Real-Time Multiple Sound Source Localization and Counting Using a Circular Microphone Array

Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, Athanasios Mouchtaris

Real-Time Multiple Sound Source Localization and Counting using a Circular Microphone Array

Despoina Pavlidi, *Student Member, IEEE,* Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris, *Member, IEEE*

***Abstract*—In this work, a multiple sound source localization and counting method is presented, that imposes relaxed sparsity constraints on the source signals. A uniform circular microphone array is used to overcome the ambiguities of linear arrays, however the underlying concepts (sparse component analysis and matching pursuit-based operation on the histogram of estimates) are applicable to any microphone array topology. Our method is based on detecting time-frequency (TF) zones where one source is dominant over the others. Using appropriately selected TF components in these "single-source" zones, the proposed method jointly estimates the number of active sources and their corresponding directions of arrival (DOAs) by applying a matching pursuit-based approach to the histogram of DOA estimates. The method is shown to have excellent performance for DOA estimation and source counting, and to be highly suitable for real-time applications due to its low complexity. Through simulations (in various signal-to-noise ratio conditions and reverberant environments) and real environment experiments, we indicate that our method outperforms other state-of-the-art DOA and source counting methods in terms of accuracy, while being significantly more efficient in terms of computational complexity.**

***Index Terms*—direction of arrival estimation, matching pursuit, microphone array signal processing, multiple source localization, real-time localization, source counting, sparse component analysis**

EDICS: AUD-LMAP:Loudspeaker and Microphone Array Signal Processing

I. Introduction

DIRECTION of arrival (DOA) estimation of audio sources is a natural area of research for array signal processing, and one that has had a lot of interest over recent decades [1]. Accurate estimation of the DOA of an audio source is a key element in many applications. One of the most common is in teleconferencing, where the knowledge of the location of a speaker can be used to steer a camera, or to enhance the capture of the desired source with beamforming, thus avoiding the need for lapel microphones. Other applications include event detection and tracking, robot movement in an unknown environment, and next generation hearing aids [2]–[5].

The focus in the early years of research in the field of DOA estimation was mainly on scenarios where a single audio source was active. Most of the proposed methods were based on the time difference of arrival (TDOA) at different microphone pairs, with the Generalized Cross-Correlation PHAse Transform (GCC-PHAT) being the most popular [6]. Improvements to the TDOA estimation problem—where both the multipath and the so-far unexploited information among multiple microphone pairs were taken into account—were proposed in [7]. An overview of TDOA estimation techniques can be found in [8].

Localizing multiple, simultaneously active sources is a more difficult problem. Indeed, even the smallest overlap of sources—caused by a brief interjection, for example—can disrupt the localization of the original source. A system that is designed to handle the localization of multiple sources sees the interjection as another source that can be simultaneously captured or rejected as desired. An extension to the GCC-PHAT algorithm was proposed in [9] that considers the second peak as an indicator of the DOA of a possible second source. One the first methods capable of estimating DOAs of multiple sources is the well-known MUSIC algorithm and its wide-band variations [2], [10]–[14]. MUSIC belongs to the classic family of subspace approaches, which depend on the eigen-decomposition of the covariance matrix of the observation vectors.

Derived as a solution to the Blind Source Separation (BSS) problem, Independent Component Analysis (ICA) methods achieve source separation—enabling multiple source localization—by minimizing some dependency measure between the estimated source signals [15]–[17]. The work of [18] proposed performing ICA in regions of the time-frequency representation of the observation signals under the assumption that the number of dominant sources did not exceed the number of microphones in each time-frequency region. This last approach is similar in philosophy to Sparse Component Analysis (SCA) methods [19, ch. 10]. These methods assume that one source is dominant over the others in some time-frequency windows or "zones". Using this assumption, the multiple source propagation estimation problem may be rewritten as a single-source one in these windows or zones, and the above methods estimate a mixing/propagation matrix, and then try to recover the sources. By estimating this mixing matrix and knowing the geometry of the microphone array, we may localize the sources, as proposed in [20]–[22], for

D. Pavlidi, A. Griffin, and A. Mouchtaris are with the Foundation for Research and Technology-Hellas, Institute of Computer Science (FORTH-ICS), Heraklion, Crete, Greece, GR-70013 e-mail: {pavlidi, agriffin, mouchtar}@ics.forth.gr.

D. Pavlidi and A. Mouchtaris are also with the University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-71409.

M. Puigt is with the Université Lille Nord de France, ULCO, LISIC, Calais, France, FR-62228 e-mail: matthieu.puigt@lisic.univ-littoral.fr. This work was performed when M. Puigt was with FORTH-ICS.

example. Most of the SCA approaches require the sources to be W-disjoint orthogonal (WDO) [23]—meaning that in each time-frequency component, at most one source is active—which is approximately satisfied by speech in anechoic environments, but not in reverberant conditions. On the contrary, other methods assume that the sources may overlap in the time-frequency domain, except in some tiny "time-frequency analysis zones" where only one of them is active (e.g., [19, p. 395], [24]). Unfortunately, most of the SCA methods and their DOA extensions are computationally intensive and therefore off-line methods (e.g., [21] and the references within). The work of [20] is a frame-based method, but requires WDO sources.

Other than accurate and efficient DOA estimation, an extremely important issue in sound source localization is estimating the number of active sources at each time instant, known as source counting. Many methods in the literature propose estimating the intrinsic dimension of the recorded data, i.e., for an acoustic problem, they perform source counting at each time instant. Most of them are based on information theoretic criteria (see [25] and the references within). In other methods, the estimation of the number of sources is derived from a large set of DOA estimates that need to be clustered. In classification, some approaches to estimating both the clusters and their number have been proposed (e.g. [26]), while several solutions specially dedicated to DOAs have been tackled in [19, p. 388], [27] and [28].

In this work, we present a novel method for multiple sound source localization using a circular microphone array. The method belongs in the family of SCA approaches, but it is of low computational complexity, it can operate in real-time and imposes relaxed sparsity constraints on the source signals compared to WDO. The methodology is not specific to the geometry of the array, and is based on the following steps: (a) finding single-source zones in the time-frequency domain [24] (i.e., zones where one source is clearly dominant over the others); (b) performing single-source DOA estimation on these zones using the method of [29]; (c) collecting these DOA estimations into a histogram to enable the localization of the multiple sources; and (d) jointly performing multiple DOA estimation and source counting through the post-processing of the histogram using a method based on matching pursuit [30]. Parts of this work have been recently presented in [22], [31], [32]. This current work presents a more detailed and improved methodology compared to our recently published results, especially in the following respects: (i) we provide a way of combining the tasks of source counting and DOA estimation using matching pursuit in a natural and efficient manner; and (ii) we provide a thorough performance investigation of our proposed approach in numerous simulation and real-environment scenarios, both for the DOA estimation and the source counting tasks. Among these results, we provide performance comparisons of our algorithm regarding the DOA estimation and the source counting performance with the main relevant state-of-art approaches mentioned earlier. More specifically, DOA estimation performance is compared to WDO-based, MUSIC-based, and frequency domain ICA-based DOA estimation methods, and source counting performance is compared to an information-theoretic method. Overall, we show that our proposed method is accurate, robust and of low computational complexity.

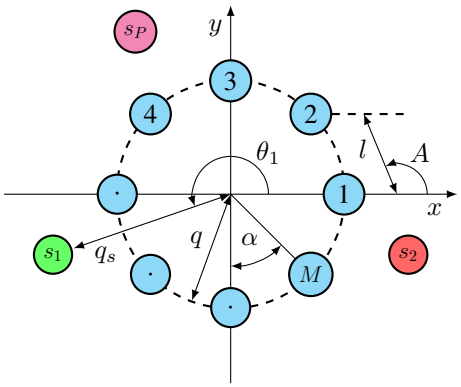

Fig. 1. Circular sensor array configuration. The microphones are numbered 1 to M and the sound sources are s_1 to s_P.

The remainder of the paper then reads as follows. We describe the considered localization and source counting problem in Section II. We then present our proposed method for joint DOA estimation and counting in Section III. In this section we also discuss additional proposed methods for source counting. We revise alternative methods for DOA estimation in Section IV. Section V provides an experimental validation of our approaches along with discussion on performance and complexity issues. Finally, we conclude in Section VI.

II. Problem Statement

We consider a uniform circular array of M microphones, with P active sound sources located in the far–field of the microphone array. Assuming the free-field model, the signal received at each microphone m_i is

$$x_i(t) = \sum_{g=1}^{P} a_{ig} s_g\left(t - t_i(\theta_g)\right) + n_i(t), \qquad i = 1, \cdots, M, \tag{1}$$

where s_g is one of the P sound sources at distance q_s from the centre of the microphone array, a_{ig} is the attenuation factor and $t_i(\theta_g)$ is the propagation delay from the g^{th} source to the i^{th} microphone. θ_g is the DOA of the source s_g observed with respect to the x-axis (Fig. 1), and $n_i(t)$ is an additive white Gaussian noise signal at microphone m_i that is uncorrelated with the source signals $s_g(t)$ and all other noise signals.

For one given source, the relative delay between signals received at adjacent microphones—hereafter referred to as microphone pair $\{m_i m_{i+1}\}$, with the last pair being $\{m_M m_1\}$—is given by [29]

$$\begin{aligned}\tau_{m_i m_{i+1}}(\theta_g) &\triangleq t_i(\theta_g) - t_{i+1}(\theta_g) \\ &= l\sin(A + \frac{\pi}{2} - \theta_g + (i-1)\alpha)/c,\end{aligned} \tag{2}$$

where α and l are the angle and distance between $\{m_i m_{i+1}\}$ respectively, A is the obtuse angle formed by the chord $m_1 m_2$ and the x-axis, and c is the speed of sound. Since the microphone array is uniform, α, A and l are given by:

$$\alpha = \frac{2\pi}{M}, \qquad A = \frac{\pi}{2} + \frac{\alpha}{2}, \qquad l = 2q\sin\frac{\alpha}{2}, \tag{3}$$

where q is the array radius. We note here that in (2) the DOA θ_g is observed with respect to the x-axis, while in [29] it is observed with respect to a line perpendicular to the chord defined by the microphone pair $\{m_1 m_2\}$. We also note that all angles in (2) and (3) are in radians.

