Multichannel Noise Reduction in the Karhunen-Loève Expansion Domain

Yesenia Lacouture-Parodi, *Member, IEEE*, Emanuël A. P. Habets, *Senior Member, IEEE*, Jingdong Chen, *Senior Member, IEEE*, and Jacob Benesty

***Abstract*—The noise reduction problem is traditionally approached in the time, frequency, or transform domain. Having a signal dependent transform has shown some advantages over the traditional signal independent transform. Recently, the single-channel noise reduction problem in the Karhunen-Loève expansion (KLE) domain has received special attention. In this paper, the noise reduction problem in the KLE domain is studied from a multichannel perspective. We present a new formulation of the problem, in which inter-channel and inter-mode correlations are optimally exploited. We derive different optimal noise reduction filters and present a set of useful performance measures within this framework. The performance of the different filters is then evaluated through experiments in which not only noise but also competing speech sources are present. It is shown that the proposed multichannel formulation is more robust to competing speech sources than the single-channel approach and that a better compromise between noise reduction and speech distortion can be obtained.**

***Index Terms*—Karhunen-Loève expansion (KLE), maximum snr filter, minimum variance distortionless response (MVDR) filter, multichannel, noise reduction, speech enhancement, tradeoff filter, wiener filter.**

I. Introduction

In MANY human-to-machine and human-to-human communication systems, such as hearing-aids, hands-free communication devices, speech recognition, or voice-controlled systems, the speech signals received by the microphones are corrupted by noise. The noise comes usually from ambient sound sources, competing/interfering speech sources and reflections. In many situations, this unwanted noise can degrade significantly the speech quality and intelligibility, which limits the usability of many communication devices. In the past decades, there has been a growing interest in the development of new techniques to improve the quality of the signals received by the microphones, which would permit a better human-to-machine and human-to-human communication. These techniques are known as noise reduction or speech enhancement techniques and even though several solutions are already available, the noise reduction problem is still a rather challenging problem in many communication applications.

Typically the noise reduction problem is approached by passing the noisy microphone signals through a linear filter in order to obtain a cleaner version of the input signal by increasing the signal-to-noise ratio (SNR) [1]. However, there is always a tradeoff between noise reduction (NR) and speech distortion (SD), since the filters might also affect the desired speech signal. Thus, it is desired to find optimal filters that not only improve the NR but at the same time preserve a reasonable quality of the desired speech signal.

The noise reduction problem is traditionally approached in either the time or frequency domain. The optimal filters are often estimated by minimizing the mean-square error (MSE) between the clean signal and its estimate. The time domain approach can be sample based, estimating one speech sample at a time [2]–[4], while the frequency domain is often formulated on a frame basis, i.e. a block of noisy speech signal is transformed into the frequency domain using the discrete Fourier transform (DFT) and then a filter is estimated and applied to the frame [5]–[10]. The frequency domain approaches are in general more flexible with respect to controlling the NR performance versus the SD, though special attention has to be paid to the aliasing distortion caused by the independent processing of subbands. The time domain approaches do not suffer from aliasing problems, but the tradeoff between NR and SD is more difficult to control and they exhibit higher computational complexity [11].

There are other domains in which the noise reduction problem can be approached. For example, the use of signal-dependent transforms has shown some advantages with regard to SD and ND [11]–[14]. Among them, the single-channel noise reduction problem in the Karhunen-Loève Expansion (KLE) domain has received special attention in the last decade [11], [15], [16]. The main difference between this method and the frequency domain methods, is that the Karhunen-Loève transform (KLT) can exactly diagonalize the signal correlation matrix, resulting in uncorrelated signal components in each subband. Thus, each subband can be processed independently while the Fourier matrix can only approximately diagonalize the noisy covariance matrix [11]. One of the main advantages of using the KLT is that if the covariance matrices are properly calculated, there is no aliasing

Manuscript received May 23, 2013; revised September 17, 2013; accepted November 13, 2013. Date of publication March 11, 2014; date of current version April 04, 2014. This work was supported by Northwestern Polytechnical University, Xi'an, China, and the International Audio Laboratories Erlangen, Germany. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Woon-Seng Gan.

Y. Lacouture-Parodi is with HUAWEI Technologies Düsseldorf GmbH, Munich Office, European Research Center, 80992 Munich, Germany (e-mail: ylacoutu@ieee.org).

E. A. P. Habets is with the International Audio Laboratories Erlangen (a joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS), 91058 Erlangen, Germany (e-mail: emanuel.habets@audiolabs-erlangen.de).

J. Chen is with the Northwestern Polytechnical University, Xi'an, 710072 Xi'an, China.

J. Benesty is with the INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada.

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/TASLP.2014.2311299

problems and that the desired speech and noise may be better separated as opposed to the frequency-domain methods [16]. A general formulation of the single-channel KLE domain approach and the design of different optimal filters has been previously proposed in [11] and [16]. In those studies, the clean speech signal is estimated from a noisy observation, which is obtained from a single microphone. It has been shown that a better noise reduction performance is achieved when properly choosing the parameters to calculate the filters.

Microphone arrays are nowadays available in many communication devices. One benefit of using more channels is that with multiple microphones, not only the temporal but also the spatial characteristics of the speech and noise sources can be exploited [3], [17], [18]. In [19], we proposed the use of multiple microphone signals to improve the performance of the optimal noise reduction Wiener filter in the KLE domain. In that study, we presented a formulation of the multichannel noise reduction problem applying a KLT to each channel. Results show that a significant improvement is obtained with respect to the single-channel case. However, by applying a different transform to each channel, the inter-channel correlations are not fully exploited. In this paper we present an extension of the proposed multichannel noise reduction problem in the KLE domain. We present a new formulation in which the inter-channel as well as the inter-mode correlations are exploited. A single KLT is applied to the joint contribution of all the channels. The obtained coefficients are then expanded into sub-coefficients, which are then treated as the coefficients corresponding to each channel. Inter-mode correlations are also exploited to take advantage of the temporal and spatial correlations contained in each sub-coefficient. Note that the proposed multichannel noise reduction in the KLE domain shares some similarities with the subspace method proposed in [14], where the correlation matrices are also diagonalized. In their subspace approach, a joint diagonalization of the noisy speech and the noise correlation matrix is done and the clean speech signal is estimated by applying a weight to the noisy eigenvectors. In our approach, on the other hand, we diagonalize only the correlation matrix of the noisy speech and estimate the clean speech signal by applying a weight to the KLE coefficients. Additionally, by expanding the KLT into sub-coefficients, we obtain inter-mode correlations which are no longer zero and are closely related to the inter-channel correlations. Thus, the proposed formulation allow us to exploit the inter-channel and inter-mode correlations in a more profound way.

This paper is organized as follows: In Section II we present the general problem statement and the signal model that is used throughout the paper. In Section III we derive the KLE in the framework of multiple microphones. The problem of multichannel noise reduction in the KLE domain and the array model is then discussed in Section IV. In Section V we recall the definitions of some useful performance measures already discussed in [16] and [19]. In Section VI we derive different optimal noise reduction filters in the KLE domain and discuss their properties and performance. In Section VII we discuss different experiments done to evaluate the performance of the filters. A summary of this study is then presented in Section VIII.

II. Signal Model

We consider the classical signal model in which a microphone array with N sensors captures a convolved source signal in some noise field. The received signals, at the discrete-time index k, are expressed as [18], [20], [21]

$$\begin{aligned} y_n(k) &= g_n(k) * s(k) + \sum_{i=1}^{I} r_{n,i}(k) * a_i'(k) + b_n(k) \\ &= x_n(k) + a_n(k) + b_n(k) \\ &= x_n(k) + v_n(k), \end{aligned} \tag{1}$$

where $n = 1, 2, \ldots, N$, $g_n(k)$ is the impulse response from the unknown desired speech source $s(k)$ to the nth microphone and $*$ denotes the convolution operation. The total additive noise at the nth microphone $v_n(k) = a_n(k) + b_n(k)$ is composed by a spatially incoherent part $b_n(k)$ and a spatially coherent part $a_n(k) = \sum_{i=1}^{I} r_{n,i}(k) * a_i'(k)$, where $r_{n,i}(k)$ is the impulse response from an unknown, undesired sound source $a_i'(k)$ to the nth microphone and I is the total number of undesired sources. We assume that the signals $x_n(k)$ and $v_n(k)$ are uncorrelated and zero mean. We assume additionally that $a_n(k)$ and $b_n(k)$ are also uncorrelated. By definition, the N signals $x_n(k)$ are coherent across the array, and so are the signals $a_n(k)$. All previous signals are considered to be real, broadband, and to simplify the development and analysis of the main ideas of this work, we further assume that they are stationary.

By processing the data by blocks of L samples, the signal model given in (1) can be put into a vector form as

$$\mathbf{y}_n(m) = \mathbf{x}_n(m) + \mathbf{v}_n(m),\ n = 1, 2, \ldots, N, \tag{2}$$

where $m \geq 0$ is the time-frame index, $\mathbf{y}_n(m) = [\, y_n(mL) \quad y_n(mL+1) \quad \cdots \quad y_n(mL+L-1)\,]^T$ is a vector of length L, superscript ^{T} denotes transpose of a vector or a matrix, and $\mathbf{x}_n(m)$ and $\mathbf{v}_n(m) = \mathbf{a}_n(m) + \mathbf{b}_n(m)$ are defined in a similar way to $\mathbf{y}_n(m)$. Let us define the stacked vector

$$\begin{aligned} \underline{\mathbf{y}}(m) &= [\, \mathbf{y}_1^T(m) \quad \mathbf{y}_2^T(m) \quad \cdots \quad \mathbf{y}_N^T(m)\,]^T \\ &= \underline{\mathbf{x}}(m) + \underline{\mathbf{v}}(m), \end{aligned} \tag{3}$$

where $\underline{\mathbf{x}}(m)$ and $\underline{\mathbf{v}}(m) = \underline{\mathbf{a}}(m) + \underline{\mathbf{b}}(m)$ are defined in a similar way to $\underline{\mathbf{y}}(m)$.

Since $x_n(k)$ and $v_n(k)$ are uncorrelated by assumption, the correlation matrix (of size $NL \times NL$) of the stacked microphone signals is

$$\begin{aligned} \mathbf{R}_{\underline{\mathbf{y}}} &= E\left[\underline{\mathbf{y}}(m)\underline{\mathbf{y}}^T(m)\right] \\ &= \mathbf{R}_{\underline{\mathbf{x}}} + \mathbf{R}_{\underline{\mathbf{v}}}, \end{aligned} \tag{4}$$

where $E[\cdot]$ denotes mathematical expectation, and $\mathbf{R}_{\underline{\mathbf{x}}} = E[\underline{\mathbf{x}}(m)\underline{\mathbf{x}}^T(m)]$ and $\mathbf{R}_{\underline{\mathbf{v}}} = E[\underline{\mathbf{v}}(m)\underline{\mathbf{v}}^T(m)]$ are the correlation matrices of $\underline{\mathbf{x}}(m)$ and $\underline{\mathbf{v}}(m)$, respectively. Note that since $a_n(k)$ and $b_n(k)$ are also uncorrelated, it follows that $\mathbf{R}_{\underline{\mathbf{v}}} = E[\underline{\mathbf{a}}(m)\underline{\mathbf{a}}^T(m)] + E[\underline{\mathbf{b}}(m)\underline{\mathbf{b}}^T(m)] = \mathbf{R}_{\underline{\mathbf{a}}} + \mathbf{R}_{\underline{\mathbf{b}}}$.

In this paper, our desired signal is designated by the clean (but convolved) speech signal received at microphone 1, namely $x_1(k) = g_1(k) * s(k)$ (obviously, any signal $x_n(k)$ could be considered as the reference). Our problem then may be stated

as follows [20]: given N mixtures of two uncorrelated signals $x_n(k)$ and $v_n(k)$, our aim is to preserve $x_1(k)$ while minimizing the contribution of the noise terms $v_n(k)$ at the array output.

