diffuse. Two important examples of diffuse noise fields are a spherically and cylindrically isotropic noise field. Assuming the noise field to be spherically or cylindrically isotropic, it can be shown that the noise correlation matrix is real, see, e.g., [22]. The method presented in [21] explicitly exploits this property and estimates therefore only the real part of the noise correlation matrix. Another method that exploits the assumption of a diffuse noise field is for example used in [23]. This method makes use of the fact that per frequency bin the noise correlation matrix for a diffuse noise field can be decomposed into a scalar, that is, the noise PSD, and a matrix that is completely determined by the noise coherence function. For the case of an ideally spherically or cylindrically isotropic noise field, the coherence function is known [22] and depends only on the inter-microphone distance. However, assuming a diffuse noise field is not always realistic and is a strong limitation of noise correlation matrix estimation.

Alternative procedures originally proposed by Frost [24] and Griffiths and Jim [25] do not estimate the noise correlation matrix directly, but estimate the filter coefficients of a linearly constrained minimum variance (LCMV) or MVDR beamformer in an adaptive manner. The algorithm proposed by Griffiths and Jim is often referred to as the generalized sidelobe canceller (GSC). Both adaptive methods employ a least-mean-square (LMS) algorithm and allow accurate estimation of the filter coefficients when the noise sources are rather stationary in both space and time. However, their performance gets degraded when noise sources tend to be more nonstationary in space and time [26].

In this paper, we present a general method for estimation of the noise correlation matrix for spatially and temporally nonstationary noise fields and an M-dimensional microphone array. The presented method exploits the fact that recently developed noise PSD estimation algorithms can track a changing noise PSD with a relatively low delay. In addition, similar as for the GSC, the proposed approach exploits the fact that given the propagation vector of the target source, a noise reference can be obtained. However, different than with the GSC the proposed method does not make use of LMS-based algorithms, but directly computes the elements of the noise correlation matrix. The proposed method does not assume the noise field to be diffuse and can estimate the noise correlation matrix also when speech is present at the frequency bin under consideration.

The remainder of this paper is organized as follows. In Section II, we present the notation and basic assumptions that we use throughout this paper. Subsequently, in Section III we present our proposed method for noise correlation estimation. In Section IV, we make an estimation error analysis of the proposed approach and present a method to reduce estimation errors. In addition, we show in Section IV that when the proposed method is combined with an MVDR-beamformer, that under certain conditions the filter coefficients are insensitive for certain estimation errors. In Section V, we present experimental results and a discussion on the proposed method and reference methods when applied in an environment with reverberation. Finally, in Section VI concluding remarks are given.

II. Notation and Basic Assumptions

In this paper, we consider a general multi-microphone setup consisting of M microphones. Each of the noisy microphone signals is windowed on a frame-by-frame basis using a window length L and window shift R, and transformed to the DFT domain, that is,

$$Y_m(k,i) = \sum_{l=0}^{L-1} Y_{m,t}(iR+l)w(l)e^{-j2\pi kl/L}$$

where $Y_m(k,i)$ denotes the noisy DFT coefficient for frequency bin k, time-frame i and microphone number m. Further, $Y_{m,t}(iR+l)$ denotes a time-domain sample for microphone-number m and $w(\cdot)$ denotes the time-domain window. In a similar way, we define the clean speech and noise DFT coefficients $S_m(k,i)$ and $N_m(k,i)$, respectively. The DFT coefficients are assumed to be random variables, indicated by uppercase letters, and their corresponding realizations are indicated by lowercase letters. Furthermore, bold-faced letters indicate the use of matrices. The speech and noise DFT coefficients are assumed to be additive, i.e.,

$$Y_m(k,i) = S_m(k,i) + N_m(k,i) \tag{1}$$

and uncorrelated, i.e.,

$$\mathrm{E}\{S_m(k,i)N_n(k,i)\} = 0 \quad \forall \quad k,i,m,n \tag{2}$$

where $\mathrm{E}\{\cdot\}$ denotes the statistical expectation operator. We assume the DFT coefficients to be independent across time and frequency, and will therefore neglect time- and frequency-indices for ease of notation. Let $\mathbf{Y} \in \mathbb{C}^M$ denote a vector containing the noisy DFT coefficients for each of the M microphones, that is $\mathbf{Y} = [Y_1, \ldots, Y_M]^T$. Similarly, we define $\mathbf{S} \in \mathbb{C}^M$ and $\mathbf{N} \in \mathbb{C}^M$ as the vectors containing the clean and noise

microphone DFT coefficients of the M microphones, respectively. It is assumed that there is a single target source whose acoustic path to the M microphones is modeled by the frequency dependent propagation vector $\mathbf{d} = [d_1, \ldots, d_M]^T$. The vector with clean speech DFT coefficients is therefore given by $\mathbf{S} = S\mathbf{d}$, where S is the clean speech DFT at the target speaker location. Altogether this leads to the following compact vector representation:

$$\mathbf{Y} = \mathbf{S} + \mathbf{N} = S\mathbf{d} + \mathbf{N}. \tag{3}$$

III. Noise Correlation Matrix Estimation

In order to estimate the noise correlation matrix, we make explicit use of the following three assumptions. At first, we assume the propagation vector $\mathbf{d}$ to be known. Second, we assume that the noise PSD per microphone is known, and finally, we make use of the assumption already expressed in (2), that is, the noise and speech DFT coefficients are assumed to be uncorrelated across time, frequency, and microphones.

The noise correlation matrix is given by

$$\begin{aligned}\mathbf{\Sigma} &= \mathrm{E}\left\{\mathbf{N}\mathbf{N}^H\right\} \\ &= \begin{pmatrix} \mathrm{E}\left\{|N_1|^2\right\} & \mathrm{E}\left\{N_1 N_2^*\right\} & \cdots & \mathrm{E}\left\{N_1 N_M^*\right\} \\ \mathrm{E}\left\{N_2 N_1^*\right\} & \mathrm{E}\left\{|N_2|^2\right\} & & \\ \vdots & & \ddots & \\ \mathrm{E}\left\{N_M N_1^*\right\} & & & \mathrm{E}\left\{|N_M|^2\right\} \end{pmatrix}\end{aligned} \tag{4}$$

where $\cdot^*$ indicates complex conjugation. Estimation of this matrix requires estimation of all elements in (4). In order to do so, we make a distinction between the diagonal and off-diagonal elements. The diagonal elements correspond to the noise PSD per microphone, which can be estimated, e.g., by [10]–[15]. Thus, with one of these estimators at hand, estimation of the noise correlation matrix reduces to estimation of the off-diagonal elements only, i.e., the cross-correlation between noise DFT coefficients at different microphones.

Based on the assumption that $\mathbf{d}$ is given, it is possible to obtain a noise reference for each microphone pair where the target signal is completely canceled. Note, that this also holds for reverberant signals when $\mathbf{d}$ describes the Fourier transforms of the room impulse responses of the M microphones.

Let $d(m)$ denote the mth element of $\mathbf{d}$. Further, let $d_{n,m} = d(n)/d(m) \in \mathbb{C}$ be defined as the scaling that needs to be applied to obtain the clean speech DFT coefficient at microphone number n from the clean speech DFT coefficient at microphone number m, that is, $S_n = d_{n,m} S_m$. However, in the case that $\mathbf{d}$ models the room impulse response, special care needs to be taken when $\mathbf{d}$ contains zeros. When only the direct path is modeled, $\mathbf{d}$ does not contain any zeros and this problem does not occur. A noise reference for element (n, m) of the correlation matrix is then obtained by

$$\begin{aligned} P_{n,m} &= Y_n - d_{n,m} Y_m && (5)\\ &= N_n - d_{n,m} N_m. && (6)\end{aligned}$$

The estimation of the cross-term (n, m) of the correlation matrix can be done by exploiting the correlation between $P_{n,m}$ and Y_m, that is,

$$\begin{aligned} \mathrm{E}\{P_{n,m} Y_m^*\} &= \mathrm{E}\{N_n Y_m^*\} - d_{n,m}\mathrm{E}\{N_m Y_m^*\} && (7)\\ &= \mathrm{E}\{N_n N_m^*\} - d_{n,m}\mathrm{E}\left\{|N_m|^2\right\} && (8)\end{aligned}$$

where, in order to go from (7) to (8), use is made of the assumption that speech and noise DFT coefficients are uncorrelated and that the speech DFT coefficients are canceled in the noise reference $P_{n,m}$. Thus, in this paper, we propose to estimate the off-diagonal elements (n, m) of the noise correlation matrix by solving $\mathrm{E}\{N_n N_m^*\}$ from (8), as

$$\mathrm{E}\{N_n N_m^*\} = \mathrm{E}\{P_{n,m} Y_m^*\} + d_{n,m}\mathrm{E}\left\{|N_m|^2\right\} \tag{9}$$

while the diagonal elements are obtained by using a single channel PSD estimator from the literature.

