Speech Communication 114 (2019) 49–59

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Nonlinear Kronecker product filtering for multichannel noise reduction

Gal Itzhak[a,*], Jacob Benesty[b], Israel Cohen[a]

[a] Andrew and Erna Viterby Faculty of Electrical Engineering, Technion – Israel Institute of Technology, Technion City, Haifa 3200003, Israel
[b] INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Montreal QC H5A 1K6, Canada

ARTICLE INFO

Keywords:
Noise reduction
Speech enhancement
Microphone arrays
Multichannel
Frequency-domain filtering
Optimal filters
Nonlinear processing

ABSTRACT

Multichannel noise reduction in the frequency domain is a fundamental problem in the areas of speech processing and speech recognition. In this paper, we address this problem and propose an alternative approach to retrieve a speech signal out of microphone array noisy observations. We focus on the spectral amplitude of the speech signal and assume that the spectral phase is less significant. The estimate of the spectral amplitude squared, that is the spectral power, is obtained by applying a complex linear filter to a modified version of the observations vector. This modified version is obtained as a Kronecker product of the complex conjugate of the observations vector and the original observations vector. The complex speech signal estimate is obtained by multiplying the spectral amplitude estimate with a complex exponential whose phase may be extracted from the minimum variance distortionless response beamformer. We present a modified optimization criterion according to which the proposed filters may be derived, and compare their performances to conventional multichannel noise reduction filters. We show that the new approach is preferable, in particular when the input signal-to-noise ratio (SNR) is low or the number of sensors is small.

1. Introduction

Many modern applications in a wide variety of areas, from speech recognition and communications to speaker identification and human-to-machine systems, are required to operate in noisy environments. Noise fields, in many cases, significantly deteriorate the speech signal quality, thus damaging the functionality of communication and speech recognition systems. The problem of enhancing speech, or reducing noise, has attracted many researchers over the years, who suggested numerous schemes and algorithms in multiple processing domains.

With the growing demand for robust noise reduction capabilities, multichannel noise reduction (MCNR) methods are often employed in order to exploit spatial information. This additional information allows, in many cases, to attain a considerable amount of noise reduction while preserving the desired signal distortionless (Benesty et al., 2009a). Often referred to as beamformers, MCNR methods may be designed and implemented in various domains.

Time-domain beamformers are the easiest to implement, as the filters are applied directly to the noisy observations, typically generating a single speech sample estimate at a time. It is also possible to estimate a vector of successive speech samples simultaneously. However, such beamformers tend to suffer from high computational complexity (Benesty and Chen, 2011; Benesty et al., 2017; Buchris et al., 2019).

Transform-domain beamformers, as in Chen et al. (2003) and Benesty et al. (2008, 2007, 2009b), are typically formulated on a frame basis. That is, the noisy signal is transformed into another domain, the optimal filter is derived and applied in the transformed domain, and subsequently the filtered observations are transformed back to the time domain using an appropriate inverse transform. Time-domain beamformers may be derived in a transform domain by appropriately adjusting the criterion for optimization (Benesty et al., 2012). Choosing an appropriate transform may be beneficial in terms of the quality of the enhanced speech and the computational complexity of the noise reduction filter application. The generalized Karhunen-Loève (KL) domain and the frequency domain, for instance, constitute common choices for these particular reasons.

The generalized KL domain is obtained by projecting noisy observations onto an orthonormal basis of eigenvectors of a speech signal correlation matrix. This projection results in uncorrelated analysis coefficients which are independently processed (Benesty et al., 2009b; Ephraim and Trees, 1995; Lacouture-Parodi et al., 2014). Moreover, it was shown (Ephraim and Trees, 1995) that by taking a signal subspace approach, KL-domain processing may separate a speech signal-plus-noise subspace from noise-only subspace. With this approach, the latter is employed to estimate the noise-only statistics and then applied on the former to form an estimate of the clean speech. Assuming the

* Corresponding author.
E-mail address: galitz@campus.technion.ac.il (G. Itzhak).

https://doi.org/10.1016/j.specom.2019.10.001
Received 21 February 2019; Received in revised form 21 August 2019; Accepted 2 October 2019
Available online 3 October 2019

estimate of the clean speech correlation matrix is accurate, it is guaranteed that no aliasing problems emerge (Benesty et al., 2009b). Nonetheless, frequency-domain beamformers (Dmochowski and Benesty, 2010; Gannot and Cohen, 2008; Tavakoli et al., 2016), which are typically implemented in the short-time Fourier transform (STFT) domain, are considered easier to employ. That is, unlike the KL domain speech signal-dependent eigenvectors, the Fourier basis functions are global. Consequently, frequency-domain beamforming does not require the overhead of estimating and diagonalizing the speech correlation matrix, yet the frequency coefficients remain uncorrelated provided that the analysis window is long enough.

Originally proposed in Capon (1969), Capon's minimum variance distortionless response (MVDR) beamformer has been investigated in theoretical studies from a variety of aspects (Souden et al., 2010). The linear MVDR, which operates directly on a vector of transformed noisy observations, is shown to be optimal in terms of the residual noise energy, under the restriction of zero desired signal distortion. Moreover, it has inspired numerous variations, e.g., the minimum power distortionless response (MPDR) (Van Trees, 2004), which avoids the estimation of the noise-only correlation matrix. This sort of flexibility, combined with its proved noise-reduction capabilities and easy-to-analyze linear nature, has made the MVDR beamformer very common in real-world applications.

Another important variation, which may also be seen as a generalization of the MVDR, is the linearly constrained minimum variance (LCMV) beamformer (Griffiths and Jim, 1982). While maintaining the desired signal distortionless in the same spatial manner as the MVDR does, the LCMV provides a convenient scheme to cope with spatial interferences by placing nulls in their respective directions. Additionally, it is optimal in the sense of the residual noise energy minimization. However, its noise reduction performance is known to be inferior in comparison to the MVDR, unless the number of sensors is significantly greater than the number of interfered directions. This limitation results from the linear nature of the two beamformers.

For classical speech analysis purposes, higher-order statistics were shown to be informative, though typically in the context of single-channel noise reduction (SCNR). In Moreno and Fonollosa (1992), the third-order statistics of noisy speech was used to determine its pitch, suggesting that unlike speech, most common noises exhibit a nearly-zero skewness. In Nemer et al. (2001), a voice activity detector (VAD) was presented assuming an underlying zero-phase harmonic representation of speech. Closed-form expressions of the third and fourth-order cumulants were derived and combined with second-order measures to yield what was demonstrated to be a more robust VAD. In Nemer et al. (2002), it was suggested to take advantage of fourth-order cumulants to estimate some widely-used parameters, such as the SNR, the speech autocorrelation and the probability of speech presence. These estimates may then feed any speech enhancement method in which they are required as input parameters.

In this paper, we present a Kronecker product (KP) approach for MCNR in the frequency domain. We propose to take advantage of higher-order statistics and apply a complex linear filter to a modified observation signal vector. The modified vector is constructed from the original noisy observations and its elements may be interpreted as the instantaneous correlation coefficients. The filtering product of the KP approach is an estimate of the desired signal spectral power, which is considered more important than the spectral phase in many applications, such as speech enhancement. The spectral phase may be extracted from a conventional beamformer, e.g., the linear MVDR. We propose a modified optimization criterion for deriving KP filters and present the KP-MVDR and the KP-LCMV. We demonstrate that when the array size is small or the SNR is low the KP filters outperform the conventional ones, provided that temporal smoothing is employed to properly estimate the correlation matrix of the modified observation signal vector.

The rest of the paper is organized as follows. In Section 2, we formulate the MCNR problem in the frequency domain. In Section 3, we review the conventional filtering approach and derive the MVDR and LCMV beamformers, which are used as benchmarks for comparison. In Section 4, we formulate the KP filtering approach and derive the KP filters based on a new optimization criterion. In Section 5, we compare the conventional and KP filters analytically through a stationary toy example. Finally, in Section 6, we evaluate the performances of the two approaches by a set of nonstationary speech signals simulations in anechoic and reverberant environments.

2. Signal model and problem formulation

Consider an array consisting of M omnidirectional microphones. The received signals at the frequency index f are expressed as (Benesty et al., 2008; 2012; Bai et al., 2014)

$$\begin{aligned} Y_m(f) &= G_m(f)S(f) + V_m(f) \\ &= X_m(f) + V_m(f),\ m = 1, 2, \dots, M, \end{aligned} \tag{1}$$

where $Y_m(f)$ is the mth microphone signal, $S(f)$ is the unknown speech source, $G_m(f)$ is the acoustic room transfer function from the position of $S(f)$ to the mth microphone, $X_m(f) = G_m(f)S(f)$ is the zero-mean speech signal which takes into account the acoustic room transfer function, and $V_m(f)$ is the zero-mean additive noise. It is assumed that $X_i(f)$ and $V_j(f)$ are uncorrelated, i.e., $E\left[X_i(f)V_j^*(f)\right] = 0,\ \forall i, j = 1, 2, \dots, M$, where $E[\cdot]$ denotes mathematical expectation and the superscript * is the complex-conjugate operator. By definition, the terms $X_m(f),\ m = 1, 2, \dots, M$ are correlated while the other terms $V_m(f),\ m = 1, 2, \dots, M$, depending on the nature of the noise, may only be partially correlated. We consider the first microphone as the reference; then, the objective of multichannel noise reduction in the frequency domain is to estimate the desired signal, $X_1(f)$, from the M observations $Y_m(f),\ m = 1, 2, \dots, M$, in the best possible way.

