A Framework for Speech Enhancement with Ad Hoc Microphone Arrays

Tavakoli, Vincent Mohammad; Jensen, Jesper Rindom; Christensen, Mads Græsbøll; Benesty, Jacob

Published in:
I E E E Transactions on Audio, Speech and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2016.2537202

Publication date:
2016

Document Version
Early version, also known as pre-print

A Framework for Speech Enhancement With Ad Hoc Microphone Arrays

Vincent Mohammad Tavakoli, *Student Member, IEEE*, Jesper Rindom Jensen, *Member, IEEE*, Mads Græsbøll Christensen, *Senior Member, IEEE*, and Jacob Benesty

***Abstract*—Speech enhancement is vital for improved listening practices. Ad hoc microphone arrays are promising assets for this purpose. Most well-established enhancement techniques with conventional arrays can be adapted into ad hoc scenarios. Despite recent efforts to introduce various ad hoc speech enhancement apparatus, a common framework for integration of conventional methods into this new scheme is still missing. This paper establishes such an abstraction based on inter and intra subarray speech coherencies. Along with measures for signal quality at the input of subarrays, a measure of coherency is proposed both for subarray selection in local enhancement approaches, and also for selecting a proper global reference when more than one subarray are used. Proposed methods within this framework are evaluated with regard to quantitative and qualitative measures, including array gains, the speech distortion ratio, the PESQ measure, and the STOI intelligibility measure. Major findings in this work are the observed changes in the superiority of different methods for certain conditions. When perceptual quality or intelligibility of the speech are the ultimate goals, there are turning points where the MVDR and the LCMV are superior to Wiener-based methods. Also, for certain scenarios, local approaches may be preferred to global ones.**

***Index Terms*—Speech enhancement, microphone array, noise reduction, multichannel, pseudo-coherence vector, ad hoc array.**

I. Introduction

NOWADAYS, smartphones and other portable (or even wearable) devices are pervasively embedded into our life by exposing their potentials and redefining our personal needs. Consequently, these devices are becoming the de facto platform for many emerging signal processing applications including speech enhancement in noisy, interfered, and reverberant environments, which is the target application of this paper. For this purpose, mobile devices can be used as nodes of an ad hoc microphone array to improve capturing the acoustic environment. Here, both the increased number of sensors in the wireless acoustic sensor network (WASN) and the extended spatial coverage potentially improve performance of the enhancement system. Although, this context can be justified as an extension to traditional distributed microphone arrays, the major mutation is the dynamic constellation of nodes which introduces new challenges that should be overcome in order to assure reliability of speech enhancement systems.

It is helpful to expound the relationship between the method of interest of this paper, which is speech enhancement based on signal pseudo-coherencies, and common approaches in speech enhancement that take advantage of the spatial selectivity of microphone arrays. These spatial filtering (beamforming) methods may be categorized based on the auxiliary parameter (spatial signature) and/or statistics used in them. A subclass of beamforming methods use direction of arrival (DOA) of acoustic wave-fronts as spatial cue, prior to steering the beampattern of the array. Fast DOA estimators, such as the broadband DOA estimator in [1], enables constrained beamforming techniques to be developed based on Capon's minimum variance distortionless response (MVDR) [2], [3], and Frost's linearly constrained minimum variance (LCMV) [4], [5]. Recently, noise reduction performance of the MVDR beamformer is studied under noisy and reverberant environments in [6], and a broadband LCMV beamformer with controllable constraints have been proposed [7]. Unfortunate for ad hoc arrays, DOA-based beamforming techniques are based on a restricting assumption, i.e., known (or even confined) array geometry. Moreover, even with the known array geometries, the mathematic expressions for proper beamforming gets more complicated when sources position in the near-field of the array that is likely probable in ad hoc arrays.

Acoustic transfer function (ATF) and relative transfer function (RTF) are other spatial fingerprints which are more useful in reverberant enclosures, since they can be used for noise removal and dereverberation, simultaneously. The multichannel Wiener filter (MWF), the generalized side-lobe canceler (GSC), and their extensions are common techniques used in ATF-based and RTF-based beamformers. A realization of speech distortion weighted multichannel Wiener filter (SDW-MWF) has been implemented based on the soft output voice activity detector (VAD) for noise reduction in hearing aids [8]. The SDW-MWF was generalized in [9] to deal with multiple speakers, and the state-of-the-art nested GSC is presented recently for joint (parallel) treatment of dereverberation and noise reduction [10]. An analysis of the SDW-MWF beamformer in reverberant environment can be found in [11]. Unfortunately, ATFs and

Manuscript received September 15, 2015; revised December 21, 2015 and February 11, 2016; accepted February 15, 2016. Date of publication March 02, 2016; date of current version April 29, 2016. This work was supported in part by the Villum Foundation for V. M. Tavakoli and in part by the Danish Council for Independent Research under Grant 1337-00084 for J. R. Jensen. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yunxin Zhao.

V. M. Tavakoli, J. R. Jensen, and M. G. Christensen are with the Audio Analysis Laboratory, Department of Architecture, Design, and Media Technology, Aalborg University, Aalborg DK-9000, Denmark (e-mail: vmt@create.aau.dk; jrj@create.aau.dk; mgc@create.aau.dk).

J. Benesty is with INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada, and also with the Department of Architecture, Design, and Media Technology, Aalborg University, Aalborg DK-9000, Denmark (e-mail: benesty@emt.inrs.ca).

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.2016.2537202

RTFs are normally very long and hard to estimate [12], which makes more complicated estimation of shorter functions, such as convolutive transfer function (CTF), inescapable [13].

Instead of directional and spatial cues (DOAs, ATFs, and RTFs), the enhancing apparatus may be parametrized to form an optimal filter based on a model of signal (or noise) without explicitly steering the beampattern of the array; however, with complex weighting, an implicit beamformer may be assumed. In a loose sense, methods in this approach are governed by interpolation and extrapolation of the time series introduced by Wiener [14] and Kolmogorov [15]. Such an optimal filter should be updated adaptively to deal with model parameter changes caused by non-stationary time-series. Various signal-adaptive filters have been proposed based on different noise and speech signal characteristics. The performance of methods based on noise statistics is essentially dependent on the accuracy of the noise estimator which varies for different noise colors, types, and stationarities. State-of-the-art noise estimation methods based on the speech presence probability (SPP) [16] and using spatially-sampled noise references [17] treat some of these challenges; however, SPP may not be accurate in multi-speaker and interfered scenarios.

For speech enhancement using the speech signal model, different characteristics may be used. Quasi-periodicity of speech is a property which has been used for designing single-channel optimal filters for speech enhancement [18], [19]. For multichannel speech enhancement, a set of inter-sensor spatial-temporal prediction (STP) transformations have been defined [20], and a common framework for noise reduction with spatial prediction (SP) in time-domain has been proposed [21]. A frequency-domain realization of SP method have been implemented and compared with the SDW-MWF which showed improved performance [22]. Although, the theory of partial coherence is well-established in optics [23]; just recently coherency of speech and interferences became a subject of interest in speech enhancement. Beamforming with trade-off between coherent and incoherent noise was proposed recently with such a view [24]. A multichannel coherence-based enhancement approach has been followed in the time-domain using non-causal filters [25], and pseudo-coherence-based MVDR filters have been proposed in short-time Fourier transformation (STFT) domain for ad hoc arrays [26].

Ad hoc microphone arrays impose more challenges than arrays with fixed geometries. Their potential applications and core challenges have been comprehensively discussed in [27]. One of such signal processing challenges is to form a robust microphone subset selection strategy. A clustered approach to blind beamforming has been developed for ASR systems to resolve this challenge [28]; however, following such a machine learning approach imposes computational and networking overheads to the ad hoc speech enhancement systems. Distributed signal processing is a solution to reduce transmission overheads by processing signals locally in sub-arrays instead of transferring all signals to a fusion center. This may also reduce the computational overheads as most algorithms include matrix inversion which has higher complexity for bigger matrices in centralized processing.

There have been several attempts to adapt most of the aforementioned approaches (with directional or spatial cues) into distributed and ad hoc microphone arrays. Bertrand et al. proposed both sequential round robin fusion [29] and simultaneous node parameter update [30], which are both suboptimal to the centralized optimization, and are applicable to determined situations. For under-determined scenarios, the concept of geometrically constrained TRINICON has been proposed in [31], which uses coarse DOA estimations to estimate RTFs. A distributed GSC is proposed in [32], which perform RTF-based filtering in local and global stages iteratively. Distributed algorithms for MVDR beamformer was proposed based on randomized gossip [33], [34]. Communication overload for multi-speaker scenarios has also been reduced using local beamforming and partial transmission channels [35]. The problem of optimal node selection, which is NP-hard in nature, has been tackled with a centralized greedy strategy in [36]. Recently, three distributed LCMV beamformers and their closeness to the optimal centralized realization have been studied in [37] in fully connected WASNs. State-of-the-art DOA-based approaches have also been proposed, for an informed parametric spatial filter [38], and for cooperative integrated noise reduction in fully connected WASNs [39].

Despite all these efforts, a well established framework taking advantage of the signal (or noise) model for optimal processing ad hoc microphone arrays is still missing. Such an abstraction is deduced in this work. The signal model herein is formed based on pseudo-coherence vectors and matrices, respectively, for enhancing speech signals in single and multiple speaker scenarios. The framework in this paper shares familiar methods with other common techniques reviewed in this section; however, this model is more insightful than others since coherency is a characteristic of the speech signals bearing all required information for optimal processing.

The rest of this paper is organized as follows. In Section II, which extends previous works in [26], [40], the theoretical framework is introduced with signal models based on inter and intra pseudo-coherence vectors and matrices. The blind estimation of coherencies is not in the scope of this paper; however, such an approach can be found in [41]. This section continues with sub-array and reference selection criteria based on the norm of pseudo-coherence vectors and the input SINR. Section III starts with derivation of MVDR and SDW-MWF beamformers within the proposed framework, both for local and global schemes. A closed-form distributed estimation of the error covariance matrix which potentially reduces network overloads is derived here. The section ends with derivation of pseudo-coherence-based LCMV and SDW-MWF in matrix form which enables joint restoration of multiple speech signals within the proposed framework. Experiments regarding the proposed framework are presented in Section IV. Firstly, the performance measures are explained, then practicality of the proposed framework is shown through experiments on multichannel audio database. The section ends with another experiment mimicking a teleconferencing set up, to compare enhancement apparatus without noise or interference estimations. The paper is finalized with conclusions in Section V.

II. Theoretical Framework

A. Signal Model and Problem Formulation

We consider the context of an ad hoc microphone array in which a set of $n \in \{1, \ldots, N\}$ randomly positioned sub-arrays are deployed in a reverberant acoustic environment. Each sub-array consists of a number of omni-directional sensors, $M(n)$, within an unknown geometry. At each time index, t, the m-th microphone of the n-th sub-array captures desired convolved source signals, $x^p_{n,m}(t)$, each contaminated with interference signals from all other speech sources, $i^p_{n,m}(t)$, and additive noise, $v_{n,m}(t)$. This can be expressed as

$$y_{n,m}(t) = x^p_{n,m}(t) + i^p_{n,m}(t) + v_{n,m}(t). \tag{1}$$

For $p \in \{1, \ldots, P\}$ distinct speakers, the convoluted desired and interference signals are defined as

$$\begin{aligned} x^p_{n,m}(t) &\triangleq g^p_{n,m}(t) * s^p(t), \\ i^p_{n,m}(t) &= \sum_{\substack{q=1 \\ q\neq p}}^{P} x^q_{n,m}(t) \triangleq \sum_{\substack{q=1 \\ q\neq p}}^{P} g^q_{n,m}(t) * s^q(t), \end{aligned}$$

where $g^p_{n,m}(t)$ and $g^q_{n,m}(t)$, respectively, are the acoustic impulse responses from the desired source, $s^p(t)$, and each competitive interfering source, $s^q(t)$, to the m-th microphone of the n-th sub-array. We assume that the acoustic impulse responses are time invariant. We also assume that the signals $x^p_{n,m}(t)$ and $v_{n,m}(t)$ are zero mean, stationary, real, broadband, and mutually uncorrelated. By definition, $x^q_{n,m}(t)$ and $x^p_{n,m}(t)$, $q \neq p$, are self-coherent across the sub-arrays, but they are not mutually-coherent since random speeches from different sources typically have different pitches and harmonics and further they do not have constant phase relationship during the same time index. The noise signal, $v_{n,m}(t)$, is typically only partially coherent across sub-arrays.

