A Steered-Response Power Algorithm Employing Hierarchical Search for Acoustic Source Localization Using Microphone Arrays

Leonardo O. Nunes, *Student Member, IEEE*, Wallace A. Martins, *Member, IEEE*, Markus V. S. Lima, *Member, IEEE*, Luiz W. P. Biscainho, *Member, IEEE*, Maurício V. M. Costa, Felipe M. Gonçalves, Amir Said, *Fellow, IEEE*, and Bowon Lee, *Senior Member, IEEE*

***Abstract*—The localization of a speaker inside a closed environment is often approached by real-time processing of multiple audio signals captured by a set of microphones. One of the leading related methods for sound source localization, the steered-response power (SRP), searches for the point of maximum power over a spatial grid. High-accuracy localization calls for a dense grid and/or many microphones, which tends to impractically increase computational requirements. This paper proposes a new method for sound source localization (called H-SRP), which applies the SRP approach to space regions instead of grid points. This arrangement makes room for the use of a hierarchical search inspired by the branch-and-bound paradigm, which is guaranteed to find the global maximum in anechoic environments and shown experimentally to also work under reverberant conditions. Besides benefiting from the improved robustness of volume-wise search over point-wise search as to reverberation effects, the H-SRP attains high performance with manageable complexity. In particular, an experiment using a 16-microphone array in a typical presentation room yielded localization errors of the order of 7 cm, and for a given fixed complexity, competing methods' errors are two to three times larger.**

***Index Terms*—Sound source localization, steered-response power, microphone array, computational complexity, hierarchical search, branch-and-bound.**

Manuscript received June 17, 2013; revised November 22, 2013; accepted June 16, 2014. Date of publication July 16, 2014; date of current version August 28, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pina Marziliano. This R&D project is a cooperation between Hewlett-Packard Brasil Ltda. and COPPE/UFRJ, being supported with resources of Informatics Law (no. 8.248, from 1991). L. O. Nunes, W. A. Martins, M. V. S. Lima, and L. W. P. Biscainho would like to thank also CAPES, CNPq, and FAPERJ agencies for funding their research work.

L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, and M. V. M. Costa are with the Signal, Multimedia, and Telecommunications Lab—DEL/Poli & PEE/COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, RJ 21941-972, Brazil (e-mail: leonardo.nunes@smt.ufrj.br; wallace.martins@smt.ufrj.br; markus.lima@smt.ufrj.br; wagner@smt.ufrj.br; mauricio.costa@smt.ufrj.br).

F. M. Gonçalves was with Signal, Multimedia, and Telecommunications Lab—DEL/Poli & PEE/COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, RJ 21941-972, Brazil. He is now with the COPPEAD, Federal University of Rio de Janeiro, Rio de Janeiro, RJ 21941-918, Brazil (e-mail: felipemg7@poli.ufrj.br).

A. Said was with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA. He is now with LG Electronics Mobile Research, San Jose, CA 98008 USA (e-mail: amir.said@lge.com).

B. Lee was with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA. He is now with the Department of Electronic Engineering, Inha University, Incheon 402-751, South Korea (e-mail: bowon.lee@inha.ac.kr).

Digital Object Identifier 10.1109/TSP.2014.2336636

I. Introduction

SOUND source localization (SSL) [1], [2] finds use in a variety of practical systems ranging from communications (e.g., teleconference systems) to medical applications (e.g., hearing aids), to mention just a few [3], [4]. In order to localize an acoustic source, one must necessarily rely on some sort of spatial information such as that supplied by a microphone array (MA).

Among the SSL techniques devised for MAs, two families of algorithms are usually the prevalent choices [2]: the first is explicitly based on the time-difference of arrival (TDoA), whereas the second relies on maximizing the steered-response power (SRP) of a beamformer. TDoA-based methods, among which the most popular techniques use the TDoAs estimated by the generalized cross-correlation (GCC) [5]–[7], require relatively few numerical operations to localize a source, as compared to SRP-based algorithms. However, the performance of TDoA-based methods is highly affected by noise and reverberation [2], [8], which might hinder their use in practical applications. In such situations, SRP-based methods, whose classical version (hereafter called C-SRP) [2], [8] is the most widely used, are more appropriate for their robustness to acoustical issues inherent to the application environment.

In order to estimate a source position, the C-SRP method is applied over a grid of predefined spatial points, which represent source location candidates. High localization accuracy can only be achieved at the cost of increasing either the number of grid points or the number of microphones (usually, the larger the number of captured sound signals, the higher the attained spatial diversity). Therefore, the burden of the point-wise search for the source position drives the computational complexity of the C-SRP algorithm, whose increase can turn real-time operation impractical, thus rendering the algorithm useless in most applications of interest [9].