IV. Estimation Error Analysis and Reduction

The expression in (9) is based on the expected values $\mathrm{E}\{P_{n,m} Y_m^*\}$ and $\mathrm{E}\left\{|N_m|^2\right\}$ that are unknown in practice and have to be estimated from the available noisy speech realizations.

The term $\mathrm{E}\left\{|N_m|^2\right\}$ is the noise PSD and can be estimated using aforementioned low-delay noise PSD estimators. However, the estimated noise PSD can be over or underestimated, which will introduce errors on the estimated cross-terms of the noise correlation matrix. These errors on the estimated noise PSD for microphone m will be denoted by $\epsilon_{\mathrm{PSD},m}$.

We use exponential smoothing to estimate the term $\mathrm{E}\{P_{n,m} Y_m^*\}$, as

$$\begin{aligned}\widetilde{\mathrm{E}}\{P_{n,m}(k,i) Y_m^*(k,i)\} = (1-\alpha)\widetilde{\mathrm{E}}\{P_{n,m}(k,i-1) Y_m^*(k,i-1)\} \\ + \alpha p_{n,m}(k,i) y_m^*(k,i)\end{aligned} \tag{10}$$

where $\widetilde{\mathrm{E}}\{\cdot\}$ denotes an estimate of $\mathrm{E}\{\cdot\}$, and $p_{n,m}$ is computed using (5) by using realizations of the noisy DFT coefficients Y_n and Y_m. The derivation of (9) relies on the assumption that speech and noise are uncorrelated. However, even when speech and noise are truly uncorrelated, estimation of $\mathrm{E}\{P_{n,m} Y_m^*\}$ based on realizations as in (10), i.e., $p_{n,m}(k,i)$ and $y_m^*(k,i)$, will give rise to a nonzero contribution due to the speech contained in $y_m^*(k,i)$. These estimation errors will be denoted by $\epsilon_{\mathrm{corr},n,m}$.

In the following section, we will show that the estimation errors $\epsilon_{\mathrm{corr},n,m}$ can be reduced by exploiting the fact that the noise correlation matrix $\mathbf{\Sigma}$ is Hermitian symmetric and therefore we have that $\mathrm{E}\{N_n N_m^*\} = (\mathrm{E}\{N_m N_n^*\})^*$. The off-diagonal elements of $\mathbf{\Sigma}$ can therefore be estimated as

$$\widehat{\mathrm{E}}\{N_n N_m^*\} = \frac{\widetilde{\mathrm{E}_\epsilon}\{N_n N_m^*\} + (\widetilde{\mathrm{E}_\epsilon}\{N_m N_n^*\})^*}{2} \tag{11}$$

where $\widetilde{\mathrm{E}_\epsilon}\{\cdot\}$ denotes an estimate of $\mathrm{E}\{\cdot\}$ that also includes estimation errors $\epsilon_{\mathrm{PSD},m}$ and/or $\epsilon_{\mathrm{corr},n,m}$. Moreover, in Section IV-B we will show that when the proposed noise correlation matrix estimator according to (11) is used in combination with an MVDR beamformer, that under far and free field

conditions the final MVDR becomes completely independent of the single-channel noise PSD estimate for each microphone and therefore independent of the error ϵ_{PSD}.

A. Reduction of the Estimation Errors ϵ_{PSD} and ϵ_{corr}

Taking the estimation of $\mathrm{E}\{P_{n,m}Y_m^*\}$ based on realizations into account, we obtain a nonzero contribution $\widetilde{\mathrm{E}}\{N_nS_m^*\} - d_{n,m}\widetilde{\mathrm{E}}\{N_mS_m^*\}$, even though $\mathrm{E}\{S_m(k,i)N_n(k,i)\} = 0 \ \forall \ k,i,m,n$. Therefore, instead of

$$\widetilde{\mathrm{E}}\{P_{n,m}Y_m^*\} = \widetilde{\mathrm{E}}\{N_nN_m^*\} - d_{n,m}\widetilde{\mathrm{E}}\{|N_m|^2\} \tag{12}$$

we obtain

$$\begin{aligned}\widetilde{\mathrm{E}}\{P_{n,m}Y_m^*\} &= \widetilde{\mathrm{E}}\{N_nN_m^*\} \\ &\quad -d_{n,m}\widetilde{\mathrm{E}}\{|N_m|^2\} + \underbrace{\widetilde{\mathrm{E}}\{N_nS_m^*\} - d_{n,m}\widetilde{\mathrm{E}}\{N_mS_m^*\}}_{\epsilon_{\mathrm{corr},n,m}}.\end{aligned} \tag{13}$$

In addition, estimation of the noise PSD $\mathrm{E}\{|N_m|^2\}$ by means of a noise PSD estimator leads to an estimate

$$\widehat{\mathrm{E}}\{|N_m|^2\} = \widetilde{\mathrm{E}}\{|N_m|^2\} + \epsilon_{\mathrm{PSD},m} \tag{14}$$

where $\widetilde{\mathrm{E}}\{|N_m|^2\}$ is the noise PSD estimate that would be obtained by smoothing the noise realizations themselves, and $\epsilon_{\mathrm{PSD},m}$ the estimation error with respect to $\widetilde{\mathrm{E}}\{|N_m|^2\}$. If we take the estimated expected values $\widetilde{\mathrm{E}}\{P_{n,m}Y_m^*\}$ from (13) and $\widehat{\mathrm{E}}\{|N_m|^2\}$ from (14) into account and substitute them into (9), we obtain

$$\widetilde{\mathrm{E}_\epsilon}\{N_nN_m^*\} = \widetilde{\mathrm{E}}\{N_nN_m^*\} + \epsilon_{\mathrm{corr},n,m} + d_{n,m}\epsilon_{\mathrm{PSD},m}. \tag{15}$$

In order to reduce the presence of $\epsilon_{\mathrm{corr},n,m}$ and $\epsilon_{\mathrm{PSD},m}$ we can make use of the Hermitian symmetric property of $\mathbf{\Sigma}$. Let us therefore consider the estimate of the complex conjugate of $\mathrm{E}\{N_nN_m^*\}$, that is

$$\widetilde{\mathrm{E}_\epsilon}\{N_mN_n^*\} = \widetilde{\mathrm{E}}\{N_mN_n^*\} + \epsilon_{\mathrm{corr},m,n} + d_{m,n}\epsilon_{\mathrm{PSD},n}. \tag{16}$$

Let $\Im(\cdot)$ denote the imaginary part. Exploiting Hermitian symmetry by computing $\widehat{\mathrm{E}}\{N_nN_m^*\}$ by means of (11) we obtain (17)-(18), shown at the bottom of the page, with $j = \sqrt{-1}$. From (18) we see that the real parts of the two terms in $\epsilon_{\mathrm{corr},n,m}$ in $\widetilde{\mathrm{E}_\epsilon}\{N_nN_m^*\}$ are completely removed when estimating $\mathrm{E}\{N_nN_m^*\}$ by means of (11).

To further reduce the effect of the imaginary part of ϵ_{corr} on the estimate $\widehat{\mathrm{E}}\{N_nN_m^*\}$, the smoothing parameter α can be increased during periods where the signal-to-noise ratio (SNR) is rather high.