Among all microphones in the n-th sub-array, one specific microphone, b^p-th, captures the best clean (but convoluted) desired speech signal for speaker p. This node is called the local reference microphone for the p-th speaker, and the desired signal captured at this microphone is called the local reference signal for speaker p at sub-array n, i.e., $x^p_{n,b^p}(t)$. From this point forward, we remove the redundant superscript p for b, and designate the local reference signal with $x^p_{n,b}(t)$. For each speaker, a set of N local reference signals exists for the whole ad hoc array, $\{x^p_{1,b}(t), x^p_{2,b}(t), \ldots, x^p_{N,b}(t)\}$. Then, many questions arise. Which microphone of each sub-array represents the best local reference signal? Which one of these local reference signals should be granted as the global reference signal for the whole ad hoc array? Which one is the best and in which term? etc. In the rest, we will try to formulate the problem in the best way we can in order to be able to answer some of these fundamental questions.

Using the short-time Fourier transform (STFT), (1) can be rewritten in the time-frequency domain as

$$Y_{n,m}(k,l) = X^p_{n,m}(k,l) + I^p_{n,m}(k,l) + V_{n,m}(k,l), \tag{2}$$

where $Y_{n,m}(k,l)$, $X^p_{n,m}(k,l)$, $I^p_{n,m}(k,l)$, and $V_{n,m}(k,l)$ are the STFT-domain representations of $y_{n,m}(t)$, $x^p_{n,m}(t)$, $i^p_{n,m}(t)$, and $v_{n,m}(t)$, respectively, at time frame l and frequency bin $k \in \{0, \ldots, K-1\}$. From this point forward, whenever there is no ambiguity, we omit the time frame and frequency bin indices for the sake of readability.

Assuming a sufficiently long analysis window, the following equations hold:

$$\begin{aligned} X^p_{n,m} &= G^p_{n,m}(k) S^p, \\ I^p_{n,m} &= \sum_{\substack{q=1 \\ q\neq p}}^{P} X^q_{n,m} = \sum_{\substack{q=1 \\ q\neq p}}^{P} G^q_{n,m}(k) S^q, \end{aligned}$$

where $G^p_{n,m}(k)$ is the acoustic transfer function between the source p and the m-th microphone of sub-array n.

It is more convenient to write the M STFT-domain microphone signals of the n-th sub-array in a vector notation:

$$\mathbf{y}_n = \mathbf{x}^p_n + \mathbf{i}^p_n + \mathbf{v}_n = \mathbf{d}^p_n(k) X^p_{n,b} + \mathbf{e}^p_n, \tag{3}$$

where $X^p_{n,b}$ is the local reference signal in the STFT-domain, stacked signals at the n-th sub-array are

$$\begin{aligned} \mathbf{y}_n &= \begin{bmatrix} Y_{n,1} & Y_{n,2} & \cdots & Y_{n,M(n)} \end{bmatrix}^T, \\ \mathbf{x}^p_n &= \begin{bmatrix} X^p_{n,1} & X^p_{n,2} & \cdots & X^p_{n,M} \end{bmatrix}^T = \mathbf{g}^p_n(k) S^p, \\ \mathbf{i}^p_n &= \begin{bmatrix} I^p_{n,1} & I^p_{n,2} & \cdots & I^p_{n,M} \end{bmatrix}^T = \sum_{\substack{q=1 \\ q\neq p}}^{P} \mathbf{x}^q_n \\ \mathbf{v}_n &= \begin{bmatrix} V_{n,1} & V_{n,2} & \cdots & V_{n,M(n)} \end{bmatrix}^T, \\ \mathbf{e}^p_n &= \mathbf{i}^p_n + \mathbf{v}_n, \end{aligned}$$

the transcript $[\cdot]^T$ denotes the transpose operator, and the stacked acoustic transfer functions for desired and interfering speech sources are ($\forall p \in \{1, \ldots, P\}$)

$$\mathbf{g}^p_n(k) = \begin{bmatrix} G^p_{n,1}(k) & \cdots & G^p_{n,M(n)}(k) \end{bmatrix}^T.$$

The stacked relative transfer functions, respectively, are

$$\mathbf{d}^p_n(k) = \begin{bmatrix} \frac{G^p_{n,1}(k)}{G^p_{n,b}(k)} & \cdots & \frac{G^p_{n,M(n)}(k)}{G^p_{n,b}(k)} \end{bmatrix}^T = \frac{\mathbf{g}^p_n(k)}{G^p_{n,b}(k)},$$

where it is assumed that $G^p_{n,b}(k) \neq 0$.

Expression (3) depends explicitly on the local reference signal, $X^p_{n,b}$, so that it is an appropriate signal model for our goal. The vector $\mathbf{d}^p_n(k)$ conveys relative delay and decay among signals from source p (and its images in a reverberant environment) captured by sensors in the sub-array n; therefore, may be regarded as a generalized steering vector for this sub-array towards the p-th speaker.

It can be verified [40] that a more interesting way to write (3) is

$$\mathbf{y}_n = \boldsymbol{\rho}_{\mathbf{x}^p_n, X^p_{n,b}} X^p_{n,b} + \mathbf{e}^p_n, \tag{4}$$

where

$$\boldsymbol{\rho}_{\mathbf{x}^p_n, X^p_{n,b}} = \frac{E\left[\mathbf{x}^p_n X^{p*}_{n,b}\right]}{E\left[\left|X^p_{n,b}\right|^2\right]} \approx \mathbf{d}^p_n(k)$$

is the intra-array pseudo-coherence vector [of length $M(n)$] between $\mathbf{x}_n^p$ and the local reference signal $X_{n,b}^p$, with $E[\cdot]$ and superscript * denoting mathematical expectation and complex conjugate, respectively. Notice that the component of the vector $\boldsymbol{\rho}_{\mathbf{x}_n^p, X_{n,b}^p}$ corresponding to the reference microphone is always equal to 1.

Statistically speaking, the equality $\boldsymbol{\rho}_{\mathbf{x}_n^p, X_{n,b}^p} = \mathbf{d}_n^p(k)$ never holds exactly, unless the STFT analysis window is infinitely long; however, $\boldsymbol{\rho}_{\mathbf{x}_n^p, X_{n,b}^p}$ converges to $\mathbf{d}_n^p(k)$, through averaging among consequent finite analysis windows, because its stationarity is determined solely by the geometry and not by the quasi-stationary nature of speech. Furthermore, (4) is much more insightful than (3) since the pseudo-coherence vector (as the quotient of inner product and power spectral density of speech signals) captures much better the acoustic environment and provides an auxiliary norm by which sub-arrays can be ranked in the ad hoc microphone array. Therefore, from now on, only the model given in (4) will be used.

Let r^p be the array index corresponding to the best reference signal (the global reference for the p-th speaker), i.e., $X_{r^p,b}^p$. From this point forward, the redundant superscript p to r, is removed, and the global reference signal is designated with $X_{r,b}^p$. In theory, it is always possible to write (4) as a function of this selected reference signal, i.e.,

$$\mathbf{y}_n = \boldsymbol{\rho}_{\mathbf{x}_n^p, X_{r,b}^p} X_{r,b}^p + \mathbf{e}_n^p, \tag{5}$$

where

$$\boldsymbol{\rho}_{\mathbf{x}_n^p, X_{r,b}^p} = \frac{E\left[\mathbf{x}_n^p X_{r,b}^{p*}\right]}{E\left[\left|X_{r,b}^p\right|^2\right]}$$

is the inter-array pseudo-coherence vector [of length $M(n)$] between $\mathbf{x}_n^p$ and the global reference signal $X_{r,b}^p$.

The covariance matrix of $\mathbf{y}_n$ can be expressed as

$$\boldsymbol{\Phi}_{\mathbf{y}_n} = E\left[\mathbf{y}_n \mathbf{y}_n^\dagger\right] = \boldsymbol{\Phi}_{\mathbf{x}_n^p} + \boldsymbol{\Phi}_{\mathbf{e}_n^p}, \tag{6}$$

where the transcript † denotes the transpose-conjugate operator, $\boldsymbol{\Phi}_{\mathbf{x}_n^p}$ is the covariance matrix (whose rank is equal to 1) of $\mathbf{x}_n^p$, and $\boldsymbol{\Phi}_{\mathbf{e}_n^p}$ is the covariance matrix of the composition of competitive speeches plus noise, called the error signal, $\mathbf{e}_n^p$. From (4), we deduce that the covariance matrix of $\mathbf{x}_n^p$ is

$$\boldsymbol{\Phi}_{\mathbf{x}_n^p} = \phi_{X_{r,b}^p} \boldsymbol{\rho}_{\mathbf{x}_n^p, X_{r,b}^p} \boldsymbol{\rho}_{\mathbf{x}_n^p, X_{r,b}^p}^\dagger, \tag{7}$$

where $\phi_{X_{r,b}^p} = E\left[\left|X_{r,b}^p\right|^2\right]$ is the variance of $X_{r,b}^p$. Furthermore, the covariance matrix of $\mathbf{e}_n^p$ can be decomposed into

$$\boldsymbol{\Phi}_{\mathbf{e}_n^p} = E\left[\mathbf{e}_n^p \mathbf{e}_n^{p\,\dagger}\right] = \boldsymbol{\Phi}_{\mathbf{v}_n} + \boldsymbol{\Phi}_{\mathbf{i}_n^p} = \boldsymbol{\Phi}_{\mathbf{v}_n} + \sum_{\substack{q=1 \\ q \neq p}}^{P} \boldsymbol{\Phi}_{\mathbf{x}_n^q}, \tag{8}$$

where $\boldsymbol{\Phi}_{\mathbf{v}_n} = E\left[\mathbf{v}_n \mathbf{v}_n^\dagger\right]$ is the covariance matrix [whose rank is assumed to be equal to $M(n)$] of $\mathbf{v}_n$, and $\boldsymbol{\Phi}_{\mathbf{x}_n^q}$ is the covariance matrix of the q-th interfering speech, $\mathbf{x}_n^q$.

Temporal smoothing is required in practice to obtain statistically valid estimates for covariance matrices in (6)–(8). For example, the covariance matrix for $\mathbf{e}_n^p$ may be recursively smoothed with a forgetting factor, $0 \leq \gamma \leq 1$, as

$$\begin{aligned}\boldsymbol{\Phi}_{\mathbf{e}_n^p}(l) = (1-\gamma)\boldsymbol{\Phi}_{\mathbf{e}_n^p}(l-1) + \gamma \mathbf{v}_n(l)\mathbf{v}_n^\dagger(l) \\ + \gamma \sum_{\substack{q=1 \\ q \neq p}}^{P} \phi_{X_{r,b}^q}(l) \boldsymbol{\rho}_{\mathbf{x}_n^q, X_{r,b}^q} \boldsymbol{\rho}_{\mathbf{x}_n^q, X_{r,b}^q}^\dagger,\end{aligned} \tag{9}$$

assuming that pseudo-coherence vectors are independent of time-frame (for a stationary geometry).

