This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831

Multiple Sound Source Location Estimation in Wireless Acoustic Sensor Networks using DOA estimates: The Data-Association Problem

Anastasios Alexandridis, *Student Member, IEEE,* and Athanasios Mouchtaris, *Member, IEEE*

***Abstract*—In this work, we consider the data-association problem for the localization of multiple sound sources in a wireless acoustic sensor network (WASN), where each node is a microphone array, using direction of arrival (DOA) estimates. The data-association problem arises because the central node that receives the multiple DOA estimates from the nodes cannot know to which source they belong. Hence, the DOAs from the different nodes that correspond to the same source must be found in order to perform accurate localization. We present a method to identify the correct association of DOAs to the sources and thus accurately estimate their locations. Our method results in high association and localization accuracy in realistic scenarios with missed detections, reverberation, noise, and moving sources and outperforms other recently proposed methods. It also incorporates a bitrate reduction scheme in order to keep the amount of information that needs to be transmitted in the network at low levels without affecting performance.**

***Index Terms*—data-association, localization, direction of arrival estimates, microphone arrays, wireless acoustic sensor networks**

I. Introduction

ENABLING machines to estimate the locations of the active sound sources from their emanating acoustic signals has always been an attractive problem in the research community. Inference of location information is crucial in many applications, such as wildlife monitoring [1], [2] and speech enhancement for robust signal acquisition [3].

When multiple acoustic sensors are used to monitor an acoustic environment, we usually refer to this setup as a Wireless Acoustic Sensor Network (WASN) [4]. The nodes may consist of a single microphone or a microphone array, i.e., multiple microphones arranged in a pre-defined geometry, and are distributed at different locations in the monitored area. The nodes are equipped with a processing unit to perform signal processing operations, and a wireless communication module to communicate with other nodes or a central node which is known as the “fusion center”.

Generally, in such a setup, the localization of the acoustic sources can be performed at the fusion center, with information received by the nodes and a suitable model that relates this information with the location of the source(s) of interest. In the literature, different sources of information have been studied, such as the energy [5], [6], temporal [7], [8] and/or directional features [9]–[12], and spatial likelihood functions such as the steered response power (SRP) function [13]–[15]. The reader is referred to [16] for a review of localization approaches using various types of information from the nodes.

However, practical considerations about the sensor network itself, such as the wireless nature of the nodes and their limited processing capabilities, as well as possible requirements for real-time processing, pose several limitations and challenges that must be taken into account. Due to their limited processing power, the nodes cannot carry out very complex and computationally intensive operations while restrictions in the bandwidth usage limit the amount of information that can be transmitted in the network. Finally, since the nodes operate individually, the acquired audio signals at different nodes will not be synchronized.

When each node is a microphone array, it can estimate and transmit direction-of-arrival (DOA) estimates of the active sound sources. DOA estimates describe the direction from which sound is propagating with respect to a node in each time instant. The location of a source can be estimated at the fusion center by fusing the DOA measurements, also known as *bearing* measurements. Although such approaches require increased computational complexity in the nodes—to perform the DOA estimation—, they attain low bandwidth usage as only DOA estimates need to be transmitted. Also, since the DOA estimation is carried out in each node individually, the audio signals at different nodes need not be perfectly synchronized. Finally, the variety of broadband DOA estimation methods for acoustic sources available in the literature makes it easy to obtain such estimates: several methods have been proposed such as the broadband MUSIC [17] and ESPRIT [18] algorithm, methods based on Independent Component Analysis (ICA) [19] and Sparse Component Analysis (SCA) [20].

In the single source case, the location of a source can be estimated by the intersection of lines emanating from the nodes' locations at the direction of the nodes' estimated DOA, as illustrated in Fig. 1, a method which is called *triangulation*. Since the DOA estimates will be contaminated by noise, several approaches have been proposed to tackle this estimation problem, including the Stansfield estimator [21], the orthogonal vectors (OV) estimator [22], the single-source grid-based (GB) method [9], [23], maximum likelihood nonlinear estimators [24]–[28], approaches based on instrumental

This research has been partly funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 644283, Project LISTEN.

A. Alexandridis and A. Mouchtaris are with the Foundation for Research and Technology- Hellas, Institute of Computer Science (FORTH-ICS), Heraklion, Crete GR-70013, Greece, and also with the Department of Computer Science, University of Crete, Heraklion, Crete GR-70013, Greece (e-mail: analexan@ics.forth.gr; mouchtar@ics.forth.gr)

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831

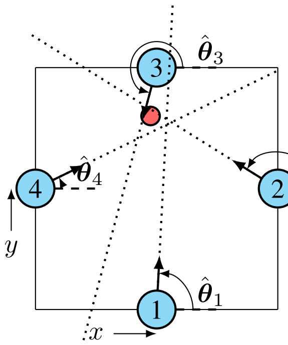

3
$\hat{\boldsymbol{\theta}}_3$
4
$\hat{\boldsymbol{\theta}}_4$
2
$\hat{\boldsymbol{\theta}}_2$
y
1
$\hat{\boldsymbol{\theta}}_1$
x

Fig. 1. Example cell with four sensor nodes (blue circles), and the estimated source location (red circle) based on triangulation of the estimated DOAs ($\hat{\boldsymbol{\theta}}_1$–$\hat{\boldsymbol{\theta}}_4$).

variables [29], [30] and total least squares [25], [31].

However, the focus of this paper is on the estimation of the locations of multiple simultaneous sound sources from their DOA estimates, which is much more complex than its single source counterpart. A fundamental problem is that the fusion center receiving the multiple DOA estimates from each node (one DOA for each detected source) cannot know to which source each DOA belongs. This is known as the *data-association problem*. The correct association of DOAs from the nodes that correspond to the same source must be found, otherwise location estimation will result in “ghost” sources, i.e., locations not corresponding to real sources.

The data-association problem is illustrated in Fig. 2 with an example in a two-node WASN with two active sound sources. The solid lines show the DOAs to the first source and the dashed lines show the DOAs to the second source. Intersecting the DOA lines from the nodes results in 4 possible source locations. When intersecting the DOAs that correspond to the same source, i.e., the two solid lines and the two dashed lines, the correct source locations are estimated (red circles). When the erroneous combination of DOAs is used, the estimation results in “ghost” sources (white circles). When the correct association of DOAs from the nodes to the sources is found, the multiple source localization problem decomposes into multiple single-source localization problems which are straightforward to solve by applying any single-source location estimator proposed in the literature to the resulted DOA associations.

Also, when multiple sources are active, some nodes may not be able to detect some sources, thus underestimating their number. As a result of such *missed detections*, the number of detected sources—and thus the number of estimated DOAs—can vary across the nodes and through time. This can occur in several situations, such as when the sources are close together in terms of their angular distance with respect to a node or when a source is located far away from a node. As illustrated in our previous work [32], the problem of missed detections occurs very often in practice and it is an important aspect to the location estimation problem which—to the best of our knowledge—has not been examined so far.

In this work, we propose a novel approach to address the data-association problem for localizing multiple sound sources using DOA estimates. Our method is based on the extraction of features at the nodes (one feature for each detected source).

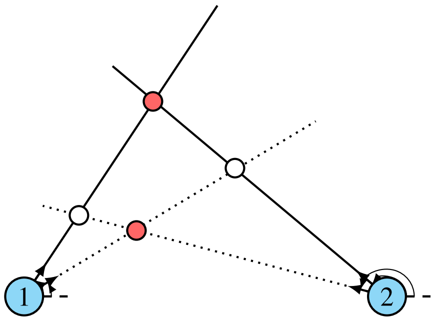


Fig. 2. Illustration of the data-association problem in a two-node WASN with two active sound sources. The four possible source locations may either be the true sources' locations (red circles) or locations of “ghost” sources (white circles) as the result of using bearing lines that do not correspond to the same source.

The association of DOAs to the sources is found by comparing the corresponding features and separating them into groups according to their similarity. We propose the use of features that describe how the frequency components of the captured signals at each node are distributed to the sources and we show that these features are robust to missed detections and noise. We also propose a greedy algorithm for the association of the DOAs based on the similarity of their corresponding features. Our algorithm can work with an arbitrary number of nodes and sources, results in high association accuracy and is computationally efficient. As the features need to be transmitted to the fusion center for the association procedure to be performed, we also study how to reduce the amount of information that needs to be transmitted. We propose a scheme that reduces the bitrate requirements of our method up to 88% without affecting its accuracy.

Parts of this work has also been presented in [32]. This current work presents an improved and more detailed methodology, especially in the following respects: (i) it takes into consideration the realistic and practical aspect of the amount of information that the method requires to be transmitted in the network and proposes solutions in order to significantly reduce the bitrate requirements without affecting performance, and (ii) it provides a more detailed performance analysis, especially in terms of the association algorithm's execution time and in terms of the method's potential to localize moving sound sources.

The remainder of this paper is organized as follows: Section II reviews the current state-of-the-art methods for the data-association problem. Section III describes the considered data-association and localization problem. The proposed method is presented in Section IV, while an approach to reduce the bitrate requirements is detailed in Section V. Section VI provides an experimental validation of our method and Section VII concludes our work.

II. State-of-the-art methods for the data-association problem

Using only the estimated DOAs from the nodes, some approaches try to address the data-association problem by

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831

enumerating all possible DOA combinations from the nodes, estimating a set of initial location estimates using all possible combinations and then deciding which of the initial location estimates correspond to the true sources' locations. In this spirit, our previously proposed grid-based method for multiple sources [9], [23] decides on the final location estimates using heuristic approaches: for each initial location estimate it measures the angular distance between the DOA combination that used to generate that estimate and the DOAs that correspond to the estimated location. The initial location estimates with the smallest distance are reported as the final location estimates.

The method in [33] incorporates the data-association to the localization procedure by designing a non-linear estimator that each time is initialized with a different estimate from the set of the initial location estimates. For each initial location estimate, the estimator is expected to converge to a location of a true source. However, as illustrated in [9], in the presence of missed detections and high noise, the performance of this approach severely degrades as the estimator converges to the locations of "ghost" sources thus not being able to correctly identify the true sources' locations.