It is more convenient to write the M frequency-domain microphone signals in a vector notation:

$$\begin{aligned} \mathbf{y}(f) &= \mathbf{g}(f)S(f) + \mathbf{v}(f) \\ &= \mathbf{x}(f) + \mathbf{v}(f) \\ &= \mathbf{d}(f)X_1(f) + \mathbf{v}(f), \end{aligned} \tag{2}$$

where

$$\begin{aligned} \mathbf{y}(f) &= \begin{bmatrix} Y_1(f) & Y_2(f) & \cdots & Y_M(f) \end{bmatrix}^T, \\ \mathbf{x}(f) &= \begin{bmatrix} X_1(f) & X_2(f) & \cdots & X_M(f) \end{bmatrix}^T \\ &= S(f)\mathbf{g}(f), \\ \mathbf{g}(f) &= \begin{bmatrix} G_1(f) & G_2(f) & \cdots & G_M(f) \end{bmatrix}^T, \\ \mathbf{v}(f) &= \begin{bmatrix} V_1(f) & V_2(f) & \cdots & V_M(f) \end{bmatrix}^T, \end{aligned}$$

the superscript ^{T} is the transpose operator, and

$$\mathbf{d}(f) = \begin{bmatrix} 1 & \frac{G_2(f)}{G_1(f)} & \cdots & \frac{G_M(f)}{G_1(f)} \end{bmatrix}^T = \frac{\mathbf{g}(f)}{G_1(f)}. \tag{3}$$

Expression (2) depends explicitly on the desired signal, $X_1(f)$; therefore, (2) is the frequency-domain signal model for noise reduction. The vector $\mathbf{d}(f)$ can be seen as the frequency-domain steering vector (Dmochowski and Benesty, 2010). This general formulation implies that we are interested in recovering the noise-free signal and not necessarily the clean speech signal.

Since $\mathbf{y}(f)$ is the sum of two uncorrelated components, its correlation matrix is

$$\begin{aligned} \mathbf{\Phi}_{\mathbf{y}}(f) &= E\left[\mathbf{y}(f)\mathbf{y}^H(f)\right] \\ &= \phi_{X_1}(f)\mathbf{d}(f)\mathbf{d}^H(f) + \mathbf{\Phi}_{\mathbf{v}}(f), \end{aligned} \tag{4}$$

where the superscript ^{H} is the conjugate-transpose operator, $\phi_{X_1}(f) = E\left[|X_1(f)|^2\right]$ is the variance of $X_1(f)$ which may also be interpreted as

the power spectral density (PSD) of the time-domain representation of $X_1(f)$ (Welch, 1967), and $\mathbf{\Phi}_{\mathbf{y}}(f) = E\left[\mathbf{y}(f)\mathbf{y}^H(f)\right]$ is the correlation matrix of $\mathbf{y}(f)$. The narrowband input SNR is given by

$$\text{iSNR}(f) = \frac{\phi_{X_1}(f)}{\phi_{V_1}(f)}, \tag{5}$$

where $\phi_{V_1}(f) = E\left[\left|V_1(f)\right|^2\right]$ is the variance of $V_1(f)$.

3. Conventional filtering approach

In the conventional filtering approach, multichannel noise reduction in the frequency domain is performed by applying a complex-valued linear filter, $\mathbf{h}(f)$, of length M, to the observation signal vector, $\mathbf{y}(f)$ (Dmochowski and Benesty, 2010; Benesty et al., 2008), i.e.,

$$\begin{aligned}\hat{X}(f) &= \mathbf{h}^H(f)\mathbf{y}(f)\\ &= X_{\text{fd}}(f) + V_{\text{rn}}(f),\end{aligned} \tag{6}$$

where the filter output, $\hat{X}(f)$, is an estimate of $X_1(f)$, $X_{\text{fd}}(f) = X_1(f)\mathbf{h}^H(f)\mathbf{d}(f)$ is the filtered desired signal, and $V_{\text{rn}}(f) = \mathbf{h}^H(f)\mathbf{v}(f)$ is the residual noise.

The two terms on the right-hand side of (6) are uncorrelated. Hence, the variance of $\hat{X}(f)$ is also the sum of two variances:

$$\begin{aligned}\phi_{\hat{X}}(f) &= \mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{y}}(f)\mathbf{h}(f)\\ &= \phi_{X_{\text{fd}}}(f) + \phi_{V_{\text{rn}}}(f),\end{aligned} \tag{7}$$

where $\phi_{X_{\text{fd}}}(f) = \phi_{X_1}(f)\left|\mathbf{h}^H(f)\mathbf{d}(f)\right|^2$ is the variance of the filtered desired signal and $\phi_{V_{\text{rn}}}(f) = \mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{h}(f)$ is the variance of the residual noise. From (7), we deduce that the narrowband output SNR is

$$\text{oSNR}[\mathbf{h}(f)] = \frac{\phi_{X_1}(f)\left|\mathbf{h}^H(f)\mathbf{d}(f)\right|^2}{\mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{h}(f)}, \tag{8}$$

which is upper bounded by Benesty et al. (2008)

$$\begin{aligned}\text{oSNR}[\mathbf{h}(f)] &\leq \phi_{X_1}(f)\mathbf{d}^H(f)\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{d}(f)\\ &= \text{oSNR}_{\text{max}}.\end{aligned} \tag{9}$$

Additionally, the narrowband output SNR gain is defined as

$$\mathcal{G}[\mathbf{h}(f)] = \frac{\text{oSNR}[\mathbf{h}(f)]}{\text{iSNR}(f)}. \tag{10}$$

A well-known and widely used example of such a filter is obtained upon minimizing the variance of the filter output or the variance of the residual noise subject to the distortionless constraint, i.e., $\mathbf{h}^H(f)\mathbf{d}(f) = 1$. This optimization results in Capon's MVDR filter (Capon, 1969), (Lacoss, 1971):

$$\begin{aligned}\mathbf{h}_{\text{MVDR}}(f) &= \frac{\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{d}(f)}{\mathbf{d}^H(f)\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{d}(f)}\\ &= \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{d}(f)}{\mathbf{d}^H(f)\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{d}(f)},\end{aligned} \tag{11}$$

which can be rewritten as (Benesty et al., 2008)

$$\mathbf{h}_{\text{MVDR}}(f) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{\Phi}_{\mathbf{y}}(f) - \mathbf{I}_M}{\text{tr}\left[\mathbf{\Phi}_{\mathbf{v}}^{-1}(f)\mathbf{\Phi}_{\mathbf{y}}(f)\right] - M}\mathbf{i}, \tag{12}$$

where $\text{tr}[\cdot]$ is the trace of a square matrix, $\mathbf{I}_M$ is the identity matrix of size $M\times M$, and $\mathbf{i}$ is the first column of $\mathbf{I}_M$. As a result, the estimate of $X_1(f)$ with the MVDR filter is

$$\begin{aligned}\hat{X}_{\text{MVDR}}(f) &= \mathbf{h}_{\text{MVDR}}^H(f)\mathbf{y}(f)\\ &= X_1(f) + \mathbf{h}_{\text{MVDR}}^H(f)\mathbf{v}(f).\end{aligned} \tag{13}$$

Another commonly used filter is the LCMV, which attempts, much like the MVDR, to minimize the variance of the residual noise. However, with the LCMV, this minimization is subject to a set of $2 \leq L \leq M$ linear constraints. The LCMV filter is usually effective in cases when some further information on the environment is known, thus allowing to a priori attenuate the output signal in noisy directions (Griffiths and Jim, 1982):

$$\mathbf{h}_{\text{LCMV}}(f) = \mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{C}(f)[\mathbf{C}^H(f)\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{C}(f)]^{-1}\boldsymbol{\beta}, \tag{14}$$

where $\mathbf{C}$ is an $M\times L$ matrix whose columns are the steering vectors in the directions of constraints, and $\boldsymbol{\beta}$ is an $L\times 1$ vector of the desired filter responses in these directions. Thus, the estimate of $X_1(f)$ with the LCMV filter is

$$\begin{aligned}\hat{X}_{\text{LCMV}}(f) &= \mathbf{h}_{\text{LCMV}}^H(f)\mathbf{y}(f)\\ &= X_1(f) + \mathbf{h}_{\text{LCMV}}^H(f)\mathbf{v}(f).\end{aligned} \tag{15}$$

4. Kronecker product filtering approach

The idea behind the new approach is to estimate the spectral power of the desired signal, i.e., $|X_1(f)|^2$, rather than the complex signal, $X_1(f)$, as it was suggested in Ephraim and Malah (1984). We can express the spectral power of $\hat{X}(f)$ defined in (6) as

$$\begin{aligned}\left|\hat{X}(f)\right|^2 &= \mathbf{h}^H(f)\mathbf{y}(f)\mathbf{y}^H(f)\mathbf{h}(f)\\ &= \text{tr}\left[\mathbf{h}(f)\mathbf{h}^H(f)\mathbf{y}(f)\mathbf{y}^H(f)\right]\\ &= \text{vec}^H\left[\mathbf{h}(f)\mathbf{h}^H(f)\right]\text{vec}\left[\mathbf{y}(f)\mathbf{y}^H(f)\right]\\ &= \left[\mathbf{h}^*(f)\otimes\mathbf{h}(f)\right]^H\left[\mathbf{y}^*(f)\otimes\mathbf{y}(f)\right]\\ &= \left[\mathbf{h}^*(f)\otimes\mathbf{h}(f)\right]^H\tilde{\mathbf{y}}(f),\end{aligned} \tag{16}$$

where vec[.] is the vectorization operation, $\otimes$ is the Kronecker product, and $\tilde{\mathbf{y}}(f) = \mathbf{y}^*(f)\otimes\mathbf{y}(f)$.