In an attempt to address this issue, several methods that modify the search process have been proposed. For instance, in [10] the authors devised a search strategy for the C-SRP based on the stochastic region contraction (SRC) algorithm, which enables the estimation of source location without necessarily evaluating the objective function associated with the C-SRP for every grid point. Similarly to the SRC method, the coarse-to-fine region contraction (CFRC) [11] tries to find the source position by progressively reducing the search space according to a set of heuristics. An improvement of the SRC method relying on particle filtering was proposed in [12]. In

addition, low-energy contributions are disregarded in the search stage of the C-SRP in [9].

All the previous methods decrease the number of arithmetical operations by avoiding the evaluation of the C-SRP objective function for every point in the search grid. This eventually means that the true source position can sometimes be missed, since there is no formal guarantee that the points which are disregarded in the search process are not good candidates for the source location. This disadvantage becomes more evident in the presence of reverberation, which induces many local maxima in the objective function of the related method.

This paper proposes a novel SRP-based localization method for a single source which is able to reduce the computational complexity of the search stage, as compared to the C-SRP, and is theoretically guaranteed to converge to the global maximum (that indicates within the partitioned 3-D search space which subregion contains the source position) under anechoic conditions—while keeping good performance in practical non-anechoic environments. The key idea is the implicit exploration of portions of the search space. The proposed search strategy, herein called *hierarchical search*, adapts the branch-and-bound (B&B) paradigm [14] to tackle the SSL problem by defining an objective function related to acoustic activity whose value computed for a given 3-D spatial region is never smaller than for any of its subregions in ideal conditions. Simulation results under realistic reverberant conditions show that the proposed method (referred as H-SRP) can achieve good accuracy with relatively low computational burden. It should be pointed out that the idea of a volumetric SRP using hierarchical search, although differently formulated, was originally proposed in [13].

The paper is organized as follows. Section II provides an overview of the proposed hierarchical search, describing how the B&B paradigm can be tailored for SSL applications; the role played by the bounding function is emphasized, as well as the conditions it must satisfy to guarantee convergence to the global maximum of the related objective function. The choice of a bounding function that meets such conditions in an anechoic environment is addressed in Section III, which also includes a brief discussion on the effects of reverberation. Section IV considers practical aspects regarding the implementation of the proposed H-SRP method, which is compared with previous work in Section V. Experiments using both artificially-generated and recorded signals are shown in Section VI. Conclusions are drawn in Section VII. The proofs of three theorems are left to Appendices A, B, and C.

Notations: The symbols $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$ denote the set of real, integer, and natural numbers, respectively. The set of nonnegative real numbers is represented by $\mathbb{R}_+$. In addition, vectors are denoted by lowercase boldface letters and $\|\cdot\|_2$ is the Euclidean norm. Given two sets $\mathcal{A}$ and $\mathcal{B}$, the notation $\mathcal{A}\backslash\mathcal{B}$ denotes the set containing the elements of $\mathcal{A}$ which are not in $\mathcal{B}$. The round$\{\cdot\}$ operator takes as an argument a real number and returns the integer number which is closest to the argument. For a function $f : \mathcal{A} \to \mathcal{B}$, and given a subset $\mathcal{A}_1 \subset \mathcal{A}$, the set $f(\mathcal{A}_1)$ is called the image of $\mathcal{A}_1$.

II. B&B-Inspired Hierarchical Search

As previously explained, most of the computational burden of the C-SRP is due to its search process, which requires the search space to be divided into a grid of points that must be visited once each. Moreover, for a given MA and a predefined sampling frequency, one can only increase the accuracy of the position estimates by increasing the number of points within the grid, i.e., turning it denser. Some methods try to circumvent this issue by avoiding visiting all grid points [9], [10]. However, since the C-SRP objective function may exhibit multiple local maxima, such methods fail to guarantee convergence to the global maximum (that determines the actual source position) from a deterministic point of view.[1]

Therefore, the following dilemma arises: in order to assure convergence to the actual source position, no grid point should be disregarded; on the other hand, exhaustive search through the grid is usually too complex for real-time applications. A natural way to address this problem is by performing an implicit exploration of the search space [13]. The branch-and-bound (B&B) paradigm [14], [15], originally developed for discrete and combinatorial optimization problems, seems to be well-tailored for this purpose since it guarantees convergence to the global maximum of its corresponding objective function.

B&B-based algorithms work with search spaces that can be divided into nested subspaces, each subspace seen as a node in a dynamically generated tree structure [14]. Rather than evaluating the underlying objective function for all possible nodes in the tree, B&B-based algorithms work with a bounding function,[2] which helps one decide how new nodes will be generated through the branching process. In summary, the main components of a general B&B algorithm are [14]: (i) selection of the node to process, (ii) bounding function calculation, and (iii) node branching. By their proper shaping (and their adaptation to the SSL problem), this paper develops the proposed *hierarchical search*.

To start with, the proposed hierarchical search considers that a node will correspond to a 3-D spatial region. In this case, the root node is the entire Euclidean search space (e.g., a meeting room). The bounding process uses a bounding function, which is proposed in Section III, that is associated with the presence (or absence) of the sound source within the given spatial region. Here, the objective function and the bounding function turn out to be the same. As for the branching process, it is simply the way a node is subdivided in order to generate new subregions (new nodes). Thus, the proposed hierarchical search operates by dividing the search space (root node) into smaller regions (nodes)—which constitutes the branching process—and then calculating for each spatial region the value of the bounding function—which constitutes the bounding process.