Other approaches tried to solve the data-association problem prior to the localization procedure. When the correct association of DOAs from the nodes to the sources is estimated beforehand, the multiple source localization problem decomposes into multiple single source localization problems, which can be conveniently solved using any of the single source location estimators which are available in the literature. In [34] the data-association problem is viewed as an assignment problem and is formulated as a statistical estimation problem which involves the maximization of the ratio of the likelihood that the measurements come from the same target to the likelihood that the measurements are false-alarms. However, the proposed solution is NP-hard when three or more nodes are considered. Some sub-optimal solutions tried to solve the same problem in pseudo-polynomial time [35], [36].

The clustering of intersections of bearing lines in scenarios with no missed detections is introduced in [37]. The motivation behind this approach is based on the observation that intersections between pairs of bearing lines that correspond to the same source will be close to each other, forming clusters around the locations of the true sources that reveal the correct DOA associations. On the contrary, intersections from bearing lines that do not belong to the same source will be randomly distributed in space. Another approach that also utilizes intersections of pairs of bearing lines in order to decide the locations of the sources is discussed in [10], but again the presence of missed detections results in significant performance drop.

The availability of additional information—apart from the DOA estimates of the detected sources—can generally lead to more efficient solutions. The method of [38] associates each detected DOA with a binary mask in the frequency domain that can be used to separate the corresponding source signal. The association of DOAs to the sources is found by comparing the binary masks across different nodes. The DOAs of the masks that correlate the most are assigned to the same source. Again, the method does not consider missed detections, while the association algorithm was designed for the limiting case of two nodes.

III. Problem statement, definitions, and assumptions

Consider a WASN with M nodes, where each node is a microphone array. In the sequel, the terms node and microphone array will be used interchangeably. The proposed method is not attached to a specific microphone array geometry or number of microphones. We assume the presence of a central node (fusion center) which is responsible for the location estimation. The nodes in the network are connected to the fusion center over wireless links.

In the acoustic environment that is monitored by the WASN, we assume that K sound sources are simultaneously active. The number of sources is assumed to be known. Each array uses a method to estimate the azimuth DOAs of the active sources in each time instant and transmits these estimates to the fusion center. Note that, although the number of sources is known, an array may not be able to detect some of the sources. We refer to these situations as *missed detections*. Missed detections can occur for several reasons: due to the challenging setup in terms of reverberation and noise, because some sources may be located close together in terms of their angular separation for an array to discriminate between them, or because they are located far away from the array. As a result, the number of DOAs each array transmits may be less than K and may also vary in time and across the arrays. In general, each array can detect *up to* K sources. However, we assume that each source is detected by *at least* one array, which is a necessary condition to find the DOA associations for all K sources.

When the correct association of DOAs to the sources is found, the location estimation can be carried out by simply applying a single source location estimator to the resulted DOA associations. To estimate the location of a source in the two-dimensional space, at least two azimuth DOA estimates are required. Thus, for localization we consider only the sources that have been detected by at least two arrays.

In such a setup, we aim to estimate the correct association of DOAs from the nodes to the arrays and the final locations of the sources.

IV. Proposed method

The main idea of the proposed method is to utilize additional information—apart from the DOA estimates—to solve the data-association problem. In this spirit, each microphone array estimates and transmits features associated with each source it detects. For the same source, such features must be "similar" across the different arrays. We denote such features as the *association features*. Apart from the design of such features, the second major part of the solution consists of the association algorithm that finds the DOA combinations whose associated features are most "similar".

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831

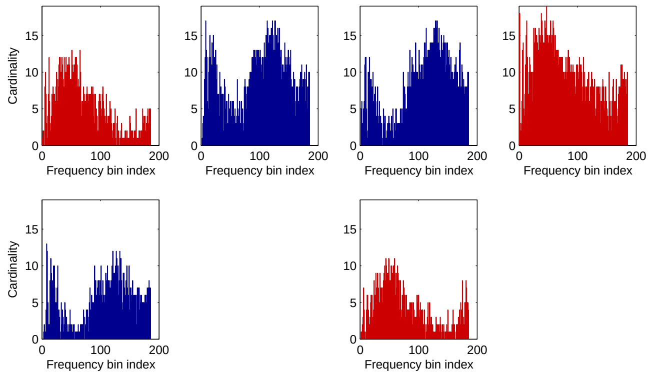


Fig. 3. Example of association features for a WASN with $M = 4$ arrays (columns) and $K = 2$ sound sources (rows). The colors indicate the association features that correspond to the same source in different arrays and are most similar to each other. In this example, array 2 and array 4 have exhibited missed detections thus detecting and estimating the corresponding association feature for only one source.

Algorithm 1 Feature computation at the mth node

Input: Frame of microphone array signals $\mathbf{X}_{\text{fr}}(\ell)$ in the frequency domain, History length B, User-defined threshold ϵ
Output: Association Features $F_{m,k}(\ell)$ for each detected source k

```
for each frequency bin $\ell$ in $L$ do
    $\phi(\ell)$ = Narrowband_DOA_Estimation($\mathbf{X}_{fr}(\ell)$)
end for
$(\hat{\boldsymbol{\theta}}, \hat{K})$ = Broadband_DOA_Estimation($\mathbf{X}_{fr}$)
$F_{m,k}(\ell) = \mathbf{0}, k = 1, \ldots, \hat{K}$
for each frame $\tau'$ between the current frame $\tau$ and $B$ previous frames do
    for each frequency bin $\ell$ in $L$ do
        $k \leftarrow \arg\min_p \left( A(\phi_{\tau'}(\ell), \hat{\theta}_p \right)$
        if $A\left(\phi_{\tau'}(\ell), \hat{\theta}_k\right) < \epsilon$ then
            $F_{m,k}(\ell) \leftarrow F_{m,k}(\ell) + 1$
        end if
    end for
end for
```

A. *In-node feature extraction*

The feature extraction is based on the assumption that each time-frequency bin belongs to at most one source. The assumption that only one source is dominant in each time-frequency bin is known as the W-disjoint orthogonality (WDO) assumption and has been shown to be valid especially for speech signals [39], [40]. The feature computation is based on the estimation of narrowband DOA estimates in each time-frequency bin and the assignment of each bin to a source based on the corresponding estimated DOA. The features are computed for each time-frame and represent the number of times each frequency bin was assigned to a source in the last B frames, where B refers to the *history length* or *block size*. Each feature is associated to a source and provides an estimate of the distribution of the frequency components to that source.

The microphone array signals at each array m are transformed into the Short-Time Fourier Transform (STFT) domain, resulting in the signals $X_{m,i}(\tau, \ell)$ where i is the microphone index and τ, ℓ denote the time-frame and frequency bin index respectively. In the sequel, we omit m and τ as the procedure is repeated in each array for each time-frame. We also denote as L the set of frequency bins ℓ up to a maximum frequency ℓ_{max}. We set ℓ_{max} to the spatial-aliasing cutoff frequency which depends on the array geometry and describes the frequency up to which reliable DOA estimates can be found.

In each frequency $\ell \in L$ we estimate a DOA, resulting in the narrowband DOA estimates $\phi(L)$. For the narrowband DOA estimation any method available in the literature can be utilized. Also, a broadband DOA estimation method estimates the number of detected sources $\hat{K}$ and their corresponding DOAs $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \ldots, \hat{\theta}_{\hat{K}}\}$. This can be achieved with an arbitrary broadband DOA estimation method, e.g., by processing the narrowband DOA estimates with a matching-pursuit algorithm as proposed in [20], [41].

Then, the frequencies in L are assigned to the detected sources. The assignment is based on the DOAs of the sources $\hat{\boldsymbol{\theta}}$ at the current frame and the narrowband DOA estimates in each frequency bin $\phi_{\tau'}(\ell)$ for each frame τ' between the current and B previous frames. A frequency bin $\ell \in L$ is assigned to source p (with corresponding direction $\hat{\theta}_p$) if the following two conditions are met:

$$A(\phi_{\tau'}(\ell), \hat{\theta}_p) < A(\phi_{\tau'}(\ell), \hat{\theta}_q), \quad \forall q \neq p, \tag{1}$$

$$A(\phi_{\tau'}(\ell), \hat{\theta}_p) < \epsilon, \tag{2}$$

where $A(X, Y)$ denotes an angular distance function that returns the difference between X and Y in the range of $[0, \pi]$ (see Appendix for details on its computation). In other words, Eqs. (1) and (2) suggest that a given frequency bin is assigned to the source whose DOA is closest to the estimated DOA at that bin, as long as that distance does not exceed a pre-defined threshold ϵ. When Eq. (2) is not satisfied, the given frequency bin is rejected and not assigned to any of the sources.

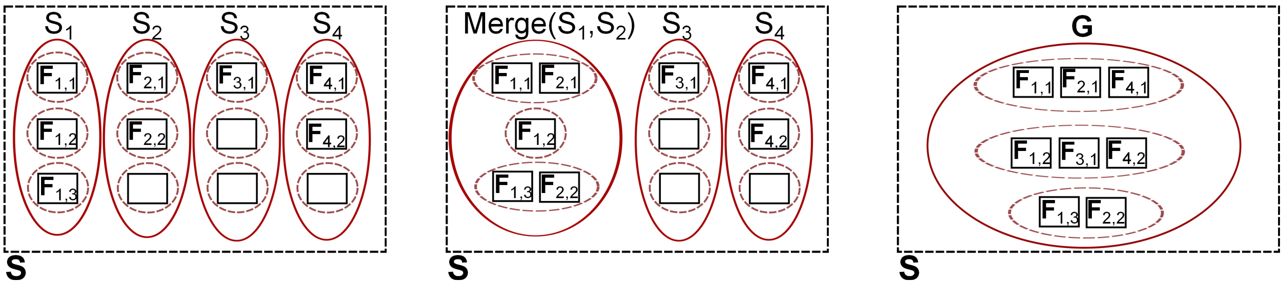
Since the assignment is carried out for the frequency bins for the current and B previous frames, a histogram can be formed for each detected source that counts how many times each frequency bin was assigned to that source. Note that since B frames are considered, a frequency bin can be assigned to a source up to B times. These histograms constitute the proposed association features which are transmitted to the fusion center together with the estimated DOAs for the detected sources. The proposed feature extraction procedure is presented in Algorithm 1.