In addition to the error reduction of $\epsilon_{\mathrm{corr},n,m}$ when computing $\widehat{\mathrm{E}}\{N_nN_m^*\}$ by means of (11), also the errors due to $\epsilon_{\mathrm{PSD},m}$ get reduced depending on the values of $|d_{n,m}|$ and $|d_{m,n}|$. Assuming that the estimation errors $\epsilon_{\mathrm{PSD},m}$ and $\epsilon_{\mathrm{PSD},n}$ are realizations of two uncorrelated zero-mean random processes $\mathcal{V}_m$ and $\mathcal{W}_n$, respectively, it follows from (18) that the variance reduction of noise PSD estimation errors becomes a factor

$$\frac{4}{1+\frac{|d_{n,m}|^2\mathrm{E}\{\epsilon_{\mathrm{PSD},m}^2\}}{|d_{m,n}|^2\mathrm{E}\{\epsilon_{\mathrm{PSD},n}^2\}}}. \tag{19}$$

B. Reduced Estimation Errors With the MVDR

Let the MVDR filter coefficients $\mathbf{w}$ be given by

$$\mathbf{w} = \frac{\mathbf{\Sigma}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Sigma}^{-1}\mathbf{d}}. \tag{20}$$

Making use of the proposed method for noise correlation matrix estimation, it can be shown that the MVDR-filter coefficients $\mathbf{w}$ become independent of the noise PSD, and therefore independent of the estimation error $\epsilon_{\mathrm{PSD},m}$. This holds in general when the noise correlation matrix is computed based on (9). When the noise correlation matrix is computed based on (11), i.e., by exploiting Hermitian symmetry in order to reduce estimation errors $\epsilon_{\mathrm{corr},m,n}$, this does not hold anymore in general, but can be shown to still hold under the assumptions of a far and free field.

We will first consider the general situation when the noise correlation matrix is computed based on (9) only, followed by the situation where the noise correlation matrix is estimated based on (11) in order to reduce the errors $\epsilon_{\mathrm{corr},m,n}$.

For notational convenience, we denote the noise PSD $\mathrm{E}\{|N_m|^2\}$ at microphone m by σ_m^2. Using the expression derived in (9), the noise correlation matrix is given by (21), shown

$$\begin{aligned}\widehat{\mathrm{E}}\{N_nN_m^*\} &= \widetilde{\mathrm{E}}\{N_nN_m^*\} + \frac{\widetilde{\mathrm{E}}\{N_nS_m^*\} - d_{n,m}\widetilde{\mathrm{E}}\{N_mS_m^*\} + \widetilde{\mathrm{E}}\{N_m^*S_n\} - d_{m,n}^*\widetilde{\mathrm{E}}\{N_n^*S_n\}}{2} \\ &\quad + \frac{d_{n,m}\epsilon_{\mathrm{PSD},m} + d_{m,n}^*\epsilon_{\mathrm{PSD},n}}{2}\end{aligned} \tag{17}$$

$$= \widetilde{\mathrm{E}}\{N_nN_m^*\} + d_{m,n}^*\Im(\widetilde{\mathrm{E}}\{N_nS_n^*\})j + d_{n,m}\Im(\widetilde{\mathrm{E}}\{N_m^*S_m\})j + \frac{d_{n,m}\epsilon_{\mathrm{PSD},m} + d_{m,n}^*\epsilon_{\mathrm{PSD},n}}{2} \tag{18}$$

at the bottom of the page. Let $\mathbf{A} \in \mathbb{C}^{M\times M}$ and $\mathbf{B} \in \mathbb{C}^{M\times M}$ be two matrices defined as

$$\mathbf{A} = \begin{pmatrix} 0 & \mathrm{E}\{P_{1,2}Y_2^*\} & \cdots & \mathrm{E}\{P_{1,M}Y_M^*\} \\ \mathrm{E}\{P_{2,1}Y_1^*\} & 0 & & \\ \vdots & & \ddots & \\ \mathrm{E}\{P_{M,1}Y_1^*\} & & & 0 \end{pmatrix} \tag{22}$$

and

$$\mathbf{B} = \begin{pmatrix} \sigma_1^2 & d_{1,2}\sigma_2^2 & \cdots & d_{1,M}\sigma_M^2 \\ d_{2,1}\sigma_1^2 & \sigma_2^2 & & \\ \vdots & & \ddots & \\ d_{M,1}\sigma_1^2 & & & \sigma_M^2 \end{pmatrix} \tag{23}$$

such that

$$\mathbf{\Sigma} = \mathbf{A} + \mathbf{B}. \tag{24}$$

Further, let $\mathbf{e}$ be defined as the element-wise inverse of the propagation vector $\mathbf{d}$ multiplied by the noise PSD per microphone, that is $\mathbf{e} = \left[\sigma_1^2/d_1, \ldots, \sigma_M^2/d_M\right]^T$. This allows us to write

$$\mathbf{B} = \mathbf{d}\mathbf{e}^T. \tag{25}$$

Using the decomposition of $\mathbf{\Sigma}$ into matrices $\mathbf{A}$ and $\mathbf{B}$ we can rewrite the numerator of (20) using the Sherman–Morrison–Woodbury formula [27] as

$$\mathbf{\Sigma}^{-1}\mathbf{d} = (\mathbf{A}+\mathbf{B})^{-1}\,\mathbf{d} \tag{26}$$

$$= \left(\mathbf{A} + \mathbf{d}\mathbf{e}^T\right)^{-1}\mathbf{d} \tag{27}$$

$$= \frac{\mathbf{A}^{-1}\mathbf{d}}{1+\mathbf{e}^T\mathbf{A}^{-1}\mathbf{d}}. \tag{28}$$

To interpret the result of (28) we can rewrite (28) as

$$\mathbf{\Sigma}^{-1}\mathbf{d} = \frac{|\mathbf{A}|\mathbf{A}^{-1}\mathbf{d}}{|\mathbf{A}|\,(1+\mathbf{e}^T\mathbf{A}^{-1}\mathbf{d})} \tag{29}$$

with $|\mathbf{A}|$ the determinant of $\mathbf{A}$. The term $|\mathbf{A}|\mathbf{A}^{-1}$ in the numerator of (29) can now be recognized as the adjugate of $\mathbf{\Sigma}$, i.e., $\mathrm{adj}\,(\mathbf{\Sigma}) = |\mathbf{A}|\mathbf{A}^{-1}$, and the term $|\mathbf{A}|\left(1+\mathbf{e}^T\mathbf{A}^{-1}\mathbf{d}\right)$ in the denominator of (29) can be recognized as the determinant of $\mathbf{\Sigma}$ obtained using the matrix determinant theorem from [28, Th. 18.1.1], i.e,

$$|\mathbf{\Sigma}| = |\mathbf{A} + \mathbf{d}\mathbf{e}^T| = |\mathbf{A}|(1+\mathbf{e}^T\mathbf{A}^{-1}\mathbf{d}). \tag{30}$$

In a similar way as the numerator of (20) was rewritten in (26)–(28) we can write the denominator as

$$\mathbf{d}^H\mathbf{\Sigma}^{-1}\mathbf{d} = \frac{\mathbf{d}^H\mathbf{A}^{-1}\mathbf{d}}{1+\mathbf{e}^T\mathbf{A}^{-1}\mathbf{d}}. \tag{31}$$

By substitution of the results from (28) and (31) into (20) we then obtain for the MVDR filter-coefficients

$$\mathbf{w} = \frac{\mathbf{A}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{A}^{-1}\mathbf{d}}. \tag{32}$$

The MVDR filter-coefficients have now become a function of matrix $\mathbf{A}$ only, and are therefore independent of the noise PSD σ_m^2. Thus, estimation of σ_m^2 using a single-channel noise PSD estimator like [10] or [15] is in general not necessary, and, noise PSD estimation errors will have no influence.

However, in the case that Hermitian symmetry is exploited by means of (11), the result obtained in (32) does not hold anymore in general. However, it still holds under the condition of a far and free field.