III. Karhunen-Loève Expansion (KLE)

As explained in [11], [22], [23], it may be advantageous to perform noise reduction in the KLE domain. In this section, we briefly recall the principle of the KLE which can be applied to $\underline{\mathbf{y}}(m)$, $\underline{\mathbf{x}}(m)$, or $\underline{\mathbf{v}}(m)$. In this study, we choose to apply it to $\underline{\mathbf{y}}(m)$ while the same concept was developed for $\underline{\mathbf{x}}(m)$ in [11], [22], [23] but in the single-channel case. Fundamentally, we should not expect much difference if we apply the KLE to $\underline{\mathbf{y}}(m)$ or $\underline{\mathbf{x}}(m)$ but, in the context of speech enhancement, it is preferable to apply it to the former as the corresponding covariance matrix is usually full rank, while the clean speech covariance matrix can be either rank deficient or ill-conditioned [4], [24]. Let us first diagonalize the correlation matrix $\mathbf{R}_{\underline{\mathbf{y}}}$ as follows [25]

$$\mathbf{Q}^T\mathbf{R}_{\underline{\mathbf{y}}}\mathbf{Q} = \Lambda, \tag{5}$$

where

$$\mathbf{Q} = [\,\mathbf{q}_1 \quad \mathbf{q}_2 \quad \cdots \quad \mathbf{q}_{NL}\,] \tag{6}$$

and

$$\Lambda = \text{diag}\,([\,\lambda_1 \quad \lambda_2 \quad \ldots \quad \lambda_{NL}\,]) \tag{7}$$

are, respectively, orthogonal and diagonal matrices. The orthonormal vectors $\mathbf{q}_l = [\,q_{l,1} \quad q_{l,2} \quad \ldots \quad q_{l,NL}\,]^T$, for $l = 1, \ldots, NL$, are the eigenvectors corresponding, respectively, to the eigenvalues λ_l of the matrix $\mathbf{R}_{\underline{\mathbf{y}}}$. The vector $\underline{\mathbf{y}}(m)$ can be written as a combination (expansion) of the eigenvectors of the correlation matrix $\mathbf{R}_{\underline{\mathbf{y}}}$ as follows

$$\underline{\mathbf{y}}(m) = \sum_{l=1}^{NL} c_{\underline{\mathbf{y}},l}(m)\mathbf{q}_l, \tag{8}$$

where

$$c_{\underline{\mathbf{y}},l}(m) = \mathbf{q}_l^T\underline{\mathbf{y}}(m) \tag{9}$$

are the coefficients of the expansion and l is the mode index. The representation of the vector $\underline{\mathbf{y}}(m)$ described by (8) and (9) is the Karhunen-Loève expansion (KLE) [26]. Equations (8) and (9) are, respectively, the synthesis and analysis parts of this expansion. From (9), we can verify that

$$E\left[c_{\underline{\mathbf{y}},l}(m)\right] = 0 \tag{10}$$

and

$$E\left[c_{\underline{\mathbf{y}},i}(m)c_{\underline{\mathbf{y}},j}(m)\right] = \begin{cases} \lambda_i, & i = j \\ 0, & i \neq j \end{cases}. \tag{11}$$

It can also be checked from (9) that

$$\sum_{l=1}^{NL} c_{\underline{\mathbf{y}},l}^2(m) = \left\|\underline{\mathbf{y}}(m)\right\|_2^2, \tag{12}$$

where $\|\underline{\mathbf{y}}(m)\|_2$ is the Euclidean norm of $\underline{\mathbf{y}}(m)$. The previous expression shows the energy conservation through the KLE process.

We also define

$$c_{\underline{\mathbf{x}},l}(m) = \mathbf{q}_l^T\underline{\mathbf{x}}(m), \tag{13}$$

$$c_{\underline{\mathbf{v}},l}(m) = \mathbf{q}_l^T\underline{\mathbf{v}}(m). \tag{14}$$

We can check that

$$\sum_{l=1}^{NL} c_{\underline{\mathbf{x}},l}^2(m) = \left\|\underline{\mathbf{x}}(m)\right\|_2^2, \tag{15}$$

$$\sum_{l=1}^{NL} c_{\underline{\mathbf{v}},l}^2(m) = \left\|\underline{\mathbf{v}}(m)\right\|_2^2. \tag{16}$$

From (11), we see that the inter-mode correlation of the coefficients $c_{\underline{\mathbf{y}},l}(m)$ is equal to 0. But the inter-mode correlations of the coefficients $c_{\underline{\mathbf{x}},l}(m)$ and $c_{\underline{\mathbf{v}},l}(m)$ are

$$E\left[c_{\underline{\mathbf{x}},i}(m)c_{\underline{\mathbf{x}},j}(m)\right] = \mathbf{q}_i^T\mathbf{R}_{\underline{\mathbf{x}}}\mathbf{q}_j,\ i \neq j, \tag{17}$$

$$E\left[c_{\underline{\mathbf{v}},i}(m)c_{\underline{\mathbf{v}},j}(m)\right] = \mathbf{q}_i^T\mathbf{R}_{\underline{\mathbf{v}}}\mathbf{q}_j,\ i \neq j, \tag{18}$$

which might not necessarily be equal to 0. If the noise is temporally and spatially white, the noise covariance matrix is a diagonal matrix. In this case, it can be easily shown that the inter-mode correlations are equal to 0 (assuming that the desired signal, i.e., speech, is always correlated which is usually the case).

Left multiplying both sides of (2) by $\mathbf{q}_l^T$, the time-domain signal model is transformed into the KLE domain as

$$c_{\underline{\mathbf{y}},l}(m) = c_{\underline{\mathbf{x}},l}(m) + c_{\underline{\mathbf{v}},l}(m). \tag{19}$$

Now, let us define the vector

$$\underline{\mathbf{q}}_{n,l} = [\,q_{(n-1)L+1,l} \quad q_{(n-1)L+2,l} \quad \cdots \quad q_{(n-1)L+L,l}\,]^T, \tag{20}$$

for $n = 1, \ldots, N$ and $l = 1, \ldots, LN$. It follows that

$$c_{\underline{\mathbf{y}},l}(m) = \sum_{n=1}^{N} \underline{\mathbf{q}}_{n,l}^T\mathbf{y}_n(m) = \sum_{n=1}^{N} c_{y_n,l}(m), \tag{21}$$

where $c_{y_n,l}(m) = \underline{\mathbf{q}}_{n,l}^T\mathbf{y}_n(m)$. Thus, the coefficients $c_{\underline{\mathbf{y}},l}(m)$ are a linear combination of the sub-coefficients $c_{y_n,l}(m)$. The sub-coefficient $c_{y_n,l}(m)$ can be seen as the coefficient corresponding to the nth-microphone. Applying the same expansion to $c_{\underline{\mathbf{x}},l}(m)$ and $c_{\underline{\mathbf{v}},l}(m)$ we obtain the sub-coefficients

$$c_{x_n,l}(m) = \underline{\mathbf{q}}_{n,l}^T\mathbf{x}_n(m), \tag{22}$$

$$\begin{aligned} c_{v_n,l}(m) &= \underline{\mathbf{q}}_{n,l}^T\underline{\mathbf{a}}_n(m) + \underline{\mathbf{q}}_{n,l}^T\mathbf{b}_n(m) \\ &= c_{a_n,l}(m) + c_{b_n,l}(m). \end{aligned} \tag{23}$$

The multichannel noise reduction in the KLE domain comes to the estimation of the coefficients $c_{x_1,l}(m)$, for $l = 1, 2, \ldots, NL$, from the observations $c_{y_n,l}(m)$, for $n = 1, 2, \ldots, N$. The variance of the coefficients $c_{y_n,l}(m)$ is then

$$\begin{aligned} \phi_{c_{y_n,l}} &= E\left[c_{y_n,l}^2(m)\right] \\ &= \phi_{c_{x_n,l}} + \phi_{c_{v_n,l}}, \end{aligned} \tag{24}$$

where $\phi_{c_{x_n,l}} = E[c^2_{x_n,l}(m)]$ and $\phi_{c_{v_n,l}} = \phi_{c_{a_n,l}} + \phi_{c_{b_n,l}} = E[c^2_{a_n,l}(m)] + E[c^2_{b_n,l}(m)]$ are the variances of $c_{x_n,l}(m)$ and $c_{v_n,l}(m)$, respectively. By applying the expansion in (21), we can not longer assume that the inter-mode correlations of the sub-coefficients $c_{y_n,l}(m)$ equal 0. That is $E[c_{y_n,i}(m)c_{y_n,j}(m)] \neq 0$ for $i \neq j$. Thus, in order to optimally use the coefficients, we need to exploit the inter-mode correlations. Let us define the vectors

$$\mathbf{c}_{y_n,l}(m) = [\, c_{y_n,f(l,0)}(m) \quad c_{y_n,f(l,1)}(m) \quad \cdots \quad c_{y_n,f(l,F-1)}(m) \,]^T, \tag{25}$$

where the function $f(l,p)$, $p = 0, 1, \ldots, F-1$, describes which inter-mode correlations are exploited, and F is the total number of modes that is used for that purpose. Note that if we use all modes, this function takes the form $f(l,p) = l + p$ with $F = NL$. However, as shown later, not all modes might be necessary for a near optimal performance. In the following, we use the subindex $f(l,p)$ for the sake of generality.

IV. Linear Array Model

Usually, in the time domain, the array processing or beamforming is performed by applying a temporal filter to each microphone signal and summing the filtered signals. In the KLE domain, we are going to focus on the simplest linear model for array processing, which is realized by applying a real weight to the output of each sensor and summing across the aperture, i.e.,

$$\begin{aligned} c_{z,l}(m) &= \sum_{n=1}^{N} \mathbf{h}^T_{n,l}\mathbf{c}_{y_n,l}(m) \\ &= \mathbf{h}^T_{:,l}\mathbf{c}_{y_:,l}(m) \\ &= \mathbf{h}^T_{:,l}\mathbf{c}_{x_:,l}(m) + \mathbf{h}^T_{:,l}\mathbf{c}_{v_:,l}(m) \\ &= c_{x_\mathrm{f},l}(m) + c_{v_\mathrm{rn},l}(m), \end{aligned} \tag{26}$$

where $c_{z,l}(m)$, which is an estimate of $c_{x_1,l}(m)$, is the beamformer output signal,

$$\mathbf{h}_{n,l} = [\, h_{n,f(l,0)} \quad h_{n,f(l,1)} \quad \cdots \quad h_{n,f(l,F-1)} \,]^T \tag{27}$$

is an FIR filter of length F, corresponding to the mode index l and microphone signal n and

$$\mathbf{h}_{:,l} = [\, \mathbf{h}^T_{1,l} \quad \mathbf{h}^T_{2,l} \quad \ldots \quad \mathbf{h}^T_{N,l} \,]^T \tag{28}$$

is the beamforming weight vector (of size NF), which is suitable for performing spatial filtering at the mode index l, $\mathbf{c}_{y_:,l}(m) = [\, \mathbf{c}^T_{y_1,l}(m) \quad \mathbf{c}^T_{y_2,l}(m) \quad \cdots \quad \mathbf{c}^T_{y_N,l}(m) \,]^T$ is a vector of length NF containing the observations from all sensors at time-frame index m, $\mathbf{c}_{x_:,l}(m)$ and $\mathbf{c}_{v_:,l}(m)$ are defined in a similar way to $\mathbf{c}_{y_:,l}(m)$, and $c_{x_\mathrm{f},l}(m) = \mathbf{h}^T_{:,l}\mathbf{c}_{x_:,l}(m)$ and $c_{v_\mathrm{rn},l}(m) = \mathbf{h}^T_{:,l}\mathbf{c}_{v_:,l}(m)$ are, respectively, the filtered speech signal and residual noise in the KLE domain.