Since all the arrays receive the same signals—albeit with relative phase differences—the histograms across the arrays that belong to the same source are expected to be similar. As each histogram is associated with a source's DOA, the grouping of the histograms in K groups, based on their similarity, is expected to reveal the association of DOAs from the arrays to the K sources.

An example of these association features is shown in Fig. 3 for a WASN of 4 microphone arrays with two active sound sources. The colors indicate the histograms that correspond to the same source at the different arrays. The association features (i.e., histograms) that correspond to the same source are expected to be "similar". In this example, two arrays have exhibited missed detections thus being able to detect only one source and thus estimating a single association feature.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831



(a) (b) (c)

Fig. 4. Example of the association algorithm for $M = 4$ arrays and $K = 3$ sources. (a) First, an assignment of features to K groups is created for the set of features $\mathbf{F}_m$ for each array. The empty boxes represent the empty groups, as the corresponding arrays have detected less than K sources. (b) The algorithm finds the assignment (array 1 and 2 in this example) that, when merged, produce the best score according to (5) and merges them. (c) The merging operations stop when a single assignment $\mathbf{G}$ remains.

Algorithm 2 Association Algorithm

```
Input: Features $\mathbf{F}$ , Number of Sources $K$
Output: Assignment $\mathbf{G}$
  $\mathbf{S} \leftarrow \bigcup_i CreateInitialAssignment(\mathbf{F}_i, K)$ (Fig. 4(a))
  while $|\mathbf{S}| > 1$ do (Fig. 4(b))
    $(i, j) \leftarrow \arg\min_{i,j} Score(Merge(\mathbf{S}_i, \mathbf{S}_j))$
    $\mathbf{S} \leftarrow \mathbf{S} \setminus (\mathbf{S}_i \cup \mathbf{S}_j) \cup Merge(\mathbf{S}_i, \mathbf{S}_j)$
  end while
  $\mathbf{G} \leftarrow GetFinalAssignment(\mathbf{S})$ (Fig. 4(c))
  while $\min_{p,q} Score(\mathbf{G}[F_p^i \leftrightarrow F_q^j]) < Score(\mathbf{G})$ do
    $\mathbf{G} \leftarrow \mathbf{G}[F_p^i \leftrightarrow F_q^j]$
  end while
```

B. Data-association algorithm at the fusion center

Given the estimated association features, the goal of the association algorithm, which is carried out at the fusion center, is to group them according to their similarity in K groups. We remind the reader that K is the number of sources which is assumed to be known. As each feature is associated with a DOA estimate, the grouping of the features will reveal the association of DOAs from the arrays to the sources, which is indicated by the different colors in the example of Fig. 3. In this section, we formally define the data-association problem as an assignment problem of the association features to K groups.

Let $\boldsymbol{F}$ denote the set of all association features and $\boldsymbol{F}_m$ denote the set of features $F_{m,k}$, $k = 1, \ldots, \hat{K}_m$ for all detected sources from the mth array. The problem is to find a set $\boldsymbol{G}$ that contains K groups of features, denoted as G_i, $i = 1, \ldots, K$, $G_i \in \boldsymbol{G}$, such that:

(a) features from the same array cannot be assigned to the same group,
(b) each feature must be assigned to exactly one group, and
(c) all groups contain features that are "similar" to each other.

We call the set $\boldsymbol{G}$ an assignment of features to groups. As each feature corresponds to a source's DOA, the resulting K groups provide the association of DOAs across the arrays for the K sources. We proceed by proposing and defining a way to measure the quality of an assignment.

Let D be a function measuring the dissimilarity of two features, taking values in $[0, 1]$. We define the *score* of each group G_i as the maximum pairwise dissimilarity of its contained features:

$$Score(G_i) = \max_{p,q} D(F_p^i, F_q^i), \tag{3}$$

where F_p^i denotes the pth feature of group G_i.

We define the overall score of an assignment $\boldsymbol{G}$ as the maximum score among the scores of its contained groups $G_i \in \boldsymbol{G}$:

$$Score(\boldsymbol{G}) = \max_i Score(G_i). \tag{4}$$

Our goal is to find an assignment that minimizes (4), while satisfying constraints (a) and (b) mentioned above. Thus, the solution to the data-association problem can be formally defined as:

$$\arg\min_{\boldsymbol{G}} Score(\boldsymbol{G}). \tag{5}$$

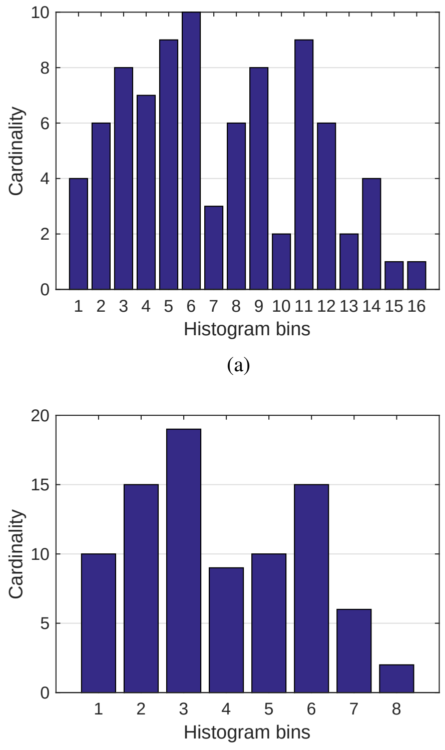
In case two assignments result in the same score, we sort the scores of their groups in descending order and compare their maximum non equal score. The motivation behind this formulation of the data-association problem is that we want to find an assignment where *all groups* contain features that are as similar as possible to each other, since the contained features in each group correspond to the same source.

The next step is to deduce an algorithm that can efficiently solve (5). A straightforward approach would be to exhaustively enumerate all possible assignments and choose the one that satisfies (5). Although it can guarantee to find the assignment with the minimum score, such a naive brute-force approach cannot be realized in practice due to its prohibitively large computational requirements, as the number of possible assignments can grow as $(K!)^M$. Next, we derive a greedy algorithm that can efficiently solve (5). The algorithm—shown in Algorithm 2—does not necessarily identify the optimal solution, but it is simple, fast, and as our experimental results indicate, it finds good solutions in practice.

First, for each set $\boldsymbol{F}_m$, $m = 1, \ldots, M$ we create an assignment $\boldsymbol{S}_m$. The assignment contains K groups, where each group contains a single feature. If the array has exhibited missed detections, thus having estimated less than K features, some groups are left empty. This procedure is illustrated in Fig. 4(a). The algorithm then tries to greedily merge those assignments $\boldsymbol{S}_m$ until only one remains. The merging is done by considering all possible ways to merge two assignments and selecting the one which produces the best possible score according to (5) (Fig. 4(b)). The possible ways to merge the

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831



(a)

(b)

Fig. 5. Example of a histogram with $N_\ell = 16$ bins (a) before and (b) after a decimation process by a decimation factor of $d = 2$.

assignments equals $K!$. This is known as the *Linear Bottleneck Assignment Problem* [42], which can be solved efficiently in polynomial time. However, when K is relatively small, a brute-force approach is often faster. The algorithm finishes when only a single assignment, $\boldsymbol{G}$, remains (Fig. 4(c)).

Then, in order to further refine the estimated assignment, we perform a second greedy step. In this step, we select two features F_p^i and F_q^j from different groups i and j and try to swap them in order to further reduce the score of $\mathbf{G}$; for brevity, we use $\mathbf{G}[F_p^i \leftrightarrow F_q^j]$ to refer to the new assignment in which F_p^i and F_q^j are interchanged. We also allow one of them to be empty or, in other words, to move a feature from one group to another. The algorithm terminates if no such pair exists. As our results in Section VI-C indicate, this second greedy step was found to result in significant performance gain.

V. Reducing transmission requirements

In terms of transmission requirements, each array must transmit the histogram (association feature) for every source it detects to the fusion center. In this section, we quantify the transmission requirements of our method and propose how to reduce the amount of information that needs to be transmitted.

The number of bins in a histogram equals the number of frequency bins that are processed. As discussed in Section IV-A, the processing for the extraction of the histograms is performed for a frequency range up to frequency $\ell_{\max}$. Let N_ℓ denote the number of frequency bins available for processing, i.e., the number of frequency bins up to frequency $\ell_{\max}$. Given a history length of B frames, the maximum cardinality of a given bin in the histogram is B. Thus the number of bits required to transmit a histogram is $\lceil N_\ell \log_2(B) \rceil$.

We propose to reduce the transmission requirements by reducing the number of bins in the histograms by performing a decimation process as follows: Let $h : \mathcal{A} \to \mathcal{B}$ be a function describing a histogram where its domain corresponds to the frequency bin indices, i.e., $\mathcal{A} = \{1, \ldots, N_\ell\}$ and its range corresponds to the cardinality of each bin, i.e., $\mathcal{B} = \{0, 1, \ldots, B\}$. A decimated version of h by a factor of d can be formed by grouping each consecutive d bins and summing their cardinalities according to:

$$h'(x) = \sum_{k=1}^{d} h\left(d(x-1) + k\right), \tag{6}$$

where $h' : \mathcal{C} \to \mathcal{D}$ with $\mathcal{C} = \{1, \ldots, \lceil \frac{N_\ell}{d} \rceil\}$ and $\mathcal{D} = \{1, \ldots, d \cdot B\}$. After the decimation process the domain of the new histogram has been shrunk by a factor of d and the range of possible values for each bin is now $d \cdot B$. The number of bits N_b required to transmit the decimated histogram depends on the history length B, the number of bins N_ℓ, and the decimation factor d through:

$$N_b = \lceil \frac{N_\ell}{d} \log_2(d \cdot B) \rceil. \tag{7}$$

The initial histogram, without any decimation, corresponds to the case where $d = 1$. An example of a histogram and its decimated version by a factor of $d = 2$ is shown in Fig. 5. In this example, the initial histogram has $N_\ell = 16$ bins and its maximum cardinality is 10, thus requiring $\lceil 16 \log_2(10) \rceil = 54$ bits for transmission. After decimation by $d = 2$ the number of bins have reduced to $\frac{N_\ell}{d} = 8$ and the maximum cardinality is now 20, thus requiring 35 bits. As we will show in Section VI-E, we can apply a decimation process by a factor of 16 to our histograms, thus significantly reducing the amount of information that needs to be transmitted in the network, without degradation in performance.