Let the noise correlation matrix $\mathbf{\Sigma}$, when making use of (11), be given by

$$\mathbf{\Sigma} = \mathbf{C} + \mathbf{D} \tag{33}$$

with $\mathbf{C} = (1/2)\left(\mathbf{A}+\mathbf{A}^H\right) \in \mathbb{C}^{M\times M}$ and $\mathbf{D} = (1/2)\left(\mathbf{B}+\mathbf{B}^H\right) \in \mathbb{C}^{M\times M}$. Making use of the Sherman–Morrison–Woodbury formula [27] along similar lines as in (26)–(32) we can write the MVDR filter coefficients $\mathbf{w}$ as

$$\mathbf{w} = \frac{\mathbf{C}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{C}^{-1}\mathbf{d}} + \frac{\frac{1}{2}\mathbf{C}^{-1}\mathbf{d}\mathbf{d}^H\mathbf{C}^{-1}\mathbf{e}^* - \frac{1}{2}\mathbf{C}^{-1}\mathbf{e}^*\mathbf{d}^H\mathbf{C}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{C}^{-1}\mathbf{d}} \tag{34}$$

where $\mathbf{e}^*$ indicates complex conjugation of the elements of $\mathbf{e}$. The second term in (34) is still a function of the noise PSD per channel via vector $\mathbf{e}$. However, under the condition of a far and free field, $d_{n,m} = d_{m,n}^*$ and $\sigma_n^2 = \sigma_m^2\ \forall n, m$. Under this condition, vector $\mathbf{e}^*$ can be written as $\mathbf{e}^* = \sigma^2\mathbf{d}$. Here σ^2 denotes the noise PSD, which is identical for all microphones in far and free field. Substituting $\mathbf{e}^* = \sigma^2\mathbf{d}$ into (34), we obtain a similar result as in (32), but now as a function of matrix $\mathbf{C}$, that is,

$$\mathbf{w} = \frac{\mathbf{C}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{C}^{-1}\mathbf{d}}. \tag{35}$$

This shows that the proposed noise correlation matrix estimator when based on (11) becomes also independent of the noise PSD under the assumption of a free and far field.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \mathrm{E}\{P_{1,2}Y_2^*\} + d_{1,2}\sigma_2^2 & \cdots & \mathrm{E}\{P_{1,M}Y_M^*\} + d_{1,M}\sigma_M^2 \\ \mathrm{E}\{P_{2,1}Y_1^*\} + d_{2,1}\sigma_1^2 & \sigma_2^2 & & \\ \vdots & & \ddots & \\ \mathrm{E}\{P_{M,1}Y_1^*\} + d_{M,1}\sigma_1^2 & & & \sigma_M^2 \end{pmatrix} \tag{21}$$

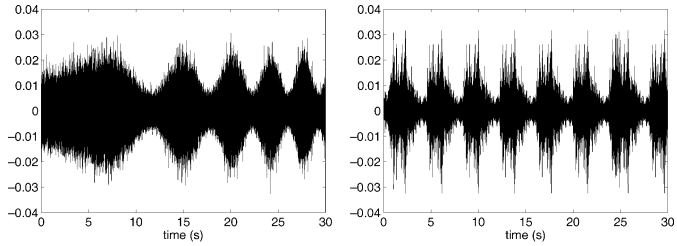

Fig. 1. Waveform of (a) modulated white Gaussian noise and (b) noise originating from a passing train.

V. Experimental Results and Discussion

For the evaluation, we apply the proposed algorithm in an MVDR beamformer. We assume that the sources are in free and far field, and thus use (35) where $\mathbf{C} = (1/2)\left(\mathbf{A} + \mathbf{A}^H\right)$, and $\mathbf{A}$ is given in (22). We thus exploit the symmetry of the noise correlation matrix, as well as the $\mathbf{A}$ matrix to reduce estimation errors as proposed in Sections IV-A and IV-B. In addition, we present in Section V-B a discussion on the influence of reverberation on the accuracy of noise correlation matrix estimation for the proposed as well as the reference methods.

A. Experimental Results

For the evaluation of the proposed and reference methods for noise correlation matrix estimation we consider an $M = 2$ microphone end-fire array with an inter-microphone distance of 1 cm. All signals are sampled using a sampling frequency of 8 kHz. Processing of the signals is done on a frame-by-frame basis using a frame-size of 256 samples, 50% overlap and a square-root Hann analysis and synthesis window. The clean speech data that is used in the experiments originates from five different female and four different male speakers from the TIMIT [29] database and has a duration of 30 seconds. As noise sources we use temporally stationary white Gaussian noise, modulated white Gaussian noise, and noise originating from a passing train. In Fig. 1(a) and (b), example time-domain waveforms of the latter two noise sources are shown, respectively. For the noise sources we use two different spatial configurations. At first a situation where the noise source moves in 30 seconds from 40° to 300°, and second a situation where the noise source alternates position between $-40°$ and 100°. Each time the noise source remains spatially stationary for one second before it switches to the other position. We refer to these two spatial configurations as configuration 1 and configuration 2, respectively.

The noise and clean speech microphone signals are generated synthetically by simulating the acoustic path from source to microphone, where a free-field situation is considered.

To compare the performance of the proposed and reference methods, we employ the estimated correlation matrices of the proposed and reference methods in an MVDR beamformer. Let $\hat{S}$ be defined as an estimate of a clean speech DFT coefficient. The MVDR beamformer can then be expressed as

$$\hat{S} = \frac{\mathbf{d}^H \boldsymbol{\Sigma}^{-1} \mathbf{Y}}{\mathbf{d}^H \boldsymbol{\Sigma}^{-1} \mathbf{d}} \tag{36}$$

where $\boldsymbol{\Sigma}$ is replaced by the noise correlation matrix estimated by the method under consideration.

For evaluation of the proposed and reference methods we use three different quality measures. At first we compute the mean squared error between the ideal MVDR beamformer response, where the noise correlation matrix is obtained using the noise only signals, and the beamformer response based on filter coefficients that are estimated using the proposed or reference methods, that is,

$$\mathrm{BR}_{\mathrm{err}} = \frac{1}{|\mathcal{Q}|} \sum_{(\phi,k,i)\in\mathcal{Q}} \|\hat{R}(\phi,k,i) - R(\phi,k,i)\|^2$$

where $\hat{R}(\phi,k,i)$ and $R(\phi,k,i)$ are the estimated and ideal beamformer responses, ϕ the direction, and $|\mathcal{Q}|$ the cardinality of the set of all (ϕ,k,i). Second, we use the segmental-SNR defined as

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{|\mathcal{P}|} \sum_{p\in\mathcal{P}} 10\log_{10}\left(\frac{\|\mathbf{s}_p\|^2}{\|\mathbf{s}_p - \hat{\mathbf{s}}_p\|^2}\right) \tag{37}$$

where $\mathbf{s}_p$ and $\hat{\mathbf{s}}_p$ denote a clean and enhanced time-domain signal frame, and $\mathcal{P}$ an index set to denote all clean speech frames with energy within 35 dB of the maximum clean speech frame energy.

Finally, to get an indication of the influence of noise correlation matrix estimation on intelligibility, we use the recently developed short-time objective intelligibility (STOI) measure [30]. The STOI measure computes an average correlation coefficient $-1 \leq d_{STOI} \leq 1$ which is monotonically related to the average intelligibility of the sentence in question.

1) Proposed Method and Reduction of Estimation Errors: In this section, we show that using the $\mathbf{C}$-matrix as proposed in (35), instead of the noise correlation matrix itself, the robustness of the MVDR beamformer is increased. The $\mathbf{C}$-matrix is given by $\mathbf{C} = (1/2)\left(\mathbf{A} + \mathbf{A}^H\right)$, and $\mathbf{A}$ is given in (22).

The proposed method for noise correlation matrix estimation depends on two expected values as indicated by (9), that are, $\mathrm{E}\{P_{n,m}(k,i)Y_m^*(k,i)\}$ and $\mathrm{E}\{|N_m|^2\}$. In addition, the diagonal elements of the cross-correlation matrix $\boldsymbol{\Sigma}$ equal the noise PSD $\mathrm{E}\{|N_m|^2\}$. Estimation of these expected values will introduce estimation errors as discussed in Section IV. As a first experiment we evaluate to which extent these estimation errors can be reduced by employing the presented modified noise correlation matrix estimator of (11) and (35), respectively.

To do so, we estimate the noise PSD $\mathrm{E}\{|N_m|^2\}$ for each microphone using the noise PSD estimator presented in [15]. The term $\mathrm{E}\{P_{n,m}(k,i)Y_m^*(k,i)\}$ in (9) is estimated by means of exponential smoothing as expressed by (10). The smoothing constant is set at $\alpha = 0.9$. Whenever the *a posteriori* SNR, i.e., $\zeta = |Y_1|^2/\widetilde{\mathrm{E}}\{|N_1|^2\}$, exceeds 7.8 dB, α is increased to $\alpha = 0.99$ for that time–frequency point to further reduce the effect of the imaginary part of ϵ_{corr} on the estimate $\widehat{\mathrm{E}}\{N_n N_m^*\}$.

