PSEUDO-COHERENCE-BASED MVDR BEAMFORMER FOR SPEECH ENHANCEMENT WITH AD HOC MICROPHONE ARRAYS

Vincent Mohammad Tavakoli[†]*, Jesper Rindom Jensen*[†]*, Mads Græsbøll Christensen*[†]*, and Jacob Benesty*[‡]

[†]Audio Analysis Lab, AD:MT
Aalborg University, Denmark
{vmt,jrj,mgc}@create.aau.dk

[‡]INRS-EMT
University of Quebec, Montreal, Canada
benesty@emt.inrs.ca

ABSTRACT

Speech enhancement with distributed arrays has been met with various methods. On the one hand, data independent methods require information about the position of sensors, so they are not suitable for dynamic geometries. On the other hand, Wiener-based methods cannot assure a distortionless output. This paper proposes minimum variance distortionless response filtering based on multichannel pseudo-coherence for speech enhancement with ad hoc microphone arrays. This method requires neither position information nor control of the trade-off used in the distortion weighted methods. Furthermore, certain performance criteria are derived in terms of the pseudo-coherence vector, and the method is compared with the multichannel Wiener filter. Evaluation shows the suitability of the proposed method in terms of noise reduction with minimum distortion in ad hoc scenarios.

Index Terms— Noise reduction, speech enhancement, distributed microphone array, STFT domain, MVDR filter

1. INTRODUCTION

Enhancing the quality of speech has been a signal processing research interest for decades. This interest focuses mostly on extracting the speech signal from a mixture of desired and unwanted signals. The unwanted part of this mixture may include competing speech, reverberant sound, and noise. Although noise reduction for the purpose of speech enhancement may be seen as an easy task compared to source separation and dereverberation, it is still a challenge. This is mainly because reducing the noise does not guarantee an improved intelligibility of desired speech since noise reduction techniques may introduce distortion in speech signals [1, 2, 3].

In the recent years, distributed microphone arrays have been widely used to tackle the speech enhancement problem [4, 5, 6], mostly in fully connected scenarios [7, 8, 9], and sometimes with multiple sources [10]. Node-specific and common constraints have been also utilized [11, 12]. One characteristic of distributed microphone arrays is randomness and dynamicity in the position of microphones. Recently, the performance of the Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) was analyzed for randomly located microphones using a uniform distribution [13]. The reliability function derived for such a scenario showed the limitations of applying SDW-MWF to randomly located microphones. Proposal of the clustered blind beamforming was one of the efforts to take randomness into account. In their work, Himawan et al. [14] used a clustering approach to group the randomly located microphones into localized clusters, which were then ranked according to their relative distance from the speaker. They compared cluster-based algorithms to the full set of microphones within different ad hoc array geometries. Unfortunately, cluster-based methods lack an explicit metric to show that their solutions are optimal in some sense. Recently, ad hoc microphone arrays have been studied in a close-talking scenario. A beamforming method based on a soft time-frequency activity mask is proposed in this paper. It has been shown that the proposed beamforming method is suboptimal but close to a centralized beamforming [15].

This paper proposes blind speech enhancement methods that are formed to act explicitly optimally with randomly positioned distributed microphone arrays, namely ad hoc microphone arrays. The aim is to answer the question when and how different beamforming techniques should be used in ad hoc scenarios. The methods are inspired by the recently proposed, single-/multi-channel, enhancement methods in [16, 17]. These methods exploit an orthogonal decomposition of the desired signal, which are then modeled using the pseudo-coherence vector. Herein, we use the pseudo-coherence vector to model the multichannel data obtained using the distributed microphone arrays. This enables us to obtain distortionless beamforming that is generally not the case for the Wiener-type beamformers for distributed microphone arrays.

The remainder of this paper is organized as follows. First, the problem of interest is defined in Section 2 by formulating the signal model and the pseudo-coherence vector along with an ad hoc microphone array. Before proposing beamforming methods suitable for ad hoc microphone arrays in Section 4, we present selection criteria based on the pseudo-coherence vector in Section 3. We derive and characterize the proposed generalized beamformer on the basis of the pseudo-coherence vector. In Section 5, we present experimental results based on Monte-Carlo simulations. Finally, Section 6 presents a discussion of the results in Section 5.