VI. Evaluation

To evaluate the proposed method we performed simulations on a square cell of a WASN with dimensions of $V = 4$ meters with $M = 4$ nodes which were configured as shown in Fig. 1. Each node was an 8-element uniform circular microphone array with 5 cm radius. The sound sources were speech recordings of 2 sec. duration, sampled at 44.1 kHz, and had equal power when placed at the center of the cell. The signal-to-noise ratio (SNR) was measured as the ratio of the power of each source signal when located at the center of the cell to the power of the noise signal. To simulate different SNR values we added white Gaussian noise at each microphone, uncorrelated with the source signals and the noise at the other microphones. Note that this framework results in different SNR at each array depending on how close the source is to the arrays.

We simulated a room of dimensions of $10 \times 10 \times 3$ meters using the Image-Source method [43] and produced signals of omnidirectional sources at different reverberation conditions.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831

TABLE I
EXPERIMENTAL PARAMETERS

parameter	notation	value
room		$10 \times 10 \times 3$ m.
WASN cell		square
WASN side length	V	4 meters
node type		8-element uniform circular array, 5 cm radius
number of nodes	M	4
framesize		2048 samples
overlapping in time		1024 samples
FFT size		2048 samples
sampling frequency	F_s	44.1 kHz
highest frequency for processing	$\ell_{\max}$	4 kHz
threshold for frequency assignment	ϵ	10°
history length (block size)	B	21 frames (0.5 sec.)
decimation factor	d	1

The WASN cell was placed at the middle of the room. Both the nodes and the sources were placed at 1.5 m. height. More specifically, the nodes were placed at $(5, 3, 1.5)$, $(7, 5, 1.5)$, $(5, 7, 1.5)$, and $(3, 5, 1.5)$. In terms of number of sources, we considered scenarios of two and three simultaneously active speakers. Each simulation was repeated 30 times and the sources were placed at different locations within the cell with independent uniform probability. For narrowband DOA estimation we used the method proposed in [44], which is designed for the uniform circular array geometry. For the estimation of the broadband DOAs of the sources in each time frame, we applied our previously proposed methodology of [20], [41]. Note that in our evaluation, we use circular microphone arrays and the DOA estimation methods employed are tailored for this specific array geometry. However, the proposed methodology is independent of the array geometry and the DOA estimation method in the sense that any DOA estimation method available in the literature can be employed to infer the narrowband and broadband DOA estimates.

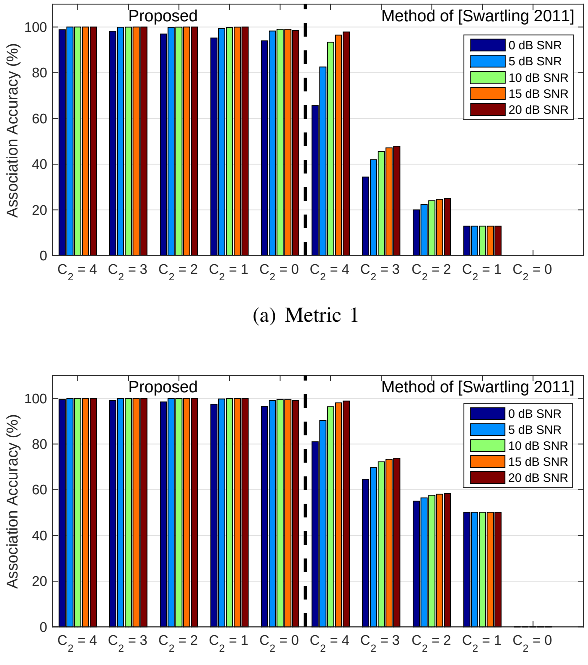
For processing we used frames of 2048 samples with 50% overlap. The FFT size was 2048. We set $\ell_{\max}$ to 4 kHz which is the spatial aliasing cutoff frequency for our given array geometry. The threshold ϵ for the frequency assignment in Eq. (2) was set to 10° and we used a history length of $B = 21$ frames, which corresponds to 0.5 seconds. As a dissimilarity measure in (3), we used the Pearson Correlation Coefficient distance which is defined as [45]:

$$D(X, Y) = \frac{1 - r_{X,Y}}{2}, \tag{8}$$

where $r_{X,Y}$ is the Pearson correlation coefficient between X and Y. Eq. (8) takes values in the range $[0, 1]$. The parameters are summarized in Table I and are used throughout our experimental evaluation, unless stated otherwise.

A. Evaluation Metrics

To measure the association accuracy we utilize two metrics. The first is denoted as *Metric 1* and measures the percentage of time frames where a correct DOA association is found. We define a DOA association as correct when *all* DOAs are assigned to the correct source from *all* the arrays.



(a) Metric 1

(b) Metric 2

Fig. 6. Data-association accuracy for two sources in an anechoic environment for different values of SNR and C_2
.

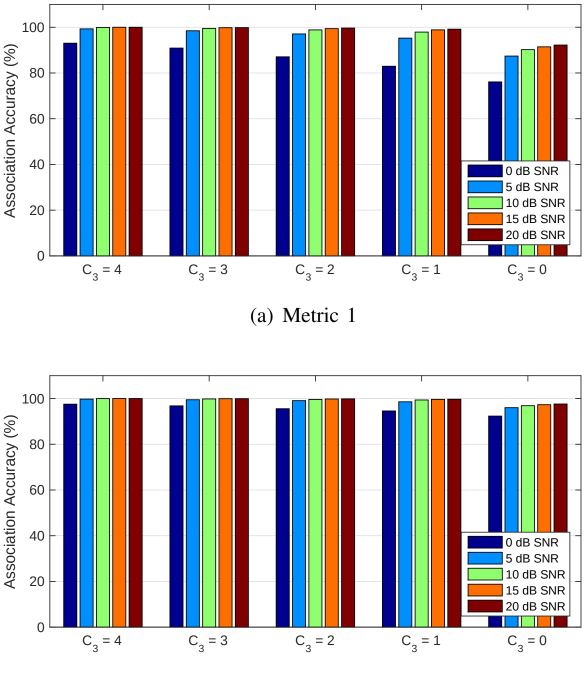
When an association is not correct, it means that some DOAs from some arrays are assigned to an erroneous source. However, although the association is erroneous there are still pairs of DOAs from different arrays that were associated correctly. As an example, let us assume that in Fig. 4(c) the DOA that corresponds to feature $F_{4,1}$ was erroneously assigned to the source that corresponds to the first group. While this association is erroneous—according to Metric 1—there are pairs of DOAs from arrays that are associated correctly, such as the pairs $(F_{1,1}, F_{2,1})$, $(F_{1,2}, F_{3,1})$,$(F_{3,1}, F_{4,2})$, $(F_{1,2}, F_{4,2})$, and so on, while other pairs are associated erroneously, such as pairs $(F_{1,1}, F_{4,1})$, $(F_{2,1}, F_{4,1})$. Of course, the more these correct pairs of DOAs are, the less impact an erroneous pair will have to the data-association and thus to the localization error. To quantify the correct "parts" of a DOA association—that it can albeit be erroneous according to the definition of Metric 1—we use our second metric (denoted as *Metric 2*), which counts the percentage of correct pairwise associations between all pairs of arrays.

B. Robustness to missed detections

First, we evaluate the efficiency of our proposed association features and our proposed association algorithm in scenarios with missed detections. We assume that the DOAs of the sources in each time-frame, i.e., vectors $\hat{\boldsymbol{\theta}}$ at each array are known. We define C_s as the number of arrays that detected s sources, i.e., $C_2 = 3$ indicates that three arrays detected two sources. To simulate missed detections, we fix C_s and remove some DOAs from some arrays until the desired value of C_s is

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

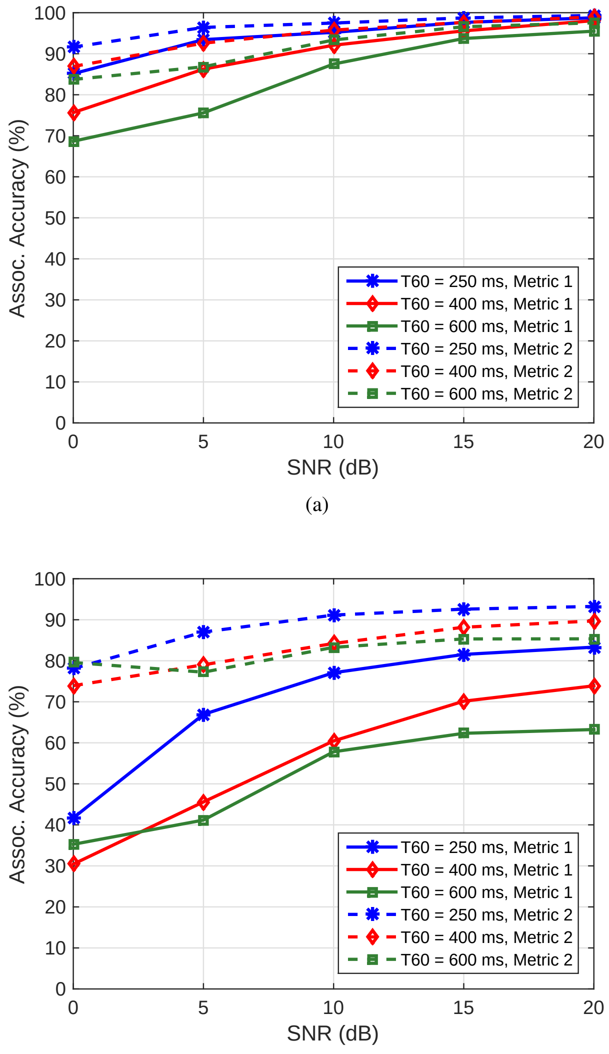
This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831



(a) Metric 1

(b) Metric 2

Fig. 7. Data association accuracy of the proposed method for three sources in an anechoic environment for different values of SNR and C_3.