In theory, we can rewrite (4) taking into account pseudo-coherence vectors for all speech signals, as

$$\mathbf{y}_n = \sum_{p=1}^{P} \boldsymbol{\rho}_{\mathbf{x}_n^p, X_{r,b}^p} X_{r,b}^p + \mathbf{v}_n = \boldsymbol{P}_n \bar{\mathbf{X}} + \mathbf{v}_n, \tag{10}$$

where

$$\boldsymbol{P}_n = \begin{bmatrix} \boldsymbol{\rho}_{\mathbf{x}_n^1, X_{r,b}^1} & \cdots & \boldsymbol{\rho}_{\mathbf{x}_n^P, X_{r,b}^P} \end{bmatrix}$$

is the intra-array pseudo-coherence matrix of size $M(n) \times P$ for the n-th sub-array composed of self-coherence vectors for all speech signals at this sub-array, and

$$\bar{\mathbf{X}} = \begin{bmatrix} X_{r,b}^1 & \cdots & X_{r,b}^P \end{bmatrix}^T$$

is the reference vector of length P, which contains global reference signals for all speakers. Notice that elements of this vector may not belong to the same sub-array, as the geometric pose of sub-arrays (proximity, orientation, etc.) are different for distinct speakers.

By stacking all vectors and matrices for the N sub-arrays, we can rewrite the signal model suitable for the multi-speaker ad hoc microphone array with received signal vector of length $M_{\text{tot}} = \sum_{n=1}^{N} M(n)$:

$$\bar{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_1^T & \cdots & \mathbf{y}_N^T \end{bmatrix}^T = \boldsymbol{P} \bar{\mathbf{X}} + \bar{\mathbf{v}}, \tag{11}$$

where

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_1^T & \cdots & \boldsymbol{P}_N^T \end{bmatrix}^T = \begin{bmatrix} \boldsymbol{\rho}_{\mathbf{x}_1^1, X_{r,b}^1} & \cdots & \boldsymbol{\rho}_{\mathbf{x}_1^P, X_{r,b}^P} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\rho}_{\mathbf{x}_N^1, X_{r,b}^1} & \cdots & \boldsymbol{\rho}_{\mathbf{x}_N^P, X_{r,b}^P} \end{bmatrix}$$

is the pseudo-coherence matrix of size $M_{\text{tot}} \times P$ for the ad hoc microphone array, and

$$\bar{\mathbf{v}} = \begin{bmatrix} \mathbf{v}_1^T & \cdots & \mathbf{v}_N^T \end{bmatrix}^T$$

is the noise vector of length M_{tot} composed of late reverberated sounds, diffused, and sensor noise components.

B. Subarray/Reference Selection Criteria

1) The Norm of Pseudo-Coherence Vector: In the previous section, we mentioned that for the p-th speaker, N distinct local references exist: $X_{1,b}^p, X_{2,b}^p, \ldots, X_{N,b}^p$. It is important to

be able to choose one desired reference signal within this set in order to estimate it correctly using a local beamformer. For the sake of comparability, we use this best local reference signal also in global beamformers, and denote it the global reference signal, $X^p_{r,b}$.

The intra-array pseudo-coherence vector, $\boldsymbol{\rho}_{\mathbf{x}^p_n, X^p_{n,b}}$, tells us how much $X^p_{n,b}$ is coherent with the other convolved desired signals, $X^p_{n,m}$, $m = \{1 : M(n)\} \setminus \{b\}$, of the n-th sub-array. Let us define the intra-array quantity:

$$\aleph^p_n = \left|\left|\boldsymbol{\rho}_{\mathbf{x}^p_n, X^p_{n,b}}\right|\right|^2_2 = \boldsymbol{\rho}^{\dagger}_{\mathbf{x}^p_n, X^p_{n,b}} \boldsymbol{\rho}_{\mathbf{x}^p_n, X^p_{n,b}}. \tag{12}$$

We select b to be the closest microphone in the n-th sub-array to the p-th speaker, i.e.,

$$b^p = \arg\max_m |X^p_{n,m}|. \tag{13}$$

Then, we always have $1 \leq \aleph^p_n \leq M(n)$. The worst scenario is when $\aleph^p_n$ is close to 1, which means that the sub-array n captures almost no desired speech.

It is clear that for two sub-arrays i and j, a value of $\aleph^p_i$ greater than a value of $\aleph^p_j$ means that the desired speech signal is captured better by the sub-array i than the sub-array j. There are several geometric reasons for this, including proximity to the desired speaker, better orientation, number of nodes within sub-array, physical extent of the sub-array, etc. As a result, we should try to recover $X^p_{i,b}$ rather than $X^p_{j,b}$.

A global optimum selection is obtained when we estimate or recover $X^p_{r,b}$, where r^p is chosen to maximize the norm of intra-array pseudo-coherence vector, $\aleph^p_n$, such that

$$r^p = \arg\max_n \aleph^p_n. \tag{14}$$

In other words, $X^p_{r,b}$ is our global reference signal that we will try to estimate with a beamforming algorithm.

Now that we have the global desired signal, it is of great importance to quantify how much the sub-arrays (other than the one containing the global reference signal, i.e., r^p) can contribute to noise reduction. For that, we can define the inter-sub-array quantity

$$\aleph^p_{n|r} = \left|\left|\boldsymbol{\rho}_{\mathbf{x}^p_n, X^p_{r,b}}\right|\right|^2_2. \tag{15}$$

We always have

$$0 \leq \aleph^p_{n|r} \leq \aleph^p_r.$$

The worst scenario is when $\aleph^p_{n,b}$ is close to zero, which means that array n will have little or no positive contribution in the estimation of $X^p_{r,b}$. The measure in (12) tells us how much array n can "hear" the reference signal, $X^p_{r,b}$.

2) The Input SINR: One fundamental measure in speech enhancement is the averaged (narrowband) input signal-to-interference-plus-noise ratio (SINR) for the p-th speaker at the n-th sub-array with a local reference, using (8):

$$\text{iSINR}^p_n = \frac{\text{tr}\left[\boldsymbol{\Phi}_{\mathbf{x}^p_n}\right]}{\text{tr}\left[\boldsymbol{\Phi}_{\mathbf{e}^p_n}\right]} = \frac{\aleph^p_n \phi_{X^p_{n,b}}}{\sum\limits_{\substack{q=1 \\ q\neq p}}^{P} \text{tr}\left[\boldsymbol{\Phi}_{\mathbf{x}^q_n}\right] + \text{tr}\left[\boldsymbol{\Phi}_{\mathbf{v}_n}\right]}, \tag{16}$$

where tr$[\cdot]$ denotes the trace of a square matrix.

Another interesting way to choose the global reference signal is the following:

$$r'^p = \arg\max_n \text{iSINR}^p_n. \tag{17}$$

In this case, we estimate $X^p_{r',b}$. The ideal case is when $r^p = r'^p$, which means that both criteria are fulfilled, but in general, $r^p \neq r'^p$, and it is not clear at this point which criterion should be used to find the desired signal.

Theoretically speaking, (16) can also be recalculated with respect to the global reference for speaker p as

$$\text{iSINR}^p_{n|r'} = \frac{\aleph^p_{n|r'} \phi_{X^p_{r',b}}}{\sum\limits_{\substack{q=1 \\ q\neq p}}^{P} \text{tr}\left[\boldsymbol{\Phi}_{\mathbf{x}^q_n}\right] + \text{tr}\left[\boldsymbol{\Phi}_{\mathbf{v}_n}\right]}, \tag{18}$$

where $\phi_{X^p_{r,b}}$ is the variance of $X^p_{r,b}$, and

$$0 \leq \text{iSINR}^p_n \leq \text{iSINR}^p_{n|r'}.$$

The averaged (narrowband) input SINR with all the distributed arrays is defined as

$$\text{iSINR}^p = \frac{\sum\limits_{n=1}^{N} \aleph_n \phi_{X^p_{n,b}}}{\sum\limits_{n=1}^{N} \text{tr}\left[\boldsymbol{\Phi}_{\mathbf{e}^p_n}\right]}. \tag{19}$$

It can be shown that

$$0 \leq \text{iSINR}^p \leq \text{iSINR}^p_n.$$

III. Enhancement Techniques

A. Pseudo-Coherence-Based MVDR Beamforming

In this section, we consider the concept of the MVDR beamforming [2], [42] for noise reduction with ad hoc microphone arrays in the presented framework. We establish two local schemes with best input SINR and best output SINR, and formulate the global scheme, which in theory is superior to both local schemes. Single speaker scenarios have been introduced in [26]; therefore, we complement the subsection with arguments and techniques, such as round-robin error covariance matrix update for multi-speaker scenario.

1) Best Input SINR Subarray: The easiest way to recover a desired signal, $X^p_{r',b}$, with an ad hoc microphone array is to select the sub-array with the best input SINR, indexed with $r'p$ obtained in (17), and subsequently ignoring all other sub-arrays. In this case, the beamformer output for the p-th speech signal is

$$Z^p = \mathbf{h}^{p\dagger}_{r'} \mathbf{y}_{r'^p},$$

where $\mathbf{h}^p_{r'}$ is a complex filter of length $M(n)$ containing all the complex gains applied to the microphone outputs of the array r'^p at each time-frequency bin.

Distortionless noise reduction for the p-th speech signal can be obtained by minimizing the variance of the beamformer output, Z^p, constrained with the preservation of the desired (reference) signal, $X^p_{r',b}$, subject to the narrowband weights, $\mathbf{h}^p_{r'}$. This can be formulated in the following constrained optimization problem:

$$\min_{\mathbf{h}^p_{r'}} \quad \mathbf{h}^{p\dagger}_{r'} \mathbf{\Phi}_{\mathbf{y}_{r'p}} \mathbf{h}^p_{r'} \quad \text{s.t.} \quad \mathbf{h}^{p\dagger}_{r'} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}} = 1,$$

Using the method of Lagrange multipliers, it is trivial to show that the MVDR optimal filter is equal to

$$\mathbf{h}^p_{r'} = \frac{\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}}}{\boldsymbol{\rho}^{\dagger}_{\mathbf{x}^p_{r'}, X^p_{r',b}} \mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}}}. \tag{20}$$

The explicit dependence of the above filter on $\boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}}$ can be eliminated to obtain

$$\begin{aligned} \mathbf{h}^p_{r'} &= \frac{\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}} \mathbf{\Phi}_{\mathbf{x}^p_{r'}}}{\text{tr}\left[\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}} \mathbf{\Phi}_{\mathbf{x}^p_{r'}}\right]} \mathbf{c}^p_{M,b} = \frac{\mathbf{C}_M - \mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}} \mathbf{\Phi}_{\mathbf{e}^p_{r'}}}{M - \text{tr}\left[\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}} \mathbf{\Phi}_{\mathbf{e}^p_{r'}}\right]} \mathbf{c}^p_{M,b} \\ &= \frac{\mathbf{C}_M - \sum_{\substack{q=1 \\ q\neq p}}^{P} \mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}} \mathbf{\Phi}_{\mathbf{x}^q_{r'}} - \mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}} \mathbf{\Phi}_{\mathbf{v}^p_{r'}}}{M - \sum_{\substack{q=1 \\ q\neq p}}^{P} \text{tr}\left[\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}} \mathbf{\Phi}_{\mathbf{x}^q_{r'}}\right] - \text{tr}\left[\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}} \mathbf{\Phi}_{\mathbf{v}^p_{r'}}\right]} \mathbf{c}^p_{M,b}, \end{aligned} \tag{21}$$

where $\mathbf{c}^p_{M,b}$ is the b^p-th column of the $M(n) \times M(n)$ identity matrix, $\mathbf{C}_M$, corresponding to the reference microphone.

If $\mathbf{\Phi}_{\mathbf{y}_{r'p}}$ (full rank) is temporally smoothed such that

$$\mathbf{\Phi}_{\mathbf{y}_{r'p}}(l) = (1-\gamma)\mathbf{\Phi}_{\mathbf{y}_{r'p}}(l-1) + \gamma \mathbf{y}_{r'p}(l)\mathbf{y}^{\dagger}_{r'p}(l),$$

the following Theorem would be useful in calculating $\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}}$.