The role of the bounding function is to allow parts of the related search space to be implicitly evaluated, i.e., the whole Euclidean search space can be explored without explicitly visiting each point. Mathematically, the proposed hierarchical search relies on an appropriate bounding function $b : \mathbb{V} \longrightarrow \mathbb{R}_+$, in which $\mathbb{V} \subset \mathcal{P}(\mathcal{F})$ is a family of sets over the full search space $\mathcal{F} \subset \mathbb{R}^3$, with $\mathcal{P}(\mathcal{F})$ denoting the power set[3] of $\mathcal{F}$. An ancillary technical condition is that the elements of $\mathbb{V}$ are compact

[1]In fact, the term "*the* global maximum" is loose, since there may even exist more than one global maximum, depending on the chosen array geometry and the acoustical characteristics of the environment.

[2]This name comes from the fact that this function is always smaller than or equal to the objective function of the related problem.

[3]The power set of a set X is the set of all subsets of X.

and connected sets (e.g., cuboids). The hierarchical search sequentially subdivides the search space $\mathcal{F}$ into subregions up to a predefined minimum "size" $V_{\min} \in \mathbb{R}_+$, which is related to the stopping criteria of the search algorithm.[4] During this process, the bounding function plays a key role in discarding sets that do not require any further subdivision. Using $\mathbf{L}$ to represent a list of pairs $[\mathcal{V}, b(\mathcal{V})]$ of subregions $\mathcal{V} \in \mathbb{V}$ to be evaluated, together with their related bounding values $b(\mathcal{V}) \geq 0$, and with $b_{\max}$ being the maximum bounding value found at a given iteration of the algorithm and $\mathcal{V}^*$ being its corresponding subregion (i.e., $b(\mathcal{V}^*) = b_{\max}$), the proposed hierarchical search is as follows.

1) (Initialization) Let $\mathbf{L} \leftarrow \{[\mathcal{F}, b(\mathcal{F})]\}$, $b_{\max} \leftarrow -1$, and $\mathcal{V}^* \leftarrow \emptyset$.
2) If $\mathbf{L} = \emptyset$ then stop: the search is complete and $\mathcal{V}^*$ is the subregion of size not larger than $V_{\min}$ with the largest bounding value.
3) Sample one element $[\mathcal{V}, b(\mathcal{V})]$ from list $\mathbf{L}$, and let $\mathbf{L} \leftarrow \mathbf{L} \setminus \{[\mathcal{V}, b(\mathcal{V})]\}$.
4) (Bound) If $b(\mathcal{V}) < b_{\max}$ then go to Step 2.
5) If size$(\mathcal{V}) \leq V_{\min}$, then let $b_{\max} \leftarrow b(\mathcal{V})$ and $\mathcal{V}^* \leftarrow \mathcal{V}$, and go to Step 2. Otherwise, go to Step 6.
6) (Branch) Divide $\mathcal{V}$ into D distinct subregions $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_D$, such that their interiors are disjoint and $\bigcup_{i=1}^{D} \mathcal{V}_i = \mathcal{V}$, and then compute $b(\mathcal{V}_1), b(\mathcal{V}_2), \ldots, b(\mathcal{V}_D)$.
7) Let $\mathbf{L} \leftarrow \mathbf{L} \cup \{[\mathcal{V}_1, b(\mathcal{V}_1)], [\mathcal{V}_2, b(\mathcal{V}_2)], \ldots, [\mathcal{V}_D, b(\mathcal{V}_D)]\}$ and go to Step 2.

Fig. 1 illustrates how the hierarchical search operates. The first step of the algorithm consists of sub-dividing the search space and computing the bounding function for each subregion (Fig. 1(a)). Then, the subregion with the largest bounding value is selected and further subdivided until the region of size $V_{\min}$ is reached (Fig. 1(b)). Finally, in order to guarantee that the current maximum value corresponds to the global maximum, any other region that has a bounding value larger than (or equal to) that of the current maximum must be subdivided until a new maximum is found or all bounding values are lower than that of the current maximum (Fig. 1(c)). Alternatively, the algorithm can be represented in a tree-format where each node is a subdivision of the search space (in the case of the example where each region is subdivided into two regions, one would have a binary tree). In this case, the global maximum is found when all leaves of the tree have bounding values lower than that of the current maximum.

From the illustration presented in the previous paragraph, it is possible to note that the potential of the hierarchical search in reducing the complexity of the search stage relies on its capability of discarding large regions without further exploration/subdivision.