(a)

(b)

Fig. 8. Data-association accuracy of the proposed method for different SNR values and reverberation conditions for (a) two and (b) three active sound sources.

reached. The removed DOAs as well as the arrays that exhibit the missed detections are selected at random in every frame, under the constraint that each source must be detected by at least one array (Section III).

Fig. 6 depicts the data-association accuracy (using the two aforementioned metrics) for an anechoic scenario of two active sound sources for all possible values of C_2, i.e., the number of arrays that detected the two sources. For comparison, Fig. 6 also presents the results using the association features proposed in [38] (denoted as [Swartling 2011] in the figure). These features were modified to work with circular microphone arrays. For association, we applied our proposed association algorithm on the features extracted from [38], as the association algorithm proposed in [38] works only for the case of two arrays. From Fig. 6 it is evident that our approach is robust to missed detections, achieving more than 90% accuracy for all SNR cases and all values of C_2.

On the other hand, the association features proposed in [38] are less robust to noise and missed detections. A severe performance degradation is evident, especially when missed detections are present ($C_2 < 4$). A key reason for that is the fact that the association features were not designed to handle missed detections: when a source is not detected, the method of [38] erroneously assigns its frequencies to the other sources, thus degrading the association performance. Our proposed method avoids such erroneous assignments through the use of Eq. (2). It is noteworthy that our proposed approach can accurately find the correct association of DOAs to the sources, even in the extreme case where all arrays detected only one source, i.e., $C_2 = 0$. Finally, the features of [38] cannot handle the case where $C_2 = 0$ (the corresponding area in Fig. 6 is left blank). In this case, the association features of [38] cannot provide any useful information in order to estimate the correct association of DOAs to the sources.

Fig. 7 depicts the association accuracy, using Metric 1 and 2, for an anechoic scenario of three active sound sources and different values of SNR and C_3, i.e., the number of arrays that detected three sources. For each value of C_3 the figure presents the mean association accuracy over all possible combinations of C_2 and C_1. Again, the robustness of the proposed approach to missed detections is evident: our method achieves high accuracy for all SNR values even in the case where missed detections are so prominent that none of the arrays detected three sources, i.e., $C_3 = 0$.

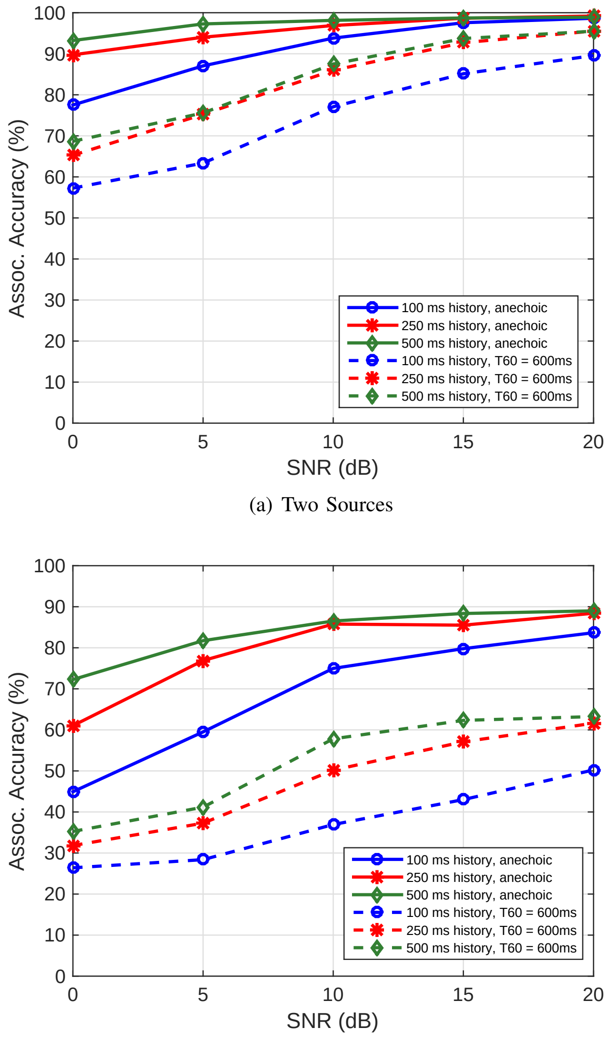
C. Data association algorithm

We now demonstrate the effectiveness of our data association approach to more realistic scenarios with reverberation, where the DOAs of the sources in each time frame are

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.

The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831



(a) Two Sources

(b) Three Sources

Fig. 9. Data association accuracy of the proposed method, using Metric 1, for different values of the history length in anechoic and reverberant conditions.

estimated using the method of [20], [41]. The data-association accuracy for two and three active sound sources in scenarios with reverberation time $T_{60} = 250$, 400, and 600 ms is shown in Fig. 8. It can be observed that, while the association accuracy (Metric 1) decreases with increasing reverberation time, most of the DOAs between pairs of arrays (Metric 2) are still associated correctly both for the two and three sources case. This indicates that while association errors in frames occur more often, most DOA pairs are assigned to the correct source.

In contrast to the results in Section VI-B, the values of C_2 and C_3 now vary in each time frame, as the DOAs of the sound sources are now estimated. Thus, C_2 and C_3 depend on how many sources the DOA estimation method was able to detect at each array at each time frame. To quantify how often missed detections occur, we counted in how many frames each value of C_2 and C_3 occurs. We observed that for the two sources case, approximately in only 12% of the frames all four arrays detected two sources, in 21% of the frames $C_2 = 3$, in 33% of the frames $C_2 = 2$, in 19% of the frames $C_2 = 1$, and in 15% of the frames all arrays detected only one source, i.e., $C_2 = 0$. The problem of missed detections becomes even more evident in the three sources case where in approximately 62% of the frames none of the arrays detected three sources ($C_3 = 0$), in 34% of the frames only one array detected three sources ($C_3 = 1$), and in only 5% of the frames the value of C_3 is greater than one. These numbers not only reveal once again the robustness of our method to missed detections, but also highlight the importance of the association method to take into account missed detections, as they occur very often in practice.

Moreover, in Fig. 9 we demonstrate how the history length B can affect the association accuracy. We consider two and three active sound sources in anechoic conditions and in a scenario with reverberation time $T_{60} = 600$ ms, and we plot the association accuracy (using Metric 1), versus SNR for history lengths of 0.1, 0.25, and 0.5 seconds. In the two sources case, the performance is improved when increasing the history length from 0.1 seconds to 0.25 and 0.5 seconds, especially in the reverberant scenario. However, the performance when using a history length of 0.25 and 0.5 seconds is very similar. In general, there is an obvious performance improvement—especially for the three sources case—as the history length increases, both in anechoic and reverberant conditions, showing that increasing the history length makes the association features more robust to noise and reverberation. However, increasing the history length also increases the latency of the system, in turn decreasing responsiveness.

Finally, we evaluate the ability of our proposed association algorithm to find the optimal solution, according to the problem we defined in Section IV-B. As a greedy algorithm, our proposed data-association algorithm is not guaranteed to find the optimal solution to the problem defined by Eq. (5). As described in Section IV-B, the optimal solution can be guaranteed by exhaustively testing all possible assignments and choosing the one with the minimum score according to (5). We call this approach the brute-force version of the association algorithm. This version is impractical as the number of assignments under test is prohibitively large when the number of sources and nodes increases.

To demonstrate the ability of our greedy association algorithm to solve (5), we compare the solutions it provides with the optimal solution derived by the brute force version of the algorithm where all possible assignments are examined. Also, in order to quantify the performance gain of each of the two steps of our data-association algorithm, we include in the comparison another version of our algorithm where only the first greedy step is performed. The comparison was made on scenarios of two and three active sound sources, on all 30 source configurations, different SNR and reverberation conditions (SNR ranging from 0 dB to 20 dB with a step of 5 dB, and reverberation times of $T_{60} = 250, 400, 600$ ms). We observed that our greedy algorithm was able to find the optimal solution in 97.22% of the time frames for the two sources and in 84.31% of the time frames for the three sources cases. Moreover, even in the cases where the optimal solution cannot be found, the association accuracy results in Fig. 8

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831

TABLE II
MEAN AND STANDARD DEVIATION (IN MILLISECONDS) FOR THE EXECUTION TIME OF THE GREEDY AND BRUTE FORCE VERSION OF OUR DATA-ASSOCIATION ALGORITHM FOR TWO AND THREE ACTIVE SOUND SOURCES

Two Sources

	Greedy	Brute Force
Mean	3.37 ms	1.93 ms
Std	0.69 ms	0.43 ms

Three Sources

	Greedy	Brute Force
Mean	7.58 ms	181.90 ms
Std	2.38 ms	28.97 ms

reveal that our greedy algorithm can still find good solutions. In contrast, the version of our algorithm that performs only the first greedy step was able to find the optimal solution in only 66.43% of the frames for the two sources case and in only 28.52% of the time frames for the three sources case. Thus, a significant performance gain is achieved due to the second greedy step of the proposed data-association algorithm.

To quantify the performance gain of our greedy association algorithm in terms of computation time, Table II compares the mean execution times of the greedy and brute force versions. The execution times were measured in MATLAB on a Windows desktop PC with a Core i7 CPU running at 3.4 GHz with 16 GB RAM. Note that while the absolute execution times may be highly dependent on the machine and the programming language, we are only interested here in the relative times between the two versions of the algorithm. It can be observed that, while the brute force version is computationally more efficient for the case of two active sound sources, it becomes impractical for the case of three active sound sources due to the high number of possible assignments that it needs to test. On the other hand, the execution time of the greedy version in the three sources case remains in the same order of magnitude compared to the two sources one.

D. Location estimation accuracy

When the correct association of DOAs to the sources is found, the locations of the sound sources can be estimated by applying a single source location estimator to the corresponding DOA associations. In this section, we evaluate the localization accuracy of our proposed approach and compare it with the use of the association features extracted from [38] (denoted as [Swartling [2011]). For comparison, we also include the localization performance of our multiple source grid-based estimator, which we proposed in [9] (denoted as [Griffin 2015]). This estimator infers a location for every possible DOA combination from the arrays and on a second step decides which locations correspond the true sources' locations, using no additional information apart from the DOA estimates.