2. PROBLEM FORMULATION

In this paper, we consider the case of a distributed (ad hoc) microphone array in a reverberant acoustic environment. We assume that N identical microphone arrays are distributed randomly. We also assume that each array consists of M omnidirectional microphones. In this environment, the ad hoc sensor arrays capture a desired convoluted speech source signal in some noise field. The signal picked up by the mth microphone of the nth array can be expressed as [18, 19]

$$y_{n,m}(t') = x_{n,m}(t') + v_{n,m}(t'), \quad (1)$$

This work was supported in part by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF 1337-00084.

where $x_{n,m}(t') = g_{n,m}(t') * s(t')$, $s(t')$ is the speech signal at time t', $g_{n,m}(t')$ is the acoustic impulse response from the source location to the mth microphone of the nth array, and $v_{n,m}(t')$ is the additive noise to the microphone m of array n. We assume that the acoustic impulse responses are time invariant. Also $x_{n,m}(t')$ and $v_{n,m}(t')$ are assumed to be uncorrelated, zero mean, stationary, real, and broadband. Moreover, our assumption is that the convolved speech signals $x_{n,m}(t')$ are coherent across all microphones in the ad hoc array while the noise signal $v_{n,m}(t')$ is only partially coherent across the microphone arrays.

Without loss of generality, we can assign the clean (but convoluted) signal at the first microphone of each array, namely $x_{1,1}(t'), x_{2,1}(t'), ..., x_{N,1}(t')$, as the reference signal for that array. Many fundamental questions arise in the context of distributed (ad hoc) microphone arrays. Which one of these references should be estimated? Which one is the best and in what sense? How can distributed arrays help in the estimation? In the rest of this section, we formulate the problem in a way that allows us to answer these fundamental questions.

Assuming a sufficiently long analysis window, taking the STFT from (1) yields

$$Y_{n,m}(k,t) = X_{n,m}(k,t) + V_{n,m}(k,t), \tag{2}$$

where $X_{n,m} = G_{n,m}(k,t)S(k,t)$, $k \in \{0, ..K-1\}$ specifies the frequency bin, and t is the time frame index.

By stacking STFT-domain signals of M microphones at the nth array, we can write

$$\begin{aligned}\mathbf{y}_n(k,t) &= \mathbf{g}_n(k)S(k,t) + \mathbf{v}_n(k,t) = \mathbf{x}_n(k) + \mathbf{v}_n(k,t)\\ &= \mathbf{d}_n(k)X_{n,1}(k,t) + \mathbf{v}_n(k,t),\end{aligned} \tag{3}$$

where $\mathbf{y}_n(k,t)$, $\mathbf{v}_n(k,t)$, $\mathbf{x}_n(k,t)$ are stacked versions of the respective STFT-domain signals, $\mathbf{g}_n(k) = [G_{n,1}(k), ..., G_{n,M}(k)]^T$, $\mathbf{d}_n(k) = \frac{\mathbf{g}_n(k)}{G_{n,1}(k)}$, and the transcript ^{T} denotes the transpose operator. Here, it is assumed that $G_{n,1}(k) \neq 0$. Expression (3) depends explicitly on the reference signal, $X_{n,1}(k,t)$; as a result, (3) is the STFT-domain signal model for noise reduction. The vector $\mathbf{d}_n(k)$ is obviously the STFT-domain steering vector for noise reduction corresponding to the nth array.

It can be verified [20] that a more interesting way to write (3) is

$$\mathbf{y}_n(k,t) = \boldsymbol{\rho}_{\mathbf{x}_n X_{n,1}}(k,t)X_{n,1}(k,t) + \mathbf{v}_n(k,t), \tag{4}$$

$$\boldsymbol{\rho}_{\mathbf{x}_n X_{n,1}}(k,t) = \frac{E\left[\mathbf{x}_n(k,t)X_{n,1}^*(k,t)\right]}{E\left[|X_{n,1}(k,t)|^2\right]} \approx \mathbf{d}_n(k) \tag{5}$$

where $\boldsymbol{\rho}_{\mathbf{x}_n X_{n,1}}(k,t)$ is the pseudo-coherence vector of length M between $\mathbf{x}_n(k,t)$ and $X_{n,1}(k,t)$, with $E[\cdot]$ and * denoting mathematical expectation and complex conjugate.