In this section the fundamentals of the proposed hierarchical search were shown, but the bounding function was not stated. So far, any function $b : \mathbb{V} \longrightarrow \mathbb{R}_+$ satisfying both the following properties is sufficient to guarantee convergence to the global maximum:

1) If $\mathcal{V}_i \in \mathbb{V}$ is a leaf that contains the source and $\mathcal{V}_j \in \mathbb{V}$ is a leaf that does not, then $b(\mathcal{V}_i) \geq b(\mathcal{V}_j)$.
2) If $\mathcal{V}_i \in \mathbb{V}$ is a subregion of $\mathcal{V}_j \in \mathbb{V}$, then $b(\mathcal{V}_i) \leq b(\mathcal{V}_j)$.

In the next section, a function that meets these properties will be proposed.

[4]More information on how this quantity is chosen is given in Section IV.D.

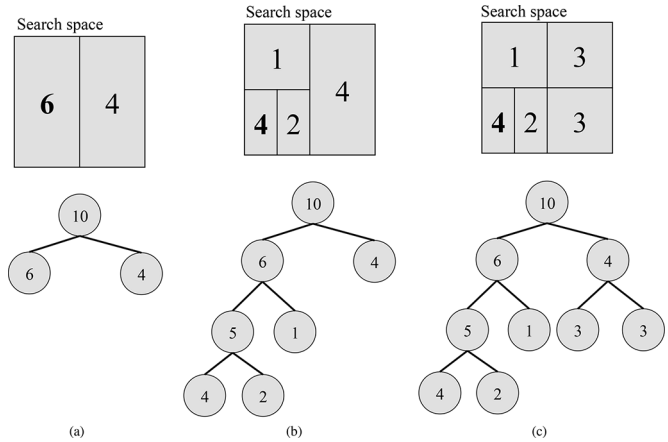

Fig. 1. Illustration of three different stages of the B&B algorithm for a 2-D search space. An equivalent tree-representation of each stage is also provided. The nodes are visited according to a "best first search" strategy. The numbers inside each region/node represent the bounding value (output of the bounding function) for the given region.

III. Bounding Function

Before describing the proposed bounding function, it is worth first presenting some known properties of the C-SRP method that naturally lead to the proposed bounding function. Thus, the idea here is to start by showing how the C-SRP technique solves the problem of localizing a source which emits a signal that is maximally concentrated in both space and time domains in an anechoic environment. It turns out that the interpretation of this simple toy example will be very instructive for further developments. After this discussion, it will be shown how the C-SRP method can be modified to deal with acoustic signals which are not concentrated in time, such as wideband/speech signals. Such a presentation ordering will point out that "counting TDoAs" is indeed the key aspect for the construction of the bounding function of the proposed hierarchical search.

A. C-SRP

The C-SRP method steers a microphone array beam to many locations searching for the acoustic source position. This search is based on maximizing the power of the output signal of a beamformer. Hence, the C-SRP maximizes the following objective function:

$$W(\mathbf{x}) \triangleq \sum_{n \in \mathbb{Z}} \left| \sum_{m=0}^{M-1} s_m[n + k_m(\mathbf{x})] \right|^2, \qquad (1)$$

where $\mathbf{x} = [x \;\; y \;\; z]^T \in \mathbb{R}^{3\times 1}$ denotes a candidate for the acoustic source position and $s_m[n]$ is the possibly filtered version of the discrete-time signal acquired by the mth microphone, for $m \in \{0, 1, \ldots, M-1\}$, where $M \in \mathbb{Z}$ denotes the number of microphones in the array. In addition, $k_m(\mathbf{x}) \in \mathbb{Z}$ would be the time-lag due to the propagation from the source position $\mathbf{x}$ to the mth microphone.

It is possible to rewrite (1) as follows [8]:

$$W(\mathbf{x}) = \frac{1}{2\pi} \sum_{m_1=0}^{M-1} \sum_{m_2=0}^{M-1} \int_{-\pi}^{\pi} S_{m_1}(\mathrm{e}^{\mathrm{j}\omega}) \times S^*_{m_2}(\mathrm{e}^{\mathrm{j}\omega}) \mathrm{e}^{\mathrm{j}\omega k_{m_1,m_2}(\mathbf{x})} \mathrm{d}\omega, \qquad (2)$$

where $S_m(\mathrm{e}^{\mathrm{j}\omega})$ represents the discrete-time Fourier transform of $s_m[n]$, and $k_{m_1,m_2}(\mathbf{x}) = k_{m_1}(\mathbf{x}) - k_{m_2}(\mathbf{x})$ is the discrete TDoA of a signal emitted at position $\mathbf{x}$ to microphones m_1 and m_2.

Now, consider the application of the C-SRP method into two different setups, namely: localizing an impulse and localizing a wideband/speech signal both within an anechoic environment.