At time-frame index m, our desired signal is $c_{x_1,l}(m)$ (and not the whole the vector $\mathbf{c}_{x_:,l}(m)$). However, the vector $\mathbf{c}_{x_:,l}(m)$ contains both the desired signal, $c_{x_1,l}(m)$, and the components $c_{x_1,f(l,p)}(m)$ and $\mathbf{c}_{x'_n,l}$ for $p = 1, \ldots, F$ and $n' = 2 \ldots N$ respectively, which are not the desired signals but signals that are correlated with $c_{x_1,l}(m)$. Therefore, the elements $c_{x_1,f(l,p)}(m)$ and $\mathbf{c}_{x'_n,l}$ contain both a part of the desired signal and a component that we consider as an interference. This suggests that we should decompose $\mathbf{c}_{x_:,l}(m)$ into two orthogonal vectors corresponding to the part of the desired signal and interference, i.e.,

$$\begin{aligned} \mathbf{c}_{x_:,l}(m) &= c_{x_1,l}(m)\boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}} + \mathbf{c}'_{x_:,l}(m) \\ &= \mathbf{c}_{x_\mathrm{d},l}(m) + \mathbf{c}'_{x_:,l}(m), \end{aligned} \tag{29}$$

where $\mathbf{c}_{x_\mathrm{d},l}(m) = c_{x_1,l}(m)\boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}$ is a signal vector depending on the desired signal $c_{x_1,l}(m)$, $\mathbf{c}'_{x_:,l}(m) = [\, \mathbf{c}'^T_{x_1,l}(m) \quad \mathbf{c}'^T_{x_2,l}(m) \quad \cdots \quad \mathbf{c}'^T_{x_N,l}(m) \,]^T$ is the interference signal vector, $\mathbf{c}'_{x_n,l}(m) = \mathbf{c}_{x_n,l}(m) - c_{x_1,l}(m)\boldsymbol{\gamma}'_{\mathbf{c}_{x_n}}$ is the interference sub-vector for each channel, $\boldsymbol{\gamma}'_{\mathbf{c}_{x_n,l}} = [\, \gamma_{c_{x_n,f(l,0)}} \quad \gamma_{c_{x_n,f(l,1)}} \quad \cdots \quad \gamma_{c_{x_n,f(l,F-1)}} \,]^T$ is a vector with the partially normalized (with respect to $c_{x_1,l}(m)$) cross-correlation coefficients between the signals $\mathbf{c}_{x_n}(m)$ and $c_{x_1,l}(m)$, and

$$\begin{aligned} \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}} &= [\, \boldsymbol{\gamma}'^T_{c_{x_1,l}} \quad \boldsymbol{\gamma}'^T_{c_{x_2,l}} \quad \cdots \quad \boldsymbol{\gamma}'^T_{c_{x_N,l}} \,]^T \\ &= [\, 1 \quad \gamma_{c_{x_1,f(l,1)}} \quad \cdots \\ &\qquad \gamma_{c_{x_1,f(l,F-1)}} \quad \boldsymbol{\gamma}'^T_{c_{x_2,l}} \quad \cdots \quad \boldsymbol{\gamma}'^T_{c_{x_N,l}} \,]^T \\ &= \frac{E\left[c_{x_1,l}(m)\mathbf{c}_{x_:,l}(m)\right]}{E\left[c^2_{x_1,l}(m)\right]} \end{aligned} \tag{30}$$

is the partially normalized (with respect to $c_{x_1,l}(m)$) cross-correlation vector (of length NF) between $c_{x_1,l}(m)$ and $\mathbf{c}_{x_:,l}(m)$.

The vector $\boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}$ can be seen as the steering vector or direction vector since it determines the direction of the desired signal $c_{x_1,l}(m)$. This definition is a generalization of the classical steering vector [17], [27], [28] in the KLE domain.

Substituting (29) into (26), we get

$$c_{z,l}(m) = c_{x_1,l}(m)\mathbf{h}^T_{:,l}\boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}} + \mathbf{h}^T_{:,l}\mathbf{c}'_{x_:,l}(m) + \mathbf{h}^T_{:,l}\mathbf{c}_{v_:,l}(m). \tag{31}$$

We observe that the estimate of the desired signal is the sum of three terms that are mutually uncorrelated. The first one is clearly the filtered desired signal while the two others are the filtered undesired signals (interference-plus-noise). Therefore, the variance of $c_{z,l}(m)$ is

$$\begin{aligned} \phi_{c_{z,l}} &= \mathbf{h}^T_{:,l}\Phi_{\mathbf{c}_{y_:,l}}\mathbf{h}_{:,l} \\ &= \mathbf{h}^T_{:,l}\Phi_{\mathbf{c}_{x_\mathrm{d},l}}\mathbf{h}_{:,l} + \mathbf{h}^T_{:,l}\Phi_{\mathbf{c}'_{x_:,l}}\mathbf{h}_{:,l} + \mathbf{h}^T_{:,l}\Phi_{\mathbf{c}_{v_:,l}}\mathbf{h}_{:,l}, \end{aligned} \tag{32}$$

where

$$\Phi_{\mathbf{c}_{y_:,l}} = E\left[\mathbf{c}_{y_:,l}(m)\mathbf{c}^T_{y_:,l}(m)\right], \tag{33}$$

$$\begin{aligned} \Phi_{\mathbf{c}_{x_\mathrm{d},l}} &= E\left[\mathbf{c}_{x_\mathrm{d},l}(m)\mathbf{c}^T_{x_\mathrm{d},l}(m)\right] \\ &= \phi_{c_{x_1,l}}\boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}\boldsymbol{\gamma}^T_{\mathbf{c}_{x_:,l}}, \end{aligned} \tag{34}$$

$$\begin{aligned} \Phi_{\mathbf{c}'_{x_:,l}} &= E\left[\mathbf{c}'_{x_:,l}(m)\mathbf{c}'^T_{x_:,l}(m)\right] \\ &= \Phi_{\mathbf{c}_{x_:,l}} - \Phi_{\mathbf{c}_{x_\mathrm{d},l}}, \end{aligned} \tag{35}$$

$$\Phi_{\mathbf{c}_{v_:,l}} = E\left[\mathbf{c}_{v_:,l}(m)\mathbf{c}^T_{v_:,l}(m)\right], \tag{36}$$

are the correlation matrices of the vectors $\mathbf{c}_{y_{:,l}}(m)$, $\mathbf{c}_{x_{\mathrm{d}},l}(m)$, $\mathbf{c}'_{x_{:,l}}(m)$, and $\mathbf{c}_{v_{:,l}}(m)$, respectively.

The estimate of the vector $\mathbf{x}_1(m)$ would be

$$\begin{aligned}\mathbf{z}(m) &= \sum_{l=1}^{NL} c_{z,l}(m)\underline{\mathbf{q}}_{1,l} \\ &= \sum_{l=1}^{NL}\sum_{n=1}^{N} \mathbf{h}_{n,l}^T \mathbf{c}_{y_n,l}(m)\underline{\mathbf{q}}_{1,l} \\ &= \sum_{n=1}^{N}\sum_{l=1}^{NL} \underline{\mathbf{q}}_{1,l}\mathbf{h}_{n,l}^T \mathbf{Q}'^T_{n,l}\mathbf{y}_n(m) \\ &= \sum_{n=1}^{N} \mathbf{H}_{\mathrm{TD},n}\mathbf{y}_n(m), \end{aligned} \tag{37}$$

where

$$\mathbf{H}_{\mathrm{TD},n} = \sum_{l=1}^{NL} \underline{\mathbf{q}}_{1,l}\mathbf{h}_{l,n}^T \mathbf{Q}'^T_{n,l}, \tag{38}$$

for $n = 1, \ldots, N$, are the time-domain filtering matrices of size $L \times L$ and $\mathbf{Q}'_{n,l} = [\,\underline{\mathbf{q}}_{n,1} \quad \underline{\mathbf{q}}_{n,2} \quad \cdots \quad \underline{\mathbf{q}}_{n,NL}\,]$. We see from (37) how the estimation of $\mathbf{x}_1(m)$ depends on the observation vectors $\mathbf{y}_n(m)$, $n = 1, 2, \ldots, N$. The correlation matrix of $\mathbf{z}(m)$ is

$$\mathbf{R}_{\mathbf{z}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{H}_{\mathrm{TD},i} E\left[\mathbf{y}_i(m)\mathbf{y}_j^T(m)\right]\mathbf{H}_{\mathrm{TD},j}^T. \tag{39}$$

V. Performance Measures

In this section, we define some useful performance measures that allow us to study, within this framework, the different multichannel noise reduction algorithms in the KLE domain developed later in this paper. Since the signal we want to recover is the clean (but convolved) signal received at microphone 1, i.e., $x_1(k)$, the first microphone is chosen as the reference sensor.

To examine what happens in each mode, we define the mode input SNR as

$$\mathrm{iSNR}_l = \frac{\phi_{c_{x_1,l}}}{\phi_{c_{v_1,l}}} = \frac{\underline{\mathbf{q}}_{1,l}^T \mathbf{R}_{\mathbf{x}_1}\underline{\mathbf{q}}_{1,l}}{\underline{\mathbf{q}}_{1,l}^T \mathbf{R}_{\mathbf{v}_1}\underline{\mathbf{q}}_{1,l}}, \tag{40}$$

where $\mathbf{R}_{\mathbf{x}_1} = E[\mathbf{x}_1(m)\mathbf{x}_1^T(m)]$ and $\mathbf{R}_{\mathbf{v}_1} = E[\mathbf{v}_1(m)\mathbf{v}_1^T(m)]$. The fullmode input SNR is

$$\mathrm{iSNR} = \frac{\sum_{l=1}^{NL} \underline{\mathbf{q}}_{1,l}^T \mathbf{R}_{\mathbf{x}_1}\underline{\mathbf{q}}_{1,l}}{\sum_{l=1}^{NL} \underline{\mathbf{q}}_{1,l}^T \mathbf{R}_{\mathbf{v}_1}\underline{\mathbf{q}}_{1,l}} = \frac{\sigma_{x_1}^2}{\sigma_{v_1}^2}, \tag{41}$$

where $\sigma_{x_1}^2 = E[x_1^2(k)]$ and $\sigma_{v_1}^2 = E[v_1^2(k)]$ are the variances of $x_1(k)$ and $v_1(k)$, respectively.

The output SNR is the SNR after the filtering operation. The mode output SNR is defined as[1]

$$\mathrm{oSNR}(\mathbf{h}_{:,l}) = \frac{\mathbf{h}_{:,l}^T \Phi_{\mathbf{c}_{x_{\mathrm{d}},l}}\mathbf{h}_{:,l}}{\mathbf{h}_{:,l}^T \Phi_{\mathrm{in},l}\mathbf{h}_{:,l}} = \frac{\phi_{c_{x_1,l}}\left(\mathbf{h}_{:,l}^T\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}\right)^2}{\mathbf{h}_{:,l}^T \Phi_{\mathrm{in},l}\mathbf{h}_{:,l}}, \tag{42}$$

where

$$\Phi_{\mathrm{in},l} = \Phi_{\mathbf{c}'_{x_{:,l}}} + \Phi_{\mathbf{c}_{v_{:,l}}}, \tag{43}$$

is the interference-plus-noise correlation matrix. For the particular filter $\mathbf{h}_{:,l} = \mathbf{i}_1$, where $\mathbf{i}_1$ is the first column of the identity matrix $\mathbf{I}_{NF}$ of size $NF \times NF$, we have

$$\mathrm{oSNR}(\mathbf{i}_1) = \mathrm{iSNR}_l, \tag{44}$$

which means that with the identity filter $\mathbf{i}_1$, the SNR cannot be improved.