The equality $\boldsymbol{\rho}_{\mathbf{x}_n X_{n,1}}(k,t) = \mathbf{d}_n(k)$ holds only when the analysis window of the STFT is infinitely long. However, (4) is much more useful than (3) since the pseudo-coherence vector captures much better the acoustic environment than the STFT-domain steering vector, especially in the context of ad hoc microphone arrays. Therefore, in the following, the given model in (4) will only be used.

As indicated earlier in this section, one fundamental question regarding speech enhancement with ad hoc microphone arrays is to select one of the reference signals from the set of N reference signals, i.e. $\{X_{1,1}, X_{2,1}, ..., X_{N,1}\}$. Assuming that $X_{n_r,1}(k,t)$ is the selected reference signal, in theory it is always possible to write (4) as a function of this selected reference signal as

$$\mathbf{y}_n(k,t) = \boldsymbol{\rho}_{\mathbf{x}_n X_{n_r,1}}(k,t)X_{n_r,1}(k,t) + \mathbf{v}_n(k,t), \tag{6}$$

$$\boldsymbol{\rho}_{\mathbf{x}_n X_{n_r,1}}(k,t) = \frac{E\left[\mathbf{x}_n(k,t)X_{n_r,1}^*(k,t)\right]}{E\left[|X_{n_r,1}(k,t)|^2\right]}, \tag{7}$$

where $\boldsymbol{\rho}_{\mathbf{x}_n X_{n_r,1}}(k,t)$ is the pseudo-coherence vector (of length M) between $\mathbf{x}_n(k,t)$ and the reference signal $X_{n_r,1}(k,t)$.

From (3) and (4), we deduce that the correlation matrix of $\mathbf{y}_n(k,t)$ is

$$\boldsymbol{\Phi}_{\mathbf{y}_n}(k,t) = \boldsymbol{\Phi}_{\mathbf{x}_n}(k,t) + \boldsymbol{\Phi}_{\mathbf{v}_n}(k,t), \tag{8}$$

where $\boldsymbol{\Phi}_{\mathbf{x}_n}(k,t) = \phi_{X_{n_r,1}}(k,t)\boldsymbol{\rho}_{\mathbf{x}_n X_{n_r,1}}(k,t)\boldsymbol{\rho}^H_{\mathbf{x}_n X_{n_r,1}}(k,t)$ is the rank 1 correlation matrix of $\mathbf{x}_n(k,t)$ and $\boldsymbol{\Phi}_{\mathbf{v}_n}(k,t)$ is the noise correlation matrix (which rank is assumed to be M).

3. SELECTION CRITERIA

In the following subsections, we define relevant selection criteria for designing distributed beamformers.

3.1. The Norm of The Pseudo-Coherence Vector

The pseudo-coherence vector, $\boldsymbol{\rho}_{\mathbf{x}_n X_{n,1}}(k,t)$, tells us how much $X_{n,1}(k,t)$ is coherent with the other convolved speech signals $X_{n,i}(k,t)$, $i = 2, ..., M$ of the nth array. Let us define the quantity:

$$\aleph_n(k,t) = ||\boldsymbol{\rho}_{\mathbf{x}_n X_{n,1}}(k,t)||_2^2. \tag{9}$$

We always have $\aleph_n(k,t) \geq 1$. The worst scenario is when $\aleph_n(k,t)$ is close to 1, which means that array n captures almost no speech. It is clear that for two arrays i and j, a value of $\aleph_i(k,t)$ greater than a value of $\aleph_j(k,t)$ means that the speech source is closer to the array i than the array j. As a result, we should try to recover $X_{i,1}(k,t)$ rather than $X_{n,1}(k,t)$. We deduce that the desired signal that we should estimate or recover is $X_{n_r,1}(k,t)$, where n_r is chosen to maximize $\aleph_n(k,t)$. It is of great importance to quantify how much the arrays (other than the one containing the reference signal, i.e., n_r) can contribute to noise reduction. For that, we can define the quantity

$$\aleph_{n,n_r}(k,t) = ||\boldsymbol{\rho}_{\mathbf{x}_n X_{n_r,1}}(k,t)||_2^2. \tag{10}$$

We always have $0 \leq \aleph_{n,n_r}(k,t) \leq \aleph_{n_r}(k,t)$. The worst scenario is when $\aleph_{n,n_r}(k,t)$ is close to zero, which means that array n will have little or no positive contribution in the estimation of $X_{n_r,1}(k,t)$. The measure in (10) tells us how much array n can "hear" the reference signal, $X_{n_r,1}(k,t)$.