Theorem 1 (The Sherman-Morrison Formula): Suppose that $\mathbf{W}$ is an invertible square matrix and $\mathbf{u}$, $\mathbf{v}$ are vectors. Suppose furthermore that $1 + \mathbf{v}^{\dagger}\mathbf{W}^{-1}\mathbf{u} \neq 0$. Then

$$\begin{aligned} (\mathbf{W} + \mathbf{u} \otimes \mathbf{v})^{-1} &= \mathbf{W}^{-1} - \frac{\mathbf{W}^{-1}(\mathbf{u} \otimes \mathbf{v})\mathbf{W}^{-1}}{1+\lambda} \\ &= \mathbf{W}^{-1} - \frac{(\mathbf{W}^{-1}\mathbf{u}) \otimes (\mathbf{v}\mathbf{W}^{-1})}{1+\lambda}, \end{aligned}$$

where $\lambda = \mathbf{v}^{\dagger}\mathbf{W}^{-1}\mathbf{u}$, and $\mathbf{u} \otimes \mathbf{v}$ is the outer product of two vectors $\mathbf{u}$ and $\mathbf{v}$.

Proof: A proof of the Sherman-Morisson Formula using the power series expansion can be found in [43]. ■

The computational complexity of $\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}}(l)$ can be reduced from $\mathcal{O}(M(n)^3)$ to $\mathcal{O}(M(n)^2)$ using

$$\begin{aligned} \mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}}(l) &= (1-\gamma)^{-1}\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}}(l-1) \\ &\quad - \frac{\gamma \mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}}(l-1)\mathbf{y}_{r'p}(l)\mathbf{y}^{\dagger}_{r'p}(l)\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}}(l-1)}{(1-\gamma)^2 + \gamma(1-\gamma)\mathbf{y}^{\dagger}_{r'p}(l)\mathbf{\Phi}^{-1}_{\mathbf{y}_{r'p}}(l-1)\mathbf{y}_{r'p}(l)}. \end{aligned}$$

Another way to derive the MVDR beamformer is by minimizing the variance of the error signal for the p-th speech signal, $\mathbf{e}^p_{r'}$, constrained with the preservation of the desired (reference) signal, $X^p_{r',b}$, subject to the narrowband weights, $\mathbf{h}^p_{r'}$. Then, the constrained optimization problem would be

$$\min_{\mathbf{h}^p_{r'}} \quad \mathbf{h}^{p}_{r'} \mathbf{\Phi}_{\mathbf{e}^p_{r'}} \mathbf{h}^p_{r'} \quad \text{s.t.} \quad \mathbf{h}^{p\dagger}_{r'} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}} = 1,$$

which yields the closed form solution to optimal weights as

$$\begin{aligned} \mathbf{h}^p_{r'} &= \frac{\mathbf{\Phi}^{-1}_{\mathbf{e}^p_{r'}} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}}}{\boldsymbol{\rho}^{\dagger}_{\mathbf{x}^p_{r'}, X^p_{r',b}} \mathbf{\Phi}^{-1}_{\mathbf{e}^p_{r'}} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}}} = \frac{\mathbf{\Phi}^{-1}_{\mathbf{e}^p_{r'}} \mathbf{\Phi}_{\mathbf{y}_{r'p}} - \mathbf{C}_M}{\text{tr}\left[\mathbf{\Phi}^{-1}_{\mathbf{e}^p_{r'}} \mathbf{\Phi}_{\mathbf{y}_{r'p}}\right] - M} \mathbf{c}^p_{M,b} \\ &= \frac{\left(\sum_{\substack{q=1 \\ q\neq p}}^{P} \mathbf{\Phi}_{\mathbf{x}^q_{r'}} + \mathbf{\Phi}_{\mathbf{v}^p_{r'}}\right)^{-1} \mathbf{\Phi}_{\mathbf{y}_{r'p}} - \mathbf{C}_M}{\text{tr}\left[\left(\sum_{\substack{q=1 \\ q\neq p}}^{P} \mathbf{\Phi}_{\mathbf{x}^q_{r'}} + \mathbf{\Phi}_{\mathbf{v}^p_{r'}}\right)^{-1} \mathbf{\Phi}_{\mathbf{y}_{r'p}}\right] - M} \mathbf{c}^p_{M,b}. \end{aligned} \tag{22}$$

Assuming that $\mathbf{\Phi}_{\mathbf{x}^q_{r'}}$ have all rank 1, and $|\mathbf{\Phi}_{\mathbf{v}^p_{r'}}| \neq 0$, then the following theorem is useful in calculating $\mathbf{\Phi}^{-1}_{\mathbf{e}^p_{r'}}$:

Theorem 2: Let $\mathbf{A}$ and $\mathbf{A}+\mathbf{B}$ be nonsingular matrices, and let $\mathbf{B}$ have rank $P-1>0$. Let $\mathbf{B} = \mathbf{B}_1 + \cdots + \mathbf{B}_{P-1}$, where each $\mathbf{B}_q$ has rank 1, and each $\mathbf{W}_{q+1} = \mathbf{A} + \mathbf{B}_1 + \cdots + \mathbf{B}_q$ is nonsingular. Setting $\mathbf{W}_1 = \mathbf{A}$, then

$$\mathbf{W}^{-1}_{q+1} = \mathbf{W}^{-1}_q - g_q \mathbf{W}^{-1}_q \mathbf{B}_q \mathbf{W}^{-1}_q,$$

where

$$g_q = \frac{1}{1 + \text{tr}\left[\mathbf{W}^{-1}_q \mathbf{B}_q\right]}.$$

Proof: A proof of this theorem can be found in [44]. ■

By putting $\mathbf{A} = \mathbf{\Phi}_{\mathbf{v}_{r'p}}$ and

$$\begin{aligned} \mathbf{B}_1 &= \mathbf{\Phi}_{\mathbf{x}^1_{r'}}, \quad \cdots \quad , \mathbf{B}_{p-1} = \mathbf{\Phi}_{\mathbf{x}^{p-1}_{r'}}, \\ \mathbf{B}_p &= \mathbf{\Phi}_{\mathbf{x}^{p+1}_{r'}}, \quad \cdots \quad , \mathbf{B}_{P-1} = \mathbf{\Phi}_{\mathbf{x}^{P}_{r'}}, \end{aligned}$$

then the inverse of the covariance matrix for the error signal for the p-th speech signal is found recursively as

$$\mathbf{\Phi}^{-1}_{\mathbf{e}^p_{r'}} = \mathbf{W}_P.$$

Theorem 2 can be used in a round robin manner by propagating the intermediate states of error covariance matrix along all other nodes, and updating them with correction terms.

The constrained optimization problems, introduced in this subsection, are independently feasible, yet it is possible to perform them simultaneously to obtain a weight matrix, of size $M(n) \times P$ in which each column corresponds to a desired source.

2) Complete Ad Hoc Microphone Array: If, from previous criteria, we consider that all the distributed arrays can contribute to noise reduction, then they should all be used in beamforming and this solution is the optimal one. It is assumed that $X^p_{r,b}$ is found to be the best reference signal.

The beamformer output is now

$$Z^p = \bar{\mathbf{h}}^{p\dagger}\bar{\mathbf{y}},$$

where $\bar{\mathbf{h}}^{p\dagger}$ is a complex filter (of length M_{tot}) containing all the complex gains applied to the microphone outputs of all arrays at frequency bin k and

$$\bar{\mathbf{y}} = \bar{\mathbf{x}}^p + \bar{\mathbf{e}}^p = \boldsymbol{\rho}_{\bar{\mathbf{x}}^p, X^p_{r,b}} X^p_{r,b} + \bar{\mathbf{e}}^p, \tag{23}$$

with

$$\boldsymbol{\rho}_{\bar{\mathbf{x}}^p_n, X^p_{r,b}} = \frac{E\left[\bar{\mathbf{x}}^p_n X^{p*}_{r,b}\right]}{E\left[\left|X^p_{r,b}\right|^2\right]}$$

being the pseudo-coherence vector (of length M_{tot}) between $\bar{\mathbf{x}}^p$ and $X^p_{r,b}$.

The minimization of the variance of Z with distortionless constraint, $\bar{\mathbf{h}}^\dagger \boldsymbol{\rho}_{\bar{\mathbf{x}}^p_n, X^p_{r,b}} = 1$, leads to the MVDR filter:

$$\bar{\mathbf{h}}^p = \frac{\boldsymbol{\Phi}^{-1}_{\bar{\mathbf{y}}} \boldsymbol{\rho}_{\bar{\mathbf{x}}^p, X^p_{r,b}}}{\boldsymbol{\rho}^\dagger_{\bar{\mathbf{x}}^p, X^p_{r,b}} \boldsymbol{\Phi}^{-1}_{\bar{\mathbf{y}}} \boldsymbol{\rho}_{\bar{\mathbf{x}}^p, X^p_{r,b}}} = \frac{\boldsymbol{\Phi}^{-1}_{\bar{\mathbf{e}}} \boldsymbol{\rho}_{\bar{\mathbf{x}}^p, X^p_{r,b}}}{\boldsymbol{\rho}^\dagger_{\bar{\mathbf{x}}^p, X^p_{r,b}} \boldsymbol{\Phi}^{-1}_{\bar{\mathbf{e}}} \boldsymbol{\rho}_{\bar{\mathbf{x}}^p, X^p_{r,b}}}, \tag{24}$$

where $\boldsymbol{\Phi}_{\bar{\mathbf{y}}}$ is the covariance matrix of $\bar{\mathbf{y}}$, and $\boldsymbol{\Phi}_{\bar{\mathbf{e}}}$ is the covariance matrix of $\bar{\mathbf{e}}$. In light of discussions in Section III-A1, it is possible to calculate inverse of the global covariance matrix in a distributed manner, as proposed in [45].

3) Best Output SINR Subarray: The third way to recover a desired signal, $X^p_{r',b}$, with an ad hoc microphone array is to select the sub-array with the best output SINR. Firstly, we need to obtain the output SINR for the p-th speech signal for N independent sub-arrays. In this case, the n-th beamformer output for the p-th speech is

$$Z^p_n = \mathbf{h}^{p\dagger}_n \mathbf{y}_n,$$

where $\mathbf{h}^{p\dagger}_n$ is a complex filter of length $M(n)$ containing all the complex gains applied to the microphone outputs of the n-th sub-array at each time-frequency bin.

The MVDR filter is similar to the one derived in the Subsection III-A1, i.e.,

$$\mathbf{h}^p_n = \frac{\boldsymbol{\Phi}^{-1}_{\mathbf{y}_n} \boldsymbol{\rho}_{\mathbf{x}^p_n, X^p_{n,b}}}{\boldsymbol{\rho}^\dagger_{\mathbf{x}^p_n, X^p_{n,b}} \boldsymbol{\Phi}^{-1}_{\mathbf{y}_n} \boldsymbol{\rho}_{\mathbf{x}^p_n, X^p_{n,b}}} = \frac{\boldsymbol{\Phi}^{-1}_{\mathbf{e}^p_n} \boldsymbol{\rho}_{\mathbf{x}^p_n, X^p_{n,b}}}{\boldsymbol{\rho}^\dagger_{\mathbf{x}^p_n, X^p_{n,b}} \boldsymbol{\Phi}^{-1}_{\mathbf{e}^p_n} \boldsymbol{\rho}_{\mathbf{x}^p_n, X^p_{n,b}}}. \tag{25}$$

and the (narrowband) output SNR corresponding to $\mathbf{h}_n$ is

$$\text{oSINR}\left[\mathbf{h}^p_n\right] = \frac{\phi_{X^p_{n,b}}}{\mathbf{h}^{p\dagger}_n \boldsymbol{\Phi}_{\mathbf{e}^p_n} \mathbf{h}^p_n}.$$

Maximizing the output SNR with respect to the array index,

$$r''^p = \arg\max_n \text{oSINR}\left[\mathbf{h}^p_n\right], \tag{26}$$

gives us the solution we are looking for, i.e., $\mathbf{h}^p_r$, which is also a bound to the solution with best input SINR scheme.