For any two vectors $\mathbf{h}_{:,l}$ and $\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}$ and a positive definite matrix $\Phi_{\mathrm{in},\,l}$, we have

$$\left(\mathbf{h}_{:,l}^T\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}\right)^2 \le \left(\mathbf{h}_{:,l}^T\Phi_{\mathrm{in},l}\mathbf{h}_{:,l}\right)\left(\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}^T\Phi_{\mathrm{in},l}^{-1}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}\right). \tag{45}$$

Using the previous inequality in (42), we deduce an upper bound for the mode output SNR:

$$\mathrm{oSNR}(\mathbf{h}_{:,l}) \le \phi_{c_{x_1,l}}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}^T\Phi_{\mathrm{in},l}^{-1}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}. \tag{46}$$

We define the mode array gain as the ratio of the mode output SNR (after beamforming) over the mode input SNR (at the reference microphone) [27], [17], i.e.,

$$\mathcal{A}(\mathbf{h}_{:,l}) = \frac{\mathrm{oSNR}(\mathbf{h}_{:,l})}{\mathrm{iSNR}_l}. \tag{47}$$

From (46), we deduce that the maximum mode array gain is

$$\mathcal{A}_{\max,l} = \phi_{c_{v_1,l}}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}^T\Phi_{\mathrm{in},l}^{-1}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}. \tag{48}$$

We define the fullmode output SNR as

$$\mathrm{oSNR}(\mathbf{h}_{:}) = \frac{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}\left(\mathbf{h}_{:,l}^T\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}\right)^2}{\sum_{l=1}^{L}\mathbf{h}_{:,l}^T\Phi_{\mathrm{in},l}\mathbf{h}_{:,l}}. \tag{49}$$

The mode and fullmode noise reduction factors are [2], [4]

$$\xi_{\mathrm{nr}}(\mathbf{h}_{:,l}) = \frac{\phi_{c_{v_1,l}}}{\mathbf{h}_{:,l}^T\Phi_{\mathrm{in},l}\mathbf{h}_{:,l}}, \tag{50}$$

$$\xi_{\mathrm{nr}}(\mathbf{h}_{:}) = \frac{\sum_{l=1}^{NL}\phi_{c_{v_1,l}}}{\sum_{l=1}^{NL}\mathbf{h}_{:,l}^T\Phi_{\mathrm{in},l}\mathbf{h}_{:,l}}. \tag{51}$$

These factors should be lower bounded by 1 for optimal filters.

To quantify the speech distortion [2], [4], we give the mode speech distortion index

$$\begin{aligned}\upsilon_{\mathrm{sd}}(\mathbf{h}_{:,l}) &= \frac{E\left\{\left[c_{x_1,l}(m)\mathbf{h}_{:,l}^T\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}} - c_{x_1,l}(m)\right]^2\right\}}{\phi_{c_{x_1,l}}} \\ &= \left(\mathbf{h}_{:,l}^T\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}} - 1\right)^2, \end{aligned} \tag{52}$$

[1]In this study, we consider the interference as part of the noise in the definitions of the performance measures.

and the fullmode speech distortion index

$$\upsilon_{\rm sd}(\mathbf{h}_{:}) = \frac{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}\left(\mathbf{h}_{:,l}^T\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}-1\right)^2}{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}} = \frac{\sum_{l=1}^{NL}\upsilon_{\rm sd}(\mathbf{h}_{:,l})\phi_{c_{x_1,l}}}{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}}. \quad (53)$$

The speech distortion index is usually upper bounded by 1.

We can also quantify signal distortion via the mode and fullmode speech reduction factors which are defined as [22], [28]

$$\xi_{\rm sr}(\mathbf{h}_{:,l}) = \frac{\phi_{c_{x_1,l}}}{\phi_{c_{x_1,l}}\left(\mathbf{h}_{:,l}^T\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}\right)^2} = \frac{1}{\left(\mathbf{h}_{:,l}^T\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}\right)^2}, \quad (54)$$

$$\xi_{\rm sr}(\mathbf{h}_{:}) = \frac{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}}{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}\left(\mathbf{h}_{:,l}^T\,\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}\right)^2} = \frac{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}}{\sum_{l=1}^{NL}\xi_{\rm sr}^{-1}(\mathbf{h}_{:,l})\phi_{c_{x_1,l}}}. \quad (55)$$

A key observation from (52) or (54) is that the design of a noise reduction algorithm in the KLE domain that does not distort the desired signal requires the constraint

$$\mathbf{h}_{:,l}^T\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}} = 1,\ \forall l. \quad (56)$$

It can be shown that

$$\frac{\rm oSNR(\mathbf{h}_{:,l})}{\rm iSNR_l} = \frac{\xi_{\rm nr}(\mathbf{h}_{:,l})}{\xi_{\rm sr}(\mathbf{h}_{:,l})}, \quad (57)$$

$$\frac{\rm oSNR(\mathbf{h}_{:})}{\rm iSNR} = \frac{\xi_{\rm nr}(\mathbf{h}_{:})}{\xi_{\rm sr}(\mathbf{h}_{:})}. \quad (58)$$

For the multichannel case, it is also of interest to know the performance of the filters with respect to spatially coherent and incoherent noise separately. Let us first rewrite (43) as follows

$$\boldsymbol{\Phi}_{{\rm in},l} = \boldsymbol{\Phi}_{{\rm coh}',l} + \boldsymbol{\Phi}_{\mathbf{c}_{b_{:,l}}}, \quad (59)$$

where

$$\boldsymbol{\Phi}_{{\rm coh}',l} = \boldsymbol{\Phi}_{\mathbf{c}'_{x_{:,l}}} + \boldsymbol{\Phi}_{\mathbf{c}_{a_{:,l}}}, \quad (60)$$

$$\boldsymbol{\Phi}_{\mathbf{c}_{b_{:,l}}} = E\left[\mathbf{c}_{b_{:,l}}(m)\mathbf{c}_{b_{:,l}}^T(m)\right], \quad (61)$$

are the interference-plus-coherent-noise and incoherent-noise correlation matrices respectively[2]. The matrix $\boldsymbol{\Phi}_{\mathbf{c}_{a_{:,l}}} = E[\mathbf{c}_{a_{:,l}}(m)\mathbf{c}_{a_{:,l}}^T(m)]$ is the coherent-noise correlation matrix.

[2]Note that we omit the term "spatially" for simplicity.

The mode coherent and incoherent noise reduction factors are, respectively,

$$\xi_{\rm cnr}(\mathbf{h}_{:,l}) = \frac{\phi_{c_{a_1,l}}}{\mathbf{h}_{:,l}^T\boldsymbol{\Phi}_{{\rm in}',l}\mathbf{h}_{:,l}}, \quad (62)$$

$$\xi_{\rm inr}(\mathbf{h}_{:,l}) = \frac{\phi_{c_{b_1,l}}}{\mathbf{h}_{:,l}^T\boldsymbol{\Phi}_{{\rm b},l}\mathbf{h}_{:,l}}. \quad (63)$$

Using (62) and (63), we can rewrite (50) as

$$\xi_{\rm nr}(\mathbf{h}_{:,l}) = \frac{\phi_{c_{v_1,l}}}{\frac{\phi_{c_{a_1,l}}}{\xi_{\rm cnr}(\mathbf{h}_{:,l})}+\frac{\phi_{c_{b_1,l}}}{\xi_{\rm inr}(\mathbf{h}_{:,l})}}. \quad (64)$$

The full-mode coherent and incoherent noise reduction factors are, respectively,

$$\xi_{\rm cnr}(\mathbf{h}_{:,l}) = \frac{\sum_{l=1}^{NL}\phi_{c_{a_1,l}}}{\sum_{l=1}^{NL}\mathbf{h}_{:,l}^T\boldsymbol{\Phi}_{{\rm in}',l}\mathbf{h}_{:,l}}, \quad (65)$$

$$\xi_{\rm inr}(\mathbf{h}_{:,l}) = \frac{\sum_{l=1}^{NL}\phi_{c_{b_1,l}}}{\sum_{l=1}^{NL}\mathbf{h}_{:,l}^T\boldsymbol{\Phi}_{{\rm b},l}\mathbf{h}_{:,l}}. \quad (66)$$

VI. Optimal Noise Reduction Filters

In this section we derive different optimal noise reduction filters in the KLE domain. The classical noise reduction filtering techniques is formulated for the multichannel case in the KLE domain and their performance is discussed.

A. Maximum SNR Filter

The maximum SNR filter, $\mathbf{h}_{{\rm max},l}$, is obtained by maximizing the mode output SNR as defined in (42) [16]. Therefore, $\mathbf{h}_{{\rm max},l}$ is the eigenvector corresponding to the maximum eigenvalue of the matrix $\boldsymbol{\Phi}_{{\rm in},l}^{-1}\boldsymbol{\Phi}_{\mathbf{c}_{x_{\rm d},l}}$. Let us denote this eigenvalue by $\lambda_{{\rm max},l}$. Since the rank of the matrix $\boldsymbol{\Phi}_{\mathbf{c}_{x_{\rm d},l}}$ is equal to 1, we have

$$\lambda_{{\rm max},l} = {\rm tr}\left(\boldsymbol{\Phi}_{{\rm in},l}^{-1}\boldsymbol{\Phi}_{\mathbf{c}_{x_{\rm d},l}}\right) = \phi_{c_{x_1,l}}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}^T\boldsymbol{\Phi}_{{\rm in},l}^{-1}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}, \quad (67)$$

where ${\rm tr}(\cdot)$ denotes the trace of a square matrix. As a result,

$$\rm oSNR(\mathbf{h}_{{\rm max},l}) = \phi_{c_{x_1,l}}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}^T\boldsymbol{\Phi}_{{\rm in},l}^{-1}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}, \quad (68)$$

which corresponds to the maximum possible mode output SNR according to the inequality in (46). We also have

$$\mathbf{h}_{{\rm max},l} = \alpha_l\boldsymbol{\Phi}_{{\rm in},l}^{-1}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}, \quad (69)$$

where α_l is an arbitrary scaling factor different from zero. While this factor has no effect on the mode output SNR, it has on the fullmode output SNR and speech distortion (mode and fullmode). In fact, all filters derived in the rest of this paper are equivalent up to this scaling factor. These filters also try to find the respective scaling factors depending on what we optimize.

B. Mean-Square Error (MSE) Criterion

The error signal between the estimated and desired signals in the mode l is

$$\begin{aligned} e_l(m) &= c_{z,l}(m) - c_{x_1,l}(m) \\ &= \mathbf{h}_{:,l}^T \mathbf{c}_{y_:,l}(m) - c_{x_1,l}(m). \end{aligned} \tag{70}$$

This error signal can also be written as the sum of two uncorrelated error signals:

$$e_l(m) = e_{\mathrm{d},l}(m) + e_{\mathrm{r},l}(m), \tag{71}$$

where

$$\begin{aligned} e_{\mathrm{d},l}(m) &= \mathbf{h}_{:,l}^T \mathbf{c}_{x_\mathrm{d},l}(m) - c_{x_1,l}(m) \\ &= \left(\mathbf{h}_{:,l}^T \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}} - 1\right) c_{x_1,l}(m) \end{aligned} \tag{72}$$

is the speech distortion due to the filter and

$$e_{\mathrm{r},l}(m) = \mathbf{h}_{:,l}^T \mathbf{c}'_{x_:,l}(m) + \mathbf{h}_{:,l}^T \mathbf{c}_{v_:,l}(m) \tag{73}$$

represents the residual interference-plus-noise.