3.2. The Input SNR

One fundamental measure in noise reduction is the averaged (narrowband) input signal-to-noise ratio (SNR) at the nth array, which is obtained from (8):

$$\mathrm{iSNR}_n(k,t) = \frac{\mathrm{tr}\left[\boldsymbol{\Phi}_{\mathbf{x}_n}(k,t)\right]}{\mathrm{tr}\left[\boldsymbol{\Phi}_{\mathbf{v}_n}(k,t)\right]} = \frac{\aleph_n(k,t)\phi_{X_{n,1}}(k,t)}{\mathrm{tr}\left[\boldsymbol{\Phi}_{\mathbf{v}_n}(k,t)\right]}, \tag{11}$$

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

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

$$n_r' = \arg\max_n \mathrm{iSNR}_n(k,t). \tag{12}$$

In this case, we estimate $X_{n'_r,1}(k,t)$. In theory, (11) can also be rewritten as

$$\text{iSNR}_{n,n'_r}(k,t) = \aleph_{n,n'_r}(k,t)\frac{\phi_{X_{n'_r,1}}(k,t)}{\text{tr}\left[\mathbf{\Phi}_{\mathbf{v}_n}(k,t)\right]}, \quad (13)$$

where $\phi_{X_{n'_r,1}}(k,t)$ is the variance of $X_{n'_r,1}(k,t)$.

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

$$\text{iSNR}(k,t) = \frac{\sum_{n=1}^{N}\aleph_n(k,t)\phi_{X_{n,1}}(k,t)}{\sum_{n=1}^{N}\text{tr}\left[\mathbf{\Phi}_{\mathbf{v}_n}(k,t)\right]} \leq \text{iSNR}_{n'_r}(k,t). \quad (14)$$

4. PSEUDO-COHERENCE-BASED BEAMFORMING

In this section, we consider the conventional MVDR beamformer [21], [22] for noise reduction. As can be expected, there are different ways to perform beamforming, depending on the criteria discussed in the previous section.

4.1. Beamforming with Best Input SNR Array

In this subsection, we select the array with the best input SNR, i.e., n'_r obtained from (12), and all the other arrays are just ignored. In this case, the beamformer output is $Z(k,t) = \mathbf{h}^H_{n'_r}(k,t)\mathbf{y}_{n'_r}(k,t)$, where $\mathbf{h}_{n'_r}(k,t)$ is a complex filter of length M containing all the complex gains applied to the microphone outputs of the array n'_r at frequency bin k and time frame t.

By minimizing the variance of the beamformer output $Z(k,t)$ with the distortionless constraint, $\mathbf{h}_{n'_r}(k,t)\boldsymbol{\rho}_{\mathbf{x}_{n'_r}X_{n'_r,1}}(k,t) = 1$, we easily find the MVDR filter:

$$\mathbf{h}_{n'_r}(k,t) = \frac{\mathbf{\Phi}^{-1}_{\mathbf{y}_{n'_r}}(k,t)\boldsymbol{\rho}_{\mathbf{x}_{n'_r}X_{n'_r,1}}(k,t)}{\boldsymbol{\rho}^H_{\mathbf{x}_{n'_r}X_{n'_r,1}}(k,t)\mathbf{\Phi}^{-1}_{\mathbf{y}_{n'_r}}(k,t)\boldsymbol{\rho}_{\mathbf{x}_{n'_r}X_{n'_r,1}}(k,t)}. \quad (15)$$