Now, let $\tilde{\mathbf{h}}(f)$ be a complex-valued filter of length M^2 which is not necessarily of the form $\tilde{\mathbf{h}}(f) = \mathbf{h}^*(f)\otimes\mathbf{h}(f)$. Eq. (16) suggests that we can estimate $|X_1(f)|^2$ by applying $\tilde{\mathbf{h}}(f)$ to $\tilde{\mathbf{y}}(f) = \mathbf{y}^*(f)\otimes\mathbf{y}(f)$, i.e.,

$$Z(f) = \tilde{\mathbf{h}}^H(f)\tilde{\mathbf{y}}(f). \tag{17}$$

We note that by not restricting $\tilde{\mathbf{h}}(f)$ to have the Kronecker product structure of the last line of (16), we generate extra degrees of freedom which may potentially yield improved noise reduction capabilities with respect to $\mathbf{h}(f)$. When $\tilde{\mathbf{h}}(f)$ is derived, we can estimate the desired signal, $X_1(f)$, with

$$\hat{X}(f) = e^{\eta\psi(f)}\sqrt{|Z(f)|}, \tag{18}$$

where $\psi(f)$ is the desired signal estimated phase that can be obtained in any given way. Practically, $\psi(f)$ may be the phase of $\hat{X}_{\text{MVDR}}(f)$ or $\hat{X}_{\text{LCMV}}(f)$, for example. Clearly, this approach is highly nonlinear.

It should be pointed out that the concept of extending the dimension of filtering beyond the observations signal dimension was, for example, suggested in Benesty et al. (2010) in the context of single-channel noise reduction with a gain. However, the differences between the widely linear filter of Benesty et al. (2010) and the work we present here are significant. The widely linear approach is essentially linear, while the approach taken here is clearly nonlinear. Furthermore, as it can be observed from the definition of $\tilde{\mathbf{y}}(f)$, in our approach a squared-dimensional filter is applied, not to the observations vector directly, but rather to their instantaneous correlation terms. As we will show, this implies that we exploit higher-order statistics.

The expression for $\tilde{\mathbf{y}}(f)$ can be further developed

$$\begin{aligned}\tilde{\mathbf{y}}(f) &= \mathbf{y}^*(f)\otimes\mathbf{y}(f)\\ &= \left[\mathbf{x}^*(f) + \mathbf{v}^*(f)\right]\otimes\left[\mathbf{x}(f) + \mathbf{v}(f)\right]\\ &= \left|X_1(f)\right|^2\tilde{\mathbf{d}}(f) + \mathbf{x}^*(f)\otimes\mathbf{v}(f)\\ &\quad + \mathbf{v}^*(f)\otimes\mathbf{x}(f) + \tilde{\mathbf{v}}(f),\end{aligned} \tag{19}$$

where $\widetilde{\mathbf{d}}(f) = \mathbf{d}^*(f) \otimes \mathbf{d}(f)$ and $\widetilde{\mathbf{v}}(f) = \mathbf{v}^*(f) \otimes \mathbf{v}(f)$. Exploiting (19) to analyze the variance of the estimated desired signal $\hat{X}(f)$, we have

$$\begin{aligned}\phi_{\hat{X}}(f) &= E[|Z(f)|] \\ &\approx |E[Z(f)]| \\ &= \left|\widetilde{\mathbf{h}}^H(f)E\left[\widetilde{\mathbf{y}}(f)\right]\right| \\ &= \left|\phi_{X_1}(f)\widetilde{\mathbf{h}}^H(f)\widetilde{\mathbf{d}}(f) + \widetilde{\mathbf{h}}^H(f)E\left[\widetilde{\mathbf{v}}(f)\right]\right| \\ &= \left|\widetilde{\mathbf{h}}^H(f)\text{vec}\left[\mathbf{\Phi}_{\mathbf{x}}(f)\right] + \widetilde{\mathbf{h}}^H(f)\text{vec}\left[\mathbf{\Phi}_{\mathbf{v}}(f)\right]\right|.\end{aligned} \tag{20}$$

Note that according to the approximation in the second row of (20), $Z(f)$ is assumed to be real and positive.

We may define the output SNR and the output SNR gain for the KP filtering by analogy to the expressions in (8) and (10), respectively, by

$$\text{oSNR}\left[\widetilde{\mathbf{h}}(f)\right] = \frac{\left|\widetilde{\mathbf{h}}^H(f)\text{vec}\left[\mathbf{\Phi}_{\mathbf{x}}(f)\right]\right|}{\left|\widetilde{\mathbf{h}}^H(f)\text{vec}\left[\mathbf{\Phi}_{\mathbf{v}}(f)\right]\right|}, \tag{21}$$

$$\mathcal{G}\left[\widetilde{\mathbf{h}}(f)\right] = \frac{\text{oSNR}\left[\widetilde{\mathbf{h}}(f)\right]}{\text{iSNR}(f)}. \tag{22}$$

To begin with, we note that when $\widetilde{\mathbf{h}}(f) = \mathbf{h}^*(f) \otimes \mathbf{h}(f)$ Eq. (21) immediately reduces to (8) and the conventional filtering approach is obtained as a special case of the KP filtering approach. In this case the approximation in (20) is always correct, and therefore the output SNR and the output SNR gain expressions are mathematically accurate. Alternatively, when $\widetilde{\mathbf{h}}(f)$ does not follow this structure, there is no guarantee that $Z(f)$ is real and positive, and hence in such cases Eq. (21) is merely an approximation. Nonetheless, and as it will be further explained in Section 6, we make sure that $Z(f)$ is always real and positive, implying that (21) is, in fact, a reasonable output SNR approximation.

Consider the following criterion:

$$\begin{aligned}\mathcal{J}\left[\widetilde{\mathbf{h}}(f)\right] &= E\left[|Z(f)|^2\right] \\ &= \widetilde{\mathbf{h}}^H(f)\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f)\widetilde{\mathbf{h}}(f),\end{aligned} \tag{23}$$

where $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f) = E\left[\widetilde{\mathbf{y}}(f)\widetilde{\mathbf{y}}^H(f)\right]$. We would like to minimize $\mathcal{J}\left[\widetilde{\mathbf{h}}(f)\right]$ subject to the distortionless constraint, i.e., $\widetilde{\mathbf{h}}^H(f)\widetilde{\mathbf{d}}(f) = 1$. The optimal filter is given by

$$\widetilde{\mathbf{h}}_{\text{MVDR}}(f) = \frac{\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(f)\widetilde{\mathbf{d}}(f)}{\widetilde{\mathbf{d}}^H(f)\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(f)\widetilde{\mathbf{d}}(f)}, \tag{24}$$

which we refer to as the KP-MVDR filter. Note that $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f)$ is the fourth moment matrix of the noisy observations vector $\mathbf{y}(f)$ and is of size $M^2 \times M^2$. Unlike former studies, in which higher-order statistics were primarily used to obtain estimates of input parameters of speech enhancement method (Nemer et al., 2001; 2002), in this study the fourth-order statistics is directly utilized to derive the speech enhancement filter. As for the computational complexity, calculating the inverse of $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f)$, for example, by using the classic Gauss-Jordan method, would require a time complexity of $O(M^6)$ operations in comparison to the $O(M^3)$ required by the conventional $\mathbf{\Phi}_{\mathbf{y}}(f)$. Nonetheless, when M is not very big, the additional complexity is insignificant.

We can generalize this approach and minimize $\mathcal{J}\left[\widetilde{\mathbf{h}}(f)\right]$ subject to a set of linear constraints as is done with the conventional LCMV:

$$\widetilde{\mathbf{h}}^H(f)\widetilde{\mathbf{C}}(f) = \boldsymbol{\beta}^H, \tag{25}$$

where $\widetilde{\mathbf{C}}(f)$ is an $M^2 \times L$ matrix whose columns are the KP filtering steering vectors $\widetilde{\mathbf{d}}(f)$ in the directions of constraints, and $\boldsymbol{\beta}$ is the same as in (14). Then, the derivation of the KP-LCMV is straightforward

$$\widetilde{\mathbf{h}}_{\text{LCMV}}(f) = \mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(f)\widetilde{\mathbf{C}}(f)[\widetilde{\mathbf{C}}^H(f)\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(f)\widetilde{\mathbf{C}}(f)]^{-1}\boldsymbol{\beta}. \tag{26}$$

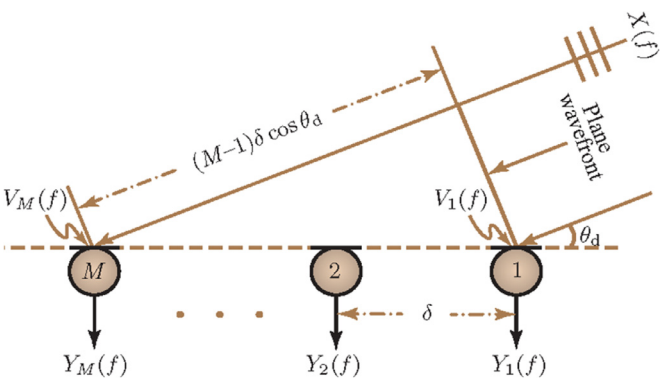

Fig. 1. A typical uniform linear array with M sensors.