We aim to estimate the number of the active sound sources, P and corresponding DOAs θ_g by processing the mixtures of source signals, x_i, and taking into account the known array geometry. It should be noted that even though we assume the free-field model, our method is shown to work robustly in both simulated and real reverberant environments.

III. Proposed Method

A. Definitions and assumptions

We follow the framework of [24] that we recall here for the sake of clarity. We partition the incoming data in overlapping time frames on which we compute a Fourier transform, providing a time-frequency (TF) representation of observations. We then define a "constant-time analysis zone", (t, Ω), as a series of frequency-adjacent TF points (t, ω). A "constant-time analysis zone", (t, Ω) is thus referred to a specific time frame t and is comprised by Ω adjacent frequency components. In the remainder of the paper, we omit t in the (t, Ω) for simplicity.

We assume the existence, for each source, of (at least) one constant-time analysis zone—said to be "single-source"—where one source is "isolated", i.e., it is dominant over the others. This assumption is much weaker than the WDO assumption [23] since sources can overlap in the TF domain except in these few single-source analysis zones. Our system performs DOA estimation and source counting assuming there is always at least one active source. This assumption is only needed for theoretical reasons and can be removed in practice, as shown in [33] for example. Additionally, any recent voice activity detection (VAD) algorithm could be used as a prior block to our system.

The core stages of the proposed method are:

1) The application of a joint-sparsifying transform to the observations, using the above TF transform.
2) The single-source constant-time analysis zones detection (Section III-B).
3) The DOA estimation in the single-source zones (Section III-C).
4) The generation and smoothing of the histogram of a block of DOA estimates (Section III-D).
5) The joint estimation of the number of active sources and the corresponding DOAs with matching pursuit (Section III-E).

B. Single-source analysis zones detection

For any pair of signals (x_i, x_j), we define the cross-correlation of the magnitude of the TF transform over an analysis zone as:

$$R'_{i,j}(\Omega) = \sum_{\omega \in \Omega} |X_i(\omega) \cdot X_j(\omega)|. \tag{4}$$

We then derive the correlation coefficient, associated with the pair (x_i, x_j), as:

$$r'_{i,j}(\Omega) = \frac{R'_{i,j}(\Omega)}{\sqrt{R'_{i,i}(\Omega) \cdot R'_{j,j}(\Omega)}}. \tag{5}$$

Our approach for detecting single-source analysis zones is based on the following theorem [24]:

Theorem 1: A necessary and sufficient condition for a source to be isolated in an analysis zone (Ω) is

$$r'_{i,j}(\Omega) = 1, \qquad \forall i, j \in \{1, \ldots, M\}. \tag{6}$$

We detect all constant-time analysis zones that satisfy the following inequality as single-source analysis zones:

$$\overline{r'}(\Omega) \geq 1 - \epsilon, \tag{7}$$

where $\overline{r'}(\Omega)$ is the average correlation coefficient between pairs of observations of adjacent microphones and ϵ is a small user-defined threshold.

C. DOA estimation in a single-source zone

Since we have detected all single-source constant time analysis zones, we can apply any known single source DOA algorithm over these zones. We propose a modified version of the algorithm in [29] and we choose this algorithm because it is computationally efficient and robust in noisy and reverberant environments [22], [29].

We consider the circular array geometry (Fig. 1) introduced in Section II. The phase of the cross-power spectrum of a microphone pair is evaluated over the frequency range of a single-source zone as:

$$G_{m_i m_{i+1}}(\omega) = \angle R_{i,i+1}(\omega) = \frac{R_{i,i+1}(\omega)}{|R_{i,i+1}(\omega)|}, \qquad \omega \in \Omega, \tag{8}$$

where the cross-power spectrum is

$$R_{i,i+1}(\omega) = X_i(\omega) \cdot X_{i+1}(\omega)^* \tag{9}$$

and * stands for complex conjugate.

We then calculate the Phase Rotation Factors [29],

$$G^{(\omega)}_{m_i \to m_1}(\phi) \triangleq e^{-j\omega\tau_{m_i \to m_1}(\phi)}, \tag{10}$$

where $\tau_{m_i \to m_1}(\phi) \triangleq \tau_{m_1 m_2}(\phi) - \tau_{m_i m_{i+1}}(\phi)$ is the difference in the relative delay between the signals received at pairs $\{m_1 m_2\}$ and $\{m_i m_{i+1}\}$, $\tau_{m_i m_{i+1}}(\phi)$ is evaluated according to (2), $\phi \in [0, 2\pi)$ in radians, and $\omega \in \Omega$.

We proceed with the estimation of the Circular Integrated Cross Spectrum (CICS), defined in [29] as

$$\text{CICS}^{(\omega)}(\phi) \triangleq \sum_{i=1}^{M} G^{(\omega)}_{m_i \to m_1}(\phi) G_{m_i m_{i+1}}(\omega). \tag{11}$$

The DOA associated with the frequency component ω in the single-source zone with frequency range Ω is estimated as,

$$\hat{\theta}_\omega = \arg \max_{0 \leq \phi < 2\pi} |\text{CICS}^{(\omega)}(\phi)|. \tag{12}$$

In each single-source zone we focus only on "strong" frequency components in order to improve the accuracy of

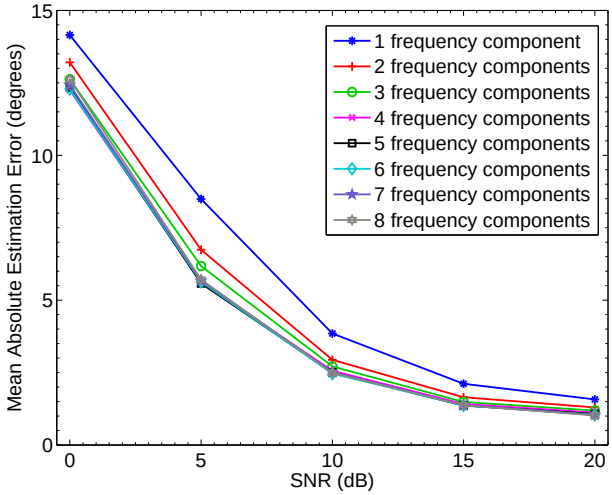

Fig. 2. DOA estimation error vs SNR in a simulated environment. Each curve corresponds to a different number of frequency components used in a single-source zone.

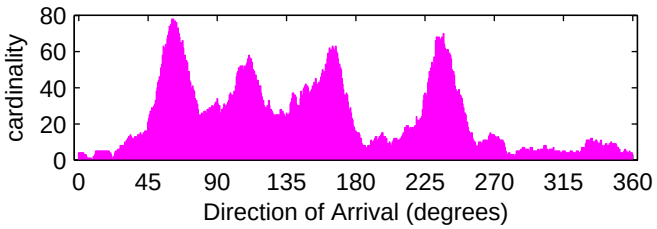

Fig. 3. Example of a smoothed histogram of four sources (speakers) in a simulated reverberant environment at 20 dB SNR.

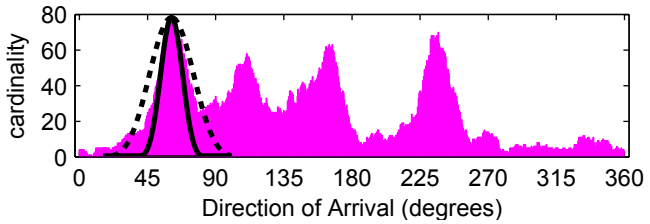

Fig. 4. A wide source atom (dashed line) and a narrow source atom (solid line) applied on the smoothed histogram of four sources (speakers).

the DOA estimation. In our previous work [22], [31], [32], we used only the $\omega_i^{\max}$ frequency, corresponding to the strongest component of the cross-power spectrum of the microphone pair $\{m_i m_{i+1}\}$ in a single-source zone, giving us a single DOA for each single-source zone. In this work we propose the use of d frequency components in each single-source zone, i.e., the use of those frequencies that correspond to the indices of the d highest peaks of the magnitude of the cross-power spectrum over all microphone pairs. This way we get d estimated DOAs from each single-source zone, improving the accuracy of the overall system.

This is illustrated in Fig. 2, where we plot the DOA estimation error versus signal to noise ratio (SNR) for various choices of d. It is clear that using more frequency bins (the terms frequency bin and frequency component are used interchangeably) leads in general to a lower estimation error. We have to keep in mind, though, that our aim is a real-time system, and increasing d increases the computational complexity.

D. Improved block-based decision

In the previous sections we described how we determine whether a constant time analysis zone is single-source and how we estimate the DOAs associated with the d strongest frequency components in a single-source zone. Once we have estimated all the local DOAs in the single-source zones (Sections III-B & III-C), a natural approach is to form a histogram from the set of estimations in a block of B consecutive time frames. Additionally, any erroneous estimates of low cardinality, due to noise and/or reverberation do not severely affect the final decision since they only add a noise floor to the histogram. We smooth the histogram by applying an averaging filter with a window of length h_N. If we denote each bin of the smoothed histogram as v, its cardinality, $y(v)$, is given by:

$$y(v) = \sum_{i=1}^{N} w\left(\frac{v \times 360^\circ/L - \zeta_i}{h_N}\right), \qquad 1 \le v \le L, \quad (13)$$

where L is the number of bins in the histogram, ζ_i is the i^{th} estimate (in degrees) out of N estimates in a block, and $w(\cdot)$ is the rectangular window of length h_N. An example of a smoothed histogram of four sources at $60^\circ, 105^\circ, 165^\circ$, and 240° at 20 dB SNR of additive white Gaussian noise is shown in Fig. 3.