B. Pseudo-Coherence-Based SDW-MWF Beamforming

In this section, we extract the pseudo-coherence-based SDW-MWF beamformers equivalents to those for the pseudo-coherence-based MVDR beamformers.

1) Best Input SINR Subarray: With the same assumptions in III-A, the optimization problem for enhancing the p-th desired speech signal using the SDW-MWF beamformer for the ad hoc microphone array can be formulated as

$$\begin{aligned} \min_{\mathbf{h}^p_{r'}} \quad & \mathbf{h}^{p\dagger}_{r'} \boldsymbol{\Phi}_{\mathbf{y}_n} \mathbf{h}^p_{r'} \\ \text{s.t.} \quad & \left| \left(\mathbf{c}^p_{M,b} - \mathbf{h}^p_{r'}\right)^\dagger \boldsymbol{\rho}_{\mathbf{x}^p, X^p_{r',b}} \right|^2 \le \sigma^2_p \phi^{-2}_{X^p_{r',b}}, \end{aligned}$$

where $\mathbf{c}^p_{M,b}$ is defined in Subsection III-A, $\epsilon_p = \sigma_p \phi^{-1}_{X^p_{r',b}}$ is the fraction of allowed narrowband distortion, and σ_p is the allowed amount for the narrowband distortion power at each time-frequency bin. A reasonable upper bound for σ_p is $\phi_{X^p_{r',b}}$, where we have the maximum allowed distortion. $\sigma_p = 0$ yields the same constraint as in the MVDR beamformer.

Then the optimal SDW-MWF weights are:

$$\mathbf{h}^p_{r'} = \left(\boldsymbol{\Phi}_{\mathbf{y}_{r'p}} + \lambda_p \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}} \boldsymbol{\rho}^\dagger_{\mathbf{x}^p_r, X^p_{r',b}}\right)^{-1} \boldsymbol{\rho}_{\mathbf{x}^p_{r'} X^p_{r',b}},$$

which can be simplified using the Sherman-Morrison formula for matrix inversion as

$$\mathbf{h}^p_{r'} = \frac{\lambda_p \boldsymbol{\Phi}^{-1}_{\mathbf{y}_{r'p}} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}}}{1 + \lambda_p \boldsymbol{\rho}^\dagger_{\mathbf{x}^p_r, X^p_{r',b}} \boldsymbol{\Phi}^{-1}_{\mathbf{y}_{r'p}} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}}}. \tag{27}$$

The real positive parameter, λ_p, controls the trade-off between noise reduction and speech distortion.

To find a relation between λ_p and σ_p, we put the filter weights into the constraint, so that we can write

$$0 \le \frac{1}{1 + \lambda_p \boldsymbol{\rho}^\dagger_{\mathbf{x}^p_r, X^p_{r',b}} \boldsymbol{\Phi}^{-1}_{\mathbf{y}_{r'p}} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}}} \le \epsilon_p \le 1,$$

which can be solved for λ_p (for the worse case) as

$$\lambda_p = \frac{1 - \epsilon_p}{\kappa_p \epsilon_p},$$

where $\kappa_p = \boldsymbol{\rho}^\dagger_{\mathbf{x}^p_r, X^p_{r',b}} \boldsymbol{\Phi}^{-1}_{\mathbf{y}_{r'p}} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}}$. The filter approaches the MVDR as $\lambda_p \to \infty$, i.e., $\epsilon_p \to 0$. The filter simplifies to the non-causal multichannel Wiener filter (MWF) if $\lambda_p = \phi_{X^p_{r',b}}$.

If the desired fraction of narrowband distortion at all time-frequency bins are equal to a broadband fraction, ϵ_{bb}, then the weights of the SDW-MWF filter are obtained from

$$\mathbf{h}^p_{r'} = \left(\frac{1 - \epsilon_{\text{bb}}}{\kappa_p}\right) \Phi^{-1}_{\mathbf{y}_{r'p}} \boldsymbol{\rho}_{\mathbf{x}^p_{r'}, X^p_{r',b}}.$$

However, the narrowband distortion power equal to a broadband limit, σ_{bb}, for all time-frequency bins is more desired, which yields

$$\mathbf{h}_{r'}^{p} = \left(\frac{1 - \sigma_{\text{bb}} \phi_{X_{r',b}^{p}}^{-1}}{\kappa_p} \right) \Phi_{\mathbf{y}_{r'p}}^{-1} \boldsymbol{\rho}_{\mathbf{x}_{r'}^{p}, X_{r',b}^{p}}.$$

Alternatively, the SDW-MWF beamformer can be obtained by minimizing the residual noise power constrained with the amount of speech distortion. The optimal weights for this problem are

$$\mathbf{h}_{r'}^{p} = \frac{\boldsymbol{\Phi}_{\mathbf{e}_{r'}^{p}}^{-1} \boldsymbol{\rho}_{\mathbf{x}_{r'}^{p}, X_{r',b}^{p}}}{\boldsymbol{\rho}_{\mathbf{x}_{r'}^{p}, X_{r',b}^{p}}^{\dagger} \boldsymbol{\Phi}_{\mathbf{e}_{r'}^{p}}^{-1} \boldsymbol{\rho}_{\mathbf{x}_{r'}^{p}, X_{r',b}^{p}} + \mu_p \phi_{X_{r',b}^{p}}^{-1}}, \tag{28}$$

where $\mu_p = \lambda_p^{-1} \phi_{X_{r',b}^{p}} = \kappa_p \sigma_{\text{bb}} (1 - \epsilon_{\text{bb}})^{-1}$. For $\mu_p = 1$, the conventional multichannel wiener filter is obtained.

2) Complete Ad Hoc Microphone Array and Best Output SINR Subarray: The SDW-MWF formulations obtained for the best input SINR sub-array can be extended to other spans, i.e., the complete ad hoc microphone array and the best output SINR sub-array, by substituting the appropriate covariance matrices and calculating weights respectively.

C. Multiple Speaker Enhancement Techniques

1) Pseudo-Coherence-Based M-LCMV Beamforming: In this section, we expand the pseudo-coherence-based beamforming scheme to form a pseudo-coherence-based LCMV filter for multi-speaker scenarios. We begin with the matrix-form signal model in (11) to establish the optimization problem.

The M-LCMV filter output vector, of length p, is

$$\bar{\mathbf{z}} = \begin{bmatrix} z^1 & \cdots & z^P \end{bmatrix}^T = \mathbf{H}^T \bar{\mathbf{y}},$$

where

$$\mathbf{H} = \begin{bmatrix} \bar{\mathbf{h}}^1 & \cdots & \bar{\mathbf{h}}^P \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1^1 & \cdots & \mathbf{h}_1^P \\ \vdots & \ddots & \vdots \\ \mathbf{h}_N^1 & \cdots & \mathbf{h}_N^P \end{bmatrix}$$

is the matrix of complex filter weights of size, $M_{\text{tot}} \times P$,

$$\mathbf{h}_n^p = \begin{bmatrix} H_{n,1}^p & \cdots & H_{n,M}^p \end{bmatrix}^T, \quad \forall p \in \{1, \ldots, P\}$$

is the complex weighting vector, of length $M(n)$, for the n-th sub-array to contribute in enhancing the p-th desired signal, and $\bar{\mathbf{h}}^p$ is the the p-th column of $\mathbf{H}$, with length of M_{tot}.

The M-LCMV filter is enhancing the p-th signal of interest by nulling the other speech signals and minimizing the variance of the p-th element in $\bar{\mathbf{z}}$. This can be written as a multi-objective constraint optimization problem, indeed a quadratic program (QP), as

$$\min_{\mathbf{H}} \quad \text{tr}\left[\mathbf{H}^{\dagger} \boldsymbol{\Phi}_{\bar{\mathbf{y}}} \mathbf{H} \right] \quad \text{s.t.} \quad \mathbf{H}^{\dagger} \boldsymbol{P} = \mathbf{C},$$

where $\mathbf{C}$ is the $P \times P$ identity matrix, corresponding to M-LCMV constraints, and

$$\boldsymbol{\Phi}_{\bar{\mathbf{y}}} = E\left[\mathbf{y}_n \mathbf{y}_n^{\dagger} \right] = \boldsymbol{P} \boldsymbol{\Phi}_{\bar{\mathbf{X}}} \boldsymbol{P}^{\dagger} + \boldsymbol{\Phi}_{\bar{\mathbf{v}}}, \tag{29}$$

where $\boldsymbol{\Phi}_{\bar{\mathbf{X}}}$ is the covariance matrix of $\bar{\mathbf{X}}$, and $\boldsymbol{\Phi}_{\bar{\mathbf{v}}}$ is the covariance matrix of the noise, $\bar{\mathbf{v}}$. Since $\boldsymbol{\Phi}_{\bar{\mathbf{y}}} \in \mathcal{S}_+$, i.e., positive semidefinite, the objective is convex quadratic which is minimized over a polyhedron. Moreover, if $\boldsymbol{\Phi}_{\bar{\mathbf{y}}} \in \mathcal{S}_{++}$, i.e., positive definite, the feasibility region of the above optimization problem is intersection of p ellipsoids.

The Lagrange function for this optimization problem can be written as

$$\boldsymbol{\mathcal{L}}(\mathbf{H}, \boldsymbol{\Lambda}) = \mathbf{H}^{\dagger} \boldsymbol{\Phi}_{\bar{\mathbf{y}}} \mathbf{H} + \left(\mathbf{C} - \mathbf{H}^{\dagger} \boldsymbol{P} \right) \boldsymbol{\Lambda} \left(\mathbf{C} - \mathbf{H}^{\dagger} \boldsymbol{P} \right)^{\dagger},$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix, $\text{diag}(\lambda^1, \ldots, \lambda^P)$, where each λ^p is the Lagrange multiplier for the constraint governing the p-th signal of interest, i.e., the p-th column of $\left(\mathbf{C} - \mathbf{H}^{\dagger} \boldsymbol{P} \right)$. Hence, (if the solution is feasible, i.e., $P \leq M_{\text{tot}}$) the solution for the M-LCMV beamformer in matrix format is

$$\mathbf{H} = \boldsymbol{\Phi}_{\bar{\mathbf{y}}}^{-1} \boldsymbol{P} \left(\boldsymbol{P}^{\dagger} \boldsymbol{\Phi}_{\bar{\mathbf{y}}}^{-1} \boldsymbol{P} \right)^{-1}. \tag{30}$$

Similar to the above, the M-LCMV filter for the ad hoc microphone array can be found by minimizing the variance of noise at the filter output; in this case, the multi-objective constraint optimization problem is

$$\min_{\mathbf{H}} \quad \text{tr}\left[\mathbf{H}^{\dagger} \boldsymbol{\Phi}_{\bar{\mathbf{v}}} \mathbf{H} \right] \quad \text{s.t.} \quad \mathbf{H}^{\dagger} \boldsymbol{P} = \mathbf{C}.$$

Following a similar approach to the above, the M-LCMV filter is found by

$$\mathbf{H} = \boldsymbol{\Phi}_{\bar{\mathbf{v}}}^{-1} \boldsymbol{P} \left(\boldsymbol{P}^{\dagger} \boldsymbol{\Phi}_{\bar{\mathbf{v}}}^{-1} \boldsymbol{P} \right)^{-1}. \tag{31}$$

2) Pseudo-Coherence-Based M-SDW-MWF Beamforming: In this section, we will derive a pseudo-coherence-based multi-speaker SDW-MWF beamformer for speech enhancement. We start with the same assumption in Section III-C1, however, here we form an inequality constrained optimization, more precisely a quadratically constrained quadratic program (QCQP), which is

$$\begin{aligned} \min_{\mathbf{H}} \quad & \text{tr}\left[\mathbf{H}^{\dagger} \boldsymbol{\Phi}_{\bar{\mathbf{y}}} \mathbf{H} \right] \\ \text{s.t.} \quad & \left(\mathbf{C} - \mathbf{H}^{\dagger} \boldsymbol{P} \right) \boldsymbol{\Phi}_{\bar{\mathbf{X}}} \left(\mathbf{C} - \mathbf{H}^{\dagger} \boldsymbol{P} \right)^{\dagger} \leq \boldsymbol{\Sigma}, \end{aligned}$$

where

$$\boldsymbol{\Sigma} = \text{diag}\begin{pmatrix} \sigma_1^2 & \sigma_2^2 & \cdots & \sigma_P^2 \end{pmatrix}$$

is controlling the amount of distortion for each desired signal. Since $\boldsymbol{\Phi}_{\bar{\mathbf{X}}} \in \mathcal{S}_+$, i.e., positive semidefinite, the objective and the set of p constraints are convex quadratic. Moreover, if $\boldsymbol{\Phi}_{\bar{\mathbf{X}}} \in \mathcal{S}_{++}$, i.e., positive definite, the feasibility region of the above optimization problem is intersection of p ellipsoids.