Obviously, we can also write the MVDR filter as

$$\mathbf{h}_{n'_r}(k,t) = \frac{\mathbf{\Phi}^{-1}_{\mathbf{v}_{n'_r}}(k,t)\boldsymbol{\rho}_{\mathbf{x}_{n'_r}X_{n'_r,1}}(k,t)}{\boldsymbol{\rho}^H_{\mathbf{x}_{n'_r}X_{n'_r,1}}(k,t)\mathbf{\Phi}^{-1}_{\mathbf{v}_{n'_r}}(k,t)\boldsymbol{\rho}_{\mathbf{x}_{n'_r}X_{n'_r,1}}(k,t)}. \quad (16)$$

The (narrowband) output SNR of this beamformer is defined as

$$\text{oSNR}\left[\mathbf{h}_{n'_r}(k,t)\right] = \frac{\phi_{X_{n'_r,1}}(k,t)}{\mathbf{h}^H_{n'_r}(k,t)\mathbf{\Phi}_{\mathbf{v}_{n'_r}}(k,t)\mathbf{h}_{n'_r}(k,t)}. \quad (17)$$

We deduce that the (narrowband) array gain is

$$\mathcal{A}\left[\mathbf{h}_{n'_r}(k,t)\right] = \text{oSNR}\left[\mathbf{h}_{n'_r}(k,t)\right]\text{iSNR}^{-1}_{n'_r}(k,t) \geq 1. \quad (18)$$

4.2. Beamforming with All Distributed Arrays

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 $X_{n_r,1}(k,t)$ is found to be the best reference signal.

The beamformer output is now $Z(k,t) = \underline{\mathbf{h}}^H(k,t)\underline{\mathbf{y}}(k,t)$, where $\underline{\mathbf{h}}^H(k,t)$ is a complex filter of length MN containing all the complex gains applied to the microphone outputs of all arrays at frequency bin k and

$$\begin{aligned}\underline{\mathbf{y}}(k,t) &= \left[\underline{\mathbf{y}}^T_1(k,t),...,\underline{\mathbf{y}}^T_1(k,t)\right]^T = \underline{\mathbf{x}}(k,t) + \underline{\mathbf{v}}(k,t)\\ &= \boldsymbol{\rho}_{\mathbf{x},X_{n_r,1}}(k,t)X_{n_r,1}(k,t) + \underline{\mathbf{v}}(k,t), \end{aligned} \quad (19)$$

with $\boldsymbol{\rho}_{\underline{\mathbf{x}}_n X_{n_r,1}}(k,t) = E\left[\underline{\mathbf{x}}_n(k,t)X^*_{n_r,1}(k,t)\right]/E\left[|X_{n_r,1}(k,t)|^2\right]$ being the pseudo-coherence vector (of length MN) between $\underline{\mathbf{x}}(k,t)$ and $X_{n_r,1}(k,t)$.

The minimization of the variance of $Z(k,t)$ with distortionless constraint, $\underline{\mathbf{h}}^H(k,t)\boldsymbol{\rho}_{\underline{\mathbf{x}}_n X_{n_r,1}}(k,t) = 1$, leads to the MVDR filter:

$$\underline{\mathbf{h}}(k,t) = \frac{\mathbf{\Phi}^{-1}_{\underline{\mathbf{y}}(k,t)}\boldsymbol{\rho}_{\underline{\mathbf{x}}X_{n_r,1}}(k,t)}{\boldsymbol{\rho}^H_{\underline{\mathbf{x}}X_{n_r,1}}(k,t)\mathbf{\Phi}^{-1}_{\underline{\mathbf{y}}}(k,t)\boldsymbol{\rho}_{\underline{\mathbf{x}}X_{n_r,1}}(k,t)}, \quad (20)$$

where $\mathbf{\Phi}_{\underline{\mathbf{y}}(k,t)}$ is the correlation matrix of $\underline{\mathbf{y}}(k,t)$.

Then, the (narrowband) output SNR and (narrowband) array gain are, respectively,

$$\text{oSNR}\left[\underline{\mathbf{h}}(k,t)\right] = \frac{\phi_{X_{n_r,1}}(k,t)}{\underline{\mathbf{h}}^H(k,t)\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,t)\underline{\mathbf{h}}(k,t)} \quad (21)$$

and $\mathcal{A}[\underline{\mathbf{h}}(k,t)] = \text{oSNR}[\underline{\mathbf{h}}(k,t)]\text{iSNR}^{-1}_{n_r}(k,t) \geq 1$.