1) Localizing an Impulse in an Anechoic Environment: Assume that a single acoustic source, located at position $\bar{\mathbf{x}} \in \mathbb{R}^{3\times 1}$, emits a pulse signal $s[n] = \delta[n]$. In this case, considering an anechoic environment and disregarding the attenuation factors associated with sound propagation, one has (based on (2))

$$W(\mathbf{x}) = \sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1} \delta[k_{m_1,m_2}(\mathbf{x}) - k_{m_1,m_2}(\bar{\mathbf{x}})], \quad (3)$$

where any discrete TDoA $k_{m_1,m_2}(\mathbf{x}) \in \mathbb{Z}$ can be rewritten as

$$k_{m_1,m_2}(\mathbf{x}) \triangleq \text{round}\left\{ f_s \frac{\|\mathbf{x}-\mathbf{x}_{m_1}\|_2 - \|\mathbf{x}-\mathbf{x}_{m_2}\|_2}{c} \right\}, \quad (4)$$

in which $\mathbf{x}_m \in \mathbb{R}^{3\times 1}$ denotes the mth microphone position, $f_s \in \mathbb{R}_+$ is the sampling frequency[5] related to the captured signals, and $c \in \mathbb{R}_+$ is the speed of sound. Note that $W(\mathbf{x}) \leq W(\bar{\mathbf{x}}) = M^2$ for any $\mathbf{x} \in \mathbb{R}^{3\times 1}$.

In this very idealized, yet instructive, setup, one can clearly see how the C-SRP method aggregates the TDoAs associated with each microphone pair in order to estimate the true source position. Indeed, (3) shows that, for each position $\mathbf{x}$ in the 3-D Euclidean space, the C-SRP method simply associates a number $W(\mathbf{x})$ that quantifies *how many actual TDoAs* (associated with the true source position $\bar{\mathbf{x}}$) *match exactly* the TDoAs computed as if the source were in position $\mathbf{x}$. In other words, counting TDoAs is the bottom line here.

2) Localizing an Acoustic Source in an Anechoic Environment—The Role of PHAT Filtering: Now, consider a more realistic type of acoustic signal $s[n]$, which can represent a speech signal, for example. In order to preserve exactly the same intuitive and useful result obtained in (3), one should modify the C-SRP objective function from (2) to the following form, which defines the C-SRP-PHAT method:

$$W_{\text{PHAT}}(\mathbf{x}) \triangleq \frac{1}{2\pi}\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\int_{-\pi}^{\pi} \Psi_{m_1,m_2}(\mathrm{e}^{\mathrm{j}\omega}) S_{m_1}(\mathrm{e}^{\mathrm{j}\omega}) \times S^*_{m_2}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{\mathrm{j}\omega k_{m_1,m_2}(\mathbf{x})}\mathrm{d}\omega, \quad (5)$$

in which the phase transform (PHAT) filter is defined as

$$\Psi_{m_1,m_2}(\mathrm{e}^{\mathrm{j}\omega}) \triangleq \frac{1}{|S_{m_1}(\mathrm{e}^{\mathrm{j}\omega})S_{m_2}(\mathrm{e}^{\mathrm{j}\omega})|}. \quad (6)$$

By using this new objective function, one arrives once more at

$$W_{\text{PHAT}}(\mathbf{x}) = \sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1} \delta[k_{m_1,m_2}(\mathbf{x}) - k_{m_1,m_2}(\bar{\mathbf{x}})]. \quad (7)$$

[5]Throughout this paper it will be assumed that the error induced by sampling is negligible, i.e., the limited time resolution does not impair the localizability of the source.

Therefore, for anechoic environments, the interpretation of the C-SRP-PHAT objective function when dealing with generic source signals is the same as that of the C-SRP associated with a spatial-temporal well-localized source signal $\delta[n]$ (i.e., counting TDoAs is the key idea).

B. H-SRP

Consider how one can implement the search for the optimal point which maximizes (5). In order to implement digitally such a process, one must somehow discretize the space, for instance, by defining a grid of 3-D points. An alternative procedure is to divide the entire search region in a finite number of 3-D compact and connected spatial regions, such as cuboids. In this case, for each microphone pair there is not necessarily only a single TDoA, but rather there may be many TDoAs associated with a given spatial region. Indeed, assuming that $\mathcal{V} \in \mathbb{V}$ is a 3-D connected spatial region, then the set $\mathcal{K}_{m_1,m_2}(\mathcal{V})$, defined as

$$\mathcal{K}_{m_1,m_2}(\mathcal{V}) \triangleq \{k \in \mathbb{Z} \,|\, k = \text{round}\{f_s\tau_{m_1,m_2}(\mathbf{x})\}, \text{ for some } \mathbf{x} \in \mathcal{V}\}, \quad (8)$$

contains all discrete TDoAs associated with the microphone pair (m_1, m_2) which are images of some point $\mathbf{x}$ within $\mathcal{V}$, with $\tau_{m_1,m_2}(\mathbf{x}) \triangleq (\|\mathbf{x}-\mathbf{x}_{m_1}\|_2 - \|\mathbf{x}-\mathbf{x}_{m_2}\|_2)/c$ standing for the continuous TDoAs.