E. DOA Estimation and Counting of Multiple Sources with Matching Pursuit

In each time frame we form a smoothed histogram from the estimates of the current frame and the $B-1$ previous frames. Once we have the histogram in the n^{th} time frame (the length-L vector, $\boldsymbol{y}_n$), our goal is to count the number of active sources and to estimate their DOAs. In our previous work, [31], [32] we performed these tasks separately, but here we combine them into a single process.

Let us go back to the example histogram of four active sources at 20 dB SNR, shown in Fig. 3. The four sources are clearly visible and similarly shaped, which inspired us to approach the source counting and DOA estimation problem as one of sparse approximation using source atoms. Thus the idea—proceeding along similar lines to matching pursuit—is to find the DOA of a possible source by correlation with a source atom, estimate its contribution and remove it. The process is then repeated until the contribution of a source is insignificant, according to some criteria. This way we can jointly estimate the number of sources and their DOAs.

We chose to model each source atom as a smooth pulse, such as that of a Blackman window, although the choice of the window did not prove to be critical. The choice of the width is key, and reasoning and experiments showed that a high *accuracy* of the method requires wide source atoms at lower SNRs and narrow source atoms at higher SNRs. Furthermore, the *resolution* of the method—the ability to discriminate between two closely spaced sources—is adversely affected as the width of the source atom increases. This suggests making the width a parameter in the estimation process, however this would come at the cost of an increase in computational

complexity—something we wish to avoid—so we chose to use fixed-width source atoms.

Further investigation revealed that a two-width method provided a good compromise between these constraints, where a narrower width is used to accurately pick the location of each peak, but a wider width is used to account for its contribution to the overall histogram and provide better performance at lower SNRs. This dual-width approach is illustrated in Fig. 4. Note that the wider width source mask is centered on the same index as the narrow one.

The correlation of the source pulse with the histogram must be done in a circular manner, as the histogram "wraps" from 359° to 0°. An efficient way to do this is to form a matrix whose rows (or columns) contain wrapped and shifted versions of the source pulse, as we now describe.

Let $\boldsymbol{b}$ be a length-Q row vector containing a length-Q Blackman window, then let $\boldsymbol{u}$ be a length-L row vector whose first Q values are populated with $\boldsymbol{b}$ and then padded with $L-Q$ zeros. Let $\boldsymbol{u}^{(k)}$ denote a version of $\boldsymbol{u}$ that has been "circularly" shifted to the right by k elements, the circular shift means that the elements at either end wrap around, and a negative value of k implies a circular shift to the left.

Choose $Q = 2Q_0 + 1$ where Q_0 is a positive integer. The maximum value of $\boldsymbol{b}$ (or equivalently $\boldsymbol{u}$) will occur at $(Q_0 + 1)$-th position. Define $\boldsymbol{c} = \boldsymbol{u}^{(-Q_0)}$. The maximum value of the length-L row vector $\boldsymbol{c}$ occurs at its first element. Let the elements of $\boldsymbol{c}$ be denoted c_i, and its energy be given by $E_c = \sum c_i^2$. Now form the matrix C, which consists of circularly shifted versions of $\boldsymbol{c}$. Specifically, the k-th row of C is given by $\boldsymbol{c}^{(k-1)}$.

As previously discussed, we need two widths of source atoms, so let $\mathrm{C_N}$ and $\mathrm{C_W}$ be matrices for the peak detection (denoted by "N" for narrow) and the masking operation (denoted by "W" for wide), respectively, with corresponding source atom widths Q_N and Q_W.

In order to estimate the number of active sources, P_n, we create $\boldsymbol{\gamma}$, a length-P_MAX vector whose elements γ_j are some predetermined thresholds, representing the relative energy of the j-th source. Our joint source counting and DOA estimation algorithm then proceeds as follows:

1) Set the loop index $j = 1$
2) Form the product $\boldsymbol{a} = \mathrm{C_N}\boldsymbol{y}_{n,j}$
3) Let the elements of $\boldsymbol{a}$ be given by a_i, find $i^* = \arg\max_i a_i$ such that i^* is further than $u_w \times L/360^\circ$ from all formerly located maximum indices, where u_w denotes a minimum offset between neighbouring sources
4) The DOA of this source is given by $(i^* - 1) \times 360^\circ/L$
5) Calculate the contribution of this source as
$$\delta_j = (\boldsymbol{c}_\mathrm{W}^{(i^*-1)})^T \frac{a_{i^*}}{E_{c_\mathrm{N}}}$$
6) If $\delta_j < \gamma_j$ go to step 10
7) Remove the contribution of this source as
$$\boldsymbol{y}_{n,j+1} = \boldsymbol{y}_{n,j} - \delta_j$$
8) Increment j
9) If $j \leq P_\mathrm{MAX}$ go to step 2
10) $\hat{P}_n = j - 1$ and the corresponding DOAs are those estimated in step 4

It should be noted that this method was developed with the goal of being computationally-efficient so that the source counting and DOA estimation could be done in real-time. By real-time we refer to the response of our system within the strict time constraint defined by the duration of a time frame. It should be clear that $\mathrm{C_N}$ and $\mathrm{C_W}$ are circulant matrices and will contain $L - Q_N$ and $L - Q_W$ zeros on each row, respectively, and both of these properties may be exploited to provide a reduced computational load.

F. Additional proposed source counting methods

In Section III-E we presented a matching pursuit-based method for source counting and described how this method can be combined in a single step with the DOA estimation of the sources. In this section we propose two alternative source counting methods, namely a Peak Search approach and a Linear Predictive Coding (LPC) approach.

1) Peak Search: In order to estimate the number of sources we perform a peak search of the smoothed histogram in the n^{th} frame (see Section III-D) in the following manner:

a) We assume that there is always at least one active source in a block of estimates. So we set $i_s = 1$, where i_s corresponds to a counter of the peaks assigned to sources so far. We also set $u_{i_s} = u_1 = \arg\max y(v)$, i.e., the histogram bin which corresponds to the highest peak of the smoothed histogram. Finally, we set the threshold $z_{i_s+1} = \max\{y(u_{i_s})/2, z_\mathrm{static}\}$, where z_static is a user-defined static threshold.
b) We locate the next highest peak in the smoothed histogram, $y(u_{i_s+1})$. If the following three conditions are simultaneously satisfied:
$$y(u_{i_s+1}) \geq z_{i_s+1} \tag{14}$$
$$u_{i_s+1} \notin \left[u_{j_s} - \frac{u_w L}{360^\circ}, u_{j_s} + \frac{u_w L}{360^\circ}\right], \qquad \forall u_{j_s} \tag{15}$$
$$j_s < (i_s + 1) \tag{16}$$
then $i_s = i_s + 1$ and $z_{i_s+1} = \max\{y(u_{i_s})/2, z_\mathrm{static}\}$. u_w is the minimum offset between neighbouring sources. (14) guaranties that the next located histogram peak is higher than the updated threshold z_{i_s+1}. (15) and (16) guarantee that the next located peak is not in the close neighbourhood of an already located peak with $j_s = 1, \ldots i_s$ and u_{j_s} all the previously identified source peaks.
c) We stop when a peak in the histogram fails to satisfy the threshold z_{i_s+1} or if the upper threshold P_MAX is reached. The estimated number of sources is $\hat{P}_n = i_s$.

We note that peak-search approaches on histograms of estimates have been proposed in literature [27]. Here, we present another perspective on these approaches by processing a smoothed histogram and by using a non-static peak threshold. In Fig. 5 we can see how the Peak Search method is applied

to a smoothed histogram where four sources are active. The black areas indicate the bins around a tracked peak of the histogram that are excluded as candidate source indicators as explained in step b).

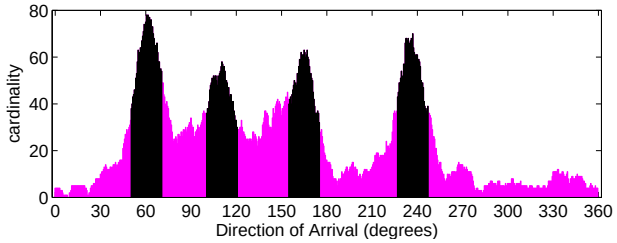

Fig. 5. Peak Search for source counting. The black areas indicate the bins around a tracked peak of the histogram that are excluded as candidate source indicators.

2) Linear Predictive Coding: Linear Predictive Coding (LPC) coefficients are widely used to provide an all-pole smoothed spectral envelope of speech and audio signals [34]. This inspired us to apply LPC to the smoothed histogram of estimates to emphasize the peaks and suppress any noisy areas. Thus, the estimated LPC envelope coincides with the envelope of the histogram. We get our estimate of $\hat{P}_n$ sources by counting the local maxima in the LPC envelope with the constraint that $\hat{P}_n \leq P_{\text{MAX}}$. In our estimation, we exclude peaks that are closer than u_w, as a minimum offset between neighbouring sources.

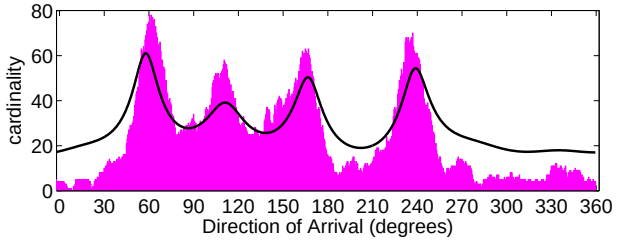

Fig. 6. LPC for source counting. The black curve corresponds to the LPC estimated envelope of the histogram.