The mode MSE criterion is then [16]

$$\begin{aligned} J\left(\mathbf{h}_{:,l}\right) &= E\left[e_l^2(m)\right] \\ &= \mathbf{h}_{:,l}^T \boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}} \mathbf{h}_{:,l} - 2\mathbf{h}_{:,l}^T \boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}\mathbf{c}_{x_:,l}} \mathbf{i}_1 + \phi_{c_{x_1,l}}, \end{aligned} \tag{74}$$

where

$$\begin{aligned} \boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}\mathbf{c}_{x_:,l}} &= E\left[\mathbf{c}_{y_:,l}(m)\mathbf{c}_{x_:,l}^T(m)\right] \\ &= E\left[\mathbf{c}_{x_:,l}(m)\mathbf{c}_{x_:,l}^T(m)\right] \\ &= \boldsymbol{\Phi}_{\mathbf{c}_{x_:,l}} \end{aligned}$$

is the cross-correlation matrix between the two signal vectors $\mathbf{c}_{y_:,l}(m)$ and $\mathbf{c}_{x_:,l}(m)$. We can rewrite the mode MSE as

$$J\left(\mathbf{h}_{:,l}\right) = J_\mathrm{d}\left(\mathbf{h}_{:,l}\right) + J_\mathrm{r}\left(\mathbf{h}_{:,l}\right),$$

where

$$\begin{aligned} J_\mathrm{d}\left(\mathbf{h}_{:,l}\right) &= E\left[e_{\mathrm{d},l}^2(m)\right] \\ &= \phi_{c_{x_1,l}} \left(\mathbf{h}_{:,l}^T \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}} - 1\right)^2 \end{aligned} \tag{75}$$

and

$$\begin{aligned} J_\mathrm{r}\left(\mathbf{h}_{:,l}\right) &= E\left[e_{\mathrm{r},l}^2(m)\right] \\ &= \mathbf{h}_{:,l}^T \boldsymbol{\Phi}_{\mathrm{in},l} \mathbf{h}_{:,l}. \end{aligned} \tag{76}$$

For the particular filter $\mathbf{h}_{:,l} = \mathbf{i}_1,\ \forall l$, we get

$$J\left(\mathbf{i}_1\right) = \phi_{c_{v_1,l}}. \tag{77}$$

C. Wiener Filter

The Wiener filter is derived by taking the gradient of the MSE, $J(\mathbf{h}_{:,l})$, with respect to $\mathbf{h}_{:,l}$ and equating the result to zero [9]:

$$\begin{aligned} \mathbf{h}_{\mathrm{W},l} &= \boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}}^{-1} \boldsymbol{\Phi}_{\mathbf{c}_{x_:,l}} \mathbf{i}_1 \\ &= \left(\mathbf{I}_N - \boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}}^{-1} \boldsymbol{\Phi}_{\mathbf{c}_{v_:,l}}\right) \mathbf{i}_1. \end{aligned} \tag{78}$$

Since

$$\boldsymbol{\Phi}_{\mathbf{c}_{x_:,l}} \mathbf{i}_1 = \phi_{c_{x_1,l}} \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}, \tag{79}$$

we can rewrite (78) as

$$\mathbf{h}_{\mathrm{W},l} = \phi_{c_{x_1,l}} \boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}}^{-1} \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}. \tag{80}$$

It can be verified that

$$\boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}} = \phi_{c_{x_1,l}} \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}} \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}^T + \boldsymbol{\Phi}_{\mathrm{in},l}. \tag{81}$$

Determining the inverse of $\boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}}$ from (81) with the Woodbury's identity

$$\boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}}^{-1} = \boldsymbol{\Phi}_{\mathrm{in},l}^{-1} - \frac{\boldsymbol{\Phi}_{\mathrm{in},l}^{-1} \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}} \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}^T \boldsymbol{\Phi}_{\mathrm{in},l}^{-1}}{\phi_{c_{x_1,l}}^{-1} + \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}^T \boldsymbol{\Phi}_{\mathrm{in},l}^{-1} \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}} \tag{82}$$

and substituting the result into (80), leads to another interesting formulation of the Wiener filter:

$$\mathbf{h}_{\mathrm{W},l} = \frac{\boldsymbol{\Phi}_{\mathrm{in},l}^{-1} \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}}{\phi_{c_{x_1,l}}^{-1} + \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}^T \boldsymbol{\Phi}_{\mathrm{in},l}^{-1} \boldsymbol{\gamma}_{\mathbf{c}_{x_:,l}}}, \tag{83}$$

that we can rewrite as

$$\begin{aligned} \mathbf{h}_{\mathrm{W},l} &= \frac{\boldsymbol{\Phi}_{\mathrm{in},l}^{-1} \boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}} - \mathbf{I}_N}{1 - N + \mathrm{tr}\left(\boldsymbol{\Phi}_{\mathrm{in},l}^{-1} \boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}}\right)} \mathbf{i}_1 \\ &= \frac{\boldsymbol{\Phi}_{\mathrm{in},l}^{-1} \boldsymbol{\Phi}_{\mathbf{c}_{x_\mathrm{d},l}}}{1 + \lambda_{\max,l}} \mathbf{i}_1. \end{aligned} \tag{84}$$

We can deduce from (83) that the mode output SNR is

$$\begin{aligned} \mathrm{oSNR}\left(\mathbf{h}_{\mathrm{W},l}\right) &= \lambda_{\max,l} \\ &= \mathrm{tr}\left(\boldsymbol{\Phi}_{\mathrm{in},l}^{-1} \boldsymbol{\Phi}_{\mathbf{c}_{y_:,l}}\right) - N \end{aligned} \tag{85}$$

and the mode speech distortion index is a clear function of the mode output SNR:

$$\upsilon_{\mathrm{sd}}\left(\mathbf{h}_{\mathrm{W},l}\right) = \frac{1}{\left[1 + \mathrm{oSNR}\left(\mathbf{h}_{\mathrm{W},l}\right)\right]^2}. \tag{86}$$

The higher is the value of $\mathrm{oSNR}(\mathbf{h}_{\mathrm{W},l})$, the less the desired signal is distorted.

It follows that

$$\mathrm{oSNR}\left(\mathbf{h}_{\mathrm{W},l}\right) \geq \mathrm{iSNR}_l, \tag{87}$$

since the Wiener filter maximizes the mode output SNR.

It is of great interest to observe that the two filters $\mathbf{h}_{\max,l}$ and $\mathbf{h}_{\mathrm{W},l}$ are equivalent up to a scaling factor. Indeed, taking

$$\alpha_l = \frac{\phi_{c_{x_1,l}}}{1+\lambda_{\max,l}} \tag{88}$$

in (69) (maximum SNR filter), we find (84) (Wiener filter).

With the Wiener filter, the mode noise reduction factor is

$$\xi_{\mathrm{nr}}\left(\mathbf{h}_{\mathrm{W},l}\right) = \frac{\left[1+\mathrm{oSNR}\left(\mathbf{h}_{\mathrm{W},l}\right)\right]^2}{\mathrm{iSNR}_l \cdot \mathrm{oSNR}\left(\mathbf{h}_{\mathrm{W},l}\right)}. \tag{89}$$

The fullmode output SNR is

$$\mathrm{oSNR}(\mathbf{h}_{\mathrm{W},:}) = \frac{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}\frac{\mathrm{oSNR}^2(\mathbf{h}_{\mathrm{W},l})}{[1+\mathrm{oSNR}(\mathbf{h}_{\mathrm{W},l})]^2}}{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}\frac{\mathrm{oSNR}(\mathbf{h}_{\mathrm{W},l})}{[1+\mathrm{oSNR}(\mathbf{h}_{\mathrm{W},l})]^2}}. \tag{90}$$

Property 6.1: With the optimal KLE-domain Wiener filter given in (78), the fullmode output SNR is always greater than or equal to the fullmode input SNR, i.e., $\mathrm{oSNR}(\mathbf{h}_{\mathrm{W},:}) \geq \mathrm{iSNR}$.

Proof: See Section VI-E. ■

D. Minimum Variance Distortionless Response (MVDR) Filter

Another important filter, proposed by Capon [29], [30], is the minimum variance distortionless response (MVDR) beamformer which is obtained by minimizing the variance of the interference-plus-noise at the beamformer output with the constraint that the desired signal is not distorted. Mathematically, this is equivalent to

$$\min_{\mathbf{h}_{:,l}} \mathbf{h}_{:,l}^T\Phi_{\mathrm{in},l}\mathbf{h}_{:,l} \quad \text{subject to} \quad \mathbf{h}_{:,l}^T\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}} = 1, \tag{91}$$

for which the solution is

$$\begin{aligned}\mathbf{h}_{\mathrm{MVDR},l} &= \frac{\phi_{c_{x_1,l}}\Phi_{\mathrm{in},l}^{-1}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}}{\lambda_{\max,l}} \\ &= \frac{\Phi_{\mathrm{in},l}^{-1}\Phi_{\mathbf{c}_{y_{:,l}}} - \mathbf{I}_N}{\mathrm{tr}\left(\Phi_{\mathrm{in},l}^{-1}\Phi_{\mathbf{c}_{y_{:,l}}}\right) - N}\mathbf{i}_1.\end{aligned} \tag{92}$$

We can rewrite the MVDR as

$$\mathbf{h}_{\mathrm{MVDR},l} = \frac{\Phi_{\mathbf{c}_{y_{:,l}}}^{-1}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}}{\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}^T\Phi_{\mathbf{c}_{y_{:,l}}}^{-1}\boldsymbol{\gamma}_{c_{x_{:,l,l}}}}. \tag{93}$$

Taking

$$\alpha_l = \frac{\phi_{c_{x_1,l}}}{\lambda_{\max,l}} \tag{94}$$

in (69) (maximum SNR filter), we find (92) (MVDR filter), showing how the maximum SNR, MVDR, and Wiener filters are equivalent up to a scaling factor. From a mode point of view, this scaling is not significant but from a fullmode point of view it can be important since speech signals are broadband in nature. Indeed, it can be shown that this scaling factor affects the fullmode output SNRs and the fullmode speech distortion indices. While the mode output SNRs of the maximum SNR, Wiener, and MVDR filters are the same, the fullmode output SNRs are not because of the scaling factor.

It is clear that we always have

$$\mathrm{oSNR}\left(\mathbf{h}_{\mathrm{MVDR},l}\right) = \mathrm{oSNR}\left(\mathbf{h}_{\mathrm{W},l}\right), \tag{95}$$

$$\upsilon_{\mathrm{sd}}\left(\mathbf{h}_{\mathrm{MVDR},l}\right) = 0, \tag{96}$$

$$\xi_{\mathrm{sr}}\left(\mathbf{h}_{\mathrm{MVDR},l}\right) = 1, \tag{97}$$

$$\xi_{\mathrm{nr}}\left(\mathbf{h}_{\mathrm{MVDR},l}\right) = \frac{\lambda_{\max,l}}{\mathrm{iSNR}_l} \leq \xi_{\mathrm{nr}}\left(\mathbf{h}_{\mathrm{W},l}\right). \tag{98}$$

The fullmode output SNR is

$$\mathrm{oSNR}(\mathbf{h}_{\mathrm{MVDR},:}) = \frac{\sum_{l=1}^{L}\phi_{c_{x_1,l}}}{\sum_{l=1}^{L}\frac{\phi_{c_{x_1,l}}}{\mathrm{oSNR}(\mathbf{h}_{\mathrm{MVDR},l})}}. \tag{99}$$

Property 6.2: With the optimal KLE-domain MVDR filter given in (92), the fullmode output SNR is always greater than or equal to the fullmode input SNR, i.e., $\mathrm{oSNR}(\mathbf{h}_{\mathrm{MVDR},:}) \geq \mathrm{iSNR}$.

Proof: See Section VI-E. ■

E. Tradeoff Filter

In the tradeoff approach, we try to compromise between noise reduction and speech distortion. Instead of minimizing the MSE to find the Wiener filter or minimizing the MSE of the residual interference-plus-noise with the constraint of no distortion to find the MVDR, we could minimize the speech distortion index with the constraint that the noise reduction factor is equal to a positive value that is greater than 1. Mathematically, this is equivalent to

$$\min_{\mathbf{h}_{:,l}} J_{\mathrm{d}}\left(\mathbf{h}_{:,l}\right) \quad \text{subject to} \quad J_{\mathrm{r}}\left(\mathbf{h}_{:,l}\right) = \beta\phi_{c_{v_1,l}}, \tag{100}$$

where $0 < \beta < 1$ to insure that we get some noise reduction. By using a Lagrange multiplier, $\mu > 0$, to adjoin the constraint to the cost function, we deduce the tradeoff filter:

$$\begin{aligned}\mathbf{h}_{\mathrm{T},\mu,l} &= \phi_{c_{x_1,l}}\left(\phi_{c_{x_1,l}}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}^T + \mu\Phi_{\mathrm{in},l}\right)^{-1}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}} \\ &= \frac{\phi_{c_{x_1,l}}\Phi_{\mathrm{in},l}^{-1}\boldsymbol{\gamma}_{\mathbf{c}_{x_{:,l}}}}{\mu+\lambda_{\max,l}},\end{aligned} \tag{101}$$

where the Lagrange multiplier, μ, satisfies $J_{\mathrm{r}}(\mathbf{h}_{\mathrm{T},\mu,l}) = \beta\phi_{c_{v_1,l}}$. However, in practice it is not easy to determine the optimal μ. Therefore, when this parameter is chosen in an ad-hoc way, we can see that for

- $\mu = 1$, $\mathbf{h}_{\mathrm{T},1,l} = \mathbf{h}_{\mathrm{W},l}$, which is the Wiener filter;
- $\mu = 0$ [replacing μ in the second line of eq. (101)], $\mathbf{h}_{\mathrm{T},0,l} = \mathbf{h}_{\mathrm{MVDR},l}$, which is the MVDR filter;
- $\mu > 1$, results in low residual noise at the expense of high speech distortion;
- $\mu < 1$, results in high residual noise and low speech distortion.

