A PARTITIONED APPROACH TO SIGNAL SEPARATION WITH MICROPHONE AD HOC ARRAYS

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

†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

In this paper, a blind algorithm is proposed for speech enhancement in multi-speaker scenarios, in which interference rejection is the main objective. Here, the ad hoc array is broken into microphone duples which are used to partition the array into local sub-arrays. The core algorithm takes advantage of differences in signal structure in each duple. A geometric mean filter is then used to merge the output signals obtained with different duples, and to form a global broadband maximum signal-to-interference ratio (SIR) enhancement apparatus. The resulting filter outputs are enhanced acoustic signals in terms of SIR, as shown with experiments.

Index Terms—Speech enhancement, blind signal separation, multichannel, microphone array, ad hoc array.

1. INTRODUCTION

Speech enhancement, through noise reduction and interference suppression, is a necessary part of many applications including, hearing aids, teleconferencing, etc. Microphone arrays have been useful means to speech enhancement for many years. Conventional array signal processing techniques assume known geometries for the array and take advantage of its manifold, often to estimate the direction of arrival (DOA) of acoustic wavefronts, such as in [1], and use it implicitly or explicitly to form beamformers to reject directional noise and interferences. The state-of-the-art minimum variance distortionless response (MVDR) beamformer [2], the controllable linearly constrained minimum variance (LCMV) beamformer [3], the informed parametric spatial filter [4], and cooperative noise reduction [5], follow this approach. For distributed and ad hoc microphone arrays, where the array geometry is unknown or the manifold is complex, enhancement apparatus based on other spatial or spectral fingerprints, such as the acoustic transfer function (ATF), the relative transfer function (RTF), and the pseudo-coherence vector, have been proposed during the past few years. Such attempts include the speech distortion weighted multichannel Wiener filter (SDW-MWF) [6, 7, 8], the nested generalized sidelobe canceler (GSC) [9], distributed GSC in local and global stages [10], geometrically constrained TRINICON [11], and the pseudo-coherence-based MVDR beamformer [12].

Without a known array geometry, speech enhancement algorithms with distributed and ad hoc microphone arrays require estimates of the statistics (usually in the form of covariance matrix) of noise, interference or desired signal. The speech presence probability (SPP) is frequently used to estimate the background noise [13, 14]; however, for restoring the interfered speech in multi-speaker scenarios, it is not sufficient. Templates of covariance matrices for desired and/or interfering signals are required in these scenarios. Unfortunately, these are not available in practical situations, which necessitate a blind speech signal separation prior to beamforming. Comprehensive review of subspace methods and specifically joint diagonalization can be found in [15]; however, algorithms which take into account the characteristics of ad hoc microphone arrays are still missing.

In this paper, such an approach is followed by dividing the ad hoc array into microphone duples, and then through searching for the signal structure inside these microphone duples and compare them using passing and rejecting masks in the short-time Fourier transform (STFT) domain. The separation process is finalized through a global geometric mean beamformer. The rest of this paper is organized as follows. The problem is formulated in Section 2, where the signal model is defined in Subsection 2.1. The signal separation process using microphone duples, array partitioning, and global geometric mean filter are introduced in Subsections 2.2, 2.3, and 2.4, respectively. Results from experiments on harmonic signals are presented in Section 3, followed by conclusions in Section 4.

2. PROBLEM FORMULATION

2.1. Signal Model

We consider the problem of separation and enhancement of \( P \) spatially-constrained acoustic sources in a reverberant envi-
ronment using an ad hoc microphone array. The ad hoc array consists of $M_{\text{tot}} \geq P$ randomly positioned omni-directional microphones. It is assumed that at least one microphone is placed closer to each acoustic source than any other source; however, no further information is available a priori on the geometry of the problem.