The resulting M-SDW-MWF filter is

$$\mathbf{H} = \left(\boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{\Phi}_{\bar{\mathbf{X}}} \boldsymbol{P}^{\dagger} + \boldsymbol{\Phi}_{\bar{\mathbf{y}}} \right)^{-1} \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{\Phi}_{\bar{\mathbf{X}}}. \tag{32}$$

From the fact that both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ are diagonal, we can simplify the weights matrix further

$$\mathbf{H} = \left(\sum_{p=1}^{P} \lambda_p \phi_{X_{r,b}^{p}} \boldsymbol{\rho}_{\bar{\mathbf{x}}^p, X_{r,b}^{p}} \boldsymbol{\rho}_{\bar{\mathbf{x}}^p, X_{r,b}^{p}}^{\dagger} + \boldsymbol{\Phi}_{\bar{\mathbf{y}}} \right)^{-1} \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{\Phi}_{\bar{\mathbf{X}}}, \tag{33}$$

where $\boldsymbol{\rho}_{\bar{\mathbf{x}}^p, X_{r,b}^{p}}$ is the p-th column of $\boldsymbol{P}$.

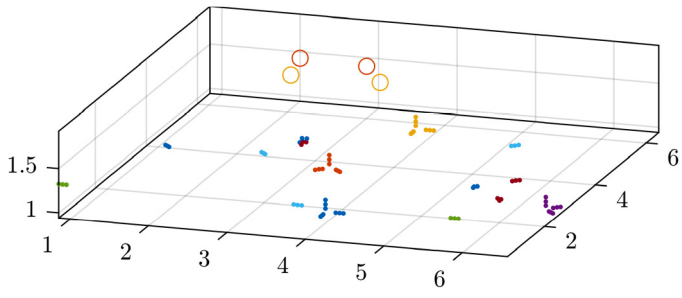


Fig. 1. Ad hoc constellations produced from SMARD.

Similar to the above, the M-SDW-MWF beamformer is obtained by minimizing the noise power at the output of beamformer with the similar constraint set. The resulting M-SDW-MWF filter is

$$\begin{aligned}\mathbf{H} &= \left(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{\Phi}_{\bar{\mathbf{X}}}\boldsymbol{P}^{\dagger} + \boldsymbol{\Phi}_{\bar{\mathbf{v}}}\right)^{-1}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{\Phi}_{\bar{\mathbf{X}}} \\ &= \left(\sum_{p=1}^{P}\lambda_p\phi_{X^p_{r,b}}\boldsymbol{\rho}_{\bar{\mathbf{x}}^p,X^p_{r,b}}\boldsymbol{\rho}^{\dagger}_{\bar{\mathbf{x}}^p,X^p_{r,b}} + \boldsymbol{\Phi}_{\bar{\mathbf{v}}}\right)^{-1}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{\Phi}_{\bar{\mathbf{X}}},\end{aligned} \tag{34}$$

which controls the trade-off between the noise reduction and speech distortion.

IV. Experiments

In this section, we study the proposed framework with experiments. Our aim is to understand more clearly the pros and cons of the framework and to compare different techniques formulated in Section III. For this reason, quantitative measures are defined in Subsection IV-A. Perceptual evaluation of speech quality (PESQ) [46] and short-time objective intelligibility measure (STOI) [47] are also used to evaluate the quality and intelligibility of the output speech signals, which are the ultimate objectives of speech enhancement. The true error covariances are used in Subsection IV-A, while the received signal covariances are used in Subsection IV-B, that give the upper and the lower bounds for the measures, respectively.

Through the rest of this section, we use recorded signals from SMARD [48] complemented with simulated data using room impulse responses obtained by the image method [49]. In our experiments, signals are down-sampled to 8000 Hz. The STFT representations are obtained using time-frames of 512 samples with 64-sample hops. As SMARD was recorded for one source at a time, its received signals in similar configurations are superposed, assuming linear response in microphones and persistent medium, i.e. fixed temperature for superposed recordings. From this, 8 constellations are obtained with 4 sets of orthogonal-linear sub-arrays, and 4 sets of linear and circular sub-arrays. Then, a subset of microphones available in each constellation is picked to form 3 sub-arrays, each contains 3 microphones, as shown in Fig. 1 with colored dots for microphones (G.R.A.S. 40AZ) and circles for superposed loud speakers (Brüel & Kjær OmniPower 4296). More details on exact locations and directions can be found in [48].

A. Performance Comparison on SMARD Constellations

By calculating the pseudo-coherence vectors, it is possible to implement and compare the enhancement techniques introduced in Section III. Blind estimation of inter and intra pseudo-coherence vectors is not in the scope of this paper, so that they are calculated from clean signals in accordance with formulations in Section II; however, we have proposed such a blind estimation approach in [41]. In practice, it is mandatory to impose limitation on time-frequency bins using a voice activity detector (VAD) for correct calculation of the norms, especially for reverberant rooms and unideal equipments, for which the signals are diminished at certain frequencies.

It is also important to take into account rank (invertibility) of the estimated error covariance matrices. To make implemented formulas from Section III robust against rank deficiencies, diagonal loading is used, which is equivalent to Tikhonov regularization in respective optimization problems. In this experiment, frequency-dependent regularization factors are used, which are equal to a small fraction (0.1%) of the long-term expected value for power spectral densities, added to the fixed level of 10^{-7} for frequencies which are heavily diminished.

The methods under study in this experiment are MVDR, MWF, and SDW-MWF (with $\mu = 5$) implemented for local and global approaches plus the global LCMV, which are respectively labeled by L-MVDR, L-MWF, L-SDW-MWF, G-MVDR, G-SDW-MWF, G-SDW-MWF, and G-LCMV. As a result of the insufficient degree of freedom, the local LCMV method fails, and is excluded.

Here, recordings for male, female, and child speakers in different constellations are used to obtain smooth charts from which reasonable conclusions can be deduced. One sample (3-5 seconds) for each speaker type is used from TSP speech audio recordings available in SMARD. SMARD constellations make it possible to compare two source positions with relatively different signal-to-interference-ratios; therefore, the free parameter in the this experiment is decided to be the amount of noise power added to the recordings at each microphone. This makes the averaging of the results be a valid approach at fixed noise levels.

1) Noise Reduction, the Output SNR: The (narrowband) output SNR of a beamformer is defined for speaker p as

$$\text{oSNR}\left[\mathbf{h}^p_r\right] = \frac{\mathbf{h}^{pH}_r\boldsymbol{\Phi}_{\mathbf{x}^p_r}\mathbf{h}^p_r}{\mathbf{h}^{pH}_r\boldsymbol{\Phi}_{\mathbf{v}_r}\mathbf{h}^p_r}.$$

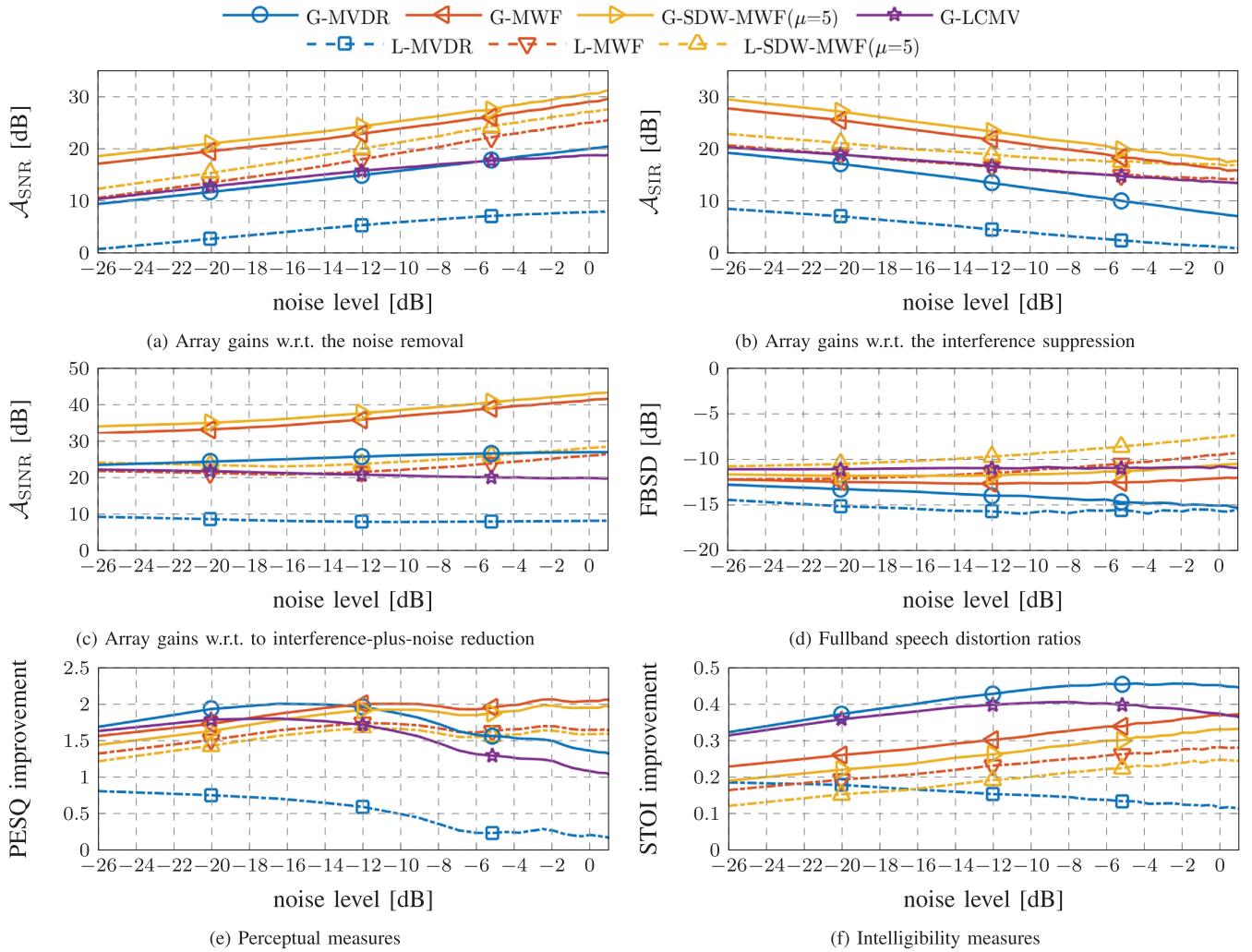
For distortionless methods, the output SNR would be

$$\text{oSNR}\left[\mathbf{h}^p_r\right] = \frac{\phi_{X^p_{r,b}}}{\mathbf{h}^{pH}_r\boldsymbol{\Phi}_{\mathbf{v}_r}\mathbf{h}^p_r}.$$

We deduce that the (narrowband) array gain for speaker p is

$$\mathcal{A}_{\text{SNR}}[\mathbf{h}^p_r] = \frac{\text{oSNR}[\mathbf{h}^p_r]}{\text{iSNR}^p_r}. \tag{35}$$

It can be shown that $\mathcal{A}[\mathbf{h}^p_r] \geq 1$. The fullband SNR array gains are obtained by firstly accumulating over all time-frequency bins and then calculating the ratio of the filtered



(a) Array gains w.r.t. the noise removal

(b) Array gains w.r.t. the interference suppression

(c) Array gains w.r.t. to interference-plus-noise reduction

(d) Fullband speech distortion ratios

(e) Perceptual measures

(f) Intelligibility measures

Fig. 2. Quantitative measures for enhancement method are shown in (a), (b), (c), and (d). Qualitative measures for these methods are shown in (e) and (f).

desired signal and the filtered noise. This measure is further averaged over all available geometries, and the results are shown in Fig. 2a. As expected, global Wiener-based methods are on top, while the local MVDR shows poor performance. The linear trend with slope less than one for all methods suggests that all beamformers are able to remove a portion (about 50%) of the excessive spatially white noise.