4.3. Beamforming with Best Output SNR Array

In this subsection, we apply N independent beamformers to the N distributed arrays. We then select the beamformer that gives the best (narrowband) output SNR. Therefore, the nth beamformer output is $Z_n(k,t) = \mathbf{h}^H_n(k,t)\mathbf{y}_n(k,t)$, where $\mathbf{h}^H_n(k,t)$ is a complex filter of length M containing all the complex gains applied to the microphone outputs of the array n at frequency bin k.

The MVDR filter is similar to the one derived in Section 4.1, i.e.,

$$\mathbf{h}_n(k,t) = \frac{\mathbf{\Phi}^{-1}_{\mathbf{y}_n}(k,t)\boldsymbol{\rho}_{\mathbf{x}_n X_{n,1}}(k,t)}{\boldsymbol{\rho}^H_{\mathbf{x}_n X_{n,1}}(k,t)\mathbf{\Phi}^{-1}_{\mathbf{y}_n}(k,t)\boldsymbol{\rho}_{\mathbf{x}_n X_{n,1}}(k,t)}. \quad (22)$$

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

$$\text{oSNR}\left[\mathbf{h}_n(k,t)\right] = \frac{\phi_{X_{n,1}}(k,t)}{\mathbf{h}^H_n(k,t)\mathbf{\Phi}_{\mathbf{v}_n}(k,t)\mathbf{h}_n(k,t)}. \quad (23)$$

Maximizing the output SNR with respect to the array index gives us the solution we are looking for, i.e., $\mathbf{h}_{n_r}(k,t)$. We deduce that the (narrowband) array gain for this method is

$$\mathcal{A}[\mathbf{h}_{n_r}(k,t)] = \text{oSNR}[\mathbf{h}_{n_r}(k,t)]\text{iSNR}^{-1}_{n_r}(k,t) \geq 1. \quad (24)$$

5. EXPERIMENTS

In this Section, we setup two experiments to evaluate the proposed beamforming techniques. Firstly, we compare the array gains for different schemes that were proposed in Section 4, namely the MVDR beamformer with best-iSNR array, with all arrays, and with best-oSNR array. Further comparison is drawn by reformulating the generalized MVDR filters in Section 4 into their respective Multichannel Wiener Filters (MWF). For example, (20) can be expressed for MWF as

$$\underline{\mathbf{h}}(k,t) = \frac{\mathbf{\Phi}^{-1}_{\underline{\mathbf{y}}(k,t)}\boldsymbol{\rho}_{\underline{\mathbf{x}}X_{n_r,1}}(k,t)}{\boldsymbol{\rho}^H_{\underline{\mathbf{x}}X_{n_r,1}}(k,t)\mathbf{\Phi}^{-1}_{\underline{\mathbf{y}}}(k,t)\boldsymbol{\rho}_{\underline{\mathbf{x}}X_{n_r,1}}(k,t) + \phi^{-1}_{X_{n_r,1}}(k,t)}.$$

In addition to the fullband array gain, we compare these methods in terms of distortion. We reformulate the multichannel distortion index defined in [18] and rewrite it in terms of the pseudo-coherence vector and complex filter weights as

$$SD = \frac{\sum_t\sum_k \phi_{X_{n_r,1}}(k,t)|(\underline{\mathbf{u}} - \underline{\mathbf{h}}(k,t))^H\boldsymbol{\rho}_{\underline{\mathbf{x}}X_{n_r,1}}(k,t)|^2}{\sum_t\sum_k \phi_{X_{n_r,1}}(k,t)},$$

where $\underline{\mathbf{u}}$ is a vector with only one nonzero element at index n_r with value one.