In order to localize the sound sources using the proposed method and the method of [38], we apply our previously proposed single-source grid-based location estimator [9] on the estimated DOA associations. To measure the localization performance we use the root-mean square error (RMSE) over all sources, all 30 different source configurations and over all frames where each source was detected by at least two arrays, which is a necessary condition to infer a location estimate for all sources. These frames represent the 90% and 63% of all frames under test for the two and three sources case respectively. As the use of the association features of [38] cannot provide a DOA association in some cases (for example when $C_2 = 0$ for the two sources case and $C_3 = 0$ for the three sources case) we consider for this method only the frames where a DOA association can be estimated. These frames represent the 83% and 36% of the total frames under test for the two and three sources case respectively. These numbers highlight again the advantage of our proposed method in terms of its ability to find a DOA association even in scenarios with severe missed detections.

Figures 10 and 11 depict the location error for various reverberation conditions and for scenarios with two and three simultaneously active sound sources. The location error when using the estimated DOAs but assuming that the correct association of DOAs to the sources is known (denoted as Perfect Association) is also included to represent the best-case scenario. As expected the localization performance degrades with increasing reverberation time. The performance of the best-case scenario with perfect associations also degrades as the DOA estimates suffer from larger noise due to the high reverberation conditions. In general, it can be observed that the proposed method always achieves the best localization performance, providing location estimates very close to the best-case especially for the higher SNR values and for both two and three active sound sources. The other two methods always perform worse than the proposed one, and their performance degradation is more evident in the three sources case.

E. Reduction in transmission requirements

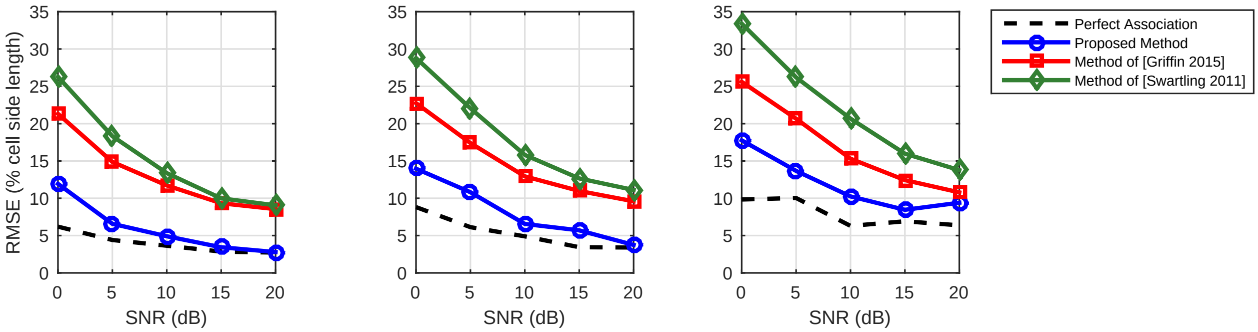
In this section, we evaluate the performance of our proposed decimation process applied to the association features in order to reduce the amount of information that needs to be transmitted by the nodes. We examine the effect of the decimation factor on the data-association and localization accuracy and investigate how much we can reduce the transmitted information without affecting performance.

Fig. 12 depicts the association accuracy using Metric 1 for a scenario of three active sound sources for different decimation factors, namely $d = 1$ (i.e., no decimation), $d = 2$, $d = 4$, $d = 16$, and $d = 32$ and for different reverberation conditions. The corresponding localization error, when the single-source grid-based method [9] is applied on the estimated DOA associations is shown in Fig. 13. It can be observed that the performance for all decimation factors up to $d = 16$ is very similar to the case where no decimation is applied (i.e., $d = 1$). The association and localization performance exhibits higher degradation when a decimation factor of $d = 32$ is used.

To quantify the gain in terms of reduction in information that must be transmitted, we can use Eq. (7) to calculate how many bits are required to transmit an association feature (i.e.,

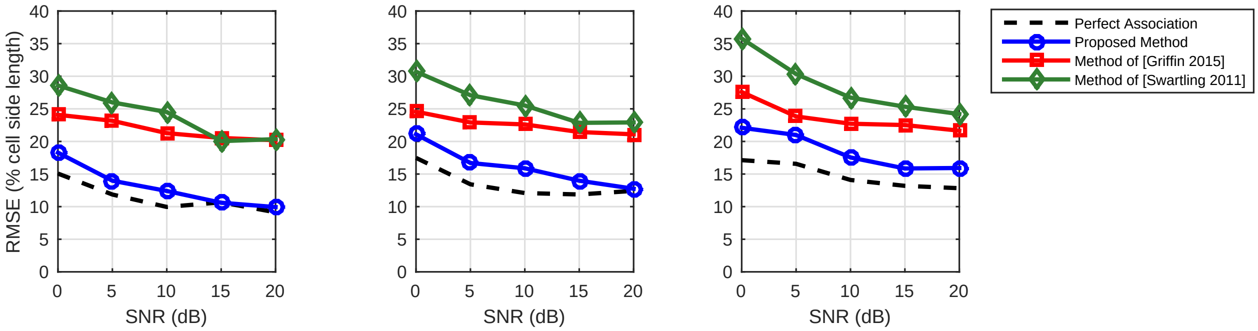
Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831



(a) $T_{60} = 250$ ms (b) $T_{60} = 400$ ms (c) $T_{60} = 600$ ms

Fig. 10. Localization error as a percentage of cell side length $V = 4$ meters for two active sound sources and different reverberation scenarios.



(a) $T_{60} = 250$ ms (b) $T_{60} = 400$ ms (c) $T_{60} = 600$ ms

Fig. 11. Localization error as a percentage of cell side length $V = 4$ meters for three active sound sources and different reverberation scenarios.

histogram) for every value of d: 813, 499, 296, and 97 bits for $d = 1$ (i.e., no decimation), $d = 2$, $d = 4$, and $d = 16$, respectively. These numbers are obtained by substituting into Eq. (7) the block size $B = 21$, and the number of bins $N_\ell = 185$, which results from an FFT size of 2048 samples, a sampling frequency of 44.1 kHz and a maximum frequency for processing $\ell_{\max} = 4$ kHz (see Table I). The results suggest that, by applying a decimation process by a factor of $d = 16$ to the estimated association features, we can reduce the amount of information that is required to transmit a histogram by almost 88%, with minor losses in the association and location estimation performance.

Finally, for the tested scenario of three simultaneously active sound sources, Table III depicts the worst-case and average bitrate requirements for a node. The worst-case corresponds to the case where all arrays detected all sources, thus requiring to transmit the maximum possible number of association features, which in the case of three sources is three. However, in a realistic situation missed detections will occur and thus the nodes will rarely need to transmit three association features. This corresponds to the average case in Table III which depicts the average bitrate over all 30 different tested source configurations, all SNRs, and reverberation times.

F. Moving sources

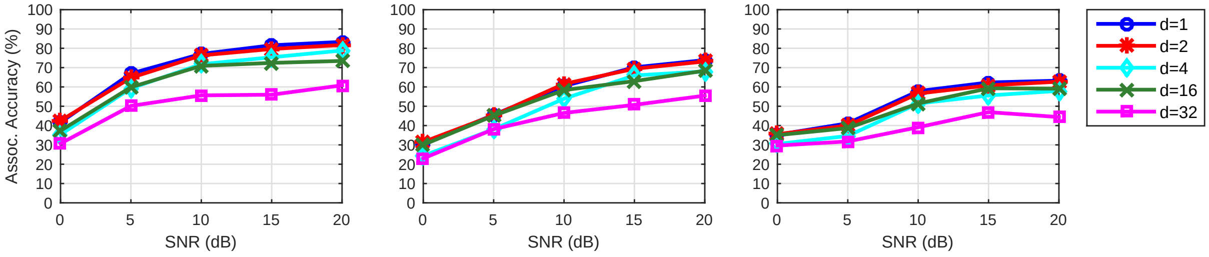
Finally, we demonstrate our method's ability to perform data-association and accurate location estimation in scenarios with moving sources. Fig. 14 depicts the location estimates for all time frames, when the single-source grid-based method [9] is applied to the estimated DOA associations for a simulated scenario with one moving and one static source (*Moving1*) at $T_{60} = 400$ ms reverberation time. The sources were 5 seconds in duration. The WASN setup and simulation parameters were the same as shown in Table I. A decimation process by a factor of $d = 16$ was applied to the association features to reduce bitrate needs. The static source was located at $(5, 4)$ meters and the moving one starts from point $(6.5, 6)$ meters and moves on a straight line to point $(3.5, 6)$ meters. Finally, Fig. 15 depicts the location estimates for a scenario of two moving sound sources (*Moving2*) at $T_{60} = 400$ ms reverberation time. The first source starts again from point $(6.5, 6)$ meters and moves on a straight line to point $(3.5, 6)$ meters, while the second one starts from point $(3.5, 4)$ meters and moves on a straight line to point $(6.5, 4)$ meters. It can be observed that in both cases,

TABLE III
WORST-CASE AND AVERAGE TRANSMISSION REQUIREMENTS, (IN KBITS PER SECOND) FOR A NODE FOR THREE ACTIVE SOURCES.

	Worst-case bitrate	Average bitrate
no decimation	105 Kbps	63 Kbps
$d = 2$	64 Kbps	39 Kbps
$d = 4$	38 Kbps	23 Kbps
$d = 16$	12 Kbps	8 Kbps

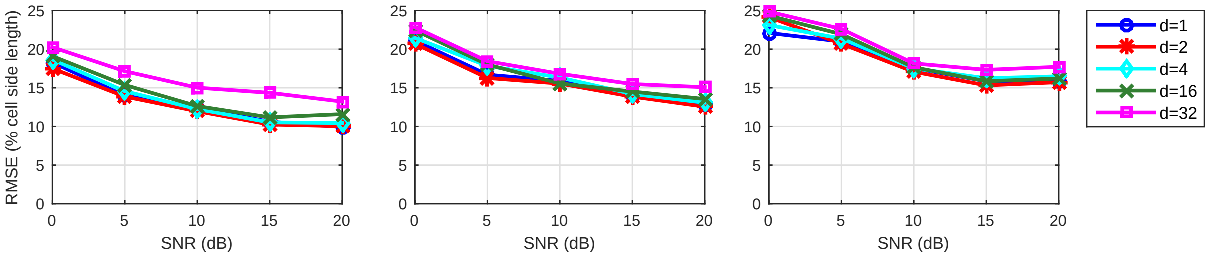
Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831



(a) $T_{60} = 250$ ms (b) $T_{60} = 400$ ms (c) $T_{60} = 600$ ms

Fig. 12. Association accuracy using Metric 1 for three active sound sources, different reverberation conditions, and different values of the decimation factor in the association features.