Note that the limited temporal resolution related to the sampling process of the acquired signals induce a limited spatial resolution over the entire search region as well. Indeed, for each integer k, there are infinitely many 3-D spatial points $\mathbf{x}$ that satisfy the relation $k = \text{round}\{f_s\tau_{m_1,m_2}(\mathbf{x})\}$, which can be expressed as

$$\left(k-\frac{1}{2}\right)\frac{1}{f_s} \leq \tau_{m_1,m_2}(\mathbf{x}) < \left(k+\frac{1}{2}\right)\frac{1}{f_s}. \quad (9)$$

In other words, all points in between the hyperboloids $\tau_{m_1,m_2}(\mathbf{x}) = \left(k \pm \frac{1}{2}\right)\frac{1}{f_s}$ cannot be distinguished, since they are mapped into the same discrete TDoA k.

Therefore, following the idea of counting TDoAs, a natural modification of (7) when dealing with a volumetric region is

$$W_{\text{PHAT}}(\mathcal{V}) \triangleq \sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\sum_{k\in\mathcal{K}_{m_1,m_2}(\mathcal{V})} \delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})], \quad (10)$$

which can be rewritten, based on (5), as

$$W_{\text{PHAT}}(\mathcal{V}) \triangleq \frac{1}{2\pi}\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\sum_{k\in\mathcal{K}_{m_1,m_2}(\mathcal{V})}\int_{-\pi}^{\pi} \Psi_{m_1,m_2}(\mathrm{e}^{\mathrm{j}\omega}) \times S_{m_1}(\mathrm{e}^{\mathrm{j}\omega})S^*_{m_2}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{\mathrm{j}\omega k}\mathrm{d}\omega. \quad (11)$$

Observe that, in (7), for each spatial *point*, one accumulates the actual TDoAs which exactly match the TDoAs of *the* referred spatial point, considering all microphone pairs. Hence, what matters for electing a candidate spatial point as the source position is the number of times the hyperboloids in (9) pass through/contain that point. On the other hand, in (10), for each spatial *region*, one accumulates the actual TDoAs that exactly match with the TDoAs of *any* spatial point within the referred region, considering all microphone pairs. Thus, the same underlying idea of the C-SRP holds here: what matters for electing a spatial region as the one containing the source position is the

number of times the hyperboloids in (9) pass through that spatial region. This is an efficient way of summing up all pieces of information that the points within the region $\mathcal{V}$ convey. Therefore, the proposed localization method sets the bounding function $b(\mathcal{V})$, described in Section II, as $b(\mathcal{V}) = W_{\rm PHAT}(\mathcal{V})$ in (11). As already mentioned in Section II, this bounding function will represent the objective function of the proposed hierarchical search.

An important question that may arise from the H-SRP objective function definition in (11) is whether the sound source is contained within the region that maximizes the H-SRP-PHAT objective function or not. The answer to this question is yes, as described in the following theorem.

Theorem 1: If $\bar{\mathbf{x}} \in \bar{\mathcal{V}} \in \mathbb{V}$, then $W_{\rm PHAT}(\mathcal{V}) \leq W_{\rm PHAT}(\bar{\mathcal{V}})$, for any $\mathcal{V} \in \mathbb{V}$.

Proof: See Appendix A. ■

Theorem 1 guarantees that the source position is not lost during the search for the regions that maximize the objective function $W_{\rm PHAT}(\mathcal{V})$. It is worth highlighting that more than one region may maximize $W_{\rm PHAT}(\mathcal{V})$, but the theorem guarantees that the volume containing the source location is always a candidate to be the winning volume in an anechoic environment. This property of the proposed objective function is one of the fundamental differences between the H-SRP and the method proposed in [13].

Another important result refers to the bounding capability of the objective function in (11), which is key to the B&B process, namely: the fact that the objective function cannot increase when one passes from a given volume to one of its subsets. This result is described in the following theorem.

Theorem 2: If $\mathcal{V}_1 \subset \mathcal{V}_2 \in \mathbb{V}$, then $W_{\rm PHAT}(\mathcal{V}_1) \leq W_{\rm PHAT}(\mathcal{V}_2)$.

Proof: See Appendix B.

C. Effects of Reverberation

In order to motivate the definition of the H-SRP objective function, only anechoic signals were considered in the former discussion of the C-SRP-PHAT method. But, what occurs when reverberation is present? There is no definite answer to this question, since it depends on the degree of reverberation. Indeed, considering that the signal acquired by the mth microphone can be written as $s_m[n] = \sum_{l=0}^{L} h_m[l]s[n-l]$, in which L is the maximum delay (including those from the reflections) for the signal $s[n]$ to arrive at any microphone in the array, and $h_m[l]$ is the lth coefficient of the multipath model of the reverberation effect between the source and the mth microphone.[6] Substituting this model into (5), one gets (for the C-SRP-PHAT objective function)

$$\begin{aligned} W_{\rm PHAT}(\mathbf{x}) &= \frac{1}{2\pi}\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\int_{-\pi}^{\pi}\Psi_{m_1,m_2}(\mathrm{e}^{\mathrm{j}\omega})|S(\mathrm{e}^{\mathrm{j}\omega})|^2 H_{m_1}(\mathrm{e}^{\mathrm{j}\omega}) \\ &\quad\times H^*_{m_2}(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{\mathrm{j}\omega k_{m_1,m_2}(\mathbf{x})}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\int_{-\pi}^{\pi}\mathrm{e}^{\mathrm{j}\left[\Theta_{m_1,m_2}(\omega)-\omega k_{m_1,m_2}(\mathbf{x})\right]}\mathrm{d}\omega, \end{aligned} \quad (12)$$