5. Analysis of a toy example

In a non reverberant environment, consider a uniform linear array (ULA) of M sensors with an interelement spacing δ (see Fig. 1) satisfying $T_s = \delta/c$, where T_s is the sampling interval and c the speed of sound in the air. Assume that a white Gaussian signal of interest, $x(t) \sim \mathcal{N}(0, f_x)$, is impinging on the array from the broadside direction, i.e., $\theta_d = 90°$, and corrupted by thermal white Gaussian noise, $v_m(t) \sim \mathcal{N}(0, f_v)$, $m = \{1, \ldots, M\}$. For simplicity, we consider the case of $M = 2$ sensors, and assume that the signals in the array are sampled $N = Tf_s$ times within the signal duration T at $f_s = 1/T_s = 16$ kHz. The correlation matrices of $\mathbf{x}(f)$ and $\mathbf{v}(f)$ are given, respectively, by

$$\mathbf{\Phi}_{\mathbf{x}}(f) = K\epsilon_x\mathbf{d}(f, \cos\theta_d)\mathbf{d}^H(f, \cos\theta_d), \tag{27}$$

$$\mathbf{\Phi}_{\mathbf{v}}(f) = K\epsilon_v\mathbf{I}_M, \tag{28}$$

where $\mathbf{d}(f, \cos\theta_d) = [1\ 1]^T$ is the desired signal steering vector and K is an appropriate scaling constant resulting from the frequency domain transform. The narrowband input SNR is given by

$$\begin{aligned}\text{iSNR}(f) &= \frac{\phi_{X_1}(f)}{\phi_{V_1}(f)} \\ &= \frac{\epsilon_x}{\epsilon_v},\end{aligned} \tag{29}$$

where $\phi_{X_1}(f)$ and $\phi_{V_1}(f)$ are the variances of the desired signal and noise at the first microphone.

It is clear that the optimal conventional MVDR filter is given by $\mathbf{h}_{\text{MVDR}}(f) = [0.5\ 0.5]^T$, which results in an output SNR gain of approximately 3dB. This gain is independent of the input SNR. For the purpose of deriving the KP-MVDR filter, we first have to evaluate its corresponding correlation matrix $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f)$. The latter is a sum of 16 $M^2 \times M^2$ matrices and therefore might be difficult to calculate in general. However, in our toy example, by recalling complex-normal distribution properties, it may be shown (as derived in the Appendix) that

$$\begin{aligned}\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f) \propto\ & 2\,\text{iSNR}\begin{bmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{bmatrix} \\ &+ \frac{1}{\text{iSNR}}\begin{bmatrix}2&0&0&1\\0&1&0&0\\0&0&1&0\\1&0&0&2\end{bmatrix} + 2\begin{bmatrix}2&1&1&1\\1&1&0&1\\1&1&0&1\\1&1&1&2\end{bmatrix},\end{aligned} \tag{30}$$

which is indeed input SNR-dependent. We note that when the input SNR is high, $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f)$ is nearly singular. This implies that regularization should be used, and that the optimal KP filter is approximately $\widetilde{\mathbf{h}}_{\text{MVDR}}(f) =$

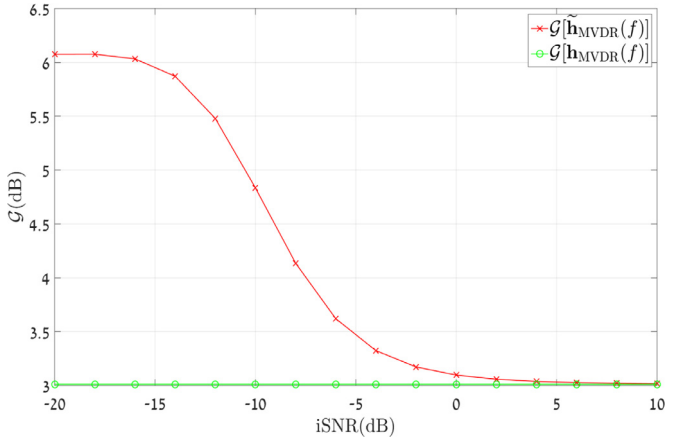

Fig. 2. Output SNR gain curve of a desired white Gaussian signal corrupted by thermal white Gaussian noise as a function of the input SNR. The array consists of $M = 2$ sensors.

$0.25\widetilde{\mathbf{d}}(f, \cos\theta_{\mathrm{d}}) = 0.25[1\ 1\ 1\ 1]^T$. Recalling that in this case $\mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{v}}(f)\right] = K\epsilon_{\mathrm{v}}[1\ 0\ 0\ 1]^T$, we have

$$\begin{aligned}\mathcal{G}\left[\widetilde{\mathbf{h}}(f)\right] &= \frac{\mathrm{oSNR}\left[\widetilde{\mathbf{h}}(f)\right]}{\mathrm{iSNR}(f)}\\ &= \frac{\left|\widetilde{\mathbf{h}}^H(f)\mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{x}}(f)\right]\right|}{\left|\widetilde{\mathbf{h}}^H(f)\mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{v}}(f)\right]\right|}\frac{\phi_{V_1}(f)}{\phi_{X_1}(f)}\\ &= \frac{1}{0.25[1\ 1\ 1\ 1][1\ 0\ 0\ 1]^T}\\ &\approx 3\mathrm{dB}, \end{aligned} \tag{31}$$

which is identical to $\mathcal{G}\left[\mathbf{h}_{\mathrm{MVDR}}(f)\right]$. However, when the input SNR is very low, the second component on the right hand side of (30) is dominant, and we have

$$\mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(f) \propto \begin{bmatrix} 2/3 & 0 & 0 & -1/3\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ -1/3 & 0 & 0 & 2/3 \end{bmatrix},$$

which implies that $\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f) = 1/8[1\ 3\ 3\ 1]^T$. Thus

$$\mathcal{G}\left[\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f)\right] = \frac{1}{1/8[1\ 3\ 3\ 1][1\ 0\ 0\ 1]^T} \approx 6\mathrm{dB}. \tag{32}$$

Fig. 2 depicts the output SNR gain curves of the conventional and KP MVDRs in the foregoing scenario as a function of the input SNR. Indeed, in high input SNRs both approaches perform equally well. However, as the input SNR decreases, the KP-MVDR yields higher SNR gain than the conventional MVDR.

It may also be informative to discuss the time complexity differences between the two approaches in the context of this example. Let us begin with the conventional MVDR. The most expensive calculation it requires in terms of computational costs is indeed the correlation matrix inversion, which requires roughly $O(M^3)$ multiplications. We assume for simplicity that the cost of M^3 multiplications is exact. Then, the calculation of the numerator, i.e., the term $\mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{d}(f)$, requires M^2 multiplications, whereas in order to compute the denominator another M multiplications must be performed. Calculating the filter takes another (real) division operation, and applying the filter to the noisy observations vector costs additional M multiplications. Thus, the total cost of generating a desired signal estimate out of the observations is roughly $M^3 + M^2 + 2M = 16$ complex multiplications, with $M = 2$, plus another real division operation. We move on to the KP-MVDR. It requires the same operations as the conventional MVDR, but rather with squared-size correlation matrix and steering vector. In addition, generating the modified observations vector $\widetilde{\mathbf{y}}(f)$ requires another M^2 multiplications, whereas the desired signal power estimate requires an additional (real) square-root operation. Thus, the total computational cost of the KP-MVDR is about $M^6 + M^4 + 3M^2 = 92$ complex multiplications, a single real division and a single real square-root operation. In practice, running the toy example code with MATLAB software on an ordinary CPU takes 250 μs to complete with the conventional MVDR and 520 μs with the KP-MVDR. Increasing the array size to $M = 7$ yields a total runtime of 350 μs with the conventional MVDR and 1000 μs with the KP-MVDR.

6. Simulations

6.1. Speech signals simulations in an anechoic environment

We are interested in examining the KP approach in more practical scenarios, i.e., with nonstationary desired signals and in the presence of spatial interferences. Therefore, the noise reduction procedure is modified as follows. The observed signals are transformed into the STFT domain using 50% overlapping time frames and a Hamming analysis window of length 512 (32 ms). We derive each of the aforementioned filters in the STFT domain. That is, the two conventional filters: $\mathbf{h}_{\mathrm{MVDR}}(f, t)$ and $\mathbf{h}_{\mathrm{LCMV}}(f, t)$; and the two KP filters: $\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f, t)$ and $\widetilde{\mathbf{h}}_{\mathrm{LCMV}}(f, t)$. The filters are applied in the STFT domain to generate estimates of the desired signal. Finally, the inverse STFT is applied to yield time-domain speech signals.