A key parameter of this approach is the order of LPC. We want to avoid a very high order that will over-fit our histogram of estimates, in turn leading to an over-estimation of the true number of sources. On the other hand, the use of a very low order risks the detection of less dominant sources (i.e., sources with less estimates in the histogram, thus lower peaks). In order to decide on an optimum LPC order, we tested a wide range of values and chose the one that gave the best results in all our considered simulation scenarios (details can be found in Section V). In Fig. 6 we plot an example LPC envelope with order 16, along with the smoothed histogram.

IV. State of the art methods for DOA estimation

In order to compare our proposed method with other algorithms, we implemented three well-studied methods, a WDO-approach [23], a wideband implementation of MUSIC [2] and the Independent Component Analysis-Generalised State Coherence Transform (ICA-GSCT) algorithm [18]. The WDO-based and the ICA-GSCT approaches were chosen since they originate from the BSS research field as does our proposed method, therefore they are similar in philosophy. The MUSIC algorithm is an extensively studied and tested algorithm for DOA estimation of multiple sources, thus it is also a well suited algorithm for comparative tests. We now provide a brief description of these methods.

A. WDO-based approach

Considering the source signals as W-disjoint orthogonal, the time-frequency representations of the signals are assumed to not overlap. So, if $S_i(t,\omega)$ and $S_j(t,\omega)$ are the TF supports of the signals $s_i(t)$ and $s_j(t)$, according to the W-disjoint orthogonality assumption [23]:

$$S_i(t,\omega)S_j(t,\omega) = 0, \qquad \forall\, t,\omega \tag{17}$$

In that sense at each TF point, (t,ω), at most one source is active and we can apply the method described in Section III-C for all (t,ω). We then form a smoothed histogram of the estimates of B consecutive frames (see Section III-D) and we apply matching pursuit (see Section III-E) to it the same way we did for the proposed method.

B. Broadband MUSIC

The MUSIC algorithm was originally proposed as a localization algorithm for narrowband signals. It is based on the covariance matrix of the observations, C_X. The sorted eigenvalues of C_X define the signal subspace, U_S and the noise subspace, U_N and the DOAs of the sources are derived from the maxima of the narrowband pseudospectrum:

$$h_{\text{narrow}}(\phi) = \frac{1}{V^H(\phi)U_N U_N^H V(\phi)}, \qquad 0 \leq \phi < 2\pi, \tag{18}$$

where $V(\phi) = [e^{-j\omega\tau_1(\phi)}, e^{-j\omega\tau_2(\phi)}, \ldots, e^{-j\omega\tau_M(\phi)}]$ is the steering vector, angle ϕ is in radians, ω is the frequency of the narrowband signals and $\tau_i(\phi)$ is the time difference of arrival of a source emitting from DOA ϕ between the i^{th} microphone and a reference point. Among the various wideband extensions that have appeared in the literature, the most popular one is comprised of estimating the narrow pseudospectrum at each frequency component of the wideband signals and deriving its wideband counterpart as the average over all frequencies [2]:

$$h_{\text{wide}}(\phi) = \frac{1}{N_b}\sum_{b=1}^{N_b} h_{\text{narrow}}(\phi), \tag{19}$$

where N_b is the number of frequency bins. Then, the DOA estimation is performed by looking for $P < M$ maxima in the final average pseudospectrum.

C. ICA-GSCT

The ICA-GSCT method can be divided into two main parts, the estimation of the mixing matrices at each frequency component and the extraction of the DOAs from the estimated mixing matrices. For the first step in our implementation we have used the Joint Approximate Diagonalization of Eigenmatrices (JADE) method [35] which exploits the fourth-order cumulants relying on the statistical independence of the sources. The

code is provided by the authors and can be found in [36], where as input we provide the STFT of the observations of B consecutive time frames. Given the mixing matrices, we then estimate the GSCT [18] which is a multivariate likelihood measure between the acoustic propagation model and the observed propagation vectors, obtained by row-wise ratios between the elements of each mixing matrix. The GSCT is given by:

$$\mathrm{GSCT}(\mathbf{T}) = \sum \mathbf{g}(E(\mathbf{T})), \tag{20}$$

where $\mathbf{T}$ is the model vector of time differences of arrival between adjacent microphones, $E(\mathbf{T})$ is the error measure between the model and the observation vectors and $\mathbf{g}(E(\mathbf{T}))$ is a non-linear monotonic function which decreases as the error measure increases. The summation in (20) takes place over all frequency components and ratios in all the columns of the mixing matrices. For non-linear function $\mathbf{g}(E(\mathbf{T}))$, we use the kernel-based one recommended by the authors of [18]

$$\mathbf{g}(E(\mathbf{T})) = \frac{1}{\omega} e^{-E^2(\mathbf{T})/(2(\omega \frac{2q\sin(\alpha/2)}{cd_K})^2)}, \tag{21}$$

where d_K is a resolution factor.

By associating each time delay vector, $\mathbf{T}$ of the propagation model to its corresponding DOA, we estimate the DOAs of P sources by looking for P local maxima of the GSCT function.

D. Computational Complexity

In order to study the computational complexity of our proposed method for DOA estimation and the above methods, we estimated the total number of operations that each method performs to derive a curve whose local maxima act as DOA indicators. More specifically, we estimated the total number of the following operations: for our proposed method and WDO, to obtain the smoothed version of the histogram of the estimates; for MUSIC, to estimate the average pseudospectrum; and for ICA-GST, to estimate the GSCT-kernel density function at each time instant. By the term "operation", we refer to any multiplication, addition or comparison, as many dedicated processors—such as DSPs—only take one cycle for each of these operations.

We present the results for a scenario with six sources in Table I. Note that for the implementation of the methods we used the same parameter values as the proposed method in order to compare them fairly. The only change was the range of frequencies of interest used for the ICA-GSCT, where instead of using frequencies up to 4000 Hz, we were constrained in the range $300 - 4000$ Hz as recommended in [18], since ICA does not behave well in terms of convergence for frequencies lower than 200 Hz. Furthermore, the resolution factor for the kernel density estimation was set to $d_K = 4$, which gave the best results for the specific simulation set-up (for more details about the parameters and their values see Section V, Table II).

Our proposed method clearly has the lowest computational complexity. MUSIC requires almost one and a half times as many operations, while WDO needs almost three times as many operations. The complexity of ICA-GSCT is much higher than all the other methods. These results were expected, since WDO follows the same procedure as the proposed method, but for all the frequency components whereas we work with d components in single-source zones only. On the other hand, MUSIC performs eigenvalue decomposition for each frequency component and averages the information from all frequency components, contributing significantly to its high complexity. However, we note that there are wideband MUSIC approaches with significantly lower complexity than the one used in this study (e.g., Section IV in [2]). These are mainly based on spherical harmonics beampattern synthesis which is still an open research problem for circular array topologies [37]–[39]. For frequency domain ICA-based methods, the estimation of the demixing matrix at each frequency bin is a cost-demanding operation. Furthermore the estimation of the GSCT function requires averaging over all frequency bins, all sources and all time frames in a block of estimates.

Note that the matching pursuit method applied to the smoothed histogram, as well as the search for maxima in the MUSIC average pseudospectrum and in the ICA-GSCT function, require an insignificant number of operations compared to the overall complexity of the methods.

TABLE I
COMPUTATIONAL COMPLEXITY

Method	number of operations
proposed method	2,638,424
WDO	10,235,565
MUSIC	3,903,280
ICA-GSCT	35,254,348

TABLE II
EXPERIMENTAL PARAMETERS

parameter	notation	value
number of microphones	M	8
sampling frequency	f_s	44100 Hz
array radius	q	0.05 m
speaker distance	q_s	1.5 m
frame size		2048 samples
overlapping in time		50%
FFT size		2048 samples
TF zones width	Ω	344 Hz
overlapping in frequency		50%
highest frequency of interest	$f_{\max}$	4000 Hz
single-source zones threshold	ϵ	0.2
frequency bins/single-source zone	d	2
number of bins in the histogram	L	720
histogram bin size		0.5°
averaging filter window length	h_N	5°
history length (block size)	B	43 frames (1 second)
narrow source atom width	Q_{N}	81
wide source atom width	Q_{W}	161
noise type	additive white Gaussian noise	

V. RESULTS AND DISCUSSION

We investigated the performance of our proposed method in simulated and real environments. In both cases we used a uniform circular array placed in the centre of each environment. All the parameters and their corresponding values can be found in Table II, unless otherwise stated.

Since the radius of the circular array is $q = 0.05$ m, the highest frequency of interest is set to $f_{\max} = 4000$ Hz in order to avoid spatial aliasing [21], [40]. Note that the final

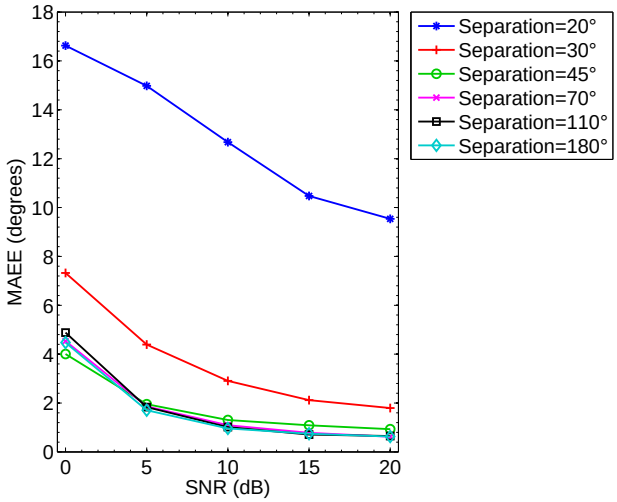

Fig. 7. DOA estimation error vs SNR for pairs of simultaneously active speakers in a simulated reverberant environment.