The signal captured at the time $t'$ with the $m$-th microphone $(m \in \{1, \ldots, M_{\text{tot}}\})$ is

$$y_m(t') = \sum_{p=1}^{P} x_m^p(t') + v_m(t'),$$

$$x_m^p(t') = g_m^p(t') \ast s^p(t'),$$

where $s^p(t')$ is the $p$-th source signal, $v_m(t')$ is the additive noise, $x_m^p(t')$ is the clean (but reverberated) received signal from source $p$, $g_m^p(t')$ is the acoustic impulse response from this source to the $m$-th microphone, and $\ast$ denotes convolution. Acoustic impulse responses are assumed to be time invariant. Furthermore, the signals $x_m^p(t')$ and $v_m(t')$ are assumed to be zero mean, stationary, real, broad-band, and uncorrelated.

Assuming a sufficiently long analysis window, (1) can be written in the time-frequency domain as

$$Y_m(k, t) = \sum_{p=1}^{P} X_m^p(k, t) + V_m(k, t),$$

$$X_m^p(k, t) = G_m^p(k) S^p(k, t),$$

where $S^p(k, t)$, $X_m^p(k, t)$, $V_m(k, t)$, and $Y_m(k, t)$ are the STFTs of $s^p(t')$, $x_m^p(t')$, $v_m(t')$, and $y_m(t')$, respectively, at the time frame $t$ and the frequency bin $k$, and $G_m^p(k)$ is the acoustic transfer function between source $p$ and the $m$-th microphone of the ad hoc array.

The ad hoc array can be partitioned into a set of $N$ sub-arrays, so that the $n$-th sub-array $(n \in \{1, \ldots, N\})$ contains $M(n)$ ≥ 1 microphones, and it is bounded by a hull which does not intersect with the bounding hull for any other sub-array. The set of all microphones in the sub-array $n$, i.e., $S_n$, provides the STFT-domain signal that can be stacked into the vectors:

$$y_n(k, t) = \text{Vect}\left(\{Y_m(k, t)\}_{m \in S_n}\right),$$

$$x_n^p(k, t) = \text{Vect}\left(\{X_m^p(k, t)\}_{m \in S_n}\right),$$

$$v_n(k, t) = \text{Vect}\left(\{V_m(k, t)\}_{m \in S_n}\right),$$

where $\text{Vect()}$ vectorises its operand in column form.

Then the stacked STFT-domain received and clean signals at the $n$-th sub-array are

$$y_n(k, t) = \sum_{p=1}^{P} x_n^p(k, t) + v_n(k, t),$$

$$x_n^p(k, t) = g_n^p(k) S^p(k, t),$$

where

$$g_n^p(k) = \text{Vect}\left(\{G_m^p(k)\}_{m \in S_n}\right),$$

is the stacked ATF for sub-array $n$.

The signal model based on the ATF in (4) can be modified further to obtain models based on the RTF and the pseudo-coherence (PC) vector. As shown in [12], the later is more informative for beamforming with sub-arrays, giving the chance to use the norm or pseudo-coherence vectors as a selection criterion. Thus, the signal model of interest in this paper is

$$y_n(k, t) = \sum_{p=1}^{P} \rho_{x_n^p, X_n^p} (k, t) X_n^p(k, t) + v_n(k, t),$$

where

$$\rho_{x_n^p, X_n^p} (k, t) = \frac{E\left[ x_n^p(k, t) X_n^p(k, t) \ast \right]}{E\left[ |X_n^p(k, t)|^2 \right]}$$

is the pseudo-coherence vector for sub-array $n$ w.r.t the reference signal $X_n^p(k, t)$, with $E[\cdot]$ and superscript $\ast$ being mathematical expectation and complex conjugate, respectively. Notice that the index $r$ is different for distinct sources. Our blind signal separation algorithm does not explicitly depend on (7); however, as discussed in Section 4, it is possible to use the proposed algorithm in conjunction with (7) to form extended filters, such as the pseudo-coherence-based MVDR beamformer in [12].