2) Interference Suppression, the Output SIR: Sometimes, the main objective of the enhancement algorithm is to suppress interferences, i.e., competitive speech signals. In such cases, the noise removal can be deliberately omitted if its effect on speech quality or ineligibility is negligible, or noise suppression can be performed as part of a separate stage, e.g., with a post filter. Then, the output SIR is the desirable quantitative measure, which is defined for speaker p as

$$\text{oSIR}\left[\mathbf{h}_r^p\right] = \frac{\mathbf{h}_r^{pH}\mathbf{\Phi}_{\mathbf{x}_r^p}\mathbf{h}_r^p}{\mathbf{h}_r^{pH}\mathbf{\Phi}_{\mathbf{i}_r^p}\mathbf{h}_r^p} = \frac{\mathbf{h}_r^{pH}\mathbf{\Phi}_{\mathbf{x}_r^p}\mathbf{h}_r^p}{\sum\limits_{\substack{q=1\\q\neq p}}^{P}\mathbf{h}_r^{pH}\mathbf{\Phi}_{\mathbf{x}_r^q}\mathbf{h}_r^p}.$$

The (narrowband) array gains for interference suppression is

$$\mathcal{A}_{\text{SIR}}[\mathbf{h}_r^p] = \frac{\text{oSIR}[\mathbf{h}_r^p]}{\text{iSIR}_r^p}. \tag{36}$$

The fullband SIR array gains are calculated for different methods in a similar manner to fullband SNR array gains. The results are shown in Fig. 2b. Here, the trend in array gains for interference suppression is decreasing as the noise level increases, suggesting that leaving the noise for a post filter has a risk of increased residual interferences. Notably, the global LCMV shows better performance than the global MVDR, which is expected from the constraint imposed on it.

3) Combined Measure, the Output SINR: Inclusively, the signal-to-interference-plus-noise-ratio is defined as

$$\text{oSINR}\left[\mathbf{h}_r^p\right] = \frac{\mathbf{h}_r^{pH}\mathbf{\Phi}_{\mathbf{x}_r^p}\mathbf{h}_r^p}{\mathbf{h}_r^{pH}\mathbf{\Phi}_{\mathbf{e}_r^p}\mathbf{h}_r^p} = \frac{\mathbf{h}_r^{pH}\mathbf{\Phi}_{\mathbf{x}_r^p}\mathbf{h}_r^p}{\mathbf{h}_r^{pH}\mathbf{\Phi}_{\mathbf{v}_r}\mathbf{h}_r^p + \sum\limits_{\substack{q=1\\q\neq p}}^{P}\mathbf{h}_r^{pH}\mathbf{\Phi}_{\mathbf{x}_r^q}\mathbf{h}_r^p},$$

and the narrowband array gains with regard to signal-to-interference-plus-noise-ratio is

$$\mathcal{A}_{\text{SINR}}[\mathbf{h}_r^p] = \frac{\text{oSINR}[\mathbf{h}_r^p]}{\text{iSINR}_r^p}. \tag{37}$$

The fullband combined array gains for SINR are calculated through dividing the accumulated filtered desired signal by the accumulated interference plus noise over all time-frequency

bins, and then smoothed over different geometries. As can be seen in Fig. 2c for combined fullband SINR array gains, the global Wiener-based methods are again superior to all other compared methods. Unlike for the SIR gain, the MVDR acts better than LCMV for SINR gain, and all distortionless methods perform equally the same (flat trend) for different levels of spatially white noise.

4) Waveform Preservation, Speech Distortion Ratio: Besides the comparison w.r.t the fullband array gains, it is essential for speech enhancement to study the speech distortion. For this, the fullband multichannel distortion index defined in [50] is reformulated in terms of the pseudo-coherence vector and complex filter weights as

$$\mathrm{FBSD}^p = \frac{\sum_l \sum_k \phi_{X^p_{r,b}} \left\| \left(\bar{\mathbf{u}} - \bar{\mathbf{h}}^p_r\right)^\dagger \boldsymbol{\rho}_{\bar{\mathbf{x}}^p, X^p_{r,b}} \right\|^2}{\sum_l \sum_k \phi_{X^p_{r,b}}}, \qquad (38)$$

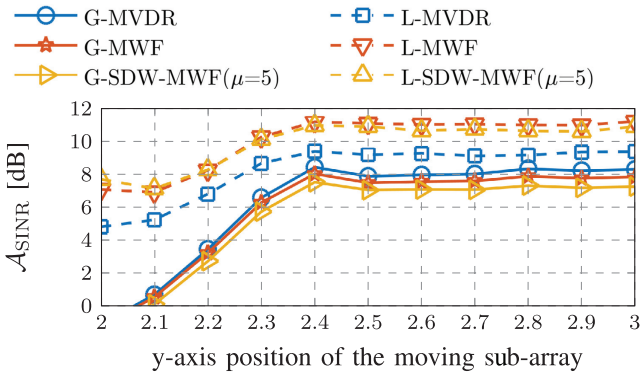
where $\bar{\mathbf{u}}$ is a vector with only one nonzero element at index r^p with value one.

The fullband speech distortion ratios are shown in Fig. 2d. The results contradict theoretical expectations.

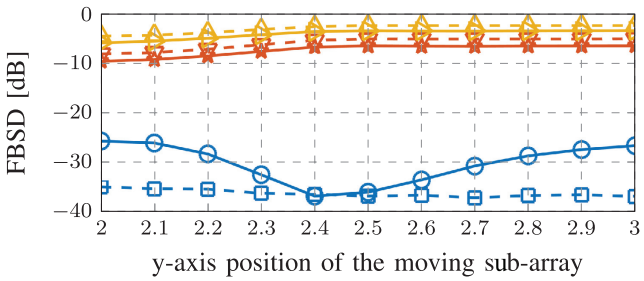
The distortionless methods did not reach much lower FBSD ratios than the Wiener-based; indeed, the global LCMV shows worse distortion than the global MWF. There are two reasons for these contradictions. Firstly, the diagonal loading impose a lower bound of -30 dB. Secondly, there is an amount of mutual coherency among different speech signals which distorts the desired signal. However, the nature of distortion in Wiener-based methods is different, so that its impact on the quality and intelligibility of speech is more, as studies confirm.

5) Perception and Intelligibility: Quantitative comparison of enhancement techniques is useful to attach a best method to a practical problem; however, it cannot assure the improvement in perception of speech or its intelligibility. Perceptual Evaluation of Speech Quality (PESQ) is a well-established measure using segmental SNRs [51] which is mapped into the range of [0,5] with the higher value predicting better perception. In this experiment, the amount of PESQ improvement in calculated from differences between PESQ measures at the input and output of the enhancement techniques. As shown in Fig. 2e, there is no unique algorithm being superior to all others for all noise levels; however, it can be deduced that the global MWF method is the best among compared methods when noise level is higher while global MVDR and LCMV methods are better approaches for lower noise level, make them perceptually better methods when the objective is to remove geometrically constrained interferences in low noise environments.

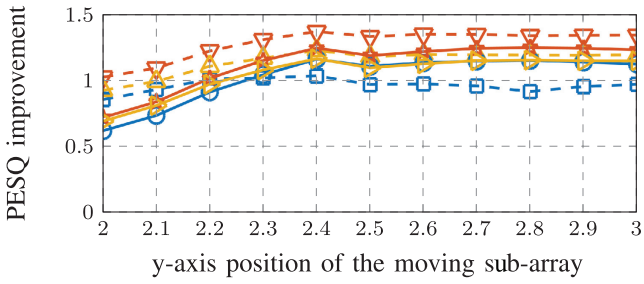
PESQ measure may still be seen as a quantitative measure rather than a qualitative one w.r.t. the intelligibility of speech, as there is no direct mapping available. Speech intelligibility can be measured with STOI in a more tangible manner. The STOI measure can vary in the range [0,1], where the higher represents the better intelligibility. The Auditory Modeling Toolbox is used here to calculate STOI measures [52]. Different methods are compared w.r.t. STOI improvement defined as the difference between STOI at the input of enhancement apparatus to the STOI at its output. As Fig. 2f shows, the STOI improvement does neither map linearly to the PESQ improvement nor to any quantitative measure. Possibly, the most important observation here is that distortionless techniques, global MVDR and LCMV methods, are superior to Wiener-based techniques (MWF) almost for every noise level. This confirms the argumentation in the previous subsection regarding the nature of distortion in different methods.



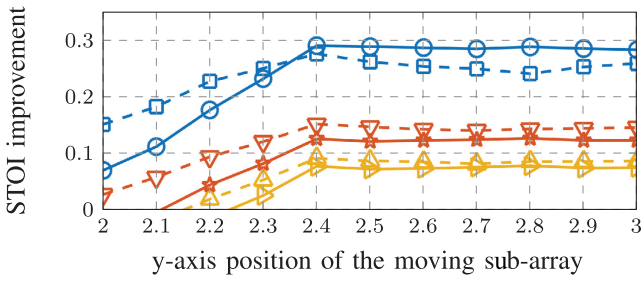
(a) Array gains w.r.t. to interference-plus-noise reduction



(b) Fullband speech distortion ratios



(c) Perceptual measures



(d) Intelligibility measures

Fig. 3. Performance of enhancement methods without the error covariance.

B. Pseudo-Coherence Enhancement Without Noise Estimation

It is time to look into the problem from a different point of view. In this experiment, the geometrical setup in [26] is reproduced with a desired speaker, two interfering speakers, and a spatially-constrained white Gaussian noise source. Three 3-element sub-arrays are available at the scene, among which one is positioned at different places from the neighborhood of the desired speaker towards the heavily noisy and interfered zone. Speech signals from TSP database are down-sampled to 8000 [Hz]. Same sentence spoken by a male, a female, and a child, is used for speakers iterated around the table. 48 Monte-Carlo iterations are used to obtain statistically valid predictions. No prior knowledge on the error covariance matrix, $\mathbf{\Phi}_{\mathbf{e}_r^p}$, is assumed, and the received signal covariance matrix is used instead, as given in (6) for $\mathbf{\Phi}_{\mathbf{y}_r^p}$.

Fig. 3 compares the results from local and global approaches for this set up. The x-axis shows position of the moving sub-array (while its role is changed from being the best input SINR sub-array to the worst one). The y-axes in this figure are the same as defined in Section IV-A. As observed in Fig. 3a, SINR array gains are lower compared to the experiment conducted in Section IV-A, and generally local methods show superiority to global ones. Speech distortions are complying with our expectations, as shown in Fig. 3b. For PESQ and STOI, as shown in Fig. 3c and Fig. 3d, global methods become closer to local ones or even get better when the moving sub-array gets further away from the desired source. The turning point for local and global MVDR w.r.t STOI is the point where all sub-arrays have approximately equal input SINRs.