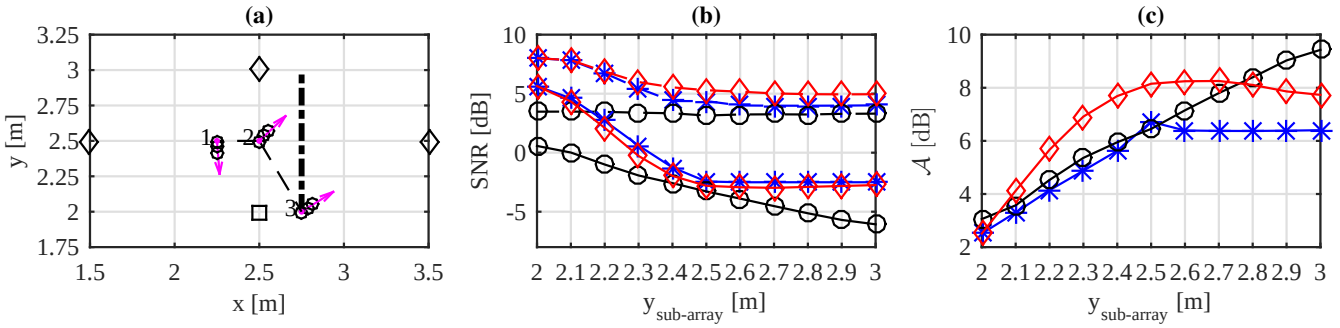

Fig. 1. (a) An instance of experiment setup. (b) Input-SNR (dashed) and Output-SNR (solid). (c) Array gains for the proposed beamformers. In (b) and (c) the MVDR beamformers with best-iSNR, all arrays, and best-oSNR are marked with stars, circles, and diamonds, respectively.

5.1. Simulations

The Dimensions of the simulated room are $5\text{m} \times 5\text{m} \times 3\text{m}$. Received signals are produced with the image method [23] in this enclosure. The ad hoc microphone array contains 3 linear sub-arrays, each consists of 3 microphones with inter-element space equal to 4.3 cm. All microphones are assumed to be omnidirectional (monopole). The desired speech signal for both experiments is played from a clean recording by sampling frequency $f_s = 8$ kHz. For STFT, the length of each time frame is set to 32 ms with 75% overlap among neighboring frames which corresponds to 8 ms hop. Averaging over 32 consecutive frames is used in estimation of pseudo-coherence vector and correlation matrices.

For the first experiment, the room is assumed to be anechoic. The source is located at $(2.5, 2.5, 2.5)$. A constrained white Gaussian noise is located at $(2.5, 2.5, 0.5)$ with the same variance as the desired speech signal. The three sub-arrays are placed randomly in the room with a volumetric uniform distribution. This geometry imposes 50% probability limit on the cases that the received clean signal blasts over the noise. Performance measures are averaged over 999 Monte-Carlo simulations. The results for the MVDR and the MWF filters are presented in Table 1. As can be seen, the proposed MVDR filters result in smaller fullband array gain compared to the MWF filters; however, the speech distortion factors for the MWF filters are very high. Table 1 offers possible benefits of the proposed MVDR filters, specifically the Best-iSNR and the Best-oSNR schemes; however, it is still not clear when each of these methods should be preferred over the others. In the next experiment, we form an experiment to analyze a practical ad hoc situation with controlled randomness.

For the second experiment, a teleconferencing scenario is assumed in the room with the same size as before but suffering from reverberation with T_{60} equal to 250 ms. The desired speaker (square) and three interfering speakers (diamonds) are positioned according to Figure 1-a. All speech signals have variance one. Spatially Gaussian white noise is added to the microphones with variance of -30 dB below the desired speech signal. Two of the ULAs are placed in a fixed position while the third ULA was moved along a line from $(2.75, 2, 1.5)$ to $(2.75, 3, 1.5)$ with 10 cm steps. The orientation of ULAs on the plane is random. Array gain and speech distortion for different locations of the third sub-array are averaged over 50 Monte-Carlo simulations. In Figure 1-b and Figure 1-c, the MVDR beamformer with best-iSNR array, all arrays, and best-oSNR array are compared. As the sub-array 3 moves farther from the desired source and closer to the interferences, the input-SNR of all three methods decreases linearly; however, the decrease stops at the mid-

Table 1. Fullband Array Gain and Speech Distortion Index for Beamforming Schemes in The First Exprement. (Values are in [dB].)