Again, we observe here as well that the tradeoff and Wiener filters are equivalent up to a scaling factor. As a result, the mode output SNR with the tradeoff filter is the same as the mode output SNR with the Wiener filter, i.e.,

$$\mathrm{oSNR}\left(\mathbf{h}_{\mathrm{T},\mu,l}\right) = \lambda_{\max,l}, \tag{102}$$

and does not depend on μ. However, the mode speech distortion index is now both a function of the variable μ and the mode output SNR:

$$\upsilon_{\rm sd}\left(\mathbf{h}_{{\rm T},\mu,l}\right)=\frac{\mu^2}{\left(\mu+\lambda_{{\rm max},l}\right)^2}. \tag{103}$$

From (103), we observe how μ can affect the desired signal.

The tradeoff filter is interesting from several perspectives since it encompasses both the Wiener and MVDR filters. It is then useful to study the fullmode output SNR and the fullmode speech distortion index of the tradeoff filter, which both depend on the variable μ.

Using (101) in (49), we find that the fullmode output SNR is

$${\rm oSNR}\left(\mathbf{h}_{{\rm T},\mu,:}\right)=\frac{\sum_{l=1}^{NL}\frac{\phi_{c_{x_1,l}}\lambda_{{\rm max},l}^2}{(\mu+\lambda_{{\rm max},l})^2}}{\sum_{l=1}^{NL}\frac{\phi_{c_{x_1,l}}\lambda_{{\rm max},l}}{(\mu+\lambda_{{\rm max},l})^2}}. \tag{104}$$

We propose the following:

Property 6.3: The fullmode output SNR of the tradeoff filter is an increasing function of the parameter μ.

Proof: The complete proof can be found in [31]. ■

From Property 6.3, we deduce that the MVDR filter gives the smallest fullmode output SNR, which is

$${\rm oSNR}\left(\mathbf{h}_{{\rm T},0,:}\right)=\frac{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}}{\sum_{l=1}^{NL}\frac{\phi_{c_{x_1,l}}}{\lambda_{{\rm max},l}}}. \tag{105}$$

We give another interesting property.

Property 6.4: We have

$$\lim_{\mu\to\infty}{\rm oSNR}\left(\mathbf{h}_{{\rm T},\mu,:}\right)=\frac{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}\lambda_{{\rm max},l}^2}{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}\lambda_{{\rm max},l}}\leq\sum_{l=1}^{NL}\lambda_{{\rm max},l}. \tag{106}$$

Proof: It can be derived from (104) [31]. ■

While the fullmode output SNR is upper bounded, it can be shown that the fullmode noise reduction factor and fullmode speech reduction factor are not. So when μ goes to infinity so are $\xi_{\rm nr}(\mathbf{h}_{{\rm T},\mu,:})$ and $\xi_{\rm sr}(\mathbf{h}_{{\rm T},\mu,:})$.

The fullmode speech distortion index is

$$\upsilon_{\rm sd}\left(\mathbf{h}_{{\rm T},\mu,:}\right)=\frac{\sum_{l=1}^{NL}\frac{\phi_{c_{x_1,l}}\mu^2}{(\mu+\lambda_{{\rm max},l})^2}}{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}}. \tag{107}$$

Property 6.5: The fullmode speech distortion index of the tradeoff filter is an increasing function of the parameter μ.

Proof: We can verify that

$$\frac{d\upsilon_{\rm sd}\left(\mathbf{h}_{{\rm T},\mu,:}\right)}{d\mu}\geq 0, \tag{108}$$

which ends the proof [31]. ■

It is clear that

$$0\leq\upsilon_{\rm sd}\left(\mathbf{h}_{{\rm T},\mu,:}\right)\leq 1,\ \forall\mu\geq 0. \tag{109}$$

Therefore, as μ increases, the fullmode output SNR increases at the price of more distortion to the desired signal.

Property 6.6: With the tradeoff filter, $\mathbf{h}_{{\rm T},\mu,l}$, the fullmode output SNR is always greater than or equal to the fullmode input SNR, i.e., ${\rm oSNR}(\mathbf{h}_{{\rm T},\mu,:})\geq{\rm iSNR},\ \forall\mu\geq 0$.

Proof: We know that

$$\lambda_{{\rm max},l}\geq{\rm iSNR}_l, \tag{110}$$

which implies that

$$\sum_{l=1}^{NL}\phi_{c_{v_1,l}}\frac{{\rm iSNR}_l}{\lambda_{{\rm max},l}}\leq\sum_{l=1}^{NL}\phi_{c_{v_1,l}}, \tag{111}$$

and hence,

$$\begin{aligned}{\rm oSNR}\left(\mathbf{h}_{{\rm T},0,:}\right)&=\frac{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}}{\sum_{l=1}^{NL}\phi_{c_{v_1,l}}\frac{{\rm iSNR}_l}{\lambda_{{\rm max},l}}}\\&\geq\frac{\sum_{l=1}^{NL}\phi_{c_{x_1,l}}}{\sum_{l=1}^{NL}\phi_{c_{v_1,l}}}={\rm iSNR}.\end{aligned} \tag{112}$$

But from Proposition 6.3, we have

$${\rm oSNR}\left(\mathbf{h}_{{\rm T},\mu,:}\right)\geq{\rm oSNR}\left(\mathbf{h}_{{\rm T},0,:}\right),\ \forall\mu\geq 0, \tag{113}$$

as a result,

$${\rm oSNR}\left(\mathbf{h}_{{\rm T},\mu,:}\right)\geq{\rm iSNR},\ \forall\mu\geq 0, \tag{114}$$

which completes the proof [31]. ■

VII. Experimental Results

In this section, we evaluate the performance of the multichannel noise reduction filters in the KLE domain. Here, we focus on the MVDR, Wiener, and tradeoff filters, and discuss the effect of different parameters in the design of the filters.

A. Simulation Environment

In the following experiments, we used an anechoic recording of a female speaker as our desired clean signal. The sampling rate of the signal was 8 kHz and the length of the signal was 35 s. The clean signal was then corrupted by a spatially coherent noise source and a spatially incoherent noise. The spatially coherent noise source consisted of an anechoic recording of a different female speaker. We used two types of spatially incoherent noises: the first one was a computer generated stationary white Gaussian noise. The second was a babble speech signal generated assuming an ideal spherical diffuse sound field [32]. Note that the latter is partially spatially coherent, which is discussed later on in the experimental results. The noisy signal is then the addition of the clean anechoic speech, the spatially incoherent and spatially coherent noise. The level of the signals was adjusted so it matched the input signal-to-incoherent-noise ratio (iSINR) and the input signal-to-coherent-noise ratio (iSCNR).

In the simulations the microphone(s) and sources were located in a room of dimensions $x=5$, $y=6$ and $z=4$ m. The room's reverberation time (RT60) was set to 0.5 s and the room impulse responses were calculated using the image method [33]. The microphone arrays were simulated to have an uniformly spaced geometry with a distance of $d=0.05$ m between microphones. Since in our noise reduction formulation we used one of the microphones as a reference to calculate the filters, the spacing between microphones should not significantly influence the performance of the noise reduction filters.

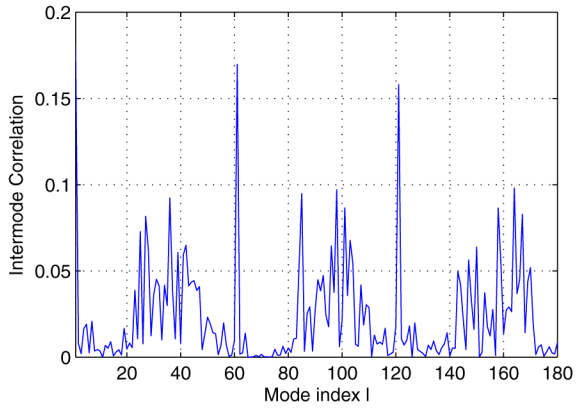

Fig. 1. First column of the inter-mode correlation $\Phi_{\mathbf{c'}_{y,:}}$ for a 5-seconds speech signal, $L = 20$ and $N = 3$.

The desired signal was simulated to be located 1 m away from the array at $40°$ azimuth and $2°$ elevation, where the point $(0°, 0°)$ is located right in front of the center of the array. The spatially coherent noise source was simulated to be located 1.5 m away from the array at $-40°$ azimuth and $-2°$ elevation.

B. Choice of modes

As mentioned in Section III, in order to fully exploit the noise reduction in the KLE domain, inter-mode correlations should be taken into account. However, not all modes are highly correlated, which suggests that a selection of the modes with high correlation is sufficient for the practical implementation. First, let us take a look at the structure of these correlations. As an example, we use an array of three microphones ($N = 3$) and a 5-seconds speech signal. For convenience, we stack all the coefficients of the three microphones in a vector of length LN^2, i.e. $\mathbf{c'}_{y,:} = [\mathbf{c}^T_{y_1,:} \quad \mathbf{c}^T_{y_2,:} \quad \mathbf{c}^T_{y_3,:}]^T$, where $\mathbf{c}_{y_n,:} = [c_{y_n,1} \quad c_{y_n,2} \quad \dots \quad c_{y_n,NL}]^T$ and $L = 20$. The inter-mode correlation matrix is thus defined as $\Phi_{\mathbf{c'}_{y,:}} = E[\,\mathbf{c'}_{y,:}(m)\mathbf{c'}_{y,:}{}^T(m)\,]$. Fig. 1 shows the magnitude of these inter-mode correlations for the first mode, i.e. first column of $\Phi_{\mathbf{c'}_{y,:}}$. It is clear from Fig. 1 that the inter-mode correlations are mostly dominated by the modes $l > L$, i.e.,

$$E\left[c_{y_n,l}(m)c_{y_n,l+p}(m)\right] \gg 0, \quad p = L, \dots, NL. \tag{115}$$

Therefore, we do not need to make use of all NL modes, but instead it is sufficient to exploit only those $F = L(N-1)+1$ modes that carry relevant information, which substantially reduces the size of the correlation matrix $\Phi_{\mathbf{c}_{y,:,l}}$ and computational complexity. We define thus

$$f(l,p) = \begin{cases} l+p+L-1 & 1 \le p \le F-1-l \\ l+p-F+1 & F-1-l < p \le F-1 \end{cases}. \tag{116}$$

This empirical selection criteria is used in the following experiments.