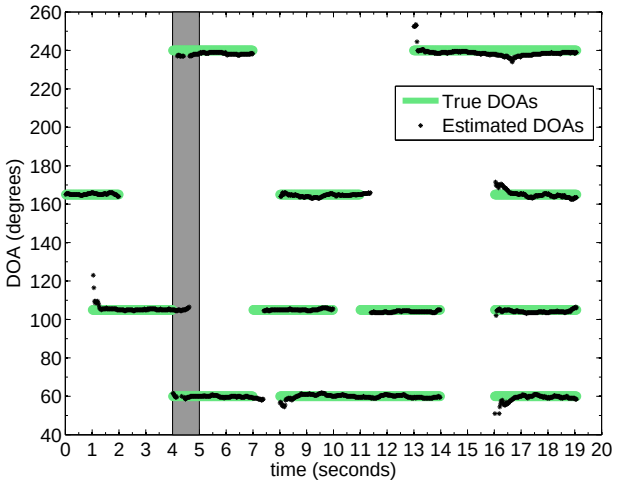

Fig. 8. Estimation of DOA of four intermittent speakers at $60°, 105°, 165°$, and $240°$ in a simulated reverberant environment with 20 dB SNR and a one-second block size. The gray-shaded area denotes an example "transition period".

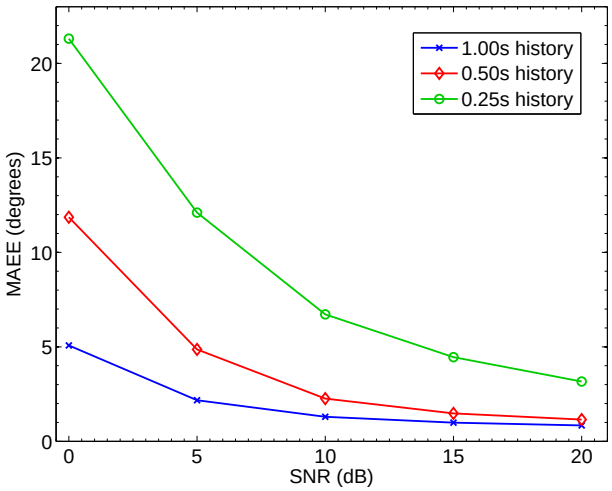

Fig. 9. DOA estimation error vs SNR for four intermittant speakers in a simulated reverberant environment.

values chosen for the source atom widths (i.e., $Q_{\mathrm{N}} = 81$ and $Q_{\mathrm{W}} = 161$) correspond to $40°$ and $80°$ respectively. However, due to the shape of the Blackman window, the effective widths are closer to $20°$ and $40°$.

A. Simulated Environment

We conducted various simulations in a reverberant room using speech recordings. We used the fast image-source method (ISM) [41], [42] to simulate a room of $6 \times 4 \times 3$ meters, characterised by reverberation time $T_{60} = 250$ ms. The uniform circular array was placed in the centre of the room, coinciding with the origin of the x and y-axis. The speed of sound was $c = 343$ m/s. In each simulation the sound sources had equal power and the signal-to-noise ratio at each microphone was estimated as the ratio of the power of each source signal to the power of the noise signal.

It must be noted that we simulated each orientation of sources in $10°$ steps around the array in order to more accurately measure the performance all around the array.

The performance of our system was measured by the mean absolute estimated error (MAEE) which measures the difference between the true DOA and the estimated DOA over all speakers, all orientations and all the frames of the source signals, unless otherwise stated.

$$\mathrm{MAEE} = \frac{1}{N_O N_F} \sum_{o,n} \frac{1}{P_n} \sum_{g} |\theta_{(o,n,g)} - \hat{\theta}_{(o,n,g)}|, \qquad (22)$$

where $\theta_{(o,n,g)}$ is the true DOA of the g^{th} speaker in the o^{th} orientation around the array in the n^{th} frame and $\hat{\theta}_{(o,f,g)}$ is the estimated DOA. N_O is the total number of different orientations of the speakers around the array, i.e., the speakers move in steps of $10°$ in each simulation, which leads to $N_O = 36$ different runs. N_F is the total number of frames after subtracting $B-1$ frames of the initialization period. We remind the reader that P_n is the number of active speakers in the n^{th} frame.

1) DOA estimation: We present and discuss our results for DOA estimation assuming known number of active sources. In our first set of simulations we investigated the spatial resolution of our proposed method, i.e., how close two sources can be in terms of angular distance while accurately estimating their DOA. Fig. 7 shows the MAEE against SNR of additive white Gaussian noise, for pairs of static, continuously active speakers for angular separations from $180°$ down to $20°$. The duration of the speech signals was approximately three seconds. Our method performs well for most separations, but the effective resolution with the chosen parameters is apparently around $30°$.

In Fig. 8 we plot an example DOA estimation of four intermittent speakers across time with the speakers at $60°, 105°, 165°$, and $240°$. Note that the estimation of each source is prolonged for some period of time after he/she stops talking or respectively is delayed when he/she starts talking. This is due to the fact that the DOA estimation at each time instant is based on a block of estimates of length B seconds ($B = 1$ second in this example). We refer to these periods as "transition periods", which we define as the time interval starting when a new or existing speaker starts or stops talking

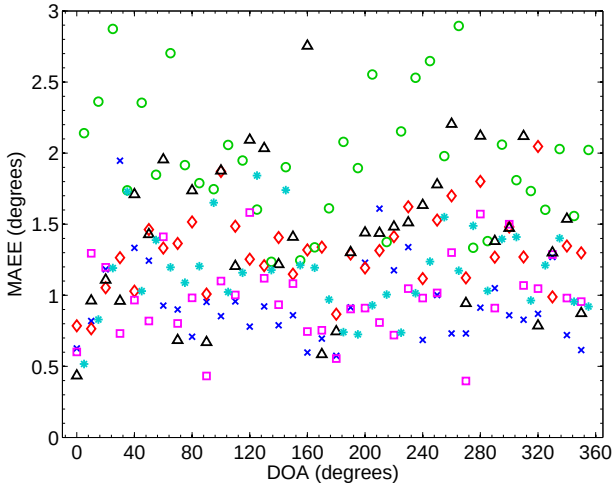

Fig. 10. DOA estimation error of six static sources versus the true DOA. Different markers correspond to different speakers.

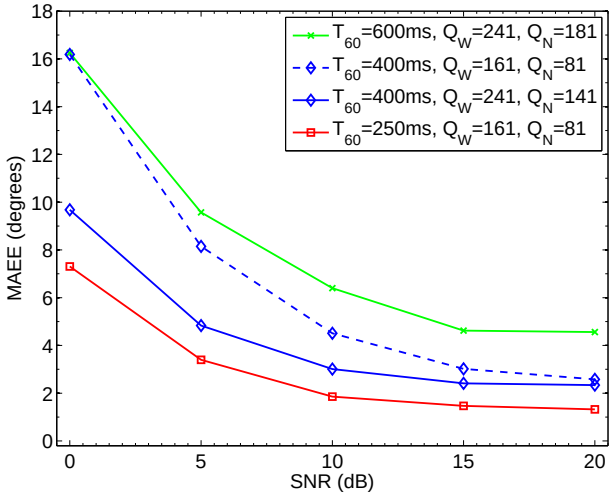

Fig. 11. DOA estimation error vs SNR for three static, continuously active speakers in a simulated environment for $T_{60} = \{250, 400, 600\}$ ms.

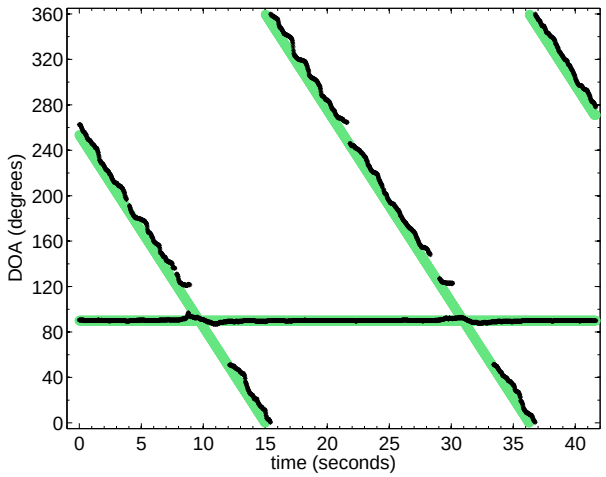

Fig. 12. Estimated DOA of one static and one moving speaker around the circular array in a simulated reverberant environment at 20 dB SNR.

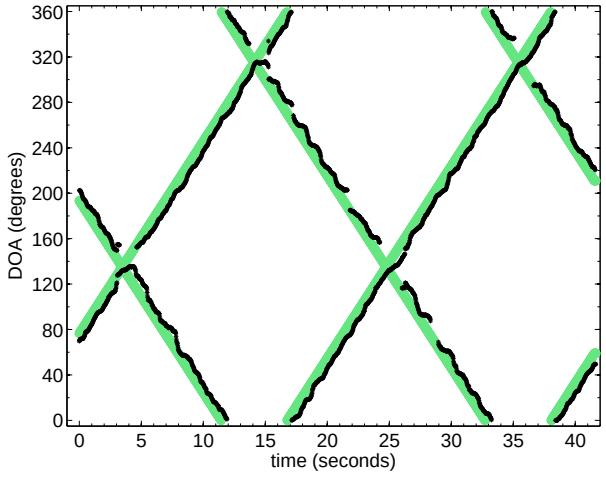

Fig. 13. Estimated DOA of two moving speakers around the circular array in a simulated reverberant environment at 20 dB SNR.

and ending B seconds later. An example of a transition period is also shown in Fig. 8 as the grey-shaded area.

We demonstrate how the size of a block of estimates affects the DOA estimation in Fig. 9. We plot the MAEE versus SNR for the four intermittent speakers scenario for block sizes—also referred to as history lengths—equal to 0.25s, 0.5s and 1s. The speakers were originally located at $0°$, $45°$, $105°$ and $180°$ and even though they were intermittent, there was a significant part of the signals where all four speakers were active simultaneously. There is an obvious performance improvement as the history length increases, as the algorithm has more data to work with in the histogram. However increasing the history also increases the latency of the system, in turn decreasing responsiveness.