	Best-iSNR		All Arrays		Best-oSNR	
	MVDR	MWF	MVDR	MWF	MVDR	MWF
$\mathcal{A}$	12.5	18.5	10.5	14.5	14	20
SD	-263.5	-6	-235	-7	-267.5	-6

point for the best-iSNR and the best-oSNR methods. The behavior of input-/output-SNR for this setup is expected as the role of the moving sub-array changes from the closest microphone to the desired signal to the closest one to the interference. Figure 1-c combines information in Figure 1-b. From the three proposed schemes, the best-oSNR MVDR filter is superior in a majority of places. However, the simplicity of the best-iSNR method compared to the two others makes it the best candidate for close talking scenarios. At last, beamforming with all arrays utilizes its multiplicity of elements for very noisy situations. It should be noted that in our derivations in previous sections and in these two simulations, we assumed the same number of elements for all sub-arrays, but the general principle of pseudo-coherence beamforming goes for other cases also. The speech distortion for this experiment is in the same scale as the MVDR filters in the first experiment.

It is important to note that in our simulations, we used (15) since it does not require estimation of the noise correlation matrix, $\mathbf{\Phi}_{\mathbf{v}_n}(k,t)$; however, there have been a rise in the output SNR and the array gain using noise statistics in (16), specifically for the beamforming with all arrays which showed an improvement of 12[dB] in this experiment (not shown in the figure).

6. DISCUSSION

Conventional distributed array processing techniques may diverge from optimality for emerging applications in ad hoc scenarios. It has been shown that methods such as activity masked beamforming are suboptimal but can result close to the centralized array processing [15]. However, it is still unclear when each method is superior to the others and should be used. In this paper, we proposed a new beamforming scheme in which the speech coherency over sub-arrays has been utilized to form different distortionless beamformers. We have shown that despite the highly random nature of the problem, it is possible to formulate useful measures and techniques to enhance the selection phase. We have also shown that the proposed methods do not require the huge transmission of inter-array data.

7. REFERENCES

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

[2] Y. Huang, J. Benesty, and J. Chen, "Analysis and comparison of multichannel noise reduction methods in a common framework," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 5, pp. 957–968, July 2008.

[3] 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.

[4] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed GSC beamforming using the relative transfer function," in *Proc. European Signal Processing Conf.*, 2012, pp. 1274–1278.

[5] Y. Zeng, R. C. Hendriks, and R. Heusdens, "Clique-based distributed beamforming for speech enhancement in wireless sensor networks," in *Proc. European Signal Processing Conf.*, 2013, pp. 1–5.

[6] 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. Intl. Workshop Acoust. Echo Noise Control*, 2012, pp. 1–4.

[7] 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, and Language Process.*, vol. 21, no. 2, pp. 343–356, Feb. 2013.

[8] 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.

[9] 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.

[10] S. Markovich-Golan, S. Gannot, and I. Cohen, "A weighted multichannel wiener filter for multiple sources scenarios," in *Electrical & Electronics Engineers in Israel, IEEE 27th Convention of*, 2012, pp. 1–5.

[11] J. Szurley, A. Bertrand, P. Ruckebusch, I. Moerman, and M. Moonen, "Greedy distributed node selection for node-specific signal estimation in wireless sensor networks," *Elsevier Signal Process.*, vol. 94, pp. 57 – 73, 2014.

[12] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Elsevier Signal Process.*, vol. 0, no. -, 2014.

[13] 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, and Language Process.*, vol. 21, no. 7, pp. 1513–1523, July 2013.

[14] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 4, pp. 661–676, May 2011.

[15] M. Taseska, S. Markovich-Golan, S. Gannot, and E. A. P. Habets, "Near-field source extraction using speech presence probabilities for ad-hoc microphone arrays," in *Proc. Intl. Workshop Acoust. Echo Noise Control*, 2014, pp. 170–174.

[16] 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.

[17] J. R. Jensen, J. Benesty, M. G. Christensen, and Sren Holdt Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 7, pp. 1948–1963, Sept. 2012.

[18] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer Topics in Signal Processing. Springer, 2008.

[19] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Digital Signal Processing - Springer-Verlag. Springer, 2001.

[20] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, SpringerBriefs in Electrical and Computer Engineering. Springer, 2011.

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

[22] R. T. Lacoss, "Data adaptive spectral analysis methods," *Geophysics*, vol. 36, pp. 661–675, Aug. 1971.

[23] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, Eindhoven, 2010.