The covariance matrix of $y_n(k, t)$ can be expressed as

$$\Phi_{y_n}(k, t) = E\left[ y_n(k, t) y_n^\dagger(k, t) \right]$$

$$= \sum_{p=1}^{P} \Phi_{x_n^p, X_n^p} (k, t) + \Phi_{v_n}(k, t),$$

where the transcript $\dagger$ denotes the Hermitian transpose operator. $\Phi_{x_n^p, X_n^p} (k, t)$ is the covariance matrix of $x_n^p(k, t)$, and $\Phi_{v_n}(k, t) = E\left[ v_n(k, t) v_n^\dagger(k, t) \right]$ is the covariance matrix of the noise, $v_n(k, t)$.

It is important to form sub-arrays in a way that a sub-array is more proximal to the $p$-th acoustic source than any other sub-array, while the rest of the sub-arrays are relatively closer to other acoustic sources. Then, many questions arise. Which microphones from the ad hoc array should be assigned to a sub-array? How does the proximal microphones to one of the sources is found and grouped into a sub-array? How is the partitioned array used in signal separation/enhancement? etc. The rest of this section investigates the answers to these questions starting by reduction of the ad hoc microphone array into duples, which are used to differentiate between sources, and then extends the speech separation and enhancement process to all microphones in the ad hoc array.
2.2. Max-SIR Signal Separation with Duples

For every pair of randomly picked microphones from the ad hoc array, designated with the ordered duple \( d = (m_1, m_2) \), the STFT-domain signals are also vectorisable. The stacked received signal at the duple \( d \) is

\[
y_d(k, t) = x_p^d(k, t) + v_p^d(k, t) + v_d(k, t),
\]

where

\[
v_p^d(k, t) = \sum_{q=1, q \neq p}^{P} x_q^d(k, t)
\]

is the interference component for the \( p \)-th acoustic source received with the duple. Without loss of generality, it can be assumed that microphone \( m_1 \) is relatively closer to source \( p \) while microphone \( m_2 \) is relatively closer to one of its interferences, so that \( m_1 \) and \( m_2 \) belong to two distinct sub-arrays. Then, the reference signal of interest that we want to separate from the mixture is \( x_{m_1}^p(k, t) \). According to the signal model in (7), microphone \( m_2 \) weakly receives the reference signal, but captures the weighted summation of interferences (at least for the one which \( m_2 \) is proximal to) much more stronger than microphone \( m_1 \). This inspires the use of complex weights, \( h_q^d(k, t) \), to recover (separate) \( x_{m_1}^p(k, t) \) from \( y_d(k, t) \), i.e.,

\[
Z_d^p(k, t) = h_q^d(k, t) y_d(k, t).
\]

The noise portion of the received signal is usually spatially white, and is also unknown at this point; therefore, the filter in (12) cannot suppress it. However, it is possible to find optimum weights for maximizing the output signal-to-interference-ratio, defined by

\[
oSIR\left[ h_q^d(k, t) \right] = \frac{h_q^t(k, t) \Phi_{x_q^d(k, t)} h_q^d(k, t)}{h_q^t(k, t) \Phi_{v_q^d(k, t)} h_q^d(k, t)}.
\]

The \( oSIR \left[ h_q^d(k, t) \right] \) in (13) is the generalized Rayleigh quotient for the generalized eigenvalue problem of the form:

\[
\Phi_{x_q^d}(k, t) h_q^d(k, t) = \lambda^p \Phi_{v_q^d}(k, t) h_q^d(k, t),
\]

where \( \Phi_{x_q^d}(k, t), \Phi_{v_q^d}(k, t) \in \mathbb{C}^{2 \times 2} \) are Hermitian matrices, and \( \Phi_{v_q} \in S_{++} \) (a positive definite matrix), then \( \Phi_{x_q} - \lambda^p \Phi_{v_q} \) is the Hermitian matrix pencil of order 2. As a result of the Courant-Fischer-Weyl min-max principle, \( h_q^d(k, t) \) is equal to the eigenvector regarding the largest eigenvalue of (14).