Aiming to highlight the consistent behaviour of our proposed method no matter where the sources are located around the array, in Fig. 10 we plot the absolute error as an average over time, separately for each of six static, simultaneously active speakers and each of 36 different orientations around the array. For the first simulation the sources were located at $0°$, $60°$, $105°$, $180°$, $250°$, and $315°$ in a simulated reverberant environment with 20 dB SNR and a one-second history. They were shifted by $10°$ for each next simulation preserving their angular separations. The duration of the speech signals was approximately 10 seconds and, as already stated, the MAEE was evaluated as the average absolute error in the estimation over time. The MAEE is always below $3°$ for any positioning of the sources around the array for all the sources.

We investigate the robustness to reverberation in Fig. 11, which shows the MAEE versus SNR for three static, continuously active speakers originally located at $0°$, $160°$, and $240°$ for reverberation time $T_{60} = \{250, 400, 600\}$ ms. For low reverberation conditions—$T_{60} = 250$ ms—the proposed method performs very well for all SNR conditions as was expected and shown in the preceding results. For medium reverberation with $T_{60} = 400$ ms and source atom widths $Q_{\mathrm{W}} = 161(80°)$ and $Q_{\mathrm{N}} = 81(40°)$ the MAEE is low for high SNR but increases rapidly for lower signal-to-noise ratios. However, by using wider pulses—i.e., $Q_{\mathrm{W}} = 241(120°)$ and $Q_{\mathrm{N}} = 141(70°)$—we can mitigate erroneous estimates due to reverberation and keep the error lower than $10°$ for all SNR values. For $T_{60} = 600$ ms—which could characterize a highly reverberant environment—the DOA estimation is effective for SNR values above 5 dB, exhibiting an MAEE lower than $7°$, when using $Q_{\mathrm{W}} = 241(120°)$ and $Q_{\mathrm{N}} = 181(90°)$. Note that increasing the source atom widths improves the DOA

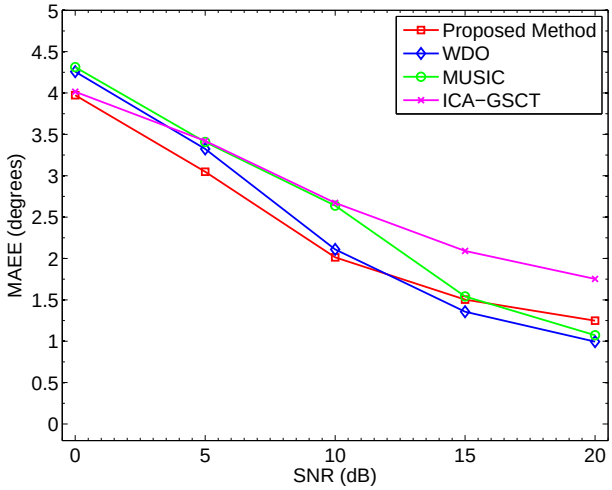

Fig. 14. DOA estimation error vs SNR for six static speakers in a simulated reverberant environment.

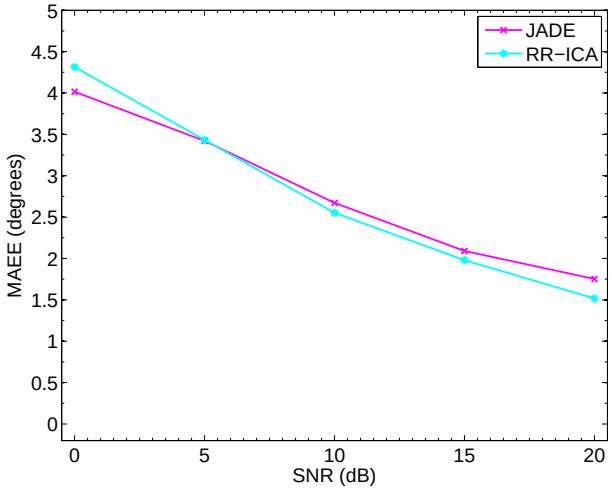

Fig. 15. DOA estimation error vs SNR for six static speakers in a simulated reverberant environment.

estimation accuracy, but also decreases the resolution of the method.

In order to investigate the tracking potential of our proposed method, we ran simulations that included moving sources. In Fig. 12 one speaker is static at 90° and the other is moving clockwise. Both speakers were males. In Fig. 13 two male speakers are moving in a circular fashion around the array. One of them is moving anticlockwise while the other is moving clockwise. We observe a consistent DOA estimation in both scenarios, even though we do not use any source labelling techniques. This preliminary simulation results, along with their real-environment experiments counterparts, indicate that the proposed method could be extended to a multiple source tracking method. The slight shift of the estimations to the right of the true DOA is due to the one-second history length. Anomalies in the DOA estimation are mainly present around the crossing points, which was expected, since the effective resolution of the proposed method is around 30° (see also Fig. 7).

2) Comparison with alternative methods: We also compared the performance of the proposed method against WDO, MUSIC, and ICA-GSCT (see Section IV). The performance of the methods was evaluated by using the MAEE over those estimates where the absolute error was found to be lower than 10°—where an estimate is considered to be successful. Along with the MAEE, we provide "success scores", i.e., percentages of estimates where the absolute error was lower than 10° (Table III to be discussed later). Since the error was very high for plenty of estimates especially at lower SNR values for some of the methods, the MAEE over all estimates was considerably affected, not allowing us to have a clear image of the performance. Furthermore, in a real system, a stable consistent behaviour—which is reflected in the "success scores"— is equally important as accuracy and computational complexity. We note that a similar method of performance evaluation was adopted in [21]. In Fig. 14 we plot the MAEE versus the SNR for six static, continuously active speakers, originally located at 0°, 60°, 105°, 180°, 250°, and 315° in a simulated reverberant environment with a one-second block size. The simulation was performed for each orientation of sources in 10° steps around the array. All four methods exhibit very good results, with an increasing performance from lower to higher SNR values. Even though the differences are small between the methods, we note that the proposed one exhibits the lowest MAEE for SNR values below 15 dB (and the highest success scores, shown in Table III to be discussed later).

Since the accuracy of the estimation of the demixing matrices (and consequently of the corresponding mixing matrices) for ICA-GSCT at each frequency bin depends on the sufficiency of the observed data—i.e., the block size—we ran the preceding simulation scenario using mixing matrices obtained with the Recursively Regularized ICA (RR-ICA) algorithm [43]. The RR-ICA algorithm exploits the consistency of demixing matrices across frequencies and the continuity of the time activity of the sources and recursively regularizes ICA. In this way, it provides improved estimates of the demixing matrices even when a short amount of data is used. We note that the code for RR-ICA is provided by the authors of [43] and can be found in [44]. The maximum number of ICA iterations was set to 20 and the natural gradient step-size to 0.1. The maximum order of the least mean square (LMS) filter was set to 10 and the corresponding step size to 0.01. These values gave the best results among various parametrizations and are in the range of values recommended in [43]. In Fig. 15 we compare the performance of ICA-GSCT using these two different methods for the estimation of the mixing matrices, i.e., the JADE algorithm and RR-ICA method. We observe that both methods exhibit good and similar results for all SNR values. We note that RR-ICA performs slightly better for SNR higher than 5 dB as was expected but did not provide a significant improvement compared to JADE for our particular simulation scenario.

In Table III we provide success scores (percentages of frames with absolute error $< 10^\circ$) for the proposed and all aforementioned methods. We observe that for an SNR of 20 dB, all methods successfully estimate the DOAs for more than 90% out of a total amount of approximately 83,000 estimates. Specifically, the proposed method along with WDO and MUSIC almost achieve score of 100%, with the proposed one

TABLE III
DOA ESTIMATION SUCCESS SCORES

Method	SNR(dB) 0	5	10	15	20
proposed	61.62%	84.07%	95.45%	99.16%	99.69%
WDO	54.96%	80.38%	95.40%	99.57%	99.94%
MUSIC	47.89%	64.82%	77.34%	92.58%	99.89%
JADE ICA-GSCT	55.44%	68.66%	80.38%	89.17%	93.90%
RR-ICA-GSCT	40.66%	57.69%	73.70%	88.04%	96.48%

being much more efficient in terms of complexity. When the SNR gets lower, the performance of the methods deteriorates, which can also be observed in Fig. 14 and 15. However, our proposed method's score is higher than the other methods for SNR values below 15 dB.

3) Source Counting results: In order to evaluate our matching pursuit-based (MP) source counting method (see Section III-E), we provide source counting results for simulation scenarios ranging from one to six static, simultaneously active sound sources in a reverberant environment with an SNR of 20 dB. In these six simulation scenarios, the smallest angular distance between sound sources was 45° and the highest was 180° while the sources were active for approximately 10 seconds, leading to roughly 14,000 source number estimations for each scenario. The thresholds vector was set to $\boldsymbol{\gamma} = [0.15, 0.14, 0.12, 0.1, 0.065, 0.065, 0.065]$ and the minimum offset between neighbouring located sources was set to $u_w = 10^\circ$. We present these results in terms of a confusion matrix in Table IV where the rows correspond to true numbers of sources and the columns correspond to the estimated ones. The method correctly estimates the number of sources more than 87% of the time for all the cases. Overall the method presents very good performance with a mean percentage of success equal to 93.52%.