We evaluate the performances of the different filters by comparing the output SNR gains. In the STFT domain, the input and output SNR expressions in (5), (8), and (21) are modified to

$$\overline{\mathrm{iSNR}}(f) = \frac{\sum_t \phi_{X_1}(f,t)}{\sum_t \phi_{V_1}(f,t)}, \tag{33}$$

$$\overline{\mathrm{oSNR}}[\mathbf{h}(f)] = \frac{\sum_t \phi_{X_1}(f,t)\left|\mathbf{h}^H(f,t)\mathbf{d}(f)\right|^2}{\sum_t \mathbf{h}^H(f,t)\mathbf{\Phi}_{\mathbf{v}}(f,t)\mathbf{h}(f,t)}, \tag{34}$$

and

$$\overline{\mathrm{oSNR}}\left[\widetilde{\mathbf{h}}(f)\right] = \frac{\sum_t \left|\widetilde{\mathbf{h}}^H(f,t)\ \mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{x}}(f,t)\right]\right|}{\sum_t \left|\widetilde{\mathbf{h}}^H(f,t)\ \mathrm{vec}\left[\mathbf{\Phi}_{\mathbf{v}}(f,t)\right]\right|}, \tag{35}$$

where $\phi_{X_1}(f,t)$ and $\phi_{V_1}(f,t)$ are the STFT-domain variances of the desired signal and noise at the first microphone, and $\mathbf{\Phi}_{\mathbf{x}}(f, t)$ and $\mathbf{\Phi}_{\mathbf{v}}(f, t)$ are the STFT-domain correlation matrices of the desired signal and noise. The average output SNR gains are given by

$$\overline{\mathcal{G}}[\mathbf{h}(f)] = \frac{\overline{\mathrm{oSNR}}[\mathbf{h}(f)]}{\overline{\mathrm{iSNR}}(f)} \tag{36}$$

and

$$\overline{\mathcal{G}}\left[\widetilde{\mathbf{h}}(f)\right] = \frac{\overline{\mathrm{oSNR}}\left[\widetilde{\mathbf{h}}(f)\right]}{\overline{\mathrm{iSNR}}(f)}, \tag{37}$$

respectively. We employ the average output SNR gain as our main performance measure.

There is another modification that should be made with the KP approach in order to obtain a reliable desired signal estimation and keep the expressions in (35) and (37) valid. We recall that for each time-frequency bin the STFT modification of (17) provides a local estimate of the desired signal spectral power:

$$Z(f,t) = \widetilde{\mathbf{h}}^H(f,t)\widetilde{\mathbf{y}}(f,t). \tag{38}$$

While it is easy to show that with both $\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f)$ and $\widetilde{\mathbf{h}}_{\mathrm{LCMV}}(f)$ this expression is real, there is no guarantee that it is strictly positive. In practice, when a desired speech signal is present, it is very likely that the inner product in (38) is indeed positive, hence yielding a valid estimate of the desired signal spectral power. This may be seen by applying

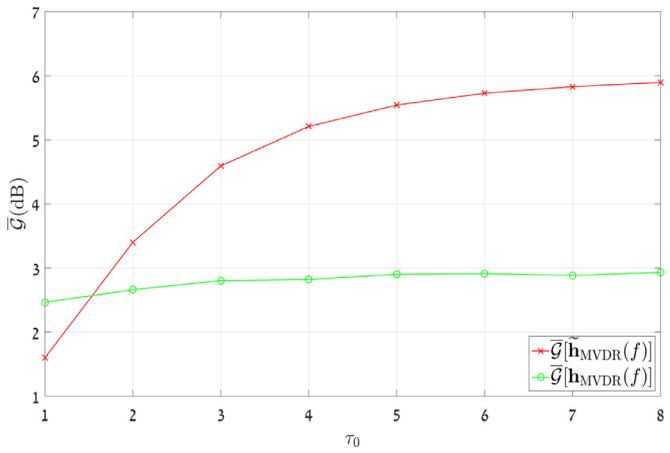

Fig. 3. Output SNR gain curve of a desired white Gaussian signal corrupted by thermal white Gaussian noise as a function of the temporal smoothing parameter τ_0. The array consists of $M = 2$ sensors and the input SNR is -20dB.

one of the KP filters to the last equality of (19) in which the first term, that is associated with the true desired signal power, is guaranteed to be positive. Nevertheless, when a desired signal is absent, this positive term is approximately zero and the power estimate might turn out to be negative. Clearly, such an estimate is non-physical and should be clipped to zero. Consequently, (38) is modified to

$$Z(f,t) = \max\left\{\widetilde{\mathbf{h}}^H(f,t)\widetilde{\mathbf{y}}(f,t), 0\right\}, \tag{39}$$

whereas when a zero estimate is obtained both the filtered desired signal and the residual noise are zeroed out, and are referred to accordingly in (35) and (37).

The correlation matrices in the STFT domain are obtained as a straightforward temporal smoothing of the appropriate instantaneous signals, i.e.,

$$\mathbf{\Phi}_{\mathbf{y}}(f,t) = \frac{1}{2\tau_0+1}\sum_{\tau=t-\tau_0}^{t+\tau_0}\mathbf{y}(f,\tau)\mathbf{y}^H(f,\tau) \tag{40}$$

and

$$\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f,t) = \frac{1}{2\tau_0+1}\sum_{\tau=t-\tau_0}^{t+\tau_0}\widetilde{\mathbf{y}}(f,\tau)\widetilde{\mathbf{y}}^H(f,\tau), \tag{41}$$

where τ_0 is a smoothing parameter indicating the temporal duration of the smoothing process. As $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(f,t)$ is considerably larger than $\mathbf{\Phi}_{\mathbf{y}}(f,t)$, the KP filter is more vulnerable to correlation estimation errors. That is, when τ_0 is not large enough the off-diagonal elements of the correlation matrices are inaccurately estimated, thus adding a significant amount of noise which corrupts the optimal filters. This behaviour is demonstrated, for example, for the stationary toy example described above in Fig. 3, in which the average output SNR gain of the KP and conventional MVDRs are plotted as a function of the smoothing parameter τ_0. Indeed, we observe that the KP-MVDR requires a considerably longer temporal smoothing in order to achieve its theoretical output SNR gain. We set $\tau_0 = 5$ for all further simulations we discuss next.

We examine the average output SNR gains of the four filters in similar settings to the previous scenario, but we employ speech signals which are taken from the TIMIT database (Darpa timit acoustic phonetic continuous speech corpus cdrom, 1993) as the desired signal, and add an interference impinging on the sensor array from the endfire direction (0°) that is three times more powerful than the background thermal white Gaussian noise. This implies that the overall noise consists of both directional and thermal noise terms. Naturally, all four filters satisfy the distortionless constraint, while the conventional and KP LCMVs also place a null at 0° ($L = 2$).

The average output SNR gains for the six combinations of $M \in \{3, 5, 7\}$ with input SNR $\in \{0, 7\}$dB for the two KP filters and their two conventional counterparts are depicted in Fig. 4. We note that the input SNR takes into account both the background noise and the interference. To begin with, we observe that the KP-LCMV achieves a significantly preferable average output SNR gain in comparison to the conventional LCMV, in particular in low frequencies. Similarly, the KP-MVDR maintains a substantial output SNR gain gap over the conventional MVDR throughout the entire frequency spectrum. However, as M and the input SNR increase the output SNR gain gap decreases- specifically in frequencies lower than 3 kHz, which are typically of high interest in speech signals. The significant output SNR gain gap in frequencies higher than 3 kHz may be explained by the lower bound in (39), which eliminates much of the residual noise with the KP filters when the desired speech signal is absent.

Next, we employ the perceptual evaluation of speech quality (PESQ) score (Rix et al., 2001) as an additional objective performance measure, which is applied on the enhanced speech in the time domain. Table 1 summarises the average PESQ score of the four filters in the aforementioned six settings combinations in addition to the PESQ score of the noisy signal in the reference microphone $Y_1(f)$. We point out that the PESQ score and the average output SNR gain in low frequencies exhibit some sort of correlation. That is, according to both measures the KP approach is shown to produce cleaner enhanced signals, particularly for low input SNRs or small arrays. For high input SNRs and large arrays, the KP approach may still be preferable, but the PESQ score and average output SNR gain differences are significantly reduced.

We end this section by comparing the spectrograms of the desired, noisy and four enhanced signals with $M = 5$ and iSNR = 0dB, which are depicted in Fig. 5. We observe that the enhanced signals with the KP approach exhibit a higher resemblance to the desired signal, with the background noise strongly attenuated. This is stressed out in particular in very low frequencies, in which the conventional LCMV, for example, amplifies the background noise while the KP-LCMV attenuates it.

6.2. Speech signals simulations in reverberant environments

In this section, we address the noise reduction performance in reverberant environments. We use a room impulse response (RIR) generator (Habets, 2014) to simulate the reverberant noise-free signal received in each of the microphones. The RIR generator is based on the image method of Allen and Berkley (1979). We point out that for the sake of verification, some of the following scenarios were repeated using the randomized image method presented in Sena et al. (2015). To begin with, we are interested in examining the desired speech signal reverberation influence on the noise reduction performance for different RIRs. Hence, the simulation settings is as follows. A $6\times6\times3$ m room contains a desired speech signal source located at $(x, y, z) = (3, 1, 1.5)$, and $M = 3$ microphones located, respectively, at (2.95, 5, 1.5), (3, 5, 1.5) and (3.05, 5, 1.5). The microphone signals contain thermal white Gaussian noise. We simulate 3 scenarios with a varying value of $T_{60} \in \{0, 250, 400\}$ msec (as defined by Sabin-Franklin's formula Pierce (1991)), and use the conventional and KP-MVDR filters to perform noise reduction. We note that the filters are derived according to the non-reverberant model with $\theta_{\mathrm{d}} = 90°$.