C. Estimation of Correlation Matrices

In order to estimate the filter coefficients, we need to calculate the correlation matrices $\mathbf{R}_{\underline{\mathbf{y}}}$, $\Phi_{\mathbf{c}_{y,:,l}}$, and $\Phi_{\mathbf{c}_{v,:,l}}$. The noisy correlation matrix can be estimated directly from the noisy signal $\underline{\mathbf{y}}(m)$ using (4) by approximating the mathematical expectation with a sample average. This sample average should be done on a short-term basis, given that speech is in practice non-stationary. In this study, we calculated $\mathbf{R}_{\underline{\mathbf{y}}}$ at each time frame m by using the most recent 40 ms of the signals received by each microphone. Additionally, in [11] it is suggested to combine the short-term sample average and a moving average to estimate the correlation matrices. At time frame m, the correlation matrix $\mathbf{R}_{\underline{\mathbf{y}}}$ is estimated by

$$\mathbf{R}_{\underline{\mathbf{y}}}(m) = \alpha_y \mathbf{R}_{\underline{\mathbf{y}}}(m-1) + (1-\alpha_y)\mathbf{R}'_{\underline{\mathbf{y}}}(m), \tag{117}$$

where α_y is a forgetting factor and $\mathbf{R}'_{\underline{\mathbf{y}}}(m) = (L/N_y)\sum_{i=m-N_y/L}^{m} \underline{\mathbf{y}}(i)\underline{\mathbf{y}}^T(i)$ is the frame correlation matrix at time frame m and N_y is the window length. The KLT $\mathbf{Q}$ is then obtained using eigenvalue decomposition. To estimate the correlation matrix $\Phi_{\mathbf{c}_{y,:,l}}$ we use the same approach as in (117), namely

$$\Phi_{\mathbf{c}_{y,:,l}}(m) = \alpha_{\mathbf{c}_y}\Phi_{\mathbf{c}_{y,:,l}}(m-1) + (1-\alpha_{\mathbf{c}_y})\mathbf{c}_{y,:,l}(m)\mathbf{c}^T_{y,:,l}(m), \tag{118}$$

where $\alpha_{\mathbf{c}_y}$ is the corresponding forgetting factor. The forgetting factors were set to $\alpha_y = 0.985$ and $\alpha_{\mathbf{c}_y} = 0.8$, which were found to be optimal in terms of noise reduction and speech distortion. A more detailed evaluation of the effect of the forgetting factors in the performance of the filters can be found in [11]. To estimate $\Phi_{\mathbf{c}_{v,:,l}}$ we would need in practice a noise estimator or a voice activity detector (VAD) to be able to compute the coefficients $\mathbf{c}_{v,:,l}$. Even though an analysis of issues concerning noise estimators or VADs would be interesting, it is out of the scope of this paper to investigate their influence on the noise reduction in the KLE domain. In this study, we are mainly interested on assessing the performance of the noise reduction filters in the KLE domain when using multiple channels compared to the single channel case. Thus, in order not to include the influence of possible errors from the noise estimator or the VAD in our experiments, we calculated the coefficients $\mathbf{c}_{v,:,l}$ directly from the noise signals. The estimation of $\Phi_{\mathbf{c}_{v,:,l}}$ is done in a similar fashion as in (118), with $\alpha_{\mathbf{c}_v} = 0.8$.

D. Experimental Results with Stationary White Gaussian Noise

In the first experiments we evaluated the performance of the filters in the presence of spatially incoherent stationary noise. The simulated noise was a computer generated white Gaussian process and the level of the signal was adjusted to control the iSINR.

Let us first take a look at the performance of the Wiener filter as a function of frame length L. Fig. 2 shows these performance results calculated for different frame lengths L and number of microphones N. In the simulated scenario, the iSINR was set to 20 dB and the iSCNR to 0 dB.

While for the single-channel case the performance does not vary with frame length, the performance improves with longer frames for the multichannel case. The improvement is particularly noticeable in the coherent noise reduction (CNR) factor, which increases with the number of microphones and shows to be the dominant factor in the overall noise reduction. The single-channel case performs better with respect to incoherent

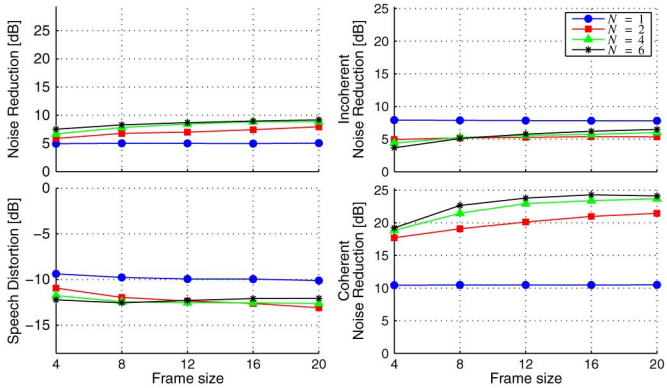

Fig. 2. Noise reduction, speech distortion, incoherent-noise reduction and coherent-noise reduction as a function of frame size L and number of microphones N. The desired speech signal is corrupted by another speech signal and stationary white Gaussian noise; $\mu = 1$, iSINR $= 20$ dB, iSCNR $= 0$ dB, RT60 $= 0.5$ s.

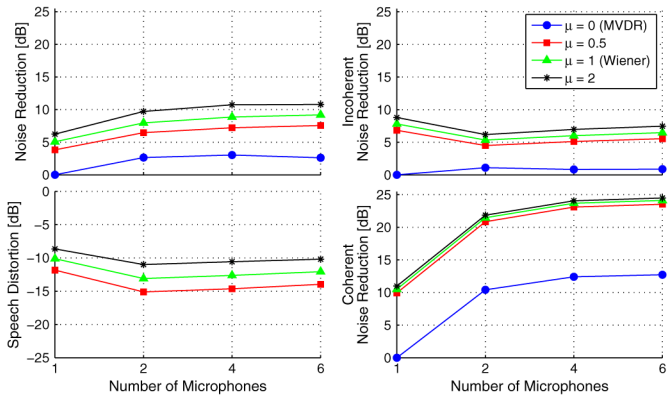

Fig. 3. Noise reduction, speech distortion, incoherent-noise reduction and coherent-noise reduction as a function of number of microphones N and filter type. The desired speech signal is corrupted by another speech signal and stationary white Gaussian noise; $L = 20$, iSINR $= 20$ dB, iSCNR $= 0$ dB, RT60 $= 0.5$ s.

noise reduction (INR) for smaller frame lengths ($L < 12$). However, for $L \geq 12$, the performance with respect to INR becomes comparable to the multichannel channel case for $N \geq 4$. The multichannel filters introduce, in general, less speech distortion than the single-channel Wiener filter.

The poor performance of the single-channel in this scenario can be attributed to the small iSCNR simulated, which implies that the noise term is generally dominated by signals with similar statistics to those of the desired signal. Given that in the single-channel scenario the spatial information is not exploited, a poor performance of the filters is expected when competing sources are dominant. In the case of multichannel setups, even though larger noise reduction and coherent noise reduction factors are obtained, less speech distortion is introduced. This suggests that the multichannel filters make a better use of the inter-channel as well as the inter-mode correlations.

Fig. 3 shows the noise reduction, speech distortion, coherent noise reduction and incoherent noise reduction for the tradeoff filter calculated for different number of microphones N and different values of the Lagrange multiplier μ. Recall that for $\mu = 1$, $\mathbf{h}_{\mathrm{T},1,l} = \mathbf{h}_{\mathrm{W},l}$, which is the Wiener filter and for $\mu = 0$, $\mathbf{h}_{\mathrm{T},0,l} = \mathbf{h}_{\mathrm{MVDR},l}$, when using the second line of Eq. (101), which is the MVDR filter. In this experiment, the iSINR and the iSCNR were also set to 20 dB and 0 dB respectively.

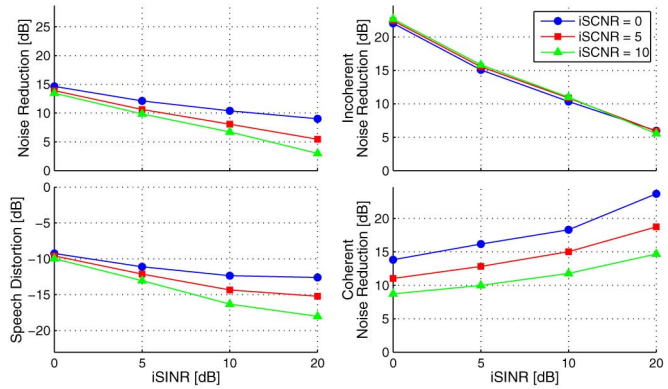

Fig. 4. Noise reduction, speech distortion, incoherent-noise reduction and coherent-noise reduction as a function of iSINR and iSCNR. The desired speech signal is corrupted by another speech signal and stationary white Gaussian noise; $\mu = 1$, $N = 4$, RT60 $= 0.5$ s, and $L = 20$.

As observed before, the speech distortion factor decreases when using multiple microphones ($N > 1$). However, a slight increase with N can be observed in this experiment. The noise reduction factor increases with number of microphones and μ, though the improvements become marginal as the number of microphones increases. The multichannel cases show again a clear improvement with respect to CNR. In the case of single-channel case, there is a better performance with respect to INR compared to the multichannel case, though the CNR factor is substantially smaller. There is also a substantial performance improvement with respect to CNR between $\mu = 0$ (MVDR) and $\mu = 0.5$. This improvement becomes then marginal for larger values of μ. As expected, the MVDR filter for the single-channel case results in no speech distortion but no noise reduction either, which can be deduced from (98) and it is in agreement with [11]. The MVDR ($\mu = 0$) filter shows in general a poor performance. This suggests that in order to significantly reduce a spatially and temporally coherent source such as a competing speaker, there must be a compromise in speech distortion.

To understand better the influence of coherent and incoherent noise sources in the performance of the filters, the third experiment tested the performance of the Wiener filter calculated for an array of 4 microphones ($N = 4$) with different iSCNR and iSINR. The frame length was set to $L = 20$. Fig. 4 shows the speech distortion, noise reduction, incoherent-noise reduction, and coherent-noise reduction factors for this experiment. As expected, the noise reduction factor increases with smaller iSCNR, while more speech distortion is introduced. From the INR we can see that the performance of the filters is rather independent of the iSCNR. As expected, the CNR factor improves with larger iSINR and smaller iSCNR.

E. Experimental Results with Spherical Isotropic Noise

In the following experiments, the performance of the noise-reduction filters in the KLE domain is evaluated in the presence of non-stationary diffuse noise as spatially incoherent noise. The non-stationary noise source was simulated using babble speech signals assuming an ideal spherical isotropic sound field [32].

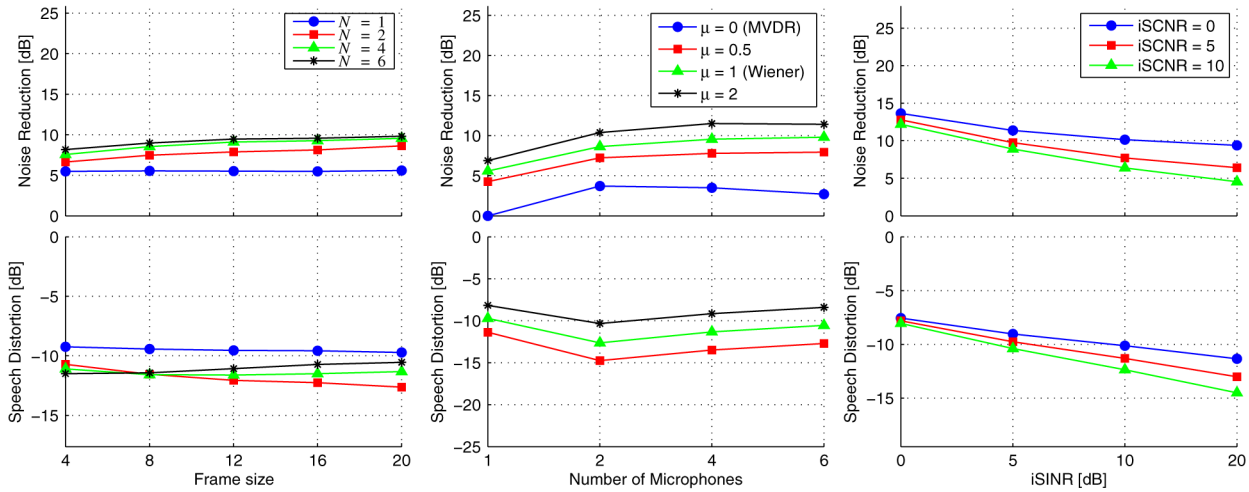

Fig. 5. Noise reduction and speech distortion as a function of: (a) frame size L and number of microphones N for $\mu = 1$, iSINR $= 20$ dB, iSCNR $= 0$ dB (b) number of microphones N and filter type for $L = 20$, iSINR $= 20$ dB, iSCNR $= 0$ dB and (c) iSINR and iSCNR for $\mu = 1$, $N = 4$, $L = 20$. The desired speech signal is corrupted by another speech signal and babble-noise; RT60 $= 0.5$ s.