We consider two different setups. At first a situation where the target source is positioned at zero degrees, i.e., endfire direction, and is degraded by white Gaussian noise in spatial configuration 1. Second, we consider the situation where the target speech source is positioned at -60 degrees and is degraded by noise originating from a passing train in spatial configuration

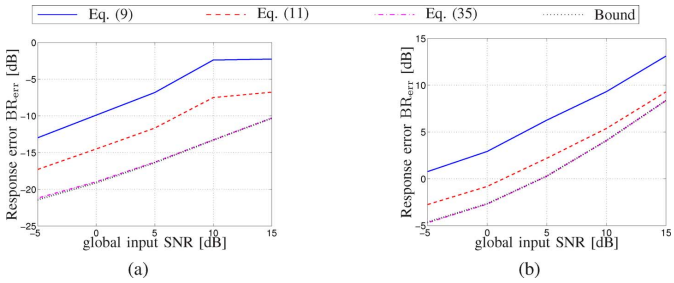

Fig. 2. (a) Results for a target source at the endfire direction (0°) and a white Gaussian noise source. The noise source moves in 30 seconds from 40° to 300°. (b) Results for a target source at −60° and a train noise source. The position of the noise source is alternating between −40° and 100°. Each time the noise source remains spatially stationary for one second before it switches to the other position.

2. In order to focus on estimation errors introduced by estimation of $\mathrm{E}\{P_{n,m}(k,i)Y_m^*(k,i)\}$ and $\mathrm{E}\{|N_m|^2\}$, we assume in this experiment that the propagation vector $\mathbf{d}$ is perfectly known such that the speech signal is completely canceled and (6) is perfectly fulfilled. We can ensure this by creating the spatial signals artificially in the short-time spectral domain. $P_{n,m}$ is then obtained using (5).

In Fig. 2, three different versions of the proposed noise correlation matrix estimator are compared in terms of the beamformer response error. In addition we compare the performance to an ideal MVDR beamformer that consists of the proposed method for noise correlation matrix estimation according to (11), but where the noise PSD $\mathrm{E}\{|N_m|^2\}$ per microphone is computed by means of exponential smoothing of the corresponding noise-only microphone signal, which is usually not available in a practical situation. This MVDR beamformer is used in order to obtain a performance bound and is referred to in Fig. 2 as *bound*. The results are given for different global, i.e., non-segmental, input SNRs.

From Fig. 2(a) and (b), we see that for the two different configurations the proposed method according to (11) improves over the original proposed method according to (9) with on average 4.7 and 3.8 dB, respectively. Computing the MVDR beamformer filter coefficients using (35) leads to a further reduction of the beamformer error response of 4.5 and 1.6 dB for the two configurations, respectively. The performance of this configuration almost coincides with the MVDR beamformer denoted by *bound* in Fig. 2. This can be explained by the fact that this version of the proposed method in combination with an MVDR beamformer overcomes all errors introduced to a wrongly estimated noise PSD when the noise PSD per microphone are identical in the far and free field case.

2) Comparison With Reference Methods: As the proposed method according to (35) has shown to effectively reduce estimation errors we continue in this section with an experimental comparison between several reference methods and this version of the proposed method for noise correlation matrix estimation.

For comparison we use three different reference methods. At first, we compare the proposed approach with the noise correlation matrix estimation based on a VAD presented in [17]. This VAD-based procedure estimates the noise correlation matrix by recursive averaging in speech pauses as indicated by a VAD. The VAD in [17] is based on a minimum statistics noise power spectral density estimate [10], where speech absence is decided, when the recursively smoothed *a posteriori* SNR is smaller than a threshold that we set at 1.2. Choosing a higher threshold might lead to a less conservative VAD with somewhat faster noise PSD update, but will also make it more likely to introduce speech leakage in the noise correlation matrix estimate. The obtained estimate of the noise correlation matrix is used in the MVDR expression in (36).

Second, we also compare the results to an MVDR beamformer, where instead of the noise correlation matrix the noisy correlation matrix is used, i.e., $\mathrm{E}\{\mathbf{Y}\mathbf{Y}^H\}$. Under ideal conditions, i.e., when $\mathbf{d}$ is known and speech and noise are uncorrelated, it can be shown using the Sherman–Morrison–Woodbury formula [27] that the MVDR based on $\mathrm{E}\{\mathbf{Y}\mathbf{Y}^H\}$ yields the same result as the MVDR based on $\mathrm{E}\{\mathbf{N}\mathbf{N}^H\}$, see, e.g., [16], [26]. In order to estimate $\mathrm{E}\{\mathbf{Y}\mathbf{Y}^H\}$, we recursively smooth realizations of $\mathbf{Y}\mathbf{Y}^H$ over time with a smoothing constant $\alpha = 0.9$.

Third, we make a comparison to the GSC [25] as summarized in [26, Table 47.2]. The GSC does not estimate the noise correlation matrix, but estimates the filter coefficients that minimize the output power constraining the target to be undistorted. The GSC is implemented based on an adaptive LMS algorithm. Comparing the GSC to the aforementioned MVDR-based methods is justified by the fact that the analytic expression of the GSC can be shown to equal the MVDR beamformer under the conditions used in this work [31].

In Figs. 3–5, comparisons are shown with the three aforementioned reference algorithms in terms of the beamformer response error, improvement in segmental SNR and STOI, respectively. The target source is in all experiments positioned at an angle $\alpha = -60$ degrees. In this section, we do not create the spatial signals artificially in the short-time spectral domain as we did in Section V-A1, but construct them by filtering the time-domain signals with the corresponding impulse response.

In order to compute the noise reference $P_{n,m}$ in (5) for element (n,m) of the noise correlation matrix the quantity $d_{n,m}$ is needed. Notice that $d_{n,m}$ is related to the impulse response that is used to construct the spatial signals. In this section, we do *not* assume we know $d_{n,m}$ exactly, i.e., the Fourier transform of the impulse response used to generate the spatial signals, as we did in Section V-A1, but only make use of the fact that we know the angle of the target source with respect to the array, i.e., −60 degrees in this case. We then construct the propagation vector $\mathbf{d}$ using the following delay-only model:

$$\mathbf{d} = \left[1, e^{-j2\pi k \cos(\alpha) d/Lc}\right]^T \tag{38}$$

with c the speed of sound. Note that, in contrast to Section V-A1, employing this $\mathbf{d}$ will not result in a perfect cancellation of the speech signal in (7), because we perform frame-by-frame processing and the impulse response can be longer than a time-frame.

From Figs. 3–5 we see that the proposed method improves all three performance measures for all three noise sources and both spatial noise configurations. The exact improvement depends on input SNR, the type of noise source, spatial configuration, and reference method. In terms of beamformer response

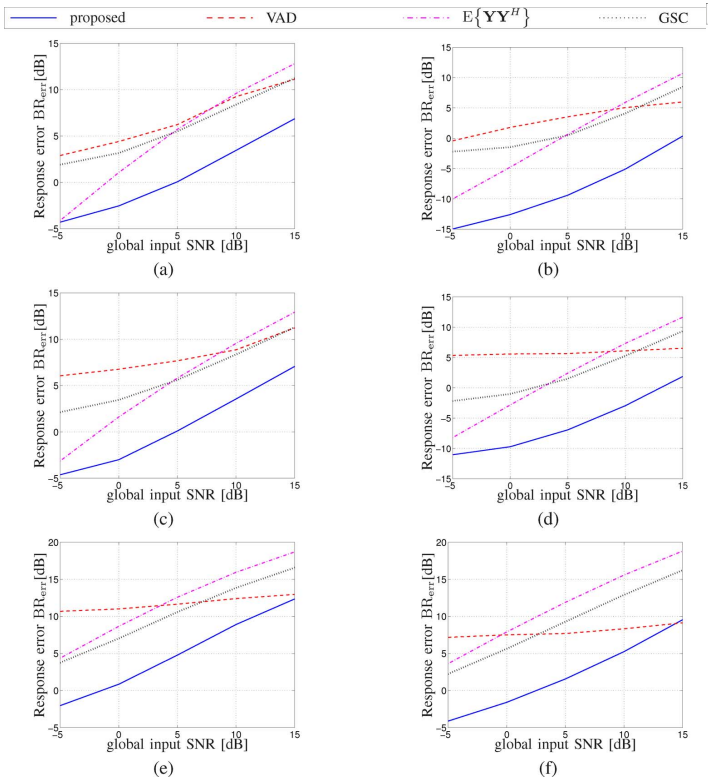

Fig. 3. Comparison in terms of beamformer response error for a target source at -60° and (a) temporally stationary white Gaussian noise and spatial configuration 1, (b) temporally stationary white Gaussian noise and spatial configuration 2, (c) modulated white Gaussian noise and spatial configuration 1, (d) modulated white Gaussian noise and spatial configuration 2, (e) passing train noise and spatial configuration 1, and (f) passing train noise and spatial configuration 2.