The simulations are carried out for both iSNR = 0 and iSNR = 7 and the same set of TIMIT speech signals used in the previous part. We compare the average PESQ scores of the time-domain enhanced speech signals to the clean and noisy reverberant signals. The results are shown in Table 2. We observe that the KP-MVDR obtains higher PESQ scores in all the foregoing scenarios, however, the performance gap is more significant for the lower values of the input SNR and T_{60}. That is, when T_{60} is high, the speech quality is mainly deteriorated by the reverberation and not by the white background noise. Hence, in such cases the PESQ score improvement due to the noise reduction is limited, and an additional dereverberation stage should be incorporated.

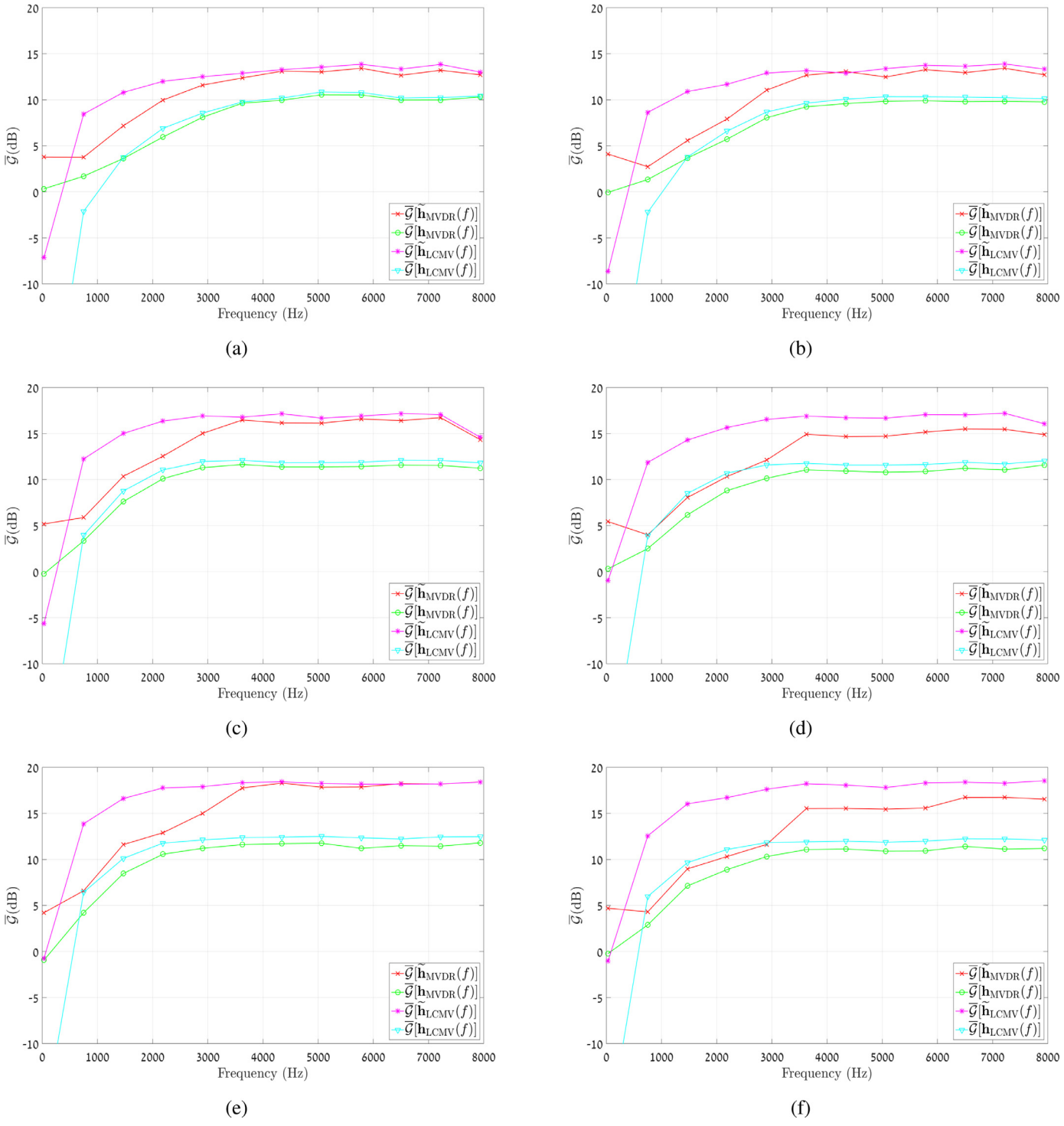

Fig. 4. Average output SNR gains as a function of the frequency for various array sizes and input SNRs. (a) iSNR = 0 dB and $M = 3$, (b) iSNR = 7 dB and $M = 3$, (c) iSNR = 0 dB and $M = 5$, (d) iSNR = 7 dB and $M = 5$, (e) iSNR = 0 dB and $M = 7$, and (f) iSNR = 7 dB and $M = 7$.

Table 1
Average PESQ scores for iSNR = 0dB and iSNR = 7dB with varying array sizes.

	iSNR = 0 dB, $M = 3$	iSNR = 0 dB, $M = 5$	iSNR = 0 dB, $M = 7$	iSNR = 7 dB, $M = 3$	iSNR = 7 dB, $M = 5$	iSNR = 7 dB, $M = 7$
$Y_1(f)$	1.502	1.502	1.502	1.982	1.982	1.982
$\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f)$	2.235	2.741	2.762	2.644	2.987	3.027
$\mathbf{h}_{\mathrm{MVDR}}(f)$	1.939	2.313	2.537	2.494	2.769	2.977
$\widetilde{\mathbf{h}}_{\mathrm{LCMV}}(f)$	2.123	2.748	2.817	2.627	3.033	3.082
$\mathbf{h}_{\mathrm{LCMV}}(f)$	1.628	2.155	2.478	2.192	2.738	3.022

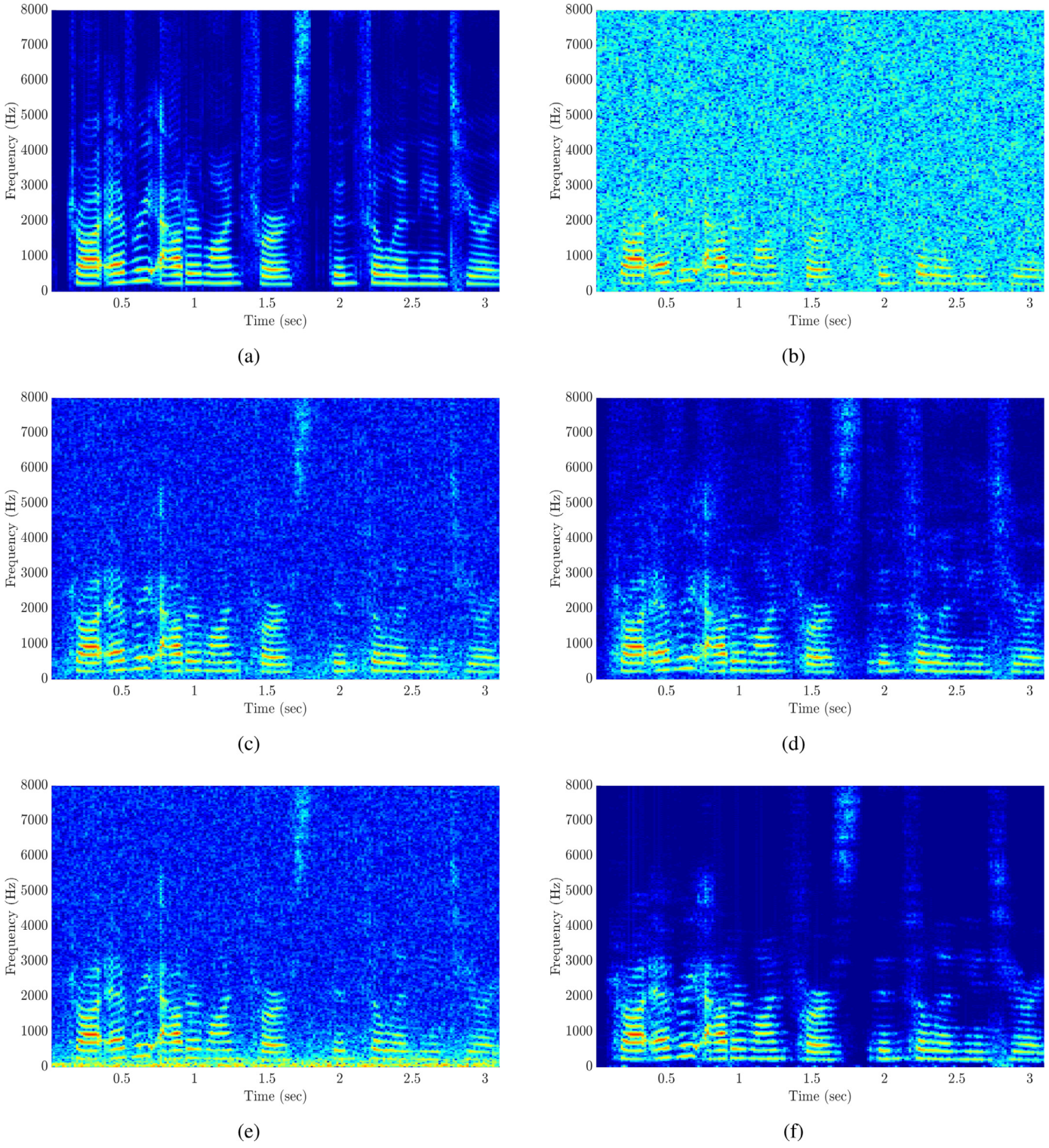

Fig. 5. Spectrograms of the desired signal, noisy input signal, and the enhanced signals in the presence of a directional interference and white thermal noise. The array size is $M = 5$ microphones and the input SNR is iSNR = 0 dB. (a) Clean desired signal, (b) noisy input signal at the first (reference) microphone, (c) enhanced signal with $\mathbf{h}_{\mathrm{MVDR}}(f)$, (d) enhanced signal with $\widetilde{\mathbf{h}}_{\mathrm{MVDR}}(f)$, (e) enhanced signal with $\mathbf{h}_{\mathrm{LCMV}}(f)$, and (f) enhanced signal with $\widetilde{\mathbf{h}}_{\mathrm{LCMV}}(f)$.