[6]This model, in practice, needs to consider large values of L in order to adequately model a reverberant room.

where $\Theta_{m_1,m_2}(\omega) \triangleq \angle H_{m_1}(\mathrm{e}^{\mathrm{j}\omega}) - \angle H_{m_2}(\mathrm{e}^{\mathrm{j}\omega})$, $H_m(\mathrm{e}^{\mathrm{j}\omega})$ is the discrete-time Fourier transform of $h_m[l]$, and $\Psi_{m_1,m_2}(\mathrm{e}^{\mathrm{j}\omega})$ is as given in (6).

Note that, in general, $\Theta_{m_1,m_2}(\omega)$ is not a linear function (with modulo 2π) of the normalized frequency ω, which means that the integral in (12) *is not* equal to a simple discrete-time pulse signal. In fact, such an integral may not have a closed-form expression. However, since the aim of the C-SRP-PHAT technique is to maximize $W_{\rm PHAT}(\mathbf{x})$ and since the integral that appears in (12) is always smaller than or equal to 2π, then this approach will elect the position $\mathbf{x}$ which yields TDoAs such that the referred integral is as close to 2π as possible for as many microphone pairs as possible. Since $\Theta_{m_1,m_2}(\omega)$ depends not only on the source and microphone positions, but also on the reverberation effects that take place, then the applicability of such a beamformer may be limited, since multipath effects might hinder an accurate estimate of the delays related to the direct paths. As a consequence, the MA may not steer its beam to the correct source location. Nevertheless, C-SRP-PHAT is still quite employed to solve source-localization problems, since it is more robust to reverberation and noise effects than TDoA-based methods that employ the generalized cross-correlation (GCC) technique [2], [8].

With respect to the proposed H-SRP-PHAT objective function, the reverberation model along with (11) yield

$$W_{\rm PHAT}(\mathcal{V}) = \frac{1}{2\pi}\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\sum_{k\in\mathcal{K}_{m_1,m_2}(\mathcal{V})}\int_{-\pi}^{\pi}\mathrm{e}^{\mathrm{j}\left[\Theta_{m_1,m_2}(\omega)-\omega k\right]}\mathrm{d}\omega. \quad (13)$$

Thus, the method will elect the spatial region whose associated TDoAs maximize the integral in (13) for as many microphone pairs as possible. In other words, the method is also sensitive to reverberation effects, which are expressed in the phase $\Theta_{m_1,m_2}(\omega)$. As regards Theorems 1 and 2, they no longer hold in the context of reverberant environments.

It is worth mentioning that, when the direct path acquired signal is much stronger than the other multipath signals, then $\Theta_{m_1,m_2}(\omega)$ can be approximated by a linear function of ω, which eventually means that the result is analogous to that obtained considering an anechoic setup.

IV. Practical Considerations

In this section, practical considerations regarding the implementation of the proposed algorithm are discussed. Firstly, the branching process, i.e., the strategy to divide the search space into subregions, is presented. Then, the algorithm employed to find the time-delay bounds of (11) and (8) is described. After that, the initialization procedure of the hierarchical search is described. The section ends with an implementation summary of the proposed algorithm. It is worth mentioning that the topics addressed in this section are also new contributions of this work which are not present in [13].

A. Branching Process

In the current implementation only cuboid regions are considered. Thus, the whole search space $\mathcal{F}$ is assumed to be a cuboid

with edges[7] of sizes l_x, l_y, and $l_z \in \mathbb{R}_+$. The first branching process subdivides the initial cuboid into $D = 8$ distinct cuboids with edges of size $\frac{l_x}{2}$, $\frac{l_y}{2}$, and $\frac{l_z}{2}$, as explained in item 6) of Section II. Every subsequent branching process further divides the chosen region into 8 cuboids, each with half the edge size of the cuboids in the previous step.