(a) $T_{60} = 250$ ms (b) $T_{60} = 400$ ms (c) $T_{60} = 600$ ms

Fig. 13. Localization error as a percentage of cell side length $V = 4$ meters for three active sound sources, different reverberation conditions, and different values of the decimation factor in the association features.

the method produces accurate and smooth location estimates, indicating its ability to localize moving sources.

To evaluate the effect of the history length in the case of moving sources, Table IV presents the location error and the DOA association accuracy, using Metric 1, for the two aforementioned scenarios (*Moving1* and *Moving2*) for different values of the history length. The results—which are in accordance with the ones in Fig. 9 for the two sources case—show that the 0.1 second history exhibits the worst performance, while the performance for the 0.25 and 0.5 seconds history is very similar, with the 0.25 seconds history exhibiting slightly higher DOA association accuracy but also slightly higher location error.

TABLE IV
LOCALIZATION ERROR AS A PERCENTAGE OF THE CELL SIDE LENGTH $V =$ 4 METERS AND DATA-ASSOCIATION ACCURACY USING METRIC 1, FOR MOVING SOURCES FOR DIFFERENT VALUES OF THE HISTORY LENGTH.

	Moving1		*Moving2*	
	Metric 1	RMSE	Metric 1	RMSE
100 ms history	84%	7.02%	82%	9.80%
250 ms history	88%	5.17%	86%	8.25%
500 ms history	87%	5.16%	84%	7.94%

Finally, we evaluate the effect of the moving source's velocity to the data-association and localization accuracy. We consider a scenario of one static source at $(5, 4)$ meters and one moving source, at $T_{60} = 400$ ms reverberation time. The moving source moves on a straight line from point $(6.5, 6)$ to point $(3.5, 6)$ and then back to point $(6.5, 6)$. We simulated different velocities for the moving source, from slow to fast walking speeds, namely $v = 2$ km/h, $v = 4$ km/h, and $v =$ 6 km/h and applied the proposed methodology to estimate the association of DOAs from the arrays to the sources and the final sources' locations. The WASN setup and simulation parameters were the same as shown in Table I and a decimation process by a factor of $d = 16$ was applied. The experiment was repeated 10 times. Fig. 16 shows the data-association accuracy using Metric 1 and the location estimation error for various SNR levels and various values for the history length. As expected, the performance degrades when the velocity of the source increases. Again, a history length of 0.5 seconds exhibits the best performance, while a history length of 0.1 seconds exhibits the worst performance. As in the previous experiments with two active sound sources, the performance when using a history length of 0.25 seconds is very similar to the one when using a history length of 0.5 seconds. It is generally evident that the speed of a moving source does not affect the choice of the history length. This can be explained by the fact that during the design of the association features, the narrowband DOA estimates in the B previous frames are compared to the broadband DOA estimates of the sources in the current frame. As a result, if the DOA of the source has significantly changed in the last B frames, the distance between the narrowband DOA and the broadband DOAs of the sources will be large and the corresponding frequencies will not be taken into account due to Eq. (2). Finally, it can be observed that the method provides satisfactory performance even in the case where the moving source moves as fast as 6 km/h, validating again its ability to localize moving sources.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831

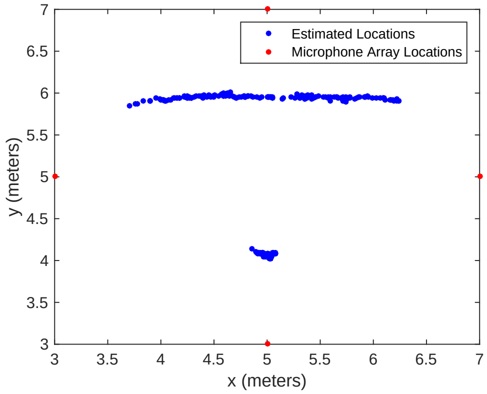


Fig. 14. Location estimates of the proposed method for all time frames for a scenario with one moving and one static sound source at $T_{60} = 400$ ms reverberation time.

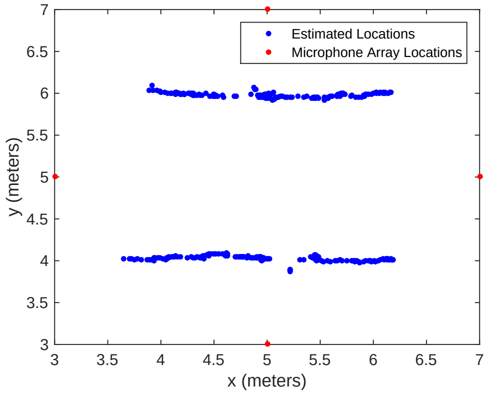
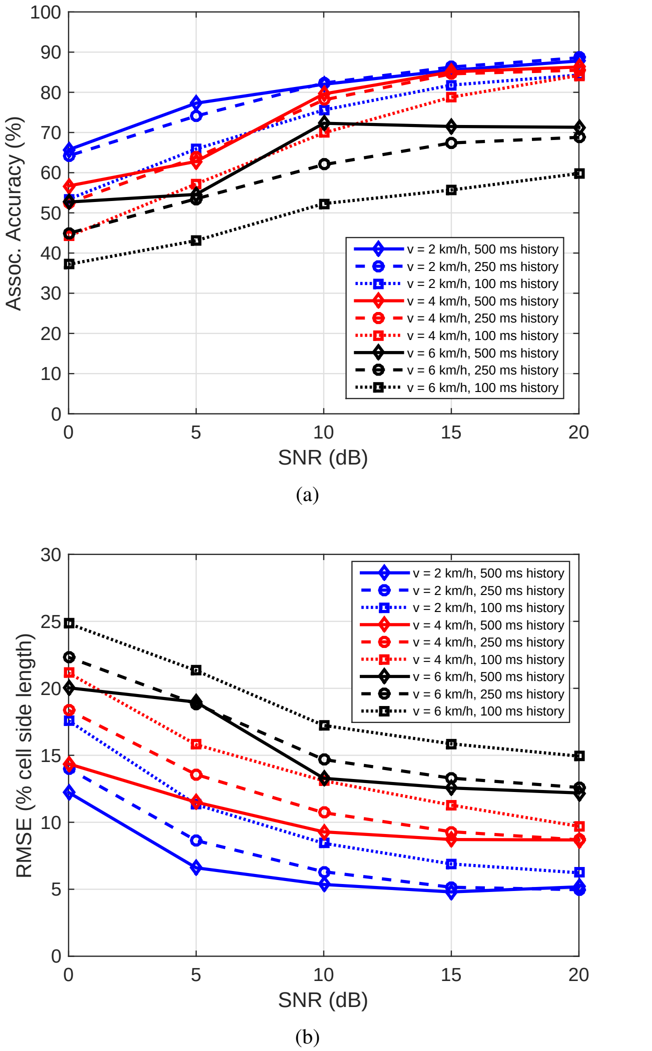


Fig. 15. Location estimates of the proposed method for all time frames for a scenario with two moving sound sources at $T_{60} = 400$ ms reverberation time.

VII. Conclusions

In this work, we considered the data-association problem for the localization of multiple sound sources using DOA estimates in a wireless acoustic sensor network where each node is a microphone array. We presented an approach that can find the correct association of DOAs from the microphone arrays to the sources with high accuracy, thus decomposing the multiple source localization problem into multiple single-source localization problems that can be easily solved using a variety of DOA-based location estimators available in the literature.

Our proposed approach utilizes additional information—apart from the DOA estimates—that consists of association features that describe how the frequency components of the captured signals are distributed to the sources. Using simulations and comparisons with other state-of-the-art methods we confirmed the efficiency of our method to accurately solve the data-association and localization problem in realistic scenarios with missed detections, reverberation, noise, and moving sources. Finally, to account for the practical limitations in the amount of information that needs to be transmitted in the network, we incorporated our method with a scheme that can reduce its bitrate needs of up to 88% without affecting its performance.



(a)

(b)

Fig. 16. (a) Data-association accuracy using Metric 1 and (b) localization error as a percentage of the cell side length $V = 4$ meters for a scenario of one static and one moving source with different velocities v and different values for the history length.

Appendix

Computation of the Angular Distance

The angular distance function returns the distance between angles X and Y in the range $[0, \pi]$. A simple way to compute the angular distance is to define the unit vectors:

$$U_x = \begin{bmatrix}\cos(X) & \sin(X)\end{bmatrix}^T, \tag{9}$$

$$U_y = \begin{bmatrix}\cos(Y) & \sin(Y)\end{bmatrix}^T \tag{10}$$

and evaluate the angle between the two vectors. The angular distance is thus given by:

$$A(X, Y) = \text{acos}(U_x^T U_y), \tag{11}$$

where $\text{acos}(\cdot)$ denotes the inverse cosine function.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831

An equivalent way to compute the angular distance, that does not involve the evaluation of trigonometric functions, is by defining the distances:

$$A_{X,Y} = (X - Y)\text{mod}(2\pi), \tag{12}$$

$$A_{Y,X} = (Y - X)\text{mod}(2\pi) \tag{13}$$

which are in the range of $[0, 2\pi]$ and taking the minimum one:

$$A(X, Y) = \min(A_{X,Y}, A_{Y,X}). \tag{14}$$

Acknowledgment

The authors would like to thank Giorgos Borboudakis for all the fruitful discussions and his important input during the development of some parts of the proposed method.