Next, we examine a more complicated set of scenarios in which in addition to the thermal white Gaussian noise, there are up to three reverberant spatial interferences. The problem settings are modified as follows. A uniform linear array of $M = 5$ microphones is located around (3, 5, 1.5) in the same room described above. The microphone array spacing is 0.05m, which implies that the total array length is 0.2m. Additionally, 3 white interference sources are placed on the $z = 1.5$ plane: U_1@(1, 5, 1.5), U_2@(1, 1, 1.5), and U_3@(5, 1, 1.5). An illustration of the $z = 1.5$ plane of the reverberant room is depicted in Fig. 6. The activity of the interference sources is set according to the 3

Table 2
Average PESQ scores for iSNR = 0 dB and iSNR = 7 dB with varying values of T_{60}. The background noise is thermal white Gaussian noise and $M = 3$.

	iSNR = 0 dB, T_{60} = 0ms	iSNR = 0 dB, T_{60} = 250ms	iSNR = 0 dB, T_{60} = 400ms	iSNR = 7 dB, T_{60} = 0ms	iSNR = 7 dB, T_{60} = 250ms	iSNR = 7 dB, T_{60} = 400ms
Noisy reverberant signal	1.87	2.04	1.98	2.34	2.33	2.2
Reverberant enhanced signal with $\mathbf{h}_{\mathrm{MVDR}}(f)$	2.19	2.18	2.13	2.64	2.41	2.3
Reverberant enhanced signal with $\tilde{\mathbf{h}}_{\mathrm{MVDR}}(f)$	2.38	2.27	2.2	2.75	2.46	2.33
Clean reverberant signal	4.5	2.72	2.43	4.5	2.72	2.43

Table 3
Average PESQ scores for the 3 aforementioned scenarios with iSNR = 0 dB and iSNR = 7 dB, T_{60} = 250msec and $M = 5$.

	iSNR = 0 dB, Scen. (a)	iSNR = 0 dB, Scen. (b)	iSNR = 0 dB, Scen. (c)	iSNR = 7 dB, Scen. (a)	iSNR = 7 dB, Scen. (b)	iSNR = 7 dB, Scen. (c)
Noisy reverberant signal	2.11	2.19	2.24	2.39	2.45	2.49
Reverberant enhanced signal with $\mathbf{h}_{\mathrm{LCMV}}(f)$	2.27	2.22	1.34	2.43	2.45	1.58
Reverberant enhanced signal with $\tilde{\mathbf{h}}_{\mathrm{LCMV}}(f)$	2.46	2.46	2.43	2.49	2.47	2.46
Reverberant enhanced signal with $\mathbf{h}_{\mathrm{MVDR}}(f)$	2.3	2.34	2.36	2.44	2.49	2.5
Reverberant enhanced signal with $\tilde{\mathbf{h}}_{\mathrm{MVDR}}(f)$	2.48	2.5	2.48	2.49	2.51	2.5
Clean reverberant signal	2.72	2.72	2.72	2.72	2.72	2.72

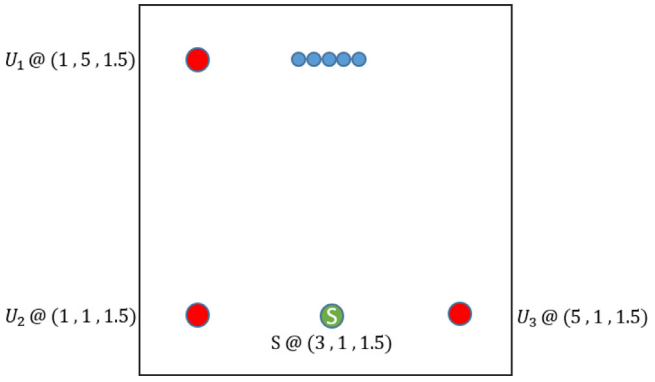

Fig. 6. An illustration of the $z = 1.5$ plane of the reverberant room.

following scenarios:

Scen. (a): only U_1 is active,
Scen. (b): only U_1 and U_2 are active,
Scen. (c): all interference sources are active.

In addition, a thermal white Gaussian noise is present. We note that in all scenarios, each interference and background noise are equally powerful. The received signal is individually simulated for each of the microphones using the RIR generator with $T_{60} = 250$ms. The set of simulations is performed for both iSNR = 0 and iSNR = 7, where the input noise, according to which the input SNR is calculated, takes into account the free field interferences and the thermal white Gaussian noise. Since in these scenarios interferences are present, it is interesting to examine the LCMV filters as well, for which the appropriate $L - 1$ null constraints are set in the direction of the direct path of the interferences. This implies that we have $L = 2$ in Scen. (a), $L = 3$ in Scen. (b), and $L = 4$ in Scen. (c). The PESQ scores of all 4 filters in the 3 scenarios and both input SNRs are shown in Table 3. We observe the following. Indeed, as the input SNR increases, the speech quality with the conventional MVDR and LCMV significantly improves. In contrast, this is not the case with the two KP filters, whose average PESQ scores remain roughly unchanged upon increasing the input SNR from iSNR = 0 to iSNR = 7. This is somewhat similar to the anechoic environment simulations of Section 6.1: The KP approach is of a better potential particularly in low input SNRs. Additionally, we note that while the conventional LCMV significantly enhances the noise in both input SNRs of Scen. (c), the KP-LCMV attains a considerable noise reduction in the low input SNR and only a slight quality deterioration in the high input SNR. Since both LCMV filters are designed to zero the direct paths of the directional interferences, we deduce that the KP-LCMV is potentially preferable in terms of white noise gain.

7. Conclusions

Conventional multichannel filters such as the MVDR and the LCMV are often used to estimate desired signals in noisy environments. Although they perform well when the input SNR is relatively high and with a large number of channels (i.e., sensors), their performances may not always be sufficient when the arrays contain only a few sensors. We have introduced a KP filtering approach to estimate the spectral power of the desired signal by exploiting higher-order statistics. We analyzed a toy example and performed a series of speech signals simulations in both anechoic and reverberant environments. We demonstrated that the proposed KP-MVDR and KP-LCMV outperform their conventional counterparts when proper temporal smoothing is employed to estimate the correlation matrix of the modified observation signal vector. This is emphasized in particular when the number of sensors is small or when the input SNR is low.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Israel Science Foundation (grant no. 576/16) and the ISF-NSFC joint research program (grant no. 2514/17).

Appendix

Derivation of (30)

Let us evaluate the KP correlation matrix $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(f)$. We have

$$\begin{aligned}\mathbf{\Phi}_{\tilde{\mathbf{y}}}(f) &= E\Big\{\big[(\mathbf{x}(f)+\mathbf{v}(f))^{*}\otimes(\mathbf{x}(f)+\mathbf{v}(f))\big]\\ &\quad\big[(\mathbf{x}(f)+\mathbf{v}(f))^{T}\otimes(\mathbf{x}(f)+\mathbf{v}(f))^{H}\big]\Big\}\\ &= E\Big\{\big[\mathbf{x}^{*}(f)\otimes\mathbf{x}(f)+\mathbf{x}^{*}(f)\otimes\mathbf{v}(f)+\mathbf{v}^{*}(f)\otimes\mathbf{x}(f)+\mathbf{v}^{*}(f)\otimes\mathbf{v}(f)\big]\\ &\quad\big[\mathbf{x}^{T}(f)\otimes\mathbf{x}^{H}(f)+\mathbf{x}^{T}(f)\otimes\mathbf{v}^{H}(f)+\mathbf{v}^{T}(f)\otimes\mathbf{x}^{H}(f)\\ &\quad+\mathbf{v}^{T}(f)\otimes\mathbf{v}^{H}(f)\big]\Big\},\end{aligned}\tag{42}$$

which is a sum of 16 $M^2\times M^2$ matrices. However, since $E[\mathbf{x}(f)]=E[\mathbf{v}(f)]=\mathbf{0}$, and $E\big[\mathbf{x}(f)\mathbf{v}^{H}(f)\big]=E\big[\mathbf{v}(f)\mathbf{x}^{H}(f)\big]=\mathbf{0}$, precisely 8 out of these 16 matrices are strictly zero. The other 8 matrices are the desired signal-only matrix, the noise-only matrix and 6 non-zero mixed matrices. We begin with the desired signal-only matrix. We recall that $M=2$, and the desired signal is normally distributed, i.e., $x_1(nT_{\mathrm{s}})=x_2(nT_{\mathrm{s}})\sim\mathcal{N}(0,\yen_{\mathrm{x}})$. Hence, its frequency domain representation is complex-normally distributed, that is

$$X_1(f)\sim\mathcal{CN}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}K\epsilon_{\mathrm{x}} & 0\\ 0 & K\epsilon_{\mathrm{x}}\end{bmatrix}\right),\tag{43}$$

where K is half of the number of bins in a single STFT frame. Then

$$\begin{aligned}&E\Big\{\big[\mathbf{x}^{*}(f)\otimes\mathbf{x}(f)\big]\big[\mathbf{x}^{T}(f)\otimes\mathbf{x}^{H}(f)\big]\Big\}\\ &\quad=E\big[|X_1(f)|^4\big]\begin{bmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{bmatrix}\\ &\quad=8\epsilon_{\mathrm{x}}^2\begin{bmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{bmatrix}.\end{aligned}$$