Notice that the simulated babble noise is spatially coherent at low frequencies. Additionally, some coherence across frames is expected due to the temporal characteristics of the speech signals. That is, the incoherent-noise correlation matrix $\Phi_{\mathbf{c}_{b,l}}$ defined in Eq. (61) will not only contain incoherent-noise components, but also coherent information. Consequently, the CNR and INR factors defined in Eq. (65) and Eq. (66) can be regarded as meaningless in this scenario. In the following experiments, we will therefore focus only on the overall NR and SD factors.

Fig. 5(a) shows the performance of the Wiener filter as a function of frame size L and number of microphones N. Similarly to the experiments with Gaussian noise, the iSINR was set to 20 dB and the iSCNR to 0 dB. Note that since in this scenario the diffuse noise is partially coherent, the actual iSCNR is expected to be smaller than the simulated one, i.e. negative and the actual iSINR larger. In spite of this, we can see that the noise reduction factors obtained are quite comparable to those of the stationary white Gaussian noise case. This supports the argument that the proposed multichannel noise reduction formulation in KLE domain is rather robust to spatially coherent sources. In the single-channel case, we do not observe a decrease in performance due to the already small iSINR. When evaluating the NR and SD factors for different number of microphones N and values of the Lagrange multiplier μ, as shown in Fig. 5(b), we can also see little difference compared to the stationary noise case.

Fig. 5(c) shows the results obtained with the Wiener filter at different iSINR and iSCNR, when using four microphones ($N = 4$) and a frame size of $L = 20$. In general, the NR factor is comparable to the stationary noise case, though in the case of iSINR = 20 dB, there is an improvement in NR when the iSCNR is larger than 5 dB. This is clearly a result of the expected decrease in the actual iSCNR, which again supports the previous observations.

VIII. Conclusions

In this paper we studied the multichannel noise reduction problem in the Karhunen-Loève expansion (KLE) domain. We derived a new formulation in which the KLT is applied to the joint contribution of multiple receivers. The KLE coefficients are then expanded into sub-coefficients, which can be seen as the coefficients corresponding to each channel. Inter-mode correlations are also utilized to fully take advantage of the spatial information contained in the input signals. Optimal noise reduction filters were derived, within this framework, and a set of useful performance measures were discussed. The filters were evaluated in the presence of undesired speech sources and spatially incoherent noise. Two spatially incoherent noise scenarios were simulated: stationary noise and non-stationary diffuse noise. Through experiments, we demonstrated that a better performance is obtained when using multiple microphones to solve the noise reduction problem in the KLE domain. The multichannel filters show to be specially robust to undesired speech sources and spatially coherent noise sources.

References

[1] J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, "Fundamentals of noise reduction," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds. Berlin, Germany: Springer-Verlag, 2008, pp. 843–872.

[2] J. Benesty, J. Chen, Y. A. Huang, and S. Doclo, "Study of the Wiener filter for noise reduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Berlin, Germany: Springer-Verlag, 2005, pp. 9–41, Signals and Communication Technology.

[3] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Berlin, Germany: Springer-Verlag, 2005.

[4] J. Chen, Y. Benesty, J. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[6] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 137–145, Apr. 1980.

[7] 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.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[9] J. Chen, J. Benesty, and Y. A. Huang, "On the optimal linear filtering techniques for noise reduction," *Speech Commun.*, vol. 49, pp. 305–316, Apr. 2007.
[10] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*. Berlin, Germany: Springer-Verlag, 2011, Springer Briefs in Electrical and Computer Engineering.
[11] J. Chen, Y. Benesty, and J. Huang, "Study of the noise-reduction problem in the Karhunen–Loève expansion domain," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 787–802, May 2009.
[12] J. Benesty, J. Chen, and Y. Huang, "Noise reduction algorithms in a generalized transform domain," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1109–1123, Aug. 2009.
[13] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, Nov. 1995.
[14] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.
[15] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 159–167, Mar. 2000.
[16] J. Benesty, J. Chen, and Y. Huang, "Speech enhancement in the karhunen–loève expansion domain," in *Synthesis Lectures on Speech and Audio Processing*. San Rafael, CA, USA: Morgan & Claypool, 2011.
[17] J. P. Dmochowski and J. Benesty, "Microphone arrays: Fundamental concepts," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds. Berlin, Germany: Springer-Verlag, Jan. 2010, ch. 11.
[18] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, M. M. Benesty, J. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer-Verlag, 2008, ch. 47, pp. 945–978.
[19] Y. Lacouture-Parodi, E. A. P. Habets, and J. Benesty, "Multichannel noise reduction Wiener filter in the Karhunen-Loève expansion domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012.
[20] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.
[21] *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001.
[22] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.
[23] J. Benesty, J. Chen, and Y. Huang, "On noise reduction in the Karhunen-Loève expansion domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2009, pp. 25–28.
[24] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.
[25] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. ed. Baltimore, MD, USA: John Hopkins Univ. Press, 1996.
[26] S. Haykin, *Adaptive Filter theory*, 4th Ed. ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.
[27] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
[28] W. Herbordt, "Combination of robust adaptive beamforming with acoustic echo cancellation for acoustic human/machine interfaces," Ph.D. dissertation, Erlangen-Nuremberg Univ. , Erlangen, Germany, 2004.
[29] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[30] R. T. Lacoss, "Data adaptive spectral analysis methods," *Geophysics*, vol. 36, pp. 661–675, 1971.
[31] M. Souden, J. Benesty, and S. Affes, "On the global output SNR of the parameterized frequency-domain multichannel noise reduction Wiener filter," *IEEE Signal Process. Lett.*, pp. 425–428, May 2010.
[32] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, pp. 2911–2917, Nov. 2008.
[33] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.

Yesenia Lacouture Parodi was born in Colombia in 1980. In 2007 she received her masters degree in Acoustics at Aalborg University, Denmark. After graduation, she enrolled as a Ph.D. student at the section of acoustics at Aalborg University and completed her degree in November 2010. During her doctoral work she carried a systematic study of binaural reproduction systems through loudspeakers, with special focus on stereo-dipoles. In 2009 (between August and December) she was a visiting researcher at the laboratory for Sound and Music Innovation Technology (SMIT) at the National Chiao-Tung University, Hsin-Chu, Taiwan. From July 2011 to June 2013 she work as a postdoctoral researcher at the International Audio Laboratories Erlangen in Germany, where she carried research work on perception-based spatial audio signal processing. In July 2013 she joined the multimedia team at the HUAWEI European research centre in Munich as a senior researcher, where she currently works on 3D audio reproduction. Her research interests include binaural techniques, psychoacoustics, perception of spatial sound, audio signal processing and immersive environments. In 2010 she received the AES 128th Convention Student Technical Paper Award.

Emanuël A. P. Habets (S'02–M'07–SM'11) received his B.Sc degree in electrical engineering from the Hogeschool Limburg, The Netherlands, in 1999, and his M.Sc and Ph.D. degrees in electrical engineering from the Technische Universiteit Eindhoven, The Netherlands, in 2002 and 2007, respectively. From March 2007 until February 2009, he was a Postdoctoral Fellow at the Technion - Israel Institute of Technology and at the Bar-Ilan University in Ramat-Gan, Israel. From February 2009 until November 2010, he was a Research Fellow in the Communication and Signal Processing group at Imperial College London, United Kingdom. Since November 2010, he is an Associate Professor at the International Audio Laboratories Erlangen (a joint institution of the University of Erlangen and Fraunhofer IIS) and a Chief Scientist for Spatial Audio Processing at Fraunhofer IIS, Germany.

His research interests center around audio and acoustic signal processing, and he has worked in particular on dereverberation, noise estimation and reduction, echo reduction, system identification and equalization, source localization and tracking, and crosstalk cancellation.

Dr. Habets was a member of the organization committee of the 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC) in Eindhoven, The Netherlands, a general co-chair of the 2013 International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in New Paltz, New York, and general co-chair of the 2014 International Conference on Spatial Audio (ICSA) in Erlangen, Germany. He is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2011-2016) and a member of the IEEE Signal Processing Society Standing Committee on Industry Digital Signal Processing Technology (2013-2015). Since 2013 he is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.

Jingdong Chen (M'99–SM'09) received the Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences in 1998.

From 1998 to 1999, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he conducted research on speech synthesis, speech analysis, as well as objective measurements for evaluating speech synthesis. He then joined the Griffith University, Brisbane, Australia, where he engaged in research on robust speech recognition and signal processing. From 2000 to 2001, he worked at ATR Spoken Language Translation Research Laboratories on robust speech recognition and speech enhancement. From 2001 to 2009, he was a Member of Technical Staff at Bell Laboratories, Murray Hill, New Jersey, working on acoustic signal processing for telecommunications. He subsequently joined WeVoice Inc. in New Jersey, serving as the Chief Scientist. He is currently a professor at the Northwestern Polytechnical University in Xi'an, China. His research interests include acoustic signal processing,

adaptive signal processing, speech enhancement, adaptive noise/echo control, microphone array signal processing, signal separation, and speech communication. Dr. Chen is currently an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, an associate member of the IEEE Signal Processing Society (SPS) Technical Committee (TC) on Audio and Acoustic Signal Processing (AASP), and a member of the editorial advisory board of the Open Signal Processing Journal. He was the Technical Program Co-Chair of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) and the Technical Program Chair of IEEE TENCON 2013, and helped organize many other conferences. He co-authored the books *Study and Design of Differential Microphone Arrays* (Springer-Verlag, 2013), *Speech Enhancement in the STFT Domain* (Springer-Verlag, 2011), *Optimal Time-Domain Noise Reduction Filters: A Theoretical Study* (Springer-Verlag, 2011), *Speech Enhancement in the Karhunen-Loève Expansion Domain* (Morgan&Claypool, 2011), *Noise Reduction in Speech Processing* (Springer-Verlag, 2009), *Microphone Array Signal Processing* (Springer-Verlag, 2008), and *Acoustic MIMO Signal Processing* (Springer-Verlag, 2006). He is also a co-editor/co-author of the book *Speech Enhancement* (Berlin, Germany: Springer-Verlag, 2005) and a section co-editor of the reference *Springer Handbook of Speech Processing* (Springer-Verlag, Berlin, 2007).

Dr. Chen received the 2008 Best Paper Award from the IEEE Signal Processing Society (with Benesty, Huang, and Doclo), the best paper award from the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2011 (with Benesty), the Bell Labs Role Model Teamwork Award twice, respectively, in 2009 and 2007, the NASA Tech Brief Award twice, respectively, in 2010 and 2009, the Japan Trust International Research Grant from the Japan Key Technology Center in 1998, the Young Author Best Paper Award from the 5th National Conference on Man-Machine Speech Communications in 1998, and the CAS (Chinese Academy of Sciences) President's Award in 1998.

Jacob Benesty was born in 1963. He received a Master degree in microwaves from Pierre & Marie Curie University, France, in 1987, and a Ph.D. degree in control and signal processing from Orsay University, France, in April 1991.

During his Ph.D. (from Nov. 1989 to Apr. 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecomunications (CNET), Paris, France. From January 1994 to July 1995, he worked at Telecom Paris University on multichannel adaptive filters and acoustic echo cancellation. From October 1995 to May 2003, he was first a Consultant and then a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ, USA. In May 2003, he joined the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, as a Professor. His research interests are in signal processing, acoustic signal processing, and multimedia communications. He is the inventor of many important technologies. In particular, he was the lead researcher at Bell Labs who conceived and designed the world-first real-time hands-free full-duplex stereophonic teleconferencing system. Also, he conceived and designed the world-first PC-based multi-party hands-free full-duplex stereo conferencing system over IP networks.

He was the co-chair of the 1999 International Workshop on Acoustic Echo and Noise Control and the general co-chair of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. He is the recipient, with Morgan and Sondhi, of the IEEE Signal Processing Society 2001 Best Paper Award. He is the recipient, with Chen, Huang, and Doclo, of the IEEE Signal Processing Society 2008 Best Paper Award. He is also the co-author of a paper for which Huang received the IEEE Signal Processing Society 2002 Young Author Best Paper Award. In 2010, he received the "Gheorghe Cartianu Award" from the Romanian Academy. In 2011, he received the Best Paper Award from the IEEE WASPAA for a paper that he co-authored with Chen.