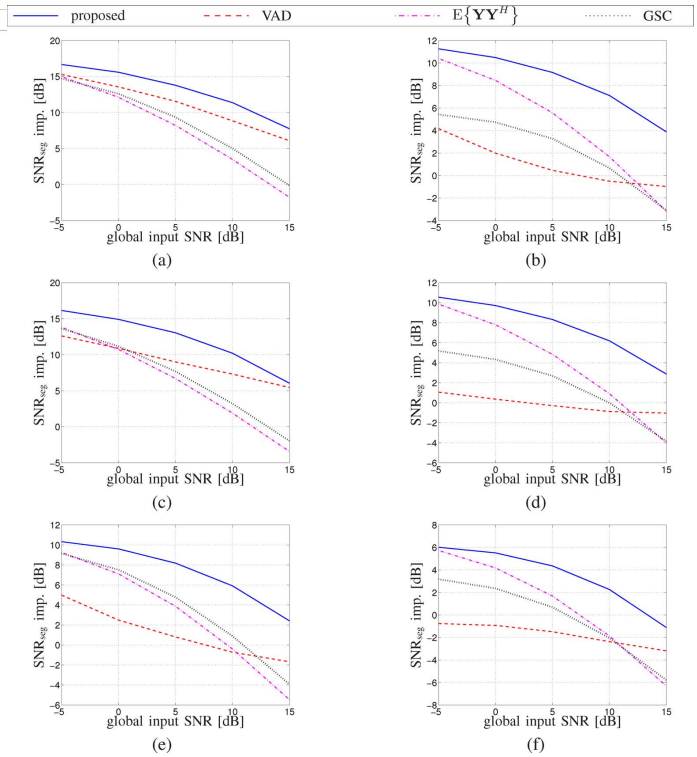

Fig. 4. Comparison in terms of segmental SNR improvement for a target source at -60° and (a) temporally stationary white Gaussian noise and spatial configuration 1, (b) temporally stationary white Gaussian noise and spatial configuration 2, (c) modulated white Gaussian noise and spatial configuration 1, (d) modulated white Gaussian noise and spatial configuration 2, (e) passing train noise and spatial configuration 1, and (f) passing train noise and spatial configuration 2.

error the improvement of the proposed method over the reference methods generally varies between 4 and 11 dB. In terms of segmental SNR, the improvement is in the order of 2 to 7 dB. In terms of STOI intelligibility improvements up to 0.1 can be observed.

B. Noise Correlation Matrix Estimation and Reverberation

Most methods for noise correlation matrix estimation, including the proposed method, assume speech and noise to be uncorrelated. Clearly, in a scenario with reverberation, speech is not completely uncorrelated with the reverberation. This justifies a further analysis of its consequences.

The MVDR beamformer depends on the propagation vector $\mathbf{d}$ and the correlation matrix $\mathbf{\Sigma}$. In free and far field, $\mathbf{d}$ can be perfectly modeled as a vector with fixed delay elements only, say τ, i.e.,

$$\mathbf{d} = [1, e^{-j2\pi k\tau/L}, \ldots, e^{-j2\pi k(M-1)\tau/L}]^T. \qquad (39)$$

For situations in an enclosure with reverberation, i.e., in a room scenario, this model is too simplistic. In that case, ideally, $\mathbf{d}$ should be replaced by the acoustic transfer function (ATF), say $\mathbf{a}$. Assuming the time frames to be sufficiently long, the microphones observe realizations of the following random vector process, that is,

$$\mathbf{Y} = S\mathbf{a} + \mathbf{N}. \qquad (40)$$

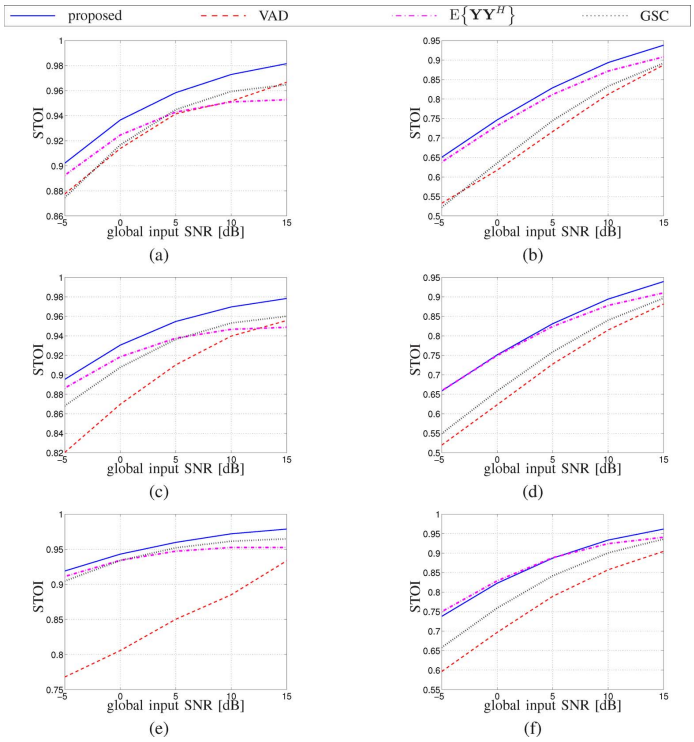

Fig. 5. Comparison in terms of STOI for a target source at -60° and (a) temporally stationary white Gaussian noise and spatial configuration 1, (b) temporally stationary white Gaussian noise and spatial configuration 2, (c) modulated white Gaussian noise and spatial configuration 1, (d) modulated white Gaussian noise and spatial configuration 2, (e) passing train noise and spatial configuration 1, and (f) passing train noise and spatial configuration 2.

Assuming $\mathbf{a}$ is known, and applying the MVDR filter coefficients $\mathbf{W} = \mathbf{\Sigma}^{-1}\mathbf{a}/\mathbf{a}^H\mathbf{\Sigma}^{-1}\mathbf{a}$ to $\mathbf{Y}$ then obviously leads to

$$\mathbf{W}^H\mathbf{Y} = S + \frac{\mathbf{a}^H\mathbf{\Sigma}^{-1}\mathbf{N}}{\mathbf{a}^H\mathbf{\Sigma}^{-1}\mathbf{a}}. \tag{41}$$

From (41) we see that due to knowing the ATF, the reverberation is canceled completely and the filter output consists of the undistorted target clean speech DFT and additive processed noise. However, knowing the ATF implies the room impulse response (RIR) to be known. Since the RIR is generally very long and difficult to estimate, simpler models like the delay-only model in (39) are often used and consequently, reverberation is not completely canceled.

Let h_m denote the RIR with respect to the position of microphone m and let $r = (i-1)R + 1$ denote the starting time-sample of the window from which DFT coefficient $S(k,r)$ originates. Assuming that the RIR does not change within a time-frame, say, it stays constant for some time, it is possible to write the DFT coefficient of the reverberant speech for microphone m as [32]

$$S_{m,\mathrm{rev}}(k,r) = \sum_{l=0}^{\infty} h_m(l)S(k,r-l). \tag{42}$$

Based on (42) it is possible to write the DFT coefficient of the reverberant speech in terms of a component consisting of the direct path only, and a component consisting of all remaining reflections, that are

$$S_{m,\mathrm{dir}}(k,r) = h_m(L-1)S(k,r-L+1) \tag{43}$$

where $h_m(L-1)$ models the response due to the direct path and

$$S_{m,\mathrm{refl}}(k,r) = \sum_{l=L}^{\infty} h_m(l)S(k,r-l) \tag{44}$$

respectively. The direct path component can be rewritten as $S_{m,\mathrm{dir}}(k,r) = d(m)S(k,r)$, where $d(m) \in \mathbb{C}$ is the mth element of the propagation vector as defined in Section II, which takes care of the delay and damping that the direct path component undergoes.

Neglecting time and frequency-bin indices for notational convenience, we can write the DFT coefficients containing the direct path component and the DFT coefficients containing reflections only for the M microphones using vector notation as $\mathbf{S}_{\mathrm{dir}} = [S_{1,\mathrm{dir}},\ldots,S_{M,\mathrm{dir}}]^T = S\mathbf{d}$ and $\mathbf{S}_{\mathrm{refl}} = [S_{1,\mathrm{refl}},\ldots,S_{M,\mathrm{refl}}]^T$, respectively.