V. Conclusion

In this work, a framework for speech enhancement with ad hoc microphone arrays is introduced based on the concept of pseudo-coherency. Various beamforming techniques are derived w.r.t. inter and intra sub-array pseudo-coherence vectors. This work extends the state-of-the-art time-domain techniques by establishing broadband beamformers. In addition, it uses the concept of speech coherency to formulate the enhancement problem for multiple speakers and deriving various beamforming techniques. Furthermore, both quantitative and qualitative measures are used in experimental studies to compare the performance of implemented methods.

According to the experimental results, the followings are concluded. Firstly, pseudo-coherence-based enhancement techniques yield performance gain w.r.t. qualitative and quantitative measures. As shown for real-life recordings, an average SINR array gain of above 40 dB is achieved using the global SDW-MWF beamformer for 9-node ad hoc arrays, while the same measure for global MVDR and LCMV beamformers is about 25 dB. It is also shown how the PESQ improvement of up to 2 levels is achieved for different noise levels. Secondly, the MVDR shows better performance w.r.t. the speech intelligibility measure (STOI); however, if another performance measure is the ultimate goal, e.g., the SIR gain, other methods may be superior. In essence, selection of the beamforming approach and decision on using whole or parts of it is a matter of the performance measure we grant as the ultimate goal. Finally, the norm of the pseudo-coherence-vector is a robust measure which can be used, alongside the input SINR, to partition the ad hoc microphone array into local and global beamformers, and to select global and local reference microphones.

References

[1] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen, "Fast joint DOA and pitch estimation using a broadband MVDR beamformer," in *Proc. Eur. Signal Process. Conf.*, Sep. 2013, pp. 1–5.

[2] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[3] J. Capon, "Probability distributions for estimators of the frequency-wavenumber spectrum," *Proc. IEEE*, vol. 58, no. 10, pp. 1785–1786, Oct. 1970

[4] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[5] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.

[6] C. Pan, J. Chen, and J. Benesty, "On the noise reduction performance of the MVDR beamformer in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 815–819.

[7] S. Karimian-Azari, J. Benesty, J. R. Jensen, and M. G. Christensen, "A broadband beamformer using controllable constraints and minimum variance," in *Proc. Eur. Signal Process. Conf.*, Sep. 2014, pp. 666–670.

[8] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Variable speech distortion weighted multichannel wiener filter based on soft output voice activity detection for noise reduction in hearing aids," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2008.

[9] S. Markovich-Golan, S. Gannot, and I. Cohen, "A weighted multichannel wiener filter for multiple sources scenarios," in *Proc. 27th Convention Elect. Electron. Eng. Israel*, 2012, pp. 1–5.

[10] O. Schwartz and S. Gannot, and E. A. P. Habets, "Nested generalized sidelobe canceller for joint dereverberation and noise reduction," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 106–110.

[11] S. Markovich-Golan, S. Gannot, and I. Cohen, "Performance of the SDW-MWF with randomly located microphones in a reverberant enclosure," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 7, pp. 1513–1523, Jul. 2013.

[12] N. D. Gaubitch, M. R. P. Thomas, and P. A. Naylor, "Subband method for multichannel least squares equalization of room transfer functions," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2007, pp. 14–17.

[13] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[14] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Cambridge, MA, USA: MIT Press, 1964.

[15] A. N. Kolmogorov, W. L. Doyle, and I. Selin, *Interpolation and Extrapolation of Stationary Random Sequences* (Russian 1941). Santa Monica, CA, USA: RAND Corporation, 1962.

[16] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2011, pp. 145–148.

[17] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.

[18] M. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[19] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[20] J. Chen, J. Benesty, and Y. Huang, "A minimum distortion noise reduction algorithm with multiple microphones," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 3, pp. 481–493, Mar. 2008.

[21] Y. Huang, J. Benesty, and J. Chen, "Analysis and comparison of multichannel noise reduction methods in a common framework," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 5, pp. 957–968, Jul. 2008.
[22] B. Cornelis, M. Moonen, and J. Wouters, "Comparison of frequency domain noise reduction strategies based on multichannel wiener filtering and spatial prediction," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2009, pp. 129–132.
[23] J. Perina, *Theory of Coherence*. London, U.K.: SNTL, 1975.
[24] E. A. P. Habets and J. Benesty, "Coherent and incoherent interference reduction using a subband tradeoff beamformer," in *Proc. Eur. Signal Process. Conf.*, Aug. 2011, pp. 481–485.
[25] J. R. Jensen, M. G. Christensen, and J. Benesty, "Multichannel signal enhancement using non-causal, time-domain filters," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7274–7278.
[26] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, and J. Benesty, "Pseudo-coherence-based MVDR beamformer for speech enhancement with ad hoc microphone arrays," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Apr. 2015, pp. 2659–2663.
[27] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. IEEE Int. Symp. Commun. Veh. Technol.*, 2011, pp. 1–6.
[28] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 661–676, May 2011.
[29] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks-part I: Sequential node updating," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5277–5291, Oct. 2010.
[30] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks-part II: Simultaneous and asynchronous node updating," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5292–5306, Oct. 2010.
[31] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann, "Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2013, pp. 1–4.
[32] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed GSC beamforming using the relative transfer function," in *Proc. Eur. Signal Process. Conf.*, 2012, pp. 1274–1278.
[33] R. Heusdens, G. Zhang, R. C. Hendriks, Y. Zeng, and W. B. Kleijn, "Distributed MVDR beamforming for (wireless) microphone networks using message passing," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2012, pp. 1–4.
[34] Y. Zeng, R. C. Hendriks, and R. Heusdens, "Clique-based distributed beamforming for speech enhancement in wireless sensor networks," in *Proc. Eur. Signal Process. Conf.*, 2013, pp. 1–5.
[35] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 2, pp. 343–356, Feb. 2013.
[36] J. Szurley, A. Bertrand, P. Ruckebusch, I. Moerman, and M. Moonen, "Greedy distributed node selection for node-specific signal estimation in wireless sensor networks," *Signal Process.*, vol. 94, pp. 57–73, 2014.
[37] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Process.*, vol. 107, pp. 4–20, 2014.
[38] O. Thiergart, M. Taseska, and E. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 2182–2196, Dec. 2014.
[39] A. Hassani, A. Bertrand, and M. Moonen, "Cooperative integrated noise reduction and node-specific direction-of-arrival estimation in a fully connected wireless acoustic sensor network," *Signal Process.*, vol. 107, pp. 68–81, 2015.
[40] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*. New York, NY, USA: Springer, 2011.
[41] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, and J. Benesty, "A partitioned approach to signal separation with ad hoc microphone arrays," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016.
[42] R. T. Lacoss, "Data adaptive spectral analysis methods," *Geophysics*, vol. 36, pp. 661–675, Aug. 1971.
[43] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C—The Art of Scientific Computing*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, Sep. 2007.
[44] K. S. Miller, "On the inverse of the sum of matrices," *Math. Mag.*, vol. 54, no. 2, pp. 67–72, 1981.
[45] Y. Zeng and R. C. Hendriks, "Distributed estimation of the inverse of the correlation matrix for privacy preserving beamforming," *Signal Process.*, vol. 107, pp. 109–122, Feb. 2015.
[46] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2001, vol. 2, pp. 749–752.
[47] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[48] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single- and multichannel audio recordings database (SMARD)," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sep. 2014, pp. 40–44.
[49] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Eindhoven, The Netherlands, Tech. Rep., 2006.
[50] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. New York, NY, USA: Springer, 2008.
[51] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[52] P. Søndergaard and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening*, J. Blauert, Ed. New York, NY, USA: Springer, 2013, pp. 33–56.

Vincent Mohammad Tavakoli (S'99) was born in Behshahr, Iran, in December 1979. He received the B.Sc. degree in telecommunications from Iran University of Science and Technology, Tehran, Iran, in 2004, and the M.Sc. degree in biomedical engineering from Tehran Polytechnic in Iran, in 2008. He is currently pursuing the Ph.D. degree at the Department of Architecture, Design, & Media Technology and the Audio Analysis Laboratory, Aalborg University, Aalborg, Denmark. In 2010, he joined the graduate program in signal processing at Blekinge Institute of Technology, Karlskrona, Sweden, and in 2012 became a Co-Lecturer for courses in signal processing and wireless communications for both undergraduate and graduate courses.

He worked for Iran Cable Manufacturing Co. and Karizan Telecom Co. in Iran. His research interests include array signal processing.

Jesper Rindom Jensen (S'09–M'12) was born in Ringkøbing, Denmark, in August 1984. He received the M.Sc. degree (*cum laude*) for completing the elite candidate education and the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 2009 and 2012, respectively. Currently, he is a Postdoctoral Researcher with the Department of Architecture, Design & Media Technology, Aalborg University, where he is also a Member of the Audio Analysis Laboratory. He has been a Visiting Researcher at the University of Quebec, INRS-EMT, Montreal, QC, Canada, and at the Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany.

His research interests include signal processing theory and methods for, e.g., microphone array and joint audio-visual signal processing. Examples of more specific research interests within this scope are enhancement, separation, localization, tracking, parametric analysis, and modeling. He has published more than 50 papers on these topics in top-tier, peer-reviewed conference proceedings and journals. Moreover, he is the coauthor of two books, namely, *Speech Enhancement: A Signal Subspace Perspective* and *Signal Enhancement With Variable Span Linear Filters*.

He is an Affiliate Member of the IEEE Signal Processing Theory and Methods Technical Committee. He was the recipient of a highly competitive Postdoc grant from the Danish Independent Research Council, as well as several travel grants from private foundations.

Mads Græsbøll Christensen (S'00–M'05–SM'11) received the M.Sc. and Ph.D. degrees from Aalborg University (AAU), Aalborg, Denmark, in 2002 and 2005, respectively. He is currently employed at the Department of Architecture, Design & Media Technology, AAU, as Professor of Audio Processing and is the Head and Founder of the Audio Analysis Lab.

He was formerly with the Department of Electronic Systems, AAU and has been held visiting positions at Philips Research Labs, ENST, UCSB, and Columbia University, New York, NY, USA. He has published 3 books and more than 150 papers in peer-reviewed conference proceedings and journals, and he has given tutorials at EUSIPCO and INTERSPEECH. His research interests include signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, enhancement, separation, and coding.

Prof. Christensen is a beneficiary of major grants from the Danish Independent Research Council, the Villum Foundation, and Innovation Fund Denmark. He is an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, a former Associate Editor of the IEEE SIGNAL PROCESSING LETTERS, and a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee. He was the recipient of several awards, including an ICASSP Student Paper Contest Award, the Spar Nord Foundations Research Prize, a Danish Independent Research Council Young Researchers Award, the Statoil Prize, and the EURASIP Early Career Award. He is also a coauthor of the paper *Sparse Linear Prediction and Its Application to Speech Processing* that received an IEEE Signal Processing Society Young Author Best Paper Award.

Jacob Benesty was born in 1963. He received the Master's degree in microwaves from Pierre & Marie Curie University, Paris, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, Orsay, France, in 1991. During the Ph.D. (from November 1989 to April 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 with Telecom Paris University, Paris, France, 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, Montreal, Quebec, Canada, as a Professor. He is also a Visiting Professor at the Technion, Haifa, Israel, an Adjunct Professor with Aalborg University, Aalborg, Denmark, and a Guest Professor with Northwestern Polytechnical University, Xi'an, China.

He is the inventor of many important technologies. In particular, he was the Lead Researcher with 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 multiparty hands-free full-duplex stereo conferencing system over IP networks. His research interests include signal processing, acoustic signal processing, and multimedia communications.

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 was the recipient (with Morgan and Sondhi) of the IEEE Signal Processing Society 2001 Best Paper Award, the IEEE Signal Processing Society 2008 Best Paper Award (with Chen, Huang, and Doclo), the Gheorghe Cartianu Award from the Romanian Academy in 2010, and the Best Paper Award from the IEEE WASPAA for a paper that he coauthored with Chen in 2011. He is also a coauthor of a paper for which Huang received the IEEE Signal Processing Society 2002 Young Author Best Paper Award.