TABLE IV
CONFUSION MATRIX FOR THE MP PROPOSED SOURCE COUNTING METHOD

		$\hat{P}$ 1	2	3	4	5	6	7
P	1	100%	0%	0%	0%	0%	0%	0%
	2	0%	100%	0%	0%	0%	0%	0%
	3	0%	3.76%	96.16%	0.08%	0%	0%	0%
	4	0%	0.42%	8.50%	88.84%	2.20%	0.04%	0%
	5	0.01%	2.23%	2.99%	0.55%	88.28%	5.76%	0.18%
	6	0.87%	2.91%	1.42%	0.17%	5.91%	87.84%	0.88%

We compared our MP proposed source counting method with our additional proposed source counting methods (see Sections III-E and III-F) and the minimum description length (MDL) information criterion [45] under the four intermittent speakers scenario, an example of which can be seen in Fig. 8. For the Peak Search method (PS), $z_{\text{static}} = 0.05 \sum_v y(v)$ and the LPC order used was 16. The thresholds for the MP were $\boldsymbol{\gamma} = [0.15, 0.14, 0.12, 0.1]$. The minimum offset between neighbouring located sources was set to $u_w = 10^\circ$ and was common for all these histogram-based methods. The MDL was estimated in the frequency domain from the STFT of the observations in blocks of B frames. In Table V we give

TABLE V
SOURCE COUNTING SUCCESS RATES EXCLUDING TRANSITION PERIODS

Method	History Length	SNR (dB) 0	5	10	15	20
MDL	0.25s	0%	0%	2.3%	15.7%	21.6%
PS	0.25s	34.7%	44.8%	60.2%	71.5%	79.1%
LPC	0.25s	25.7%	40.5%	57.0%	63.0%	64.6%
MP	0.25s	42.9%	61.5%	77.8%	84.7%	86.7%
MDL	0.5s	0%	0%	6.8%	38.8%	74.8%
PS	0.5s	44.5%	60.1%	77.5%	84.9%	88.2%
LPC	0.5s	35.5%	59.5%	73.8%	75.6%	74.2%
MP	0.5s	64.3%	84.8%	95.7%	96.7%	96.7%
MDL	1s	0%	0%	21.2%	70.8%	87.7%
PS	1s	47.3%	68.7%	83.6%	90.5%	92.7%
LPC	1s	45.4%	81.9%	85.4%	82.5%	80.1%
MP	1s	82.1%	99.2%	100%	100.0%	100.0%

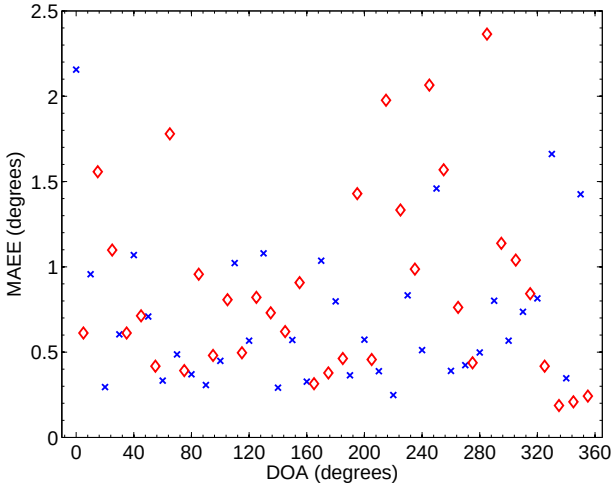

Fig. 16. DOA estimation error for two speakers separated by 45° versus the true DOA in a real environment. Each different marker corresponds to a different speaker

success rates of the source counting (percentage of frames correctly counting the number of sources) for the four methods under consideration with various history lengths and differing values of SNR. The success rates were again calculated over all orientations of the sources in 10° steps around the array (preserving the angular separations) while the transition periods were not taken into account.

We can observe similar behaviour as in Fig. 9. Longer history length leads to increased success rates for all four methods, affecting however, the responsiveness of the system. The MDL method is severely affected by noise and the amount of available data. While it achieves a high percentage of success for one-second history length and 20 dB SNR, this percentage falls dramatically as the history length is reduced and most obviously as the SNR becomes lower. For SNRs equal to 0 and 5 dB the criterion fails completely since it always responds as if there are no active sources. The matching pursuit method is clearly the best performing source counting method. Moreover, matching pursuit can be used in a single step both for the DOA estimation and the source counting (as explained in Section III-E), resulting in computational efficiency.

B. Real Environment

We conducted experiments in a typical office room with approximately the same dimensions and placement of the

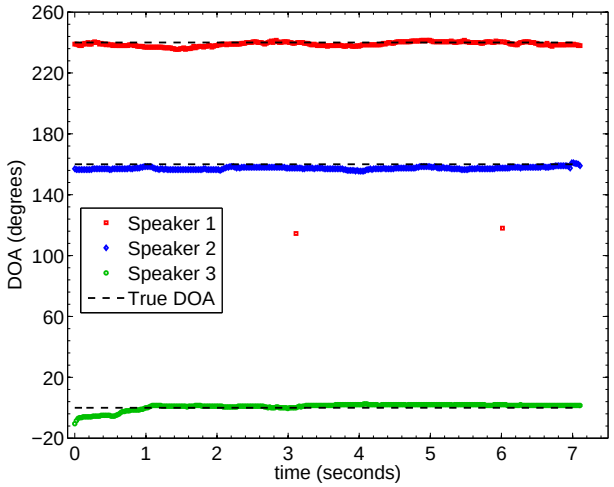

Fig. 17. Estimated DOA of 3 static speakers in a real environment.

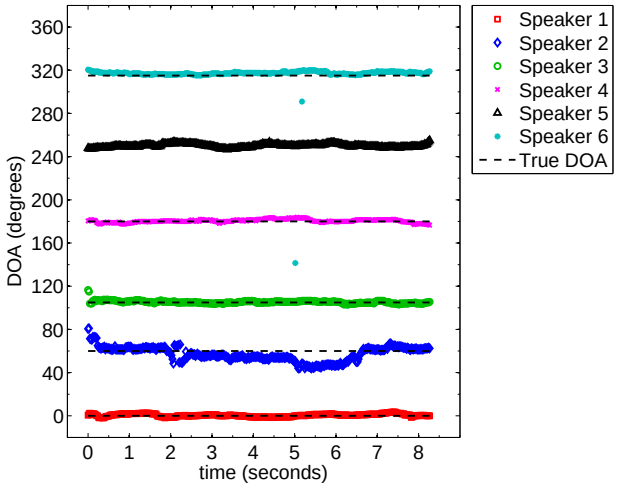

Fig. 18. Estimated DOA of six static speakers in a real environment.

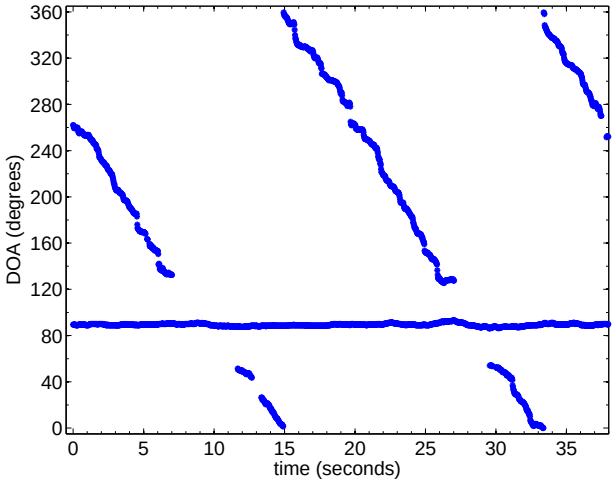

Fig. 19. Estimated DOA of one static speaker and one moving speaker around the circular array in a real environment.

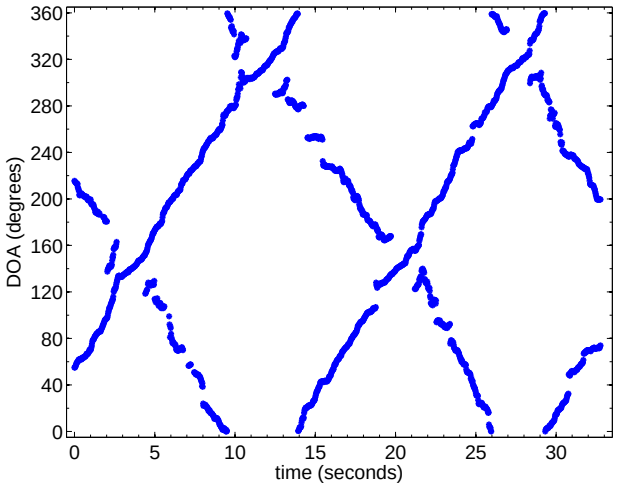

Fig. 20. Estimated DOA of two moving speakers around the circular array in a real environment.

microphone array as in the simulations and with reverberation time approximately equal to 400 ms. The algorithm was implemented in software executed on a standard PC (Intel 2.40 GHz Core 2 CPU, 2GB RAM). We used eight Shure SM93 microphones (omnidirectional) with a TASCAM US2000 8-channel USB soundcard. We measured the execution time and found it to be 55% real time (i.e., 55% of the available processing time). In the following results, some percentage of the estimated error can be attributed to the inaccuracy of the source positions.

We demonstrate the performance of our system for two simultaneously active male speakers in Fig. 16. The speakers were separated by 45° and they moved 10° in each experiment in order to test the performance all around the array. The duration of each experiment was approximately six seconds. The signal to noise ratio in the room was, on average, 15 dB. We plot the MAEE versus each different DOA, where the MAEE is evaluated as the mean absolute error in the estimation over time. The mean absolute error is lower that 2.5° for every positioning of the speakers around the array (among 36 different orientations) while for about half of the orientations, the MAEE is below 1° for both speakers.

The next experiment involved three speakers sitting around the microphone array at 0°, 160°, and 240°. The speakers at 0° and 240° were males, while the speaker at 160° was female. The signal to noise ratio in the room was also around 15 dB. In Fig. 17 we plot the estimated DOA in time. All three speakers are accurately located through the whole duration of the experiment.