Equation (14) requires estimates of covariance matrices \( \Phi_{x_q^d}(k, t) \) and \( \Phi_{v_q^d}(k, t) \) which are not available a priori; however, it is possible to use the duple \( d \) as a differential sensor, i.e., by reversing the order of microphones, and take advantage of relative proximity of \( m_1 \) and \( m_2 \) to different sources. The reverse ordered duple \( \tilde{d} = (m_2, m_1) \) is used together with duple \( d \) to form the approximate generalized eigenvalue problem of the form:

\[
\Phi_{x_{\tilde{d}}}(k, t) h_{\tilde{d}}^p(k, t) = \lambda^p \Phi_{v_{\tilde{d}}}(k, t) h_{\tilde{d}}^p(k, t),
\]

which is specifically correct when desired and interference acoustic signals occupy different frequency bins within a time-frame. Music and voiced speech are signals where it is expected that clean time-frequency bins be obtained from the weights equal to the principal eigenvector of (15).

The degree of freedom in each duple is only sufficient for rejecting the interfering source which is dominantly captured by microphone \( m_2 \) at time-frequency bin \( (k, t) \); however, it is possible to merge various duples by fixing the first element \( (m_1) \) and spanning the second one \( (m_2) \) over all microphones in sub-arrays other than the sub-array that \( m_1 \) belongs to. The merging process requires partitioning the ad hoc array based on information obtain from the max-SIR speech separation for all possible duples.

2.3. Partitioning of the Ad Hoc Array

With \( m \in \{1, \ldots, M_{\text{tot}}\} \) microphones in the ad hoc array, it is possible to form \( \frac{1}{2} M_{\text{tot}}^2 \) distinct duples. Part of the microphones may share similar geometrical constraints, such as relative closeness to one source, so that they should be grouped into a sub-array. A total number of \( N = P \) sub-arrays are required, \( \{S_n\}_{n=1}^P \), in order to separate and enhance \( P \) acoustic sources with the ad hoc array.

The set of narrowband filter outputs for distinct duples is

\[
\left\{ Z_{m_1,m_2}(k, t) \right\}_{m_1 \in \{1, \ldots, M_{\text{tot}}\}, m_2 \in \{m_1+1, \ldots, M_{\text{tot}}\}},
\]

where the superscript \( p \) is dropped since it is not clear yet which source is extracted using each duple. However, (16) provides sufficient information to find this which leads to partitioning the ad hoc array. The set in (16) can be treated as the lower triangle part of a \( M_{\text{tot}} \times M_{\text{tot}} \) matrix. The upper triangle can also be formed; however, it provides redundant information, so it is ignored. To extract the required information, the power spectral density of filter outputs at each time-frequency is compared to the power spectral density of the signal received with the first element of the duple to obtain two STFT-domain masks, namely, the passing and rejecting masks. The elements of the passing mask are equal to one for time-frequency bins that the signal is untouched up to a threshold, and the elements of the rejecting mask are equal to one for time-frequency bins that the signal is attenuated more than a different threshold. These STFT-domain masks, are then used as fingerprints to group microphones into sub-arrays based on the following criteria. The similarity of passing masks for two duples means that their first microphones are geometrically similar, and the similarity of rejecting masks for two duples means that their second microphones are geometrically similar.
The acoustic enclosure is a room of size desired speaker, and the other two as interfering sources. With reverberation time (tic sources are positioned at clean sources are sampled at enhancement apparatus regarding the problem. The acoustic however, no other prior knowledge is available to the en-

Here, the minimum number of microphones is used, i.e., one 

\[
Z_p(k, t) = \prod_{j \in \{S_1 \cup \ldots \cup S_P \setminus S_p \}} Z_{k,t}^{j}(k, t)
\]

removes all interfering signals from the received signal. This 
can be regarded as cascading notch filters to remove the interfer-
ences and to restore the desired signal.