Altogether, $\mathbf{Y} \in \mathbb{C}^{\mathbb{M}}$ can be written as

$$\mathbf{Y} = S\mathbf{d} + \mathbf{S}_{\mathrm{refl}} + \mathbf{N}. \tag{45}$$

If we assume that $\mathbf{d}$ is known instead of the ATF $\mathbf{a}$ and apply an MVDR to this noisy random vector process we obtain

$$\mathbf{W}^H\mathbf{Y} = S + \frac{\mathbf{d}^H\mathbf{\Sigma}^{-1}\mathbf{S}_{\mathrm{refl}}}{\mathbf{d}^H\mathbf{\Sigma}^{-1}\mathbf{d}} + \frac{\mathbf{d}^H\mathbf{\Sigma}^{-1}\mathbf{N}}{\mathbf{d}^H\mathbf{\Sigma}^{-1}\mathbf{d}}. \tag{46}$$

From (46) we see that opposed to (41) the speech DFT coefficients originating from the direct path are maintained undistorted, while now both the reduction of additive noise process $\mathbf{N}$, as well as reduction of the reverberation depend on $\mathbf{\Sigma}$. Ideally, the MVDR is distortionless in the direction of $\mathbf{d}$ and minimizes the variance of the filter output as much as possible. For the MVDR to be able to do so, $\mathbf{\Sigma}$ is now given by $\mathbf{\Sigma} = \mathrm{E}\{\mathbf{N}_{\mathrm{tot}}\mathbf{N}_{\mathrm{tot}}^H\} \in \mathbb{C}^{M\times M}$, where $\mathbf{N}_{\mathrm{tot}} = \mathbf{N} + \mathbf{S}_{\mathrm{refl}}$ contains all disturbances, i.e., the additive noise as well as all reflections.

Most methods for noise correlation matrix estimation, as well as the generalized sidelobe canceller, do assume that the disturbances are uncorrelated with the speech DFT coefficients. Clearly, this does not hold in the case of reverberant speech, as S_{refl} is correlated with S. This raises the question to which extent the proposed, as well as existing methods for noise correlation matrix estimation, are robust with respect to reverberation. Before presenting experimental results on noise correlation matrix estimation for a situation with room reverberation, we discuss how the proposed method and reference methods can handle reverberant speech.

1) VAD-Based Noise Correlation Matrix Estimation: VAD-based noise correlation matrix can only capture the characteristics of the noise field in speech absence. This means, on the one hand, that they cannot follow quickly changing noise fields during speech presence, i.e., the contribution of $\mathbf{N}$ to $\mathbf{\Sigma}$. On the other hand, the contribution of the reverberation $\mathbf{S}_{\mathrm{refl}}$ to $\mathbf{\Sigma}$ can only be captured at the end of speech activity in each frequency bin. However, for moderately sized rooms, the reverberation tail at the end of an utterance may often be too short to measure the contribution of $\mathbf{S}_{\mathrm{refl}}$ to the correlation matrix $\mathbf{\Sigma}$ using recursive smoothing. This problem gets even more crucial if the VAD is prone to false decisions and indicates speech absence only with a certain delay.

2) MVDR Based on the Noisy Correlation Matrix: Using the Sherman–Morrison–Woodbury formula [27], it can be shown that under certain conditions, the MVDR filter coefficients based on the noisy correlation matrix equal the MVDR filter coefficients based on the noise correlation matrix [26]

$$\mathbf{W} = \frac{\mathrm{E}\{\mathbf{Y}\mathbf{Y}^H\}^{-1}\mathbf{d}}{\mathbf{d}^H\mathrm{E}\{\mathbf{Y}\mathbf{Y}^H\}^{-1}\mathbf{d}} = \frac{\mathrm{E}\{\mathbf{N}_{\mathrm{tot}}\mathbf{N}_{\mathrm{tot}}^H\}^{-1}\mathbf{d}}{\mathbf{d}^H\mathrm{E}\{\mathbf{N}_{\mathrm{tot}}\mathbf{N}_{\mathrm{tot}}^H\}^{-1}\mathbf{d}}. \tag{47}$$

However, this relation only holds when $\mathbf{N}_{\mathrm{tot}}$ and $\mathbf{S}$ are uncorrelated. For the reverberant situation where the exact acoustic transfer function is unknown we obtain using the Sherman–Morrison–Woodbury formula

$$\begin{aligned}\mathbf{W} &= \frac{\mathrm{E}\{\mathbf{Y}\mathbf{Y}^H\}^{-1}\mathbf{d}}{\mathbf{d}^H\mathrm{E}\{\mathbf{Y}\mathbf{Y}^H\}^{-1}\mathbf{d}} \\ &= \frac{\mathrm{E}\{\mathbf{N}_{\mathrm{tot}}\mathbf{N}_{\mathrm{tot}}^H + \mathbf{S}_{\mathrm{refl}}\mathbf{S}^H + \mathbf{S}\mathbf{S}_{\mathrm{refl}}^H\}^{-1}\mathbf{d}}{\mathbf{d}^H\mathrm{E}\{\mathbf{N}_{\mathrm{tot}}\mathbf{N}_{\mathrm{tot}}^H + \mathbf{S}_{\mathrm{refl}}\mathbf{S}^H + \mathbf{S}\mathbf{S}_{\mathrm{refl}}^H\}^{-1}\mathbf{d}}.\end{aligned}$$

This shows that the relation in (47) does not hold when $\mathbf{N}_{\mathrm{tot}}$ is correlated with $\mathbf{S}$ and that the MVDR based on the noisy correlation matrix is also not robust for reverberation when the ATF $\mathbf{a}$ is unknown.

3) Generalized Sidelobe Canceller: With the GSC, the optimal filter coefficients are estimated using an adaptive LMS algorithm. This adaptive algorithm estimates the optimal filter coefficients, such that the correlation between a set of noise references and a speech reference is minimized. Presence of

reverberation in the noise reference might therefore lead to elimination of the target speech in the output of the GSC. To avoid this, an often used procedure is to employ VAD and update the filter coefficients only when speech is absent, see [33] for an analysis and references on how to limit speech distortion resulting from speech leakage in the noise references. However, using a VAD in combination with a GSC will lead to similar problems as when using a VAD to directly estimate the noise correlation matrix.

4) Proposed Method: An important aspect in the derivation of the proposed method is that for each microphone pair (n, m) a noise reference denoted by $P_{n,m}$ can be obtained that is uncorrelated to the speech DFT coefficient at microphone number m. This is expressed by (8) and is a similar assumption that underlies the derivation of the GSC. For the case that these noise references are constructed based on knowledge of $\mathbf{d}$ and not the ATF $\mathbf{a}$, these noise references will contain speech reverberation components that are correlated to the target speech DFT coefficient. Evaluation of (9) then gives rise to additional error terms in the estimate of the off-diagonal elements. Assuming speech and noise are uncorrelated and assuming a perfect noise PSD we obtain

$$\begin{aligned}&\widehat{\mathrm{E}}\{N_{n,\mathrm{tot}}N_{m,\mathrm{tot}}^{*}\}\\&\quad=\widetilde{\mathrm{E}}\{N_{n,\mathrm{tot}}N_{m,\mathrm{tot}}^{*}\}-d_{n,m}\widetilde{\mathrm{E}}\{|S_{m,\mathrm{refl}}|^{2}\}\\&\qquad+\widetilde{\mathrm{E}}\{S_{n,\mathrm{refl}}S_{m,\mathrm{dir}}^{*}\}-d_{n,m}\widetilde{\mathrm{E}}\{S_{m,\mathrm{refl}}S_{m,\mathrm{dir}}^{*}\}.\end{aligned}\quad(48)$$

Note that the error term $\widetilde{\mathrm{E}}\{|S_{m,\mathrm{refl}}|^{2}\}$ can be estimated and compensated by assuming a statistical model for the room impulse response as done in single channel speech dereverberation [34]. However, for simplicity we do not compensate for this error-term in our experiments.

5) Experimental Results in Reverberant Situation: To investigate the robustness of the proposed and reference methods with respect to reverberance we created two reverberant scenarios. The room dimensions are in both scenarios $4\times4\times4$ meter with an $M=2$ microphone array in the center of the room. The target source is positioned at one meter distance from the array at $\alpha=0$ degrees, i.e., the endfire direction. The microphones of the array have an inter-microphone distance of 1 cm. The reverberant microphone signals are created by convolving the free-field microphone signals with the room impulse responses. Subsequently, the reverberant speech signals are degraded by modulated white Gaussian noise in spatial configuration 2. Two different room impulse responses were used, one with $T_{60}=200$ ms and one with $T_{60}=300$ ms. In the simulation results we assume that the ATF is unknown, and use instead the propagation vector $\mathbf{d}$ based on a delay-only model for $M=2$ in (39).