We move on to the noise-only matrix. Since the noise distribution is normal as well, the frequency domain representations of the noise terms are complex-normally distributed, and we easily obtain

$$E\Big\{\big[\mathbf{v}^{*}(f)\otimes\mathbf{v}(f)\big]\big[\mathbf{v}^{T}(f)\otimes\mathbf{v}^{H}(f)\big]\Big\}=4\epsilon_{\mathrm{v}}^2\begin{bmatrix}2&0&0&1\\0&1&0&0\\0&0&1&0\\1&0&0&2\end{bmatrix}.$$

Similarly, the following 6 mixed matrices are given by

$$\begin{aligned}&E\Big\{\big[\mathbf{v}^{*}(f)\otimes\mathbf{x}(f)\big]\big[\mathbf{x}^{T}(f)\otimes\mathbf{v}^{H}(f)\big]\Big\}\\ &\quad=E\Big\{\big[\mathbf{x}^{*}(f)\otimes\mathbf{v}(f)\big]\big[\mathbf{v}^{T}(f)\otimes\mathbf{x}^{H}(f)\big]\Big\}\\ &\quad=\mathbf{0},\end{aligned}$$

$$\begin{aligned}&E\Big\{\big[\mathbf{x}^{*}(f)\otimes\mathbf{v}(f)\big]\big[\mathbf{x}^{T}(f)\otimes\mathbf{v}^{H}(f)\big]\Big\}\\ &\quad=4\epsilon_{\mathrm{x}}\epsilon_{\mathrm{v}}\begin{bmatrix}1&0&1&0\\0&1&0&1\\1&0&1&0\\0&1&0&1\end{bmatrix},\end{aligned}$$

$$\begin{aligned}&E\Big\{\big[\mathbf{v}^{*}(f)\otimes\mathbf{x}(f)\big]\big[\mathbf{v}^{T}(f)\otimes\mathbf{x}^{H}(f)\big]\Big\}\\ &\quad=4\epsilon_{\mathrm{x}}\epsilon_{\mathrm{v}}\begin{bmatrix}1&1&0&0\\1&1&0&0\\0&0&1&1\\0&0&1&1\end{bmatrix},\end{aligned}$$

$$\begin{aligned}&E\Big\{\big[\mathbf{x}^{*}(f)\otimes\mathbf{x}(f)\big]\big[\mathbf{v}^{T}(f)\otimes\mathbf{v}^{H}(f)\big]\Big\}\\ &\quad=4\epsilon_{\mathrm{x}}\epsilon_{\mathrm{v}}\begin{bmatrix}1&0&0&1\\1&0&0&1\\1&0&0&1\\1&0&0&1\end{bmatrix},\end{aligned}$$

and

$$\begin{aligned}&E\Big\{\big[\mathbf{v}^{*}(f)\otimes\mathbf{v}(f)\big]\big[\mathbf{x}^{T}(f)\otimes\mathbf{x}^{H}(f)\big]\Big\}\\ &\quad=4\epsilon_{\mathrm{x}}\epsilon_{\mathrm{v}}\begin{bmatrix}1&1&1&1\\0&0&0&0\\0&0&0&0\\1&1&1&1\end{bmatrix}.\end{aligned}$$

Combining all the matrices together and dividing by $4\epsilon_{\mathrm{x}}\epsilon_{\mathrm{v}}$, we obtain the expression in (30).

References

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 65 (4), 943–950.

Bai, M., IH, J., Benesty, J., 2014. Acoustic Array Systems: Theory, Implementation, and Application. Wiley-IEEE Press.

Benesty, J., Chen, J., 2011. Optimal Time-Domain Noise Reduction Filters; A Theoretical Study. Springer-Verlag, Berlin Heidelberg doi:10.1007/978-3-642-19601-0.

Benesty, J., Chen, J., Habets, E.A.P., 2012. Speech Enhancement in the STFT Domain. Springer-Verlag, Berlin Heidelberg doi:10.1007/978-3-642-23250-3.

Benesty, J., Chen, J., Huang, Y., 2008. Microphone Array Signal Processing. Springer-Verlag, Berlin Heidelberg doi:10.1007/978-3-540-78612-2.

Benesty, J., Chen, J., Huang, Y., 2010. A widely linear distortionless filter for single-channel noise reduction. IEEE Signal Process. Lett. 17 (5), 469–472. doi:10.1109/LSP.2010.2043152.

Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Noise Reduction in Speech Processing, first ed. Springer-Verlag, Berlin Heidelberg.

Benesty, J., Chen, J., Huang, Y., Dmochowski, J., 2007. On microphone-array beamforming from a mimo acoustic signal processing perspective. IEEE Trans. Audio Speech Lang. Process. 15 (3), 1053–1065. doi:10.1109/TASL.2006.885251.

Benesty, J., Chen, J., Huang, Y.A., 2009. Noise reduction algorithms in a generalized transform domain. IEEE Trans. Audio Speech Lang. Process. 17 (6), 1109–1123. doi:10.1109/TASL.2009.2020415.

Benesty, J., Cohen, I., Chen, J., 2017. Fundamentals of Signal Enhancement and Array Signal Processing. Wiley-IEEE Press.

Buchris, Y., Cohen, I., Benesty, J., 2019. On the design of time-domain differential microphone arrays. Appl. Acoust. 148, 212–222.

Capon, J., 1969. High-resolution frequency-wavenumber spectrum analysis. Proc. IEEE 57 (8), 1408–1418. doi:10.1109/PROC.1969.7278.

Chen, J., Huang, Y., Benesty, J., 2003. Filtering Techniques for Noise Reduction and Speech Enhancement. Springer, Berlin Heidelberg, pp. 129–154. doi:10.1007/978-3-662-11028-7_5.

Darpa timit acoustic phonetic continuous speech corpus cdrom, 1993.

Dmochowski, J., Benesty, J., 2010. Microphone Arrays: Fundamental Concepts. Springer, Berlin Heidelberg, pp. 199–223. doi:10.1007/978-3-642-11130-3_8.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. 32 (6), 1109–1121. doi:10.1109/TASSP.1984.1164453.

Ephraim, Y., Trees, H.L.V., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 3 (4), 251–266. doi:10.1109/89.397090.

Gannot, S., Cohen, I., 2008. Adaptive Beamforming and Postfiltering. Springer, Berlin Heidelberg, pp. 945–978. doi:10.1007/978-3-540-49127-9_47.

Griffiths, L., Jim, C., 1982. An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. 30 (1), 27–34. doi:10.1109/TAP.1982.1142739.

Habets, E. A. P.,. Rir-generator 2014. https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.

Lacoss, R.T., 1971. Data adaptive spectral analysis methods. Geophysics 36 (4), 661–675. doi:10.1190/1.1440203.

Lacouture-Parodi, Y., Habets, E.A.P., Chen, J., Benesty, J., 2014. Multichannel noise reduction in the Karhunen-Loeve expansion domain. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (5), 923–936. doi:10.1109/TASLP.2014.2311299.

Moreno, A., Fonollosa, J., 1992. Pitch determination of noisy speech using higher order statistics. Vol. 1, pp. 133–136 vol.1. doi:10.1109/ICASSP.1992.225954.

Nemer, E., Goubran, R., Mahmoud, S., 2001. Robust voice activity detection using higher-order statistics in the lpc residual domain. IEEE Trans. Speech Audio Process. 9 (3), 217–231. doi:10.1109/89.905996.

Nemer, E., Goubran, R., Mahmoud, S., 2002. Speech enhancement using fourth-order cumulants and optimum filters in the subband domain. Speech Commun. 36 (3), 219–246.

Pierce, A., 1991. Acoustics: An Introduction to Its Physical Principles and Applications. Acoustical Society of America.

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings, 2 doi:10.1109/ICASSP.2001.941023. 749–752 vol.2.

Sena, E.D., Antonello, N., Moonen, M., van Waterschoot, T., 2015. On the modeling of rectangular geometries in room acoustic simulations. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (4), 774–786. doi:10.1109/TASLP.2015.2405476.

Souden, M., Benesty, J., Affes, S., 2010. A study of the LCMV and MVDR noise reduction filters. IEEE Trans. Signal Process. 58 (9), 4925–4935. doi:10.1109/TSP.2010.2051803.

Tavakoli, V., Jensen, J., Christensen, M., Benesty, J., 2016. A framework for speech enhancement with ad hoc microphone arrays. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (6), 1038–1051.

Van Trees, H., 2004. Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory. Wiley.

Welch, P., 1967. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. 15 (2), 70–73. doi:10.1109/TAU.1967.1161901.