References

[1] D. J. Mennill, M. Battiston, D. R. Wilson, J. R. Foote, and S. M. Doucet, “Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour,” *Methods in Ecology and Evolution*, vol. 3, no. 4, pp. 704–712, 2012.

[2] H. Wang, C. E. Chen, A. Ali, S. Asgari, R. E. Hudson, K. Yao, D. Estrin, and C. Taylor, “Acoustic sensor networks for woodpecker localization,” in *SPIE Conf. on Advanced Signal Processing Algorithms, Architectures, and Implementations*, F. T. Luk, Ed., vol. 5910, 2005, pp. 80–91.

[3] M. Taseska and E. A. P. Habets, “Informed spatial filtering for sound extraction using distributed microphone arrays,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1195–1207, July 2014.

[4] A. Bertrand, “Applications and trends in wireless acoustic sensor networks: a signal processing perspective,” in *Proceedings of the 18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, IBBT, Ghent, Belgium, Nov 2011.

[5] D. Li and Y. H. Hu, “Energy-based collaborative source localization using acoustic microsensor array,” *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 985029, Dec 2003.

[6] X. Sheng and Y.-H. Hu, “Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks,” *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 44–53, Jan 2005.

[7] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, “Acoustic source localization with distributed asynchronous microphone networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. PP, no. 99, 2012.

[8] M. Compagnoni, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, “Localization of acoustic sources through the fitting of propagation cones using multiple independent arrays,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1964 –1975, Sep 2012.

[9] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, “Localizing multiple audio sources in a wireless acoustic sensor network,” *Signal Processing*, vol. 107, pp. 54 – 67, 2015, Special Issue on ad hoc microphone arrays and wireless acoustic sensor networks.

[10] A. Griffin and A. Mouchtaris, “Localizing multiple audio sources from DOA estimates in a wireless acoustic sensor network,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2013, pp. 1–4.

[11] A. Alexandridis and A. Mouchtaris, “Multiple sound source location estimation and counting in a wireless acoustic sensor network,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2015, pp. 1–5.

[12] A. Alexandridis, N. Stefanakis, and A. Mouchtaris, “Towards wireless acoustic sensor networks for location estimation and counting of multiple speakers in real-life conditions,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 6140–6144.

[13] M. Cobos, A. Marti, and J. J. Lopez, “A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling,” *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, 2011.

[14] H. Do and H. Silverman, “A Fast Microphone Array SRP-PHAT Source Location Implementation using Coarse-To-Fine Region Contraction (CFRC),” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2007, pp. 295–298.

[15] P. Aarabi, “The fusion of distributed microphone arrays for sound localization,” *EURASIP Journal of Applied Signal Processing*, vol. 2003, pp. 338–347, Jan 2003.

[16] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, “A survey of sound source localization methods in wireless acoustic sensor networks,” *Wireless Acoustic Sensor Networks and Applications*, 2017.

[17] S. Argentieri and P. Danes, “Broadband variations of the MUSIC high-resolution method for sound source localization in robotics,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS)*, Oct 2007, pp. 2009–2014.

[18] R. Roy and T. Kailath, “ESPRIT-estimation of signal parameters via rotational invariance techniques,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, Jul 1989.

[19] F. Nesta and M. Omologo, “Generalized state coherence transform for multidimensional TDOA estimation of multiple sources,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 246–260, Jan 2012.

[20] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, “Real-time multiple sound source localization and counting using a circular microphone array,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[21] R. G. Stansfield, “Statistical theory of D.F. fixing,” *Journal of the Institute of Electrical Engineers - Part IIIA: Radiocommunication*, vol. 94, no. 15, pp. 762–770, 1947.

[22] S. Nardone, A. Lindgren, and K. Gong, “Fundamental properties and performance of conventional bearings-only target motion analysis,” *IEEE Transactions on Automatic Control*, vol. 29, no. 9, pp. 775–787, Sep 1984.

[23] A. Griffin, A. Alexandridis, D. Pavlidi, and A. Mouchtaris, “Real-time localization of multiple audio sources in a wireless acoustic sensor network,” in *European Signal Processing Conference (EUSIPCO)*, Sep 2014, pp. 306–310.

[24] L. M. Kaplan, Q. Le, and N. Molnar, “Maximum likelihood methods for bearings-only target localization,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, 2001, pp. 3001–3004.

[25] K. Doğançay, “Bearings-only target localization using total least squares,” *Signal Processing*, vol. 85, no. 9, pp. 1695–1710, 2005.

[26] L. M. Kaplan and Q. Le, “On exploiting propagation delays for passive target localization using bearings-only measurements,” *Journal of the Franklin Institute*, vol. 342, no. 2, pp. 193–211, 2005.

[27] A. Bishop, B. D. O. Anderson, B. Fidan, P. Pathirana, and G. Mao, “Bearing-only localization using geometrically constrained optimization,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 45, no. 1, pp. 308–320, 2009.

[28] Z. Wang, J. Luo, and X. Zhang, “A novel location-penalized maximum likelihood estimator for bearing-only target localization,” *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6166–6181, 2012.

[29] K. Doğançay, “Passive emitter localization using weighted instrumental variables,” *Signal Processing*, vol. 84, no. 3, pp. 487 – 497, 2004.

[30] ——, “Bias compensation for the bearings-only pseudolinear target track estimator,” *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 59–68, Jan 2006.

[31] ——, “Reducing the bias of a bearings-only TLS target location estimator through geometry translations,” in *European Signal Processing Conference (EUSIPCO)*, Sep 2004, pp. 1123–1126.

[32] A. Alexandridis, G. Borboudakis, and A. Mouchtaris, “Addressing the data-association problem for multiple sound source localization using DOA estimates,” in *European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 1551–1555. Best student paper award.

[33] L. M. Kaplan, P. Molnar, and Q. Le, “Bearings-only target localization for an acoustical unattended ground sensor network,” in *Proc. SPIE*, vol. 4393, 2001, pp. 40–51.

[34] K. Pattipati, S. Deb, Y. Bar-Shalom, and R. B. Washburn, “A new relaxation algorithm and passive sensor data association,” *IEEE Transactions on Automatic Control*, vol. 37, no. 2, pp. 198–213, Feb 1992.

[35] S. Deb, M. Yeddanapudi, K. Pattipati, and Y. Bar-Shalom, “A generalized S-D assignment algorithm for multisensor-multitarget state estimation,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, no. 2, pp. 523–538, Apr 1997.

[36] R. Popp, K. Pattipati, and Y. Bar-Shalom, “m-Best S-D assignment algorithm with application to multitarget tracking,” *IEEE Transactions*

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2017.2772831

on Aerospace and Electronic Systems, vol. 37, no. 1, pp. 22–39, Jan 2001.

[37] J. Reed, C. da Silva, and R. Buehrer, "Multiple-source localization using line-of-bearing measurements: Approaches to the data association problem," in *IEEE Military Communications Conf. (MILCOM)*, Nov 2008, pp. 1–7.

[38] M. Swartling, N. Grbić, and I. Claesson, "Source localization for multiple speech sources using low complexity non-parametric source separation and clustering," *Signal Processing*, vol. 91, no. 8, pp. 1781–1788, 2011.

[39] S. Schulz and T. Herfet, "On the window-disjoint-orthogonality of speech source in reverberant humanoid scenarios," in *Proc. of DAFx–08*, 2008, pp. 241–248.

[40] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of the European Conference on Speech Communication (EUROSPEECH)*, 2003.

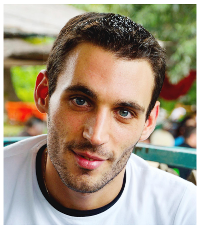
[41] A. Griffin, D. Pavlidi, M. Puigt, and A. Mouchtaris, "Real-time multiple speaker DOA estimation in a circular microphone array based on matching pursuit," in *European Signal Processing Conference (EUSIPCO)*, Aug 2012, pp. 2303–2307.

[42] G. Carpaneto and P. Toth, "Algorithm for the solution of the bottleneck assignment problem," *Computing*, vol. 27, no. 2, pp. 179–187, 1981. [Online]. Available: http://dx.doi.org/10.1007/BF02243552

[43] E. Lehmann and A. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, Aug. 2010.

[44] A. Karbasi and A. Sugiyama, "A new DOA estimation method using a circular microphone array," in *European Signal Processing Conference (EUSIPCO)*, 2007, pp. 778–782.

[45] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*, 4th ed. Academic Press, 2008.



Anastasios Alexandridis received his B.Sc. (2010) and his M.Sc. degree (2013) in Computer Science from the Computer Science Department of the University of Crete, Greece. He is currently pursuing his Ph.D. degree at the Computer Science Department of the University of Crete. Since 2009 he has been affiliated with the Institute of Computer Science at the Foundation for Research and Technology-Hellas (FORTH-ICS) as a research assistant. His research interests include audio and speech signal processing with emphasis on sound localization, wireless acoustic sensor networks, microphone arrays, and spatial and multichannel audio. He was awarded the best paper award in the European Signal Processing Conference (EUSIPCO) in 2015 for his paper "Addressing the data-association problem for multiple sound source localization using DOA estimates". Mr. Alexandridis is a student member of IEEE.

Athanasios Mouchtaris (S'02-M'04) received the Diploma degree in electrical engineering from Aristotle University of Thessaloniki, Greece, in 1997 and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA in 1999 and 2003 respectively. He is currently an Associate Professor in the Computer Science Department of the University of Crete, and an Affiliated Researcher in the Institute of Computer Science of the Foundation for Research and Technology-Hellas (FORTH-ICS), Heraklion, Crete. From 2003 to 2004 he was a Postdoctoral Researcher in the Electrical and Systems Engineering Department of the University of Pennsylvania, Philadelphia. From 2004 to 2007 he was a Postdoctoral Researcher in FORTH-ICS, and a Visiting Professor in the Computer Science Department of the University of Crete. His research interests include signal processing for audio and speech signals, with emphasis on immersive audio environments, spatial and multichannel audio, sound source localization and microphone arrays, acoustic sensor networks, user-generated content, and voice conversion. He has contributed to more than 100 publications in various journal and conference proceedings in these areas, and is co-inventor in 4 issued and 6 pending US patents. Dr. Mouchtaris is a member of IEEE.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.