In Fig. 6 the comparison between the proposed and reference methods for these two different reverberant scenarios is shown in terms of the beamformer response error and improvement in segmental SNR. It is clearly visible that the performance differences between all methods have decreased compared to the nonreverberant scenario. Also clearly visible is the fact that all methods have smaller improvement in terms of segmental SNR than in the non-reverberant case in Figs. 3(d) and 4(d). As already argued in Sections V-B1–V-B4, we can conclude that all methods in this comparison are sensitive for reverberant speech. For the spatially and temporally nonstationary noise field used in this example the proposed approach generally still improves performance, albeit to a much smaller degree than in the non-reverberant case.

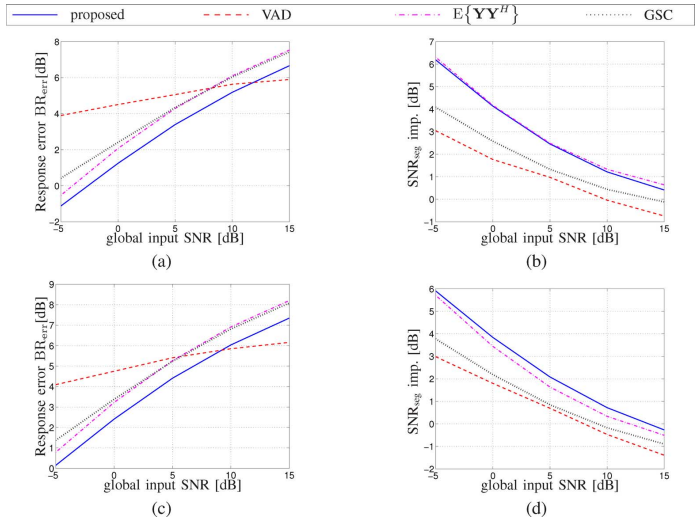

Fig. 6. Comparison in terms of (a) beamformer response error and room impulse response with $T_{60}=300$ ms and (b) segmental SNR improvement and room impulse response with $T_{60}=300$ ms (c) beamformer response error and room impulse response with $T_{60}=200$ ms and (d) segmental SNR improvement and room impulse response with $T_{60}=200$ ms.

VI. Conclusion

In this paper, we have presented and analyzed an estimator of the noise correlation matrix which is needed in many multi-channel noise reduction algorithms, e.g., the minimum variance distortionless response (MVDR) beamformer or the multi-channel Wiener filter. While in the literature it is usually proposed to estimate the noise correlation matrix during speech absence, the proposed approach can update the noise correlation matrix also during speech presence. Thus, changing noise fields can be tracked more accurately. For the proposed algorithm, the diagonal elements of the cross-correlation matrix are estimated using single-channel noise PSD estimators. The off-diagonal elements are estimated by measuring the correlation between the noisy input signal and a noise reference which can be obtained, e.g., by steering a null towards the target source.

We have shown how estimation errors can be reduced. In addition, we have shown that when the proposed noise correlation matrix estimator is applied in combination with an MVDR beamformer under far and free field conditions, that the filter coefficients become independent of the noise PSD.

We have employed the estimated noise correlation matrix in an MVDR beamformer and have evaluated its performance in terms of the beamformer response error, the segmental signal-to-noise ratio improvement, and an instrumental measure for speech intelligibility. We have shown that the proposed algorithm improves over algorithms such as the GSC, an MVDR with VAD-based noise estimation, and an MVDR that employs the correlation matrix of the noisy signal instead of the noise signal.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments that helped to improve the presentation of this work.

REFERENCES

[1] P. Loizou, *Speech Enhancement Theory and Practice*. Boca Raton, FL: CRC, 2007.
[2] K. Eneman *et al.*, "Evaluation of signal enhancement algorithms for hearing instruments," in *Proc. Eur. Signal Process. Conf.*, 2008.
[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[4] 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.
[5] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, 2005.
[6] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.
[7] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.
[8] , M. Brstein and D. Ward, Eds., *Microphone Arrays*. New York: Springer, 2001.
[9] S. Doclo, "Multi-microphone noise reduction and dereverberation techniques for speech applications," Ph.D. dissertation, Katholieke Univ. Leuven, , 2003.
[10] 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.
[11] 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.
[12] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.
[13] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1112–1123, Aug. 2008.
[14] R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "Low complexity DFT-domain noise PSD tracking using high-resolution periodograms," *Eurasip J. Adv. Signal Process.*, pp. 1–15, 2009.
[15] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Process.*, 2010, pp. 4266–4269.
[16] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Amer.*, vol. 54, no. 3, pp. 771–785, 1973.
[17] X. Zhang and Y. Jia, "A soft decision based noise cross power spectral density estimation for two-microphone speech enhancement systems," in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Process.*, 2005, vol. 1, pp. 813–816.
[18] R. L. Bouquin-Jeannès, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 484–487, Sep. 1997.
[19] M. Rahmani, A. Akbari, B. Ayad, M. Mazoochi, and M. S. Moin, "A modified coherence based method for dual microphone speech enhancement," in *Proc. IEEE Int. Conf. Signal Process. Commun.*, Nov. 2007, pp. 225–228.
[20] J. Freudenberger, S. Stenzel, and B. Venditti, "A noise PSD and cross-PSD estimation method for two-microphone speech enhancement systems," in *Proc. IEEE Workshop Statist. Signal Process.*, Sep. 2009, pp. 709–712.
[21] M. Rahmani, A. A. B. Ayad, and B. Lithgow, "Noise cross PSD estimation using phase information in diffuse noise field," *Signal Process.*, vol. 89, pp. 703–709, 2009.
[22] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.
[23] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–14, 2006.
[24] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.
[25] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.
[26] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*. Berlin, Heidelberg, Germany: Springer-Verlag, 2008.
[27] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.
[28] D. A. Harville, *Matrix Algebra From a Statistician's Perspective*. New York: Springer-Verlag, 1997.
[29] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," National Institute of Standards and Technology (NIST), 1988.
[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Process.*, 2010, pp. 4214–4217.
[31] C. W. Jim, "A comparison of two LMS constrained optimal array structures," *Proc. IEEE*, vol. 65, no. 12, pp. 1730–1731, Dec. 1977.
[32] J. S. Erkelens and R. Heusdens, "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1746–1765, Sep. 2010.
[33] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Commun.*, vol. 2007, pp. 636–656, 2007.
[34] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica*, vol. 87, pp. 359–366, 2001.

Richard C. Hendriks received the B.Sc., M.Sc., (*cum laude*), and Ph.D. (*cum laude*) degrees in electrical engineering from the Delft University of Technology (DUT), Delft, The Netherlands, in 2001, 2003, and 2008, respectively.

From 2003 to 2007, he was a Ph.D. Researcher at DUT. From 2007 to 2010, he was a Postdoctoral Researcher at DUT. Since 2010, he has been an Assistant Professor in the Multimedia Signal Processing Group, Faculty of Electrical Engineering, Mathematics, and Computer Science DUT. In the autumn of 2005, he was a Visiting Researcher at the Institute of Communication Acoustics, Ruhr-University Bochum, Bochum, Germany. From March 2008 to March 2009, he was a Visiting Researcher at Oticon A/S, Copenhagen, Denmark. His main research interests are digital speech and audio processing, including single-channel and multi-channel acoustical noise reduction, speech enhancement, and intelligibility improvement.

Timo Gerkmann (M'10) studied electrical engineering at the Universität Bremen, Bremen, Germany, and the Ruhr-Universität Bochum, Bochum, Germany. He received the Dipl.-Ing. degree and the Dr.-Ing. degree from the Ruhr-Universität Bochum, Bochum, Germany, in 2004 and 2010, respectively.

From January 2005 to July 2005, he visited Siemens CR, Princeton, NJ. Currently, he is a Postdoctoral Researcher at the Sound and Image Processing Lab, Royal Institute of Technology (KTH), Stockholm, Sweden. His main research interests are in digital speech and audio processing, including single- and multi-channel speech enhancement.