3. EXPERIMENTS

To show the applicability of the proposed algorithm, an experi-
ment with synthesized signals using room impulse responses 
is performed here. The image method is used [16] to pro-
duce 3 acoustic sources, one of which is designated as the 
desired speaker, and the other two as interfering sources. 
The acoustic enclosure is a room of size $5m \times 5m \times 3m$ 
with reverberation time ($T_{60}$) equal to 150 ms. The acous-
tic sources are positioned at $\{2.5, 0.5, 1.5\}$, $\{4.23, 3.5, 1.5\}$, 
and $\{0.77, 3.5, 1.5\}$, while the microphones are positioned at $\{2.5, 1.54, 1.5\}$, $\{3.36, 3.0, 1.5\}$, and $\{1.63, 3.0, 1.5\}$. Here, the minimum number of microphones is used, i.e., one 

microphone being relatively closer to each acoustic source; 
however, no other prior knowledge is available to the en-
hancement apparatus regarding the problem. The acoustic 
clean sources are sampled at $f_s = 8$ kHz. For the STFT, 
the length of each time frame is set to $128$ ms with $75\%$ overlap 
among neighboring frames which corresponds to $32$ ms hop. 
Recursive averaging with a forgetting factor, $\lambda = 0.2$, is used 
in estimation of covariance matrices. The desired speaker is 
iterrated for three simulations such that for the first one, $p_1$ 
is the desired signal while interferences are $p_2$ and $p_3$. All 
synthetic sources are harmonic signals with pitch and harmonics 
according to Table 1. As can be seen in this table, some frequencies are close for different sources, so that it is expected to have less enhancement (interference suppression)

| Table 1. Fundamental frequencies ($F_0$) and their harmonics ($F_{0,i}$) for three acoustic signals. |
|------------------|------------------|------------------|------------------|------------------|
| $F_0$ | $H_1$ | $H_2$ | $H_3$ | $H_4$ |
| $p_1$ | 100 | 200 | 300 | 400 | 500 |
| $p_2$ | 130 | 260 | 390 | 520 | 650 |
| $p_3$ | 170 | 340 | 510 | 680 | 850 |

or more of distortion in the enhanced signal. The results are 
shown in Fig. 1.

4. CONCLUSIONS AND FURTHER REMARKS

According to the results of the experiment in Section 3, merg-
ing microphone duples with a global geometric mean filter 
is an applicable method to separate acoustic sources with 
microphone ad hoc arrays. The results shows interference 
suppression of up to $40$ dB in a reverberant environment; 
however, for certain frequencies the interference suppression 
is as low as $10$ dB. In addition to speech enhancement, the 
proposed algorithm can also be used to provide separated 
signals suitable for estimating necessary parameters for other 


 according to Subsection 2.3, it is possible to estimate the norm, $N_{m_1,m_2}$, and the group-delay, $\tau_{m_1,m_2}$, of the pseudo-coherence vector between the received signals to estimate closeness of the microphone in terms of coherency. The pseudo-coherence vector between the received signals is a 
valid estimate of the pseudo-coherence vector between clean 
(but reverberated) signals for time-frequency bins at which the 


 the passing mask is one, which means that interferences are 
less probably exist. The norm and group delay are defined as 

\[
\begin{align*}
N_{m_1,m_2} &= \left\| \rho_{y_{m_1},y_{m_2}} \right\|^2, \\
\tau_{m_1,m_2} &= \frac{\partial \rho_{y_{m_1},y_{m_2}}}{\partial \omega},
\end{align*}
\]

where the continuous coherencies are estimated from a sparse 
set of time-frequency bins using 2-D regression. Following 
this estimation, the state of the art pseudo-coherence-based 
MVDR beamformer can be used to enhance the acoustic sig-


5.REFERENCES