In Fig. 18 we plot the estimated DOAs of six static speakers versus time. This experiment is the only one that involved loudspeakers instead of actual speakers. We used six Genelec 8050 loudspeakers that reproduced pre-recorded audio files of six continuously active, actual speakers, three males and three females positioned alternately. The loudspeakers were approximately located at 0°, 60°, 105°, 180°, 250°, and 315° at a distance of 1.5 meters from the centre of the array. The signal to noise ratio in the room was estimated at 25 dB. The DOA of all six sources is in general accurately estimated. The DOA estimation of the second speaker deviates slightly from the true DOA for some periods of time (e.g., around the sixth second of the experiment). This might be attributed to a lower energy of the signal of the particular speaker over these periods in comparison to the other speakers.

We also conducted experiments with moving sources. The scenarios followed the simulations (see Fig. 12 and 13). For these experiments, the signal to noise ratio in the room was,

on average, 20 dB. We plot the DOA estimation in Fig. 19 and 20. The DOA estimation is in general effective except for the areas around the crossing points. Nevertheless, as we stated for the corresponding simulations, our method shows the potential of localizing moving sources that cross each other.

VI. Conclusion

In this work, we presented a method for jointly counting the number of active sound sources and estimating their corresponding DOAs. Our method is based on the sparse representation of the observation signals in the TF-domain with relaxed sparsity constraints. This fact—in combination with the matching pursuit-based technique that we apply to a histogram of a block of DOA estimations—improves accuracy and robustness in adverse environments. We performed extensive simulations and real environment experiments for various numbers of sources and separations, and in a wide range of SNR conditions. In our tests, our method was shown to outperform other localization and source counting methods, both in accuracy and in computational complexity. Our proposed method is suitable for real-time applications, requiring only 55% of the available processing time of a standard PC. We implemented our method using a uniform circular array of microphones, in order to overcome the ambiguity constraints of linear topologies. However, the philosophy of the method is suitable for any microphone array topology.

Acknowledgment

The authors would like to acknowledge the anonymous reviewers for their valuable comments to improve the present work. This research was co-financed by the Marie Curie IAPP "AVID-MODE" grant within the European Commission's FP7 and by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES, Project "MUSINET".

References

[1] H. Krim and M. Viberg, "Two decades of array signal processing research - the parametric approach," *IEEE Signal Processing Magazine*, pp. 67–94, July 1996.

[2] S. Argentieri and P. Danès, "Broadband variations of the music high-resolution method for sound source localization in robotics," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, November 2007, pp. 2009–2014.

[3] T. Van den Bogaert, E. Carette, and J. Wouters, "Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna," *International Journal of Audiology*, vol. 50, no. 3, pp. 164–176, 2011.

[4] K. Nakadai, D. Matsuura, H. Kitano, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2003, pp. 1147–1152.

[5] D. Bechler, M. Schlosser, and K. Kroschel, "System for robust 3D speaker tracking using microphone array measurements," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, September 2004, pp. 2117–2122.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, August 1976.

[7] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, September 2004.

[8] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Applied Signal Processing*, pp. 1–19, 2006.

[9] D. Bechler and K. Kroschel, "Considering the second peak in the GCC function for multi-source TDOA estimation with microphone array," in *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2003, pp. 315–318.

[10] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, March 1986.

[11] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband music: Opportunities and challenges for multiple source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2007, pp. 18–21.

[12] F. Belloni and V. Koivunen, "Unitary root-music technique for uniform circular array," in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, (ISSPIT)*, December 2003, pp. 451–454.

[13] J. Zhang, M. Christensen, J. Dahl, S. Jensen, and M. Moonen, "Robust implementation of the music algorithm," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, April 2009, pp. 3037–3040.

[14] C. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS)*, October 2009, pp. 2027–2032.

[15] B. Loesch, S. Uhlich, and B. Yang, "Multidimensional localization of multiple sound sources using frequency domain ica and an extended state coherence transform," in *Proceedings of the IEEE/SP 15th Workshop on Statistical Signal Processing, (SSP)*, September 2009, pp. 677–680.

[16] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1490–1503, August 2011.

[17] H. Sawada, R. Mukai, S. Araki, and S. Malcino, "Multiple source localization using independent component analysis," in *IEEE Antennas and Propagation Society International Symposium*, vol. 4B, July 2005, pp. 81–84.

[18] 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, January 2012.

[19] P. Comon and C. Jutten, *Handbook of blind source separation: independent component analysis and applications*, ser. Academic Press. Elsevier, 2010.

[20] M. Swartling, B. Sällberg, and N. Grbić, "Source localization for multiple speech sources using low complexity non-parametric source separation and clustering," *Signal Processing*, vol. 91, pp. 1781–1788, August 2011.

[21] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, October 2011.

[22] D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris, "Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 2625–2628.

[23] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[24] M. Puigt and Y. Deville, "A new time-frequency correlation-based source separation method for attenuated and time shifted mixtures," in *Proceedings of the 8th International Workshop (ECMS and Doctoral School) on Electronics, Modelling, Measurement and Signals*, 2007, pp. 34–39.

[25] E. Fishler, M. Grosmann, and H. Messer, "Detection of signals by information theoretic criteria: general asymptotic performance analysis," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1027–1036, may 2002.

[26] G. Hamerly and C. Elkan, "Learning the k in k-means," in *Neural Information Processing Systems*. MIT Press, 2003, pp. 281–288.

[27] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proceedings of the International Workshop for Acoustics Echo and Noise Control, (IWAENC)*, 2008.

[28] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with map estimation with dirichlet prior considering spatial aliasing problem," in *Independent Component Analysis and Signal Separation*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, vol. 5441, pp. 742–750.

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

[30] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[31] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Source counting in real-time sound source localization using a circular microphone array," in *Proceedings of the IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, June 2012, pp. 521–524.

[32] 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 *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, August 2012, pp. 2303–2307.

[33] Y. Deville and M. Puigt, "Temporal and time-frequency correlation-based blind source separation methods. part i: Determined and underdetermined linear instantaneous mixtures," *Signal Processing*, vol. 87, pp. 374–407, March 2007.

[34] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[35] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, December 1993.

[36] [Online]. Available: http://math.uci.edu/

[37] H. Teutsch and W. Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2724–2736, 2006.

[38] J. Meyer and G. Elko, "Spherical harmonic modal beamforming for an augmented circular microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP) 2008*, 2008, pp. 5280–5283.

[39] T. Abhayapala and A. Gupta, "Spherical harmonic analysis of wavefields using multiple circular sensor arrays," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1655–1666, 2010.

[40] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1327–1339, May 2007.

[41] 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, pp. 1429–1439, August 2010.

[42] [Online]. Available: http://www.eric-lehmann.com/

[43] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive BSS of short mixtures by ICA recursively regularized across frequencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 624–639, 2011.

[44] [Online]. Available: http://bssnesta.webatu.com/software.html

[45] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech and Signal Processing,*, vol. 33, no. 2, pp. 387–392, 1985.

Despoina Pavlidi (S'12) received the diploma degree in Electrical and Computer Engineering in 2009 from the National Technical University of Athens (NTUA), Greece, and the M.Sc. degree in Computer Science in 2012 from the Computer Science Department of the University of Crete, Greece. She is currently pursuing the Ph.D. degree at the Computer Science Department of the University of Crete. Since 2010 she is affiliated with the Institute of Computer Science at the Foundation for Research and Technology-Hellas (FORTH-ICS) as a research assistant. Her research interests include audio signal processing, microphone arrays and sound source localization and audio coding.

Anthony Griffin received his Ph.D. in Electrical & Electronic Engineering from the University of Canterbury in Christchurch, New Zealand in 2000. He then spent three years programming DSPs for 4RF, a Wellington-based company selling digital microwave radios. He subsequently moved to Industrial Research Limited—also based in Wellington—focussing on signal processing for audio signals and wireless communications. In 2007, he joined the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS), Heraklion, Greece as a Marie Curie Fellow, where he is working on real-time audio signal processing, compressed sensing, and wireless sensor networks. He also occasionally teaches a postgraduate course in Applied DSP at the University of Crete.

Matthieu Puigt is an Associate Professor at the Université du Littoral Côte d'Opale (ULCO) since September 2012. His research activities are conducted at the Laboratoire d'Informatique, Signal et Image de la Côte d'Opale, while he is teaching at the University Institute of Technology of Saint-Omer Dunkerque, in the Industrial Engineering and Maintenance Department. He received both the Bachelor and first year of M.S. degrees in Pure and Applied Mathematics, in 2001 and 2002 respectively, from the Université de Perpignan, France. He then received the M.S. degree in Signal, Image Processing, and Acoustics, from the Université Paul Sabatier Toulouse 3, Toulouse, France, in 2003, and his Ph.D. in Signal Processing from the Université de Toulouse in 2007. From 2007 to 2009 he was a Postdoctoral Lecturer at the Université Paul Sabatier Toulouse 3 and the Laboratoire d'Astrophysique de Toulouse-Tarbes. From September 2009 to June 2010, he held an Assistant Professor position at the University for Information Science and Technology, in Ohrid, Republic of Macedonia (FYROM). From August 2010 to July 2012, he was a Marie Curie postdoctoral fellow in the Signal Processing Lab of the Institute of Computer Science of the Foundation for Research and Technology – Hellas (FORTH-ICS). Matthieu Puigt's current research interests include linear and nonlinear signal processing, time-frequency and wavelet analysis, unsupervised classification, and especially blind source separation methods and their applications to acoustics and astrophysics. He has authored or co-authored more than 15 publications in journal or conference proceedings and has served as a reviewer for several scientific journals and international conferences in these areas.

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 Assistant 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 immersive audio environments, spatial and multichannel audio, sound source localization and microphone arrays, and speech processing with emphasis on voice conversion and speech enhancement. He has contributed to more than 70 publications in various journal and conference proceedings in these areas. Dr. Mouchtaris is a member of IEEE.