B. TDoA Bounds

Given the definition of the set $\mathcal{K}_{m_1,m_2}(\mathcal{V})$ in (8), it is worth pointing out that, as $\mathcal{V} \in \mathbb{V}$ is assumed to be compact and since $\tau_{m_1,m_2}(\cdot) : \mathbb{R}^3 \longrightarrow \mathbb{R}$ is a continuous function, then the set $\{\tau_{m_1,m_2}(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{V}\}$ is a compact subset of the real line $\mathbb{R}$ and, therefore, has minimum and maximum values, $\tau^{\min}_{m_1,m_2}(\mathcal{V})$ and $\tau^{\max}_{m_1,m_2}(\mathcal{V})$, respectively. In addition, using the continuity of $\tau_{m_1,m_2}(\cdot)$ along with the assumption that $\mathcal{V}$ is also a connected set, then the image $\{f_s\tau_{m_1,m_2}(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{V}\}$ is an interval of the real line, i.e., $\{f_s\tau_{m_1,m_2}(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{V}\} = [f_s\tau^{\min}_{m_1,m_2}(\mathcal{V}), f_s\tau^{\max}_{m_1,m_2}(\mathcal{V})] \subset \mathbb{R}$. These facts eventually imply that $\mathcal{K}_{m_1,m_2}(\mathcal{V}) = \{k^{\min}_{m_1,m_2}(\mathcal{V}), k^{\min}_{m_1,m_2}(\mathcal{V}) + 1, \ldots, k^{\max}_{m_1,m_2}(\mathcal{V})\}$, in which $k^{\min}_{m_1,m_2}(\mathcal{V}) = \text{round}\{f_s\tau^{\min}_{m_1,m_2}(\mathcal{V})\}$ and $k^{\max}_{m_1,m_2}(\mathcal{V}) = \text{round}\{f_s\tau^{\max}_{m_1,m_2}(\mathcal{V})\}$. The aim of this subsection is, therefore, to show how one can determine the continuous TDoA bounds $\tau^{\min}_{m_1,m_2}(\mathcal{V})$ and $\tau^{\max}_{m_1,m_2}(\mathcal{V})$, for a given microphone pair (m_1, m_2) and for a predefined cuboid region $\mathcal{V}$.

In order to determine the minimum value of $\tau_{m_1,m_2}(\mathbf{x})$, with $\mathbf{x} \in \mathcal{V}$, one may solve the following optimization problem:[8]

$$\min_{[x\ \ y\ \ z]^T \in \mathbb{R}^{3\times 1}} \{\tau_{m_1,m_2}(x,y,z)\} \tag{14}$$

$$\text{s.t.}: \quad \delta^{\min}_x(\mathcal{V}) \le x \le \delta^{\max}_x(\mathcal{V}), \tag{15}$$

$$\delta^{\min}_y(\mathcal{V}) \le y \le \delta^{\max}_y(\mathcal{V}), \tag{16}$$

$$\delta^{\min}_z(\mathcal{V}) \le z \le \delta^{\max}_z(\mathcal{V}), \tag{17}$$

in which the limits of the inequality constraints are real numbers denoting the bounds of the edges of the cuboid $\mathcal{V}$. The solution to such optimization problem is described in Theorem 3.

Theorem 3: The minimum value $\tau^{\min}_{m_1,m_2}(\mathcal{V}) \in \mathbb{R}$ of the set $\{\tau_{m_1,m_2}(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{V}\}$, where $\mathcal{V}$ is a cuboid, is

$$\tau^{\min}_{m_1,m_2}(\mathcal{V}) = \min_{[x\ \ y\ \ z]^T \in \mathcal{S}_{m_1,m_2}(\mathcal{V})} \{\tau_{m_1,m_2}(x,y,z)\}, \tag{18}$$

in which $\mathcal{S}_{m_1,m_2}(\mathcal{V})$ is a finite set containing at most 26 points of the Euclidean space, being defined as

$$\mathcal{S}_{m_1,m_2}(\mathcal{V}) \triangleq \mathcal{S}^{\text{f}}_{m_1,m_2}(\mathcal{V}) \cup \mathcal{S}^{\text{e}}_{m_1,m_2}(\mathcal{V}) \cup \mathcal{S}^{\text{v}}_{m_1,m_2}(\mathcal{V}), \tag{19}$$

where $\mathcal{S}^{\text{f}}_{m_1,m_2}(\mathcal{V})$ is a (possibly empty) set containing at most 6 candidate solutions on the faces of $\mathcal{V}$, $\mathcal{S}^{\text{e}}_{m_1,m_2}(\mathcal{V})$ is a (possibly empty) set containing at most 12 candidate solutions in the edges of $\mathcal{V}$, and $\mathcal{S}^{\text{v}}_{m_1,m_2}(\mathcal{V})$ is the set containing 8 candidate solutions which are the vertexes of $\mathcal{V}$.

Proof: See Appendix C. ∎

The analytic definitions of the sets $\mathcal{S}^{\text{f}}_{m_1,m_2}(\mathcal{V})$ and $\mathcal{S}^{\text{e}}_{m_1,m_2}(\mathcal{V})$ are clearly stated along the proof of Theorem 3 in Appendix C. Although those definitions are not essential for the forthcoming developments, they are important to justify a simplification that is adopted in this paper, described in Corollary 1.

[7]For the sake of simplicity, it will be assumed that each edge of the cuboid is parallel to one of the axis of the Cartesian coordinate system.

[8]The formulation is analogous for the maximization problem.

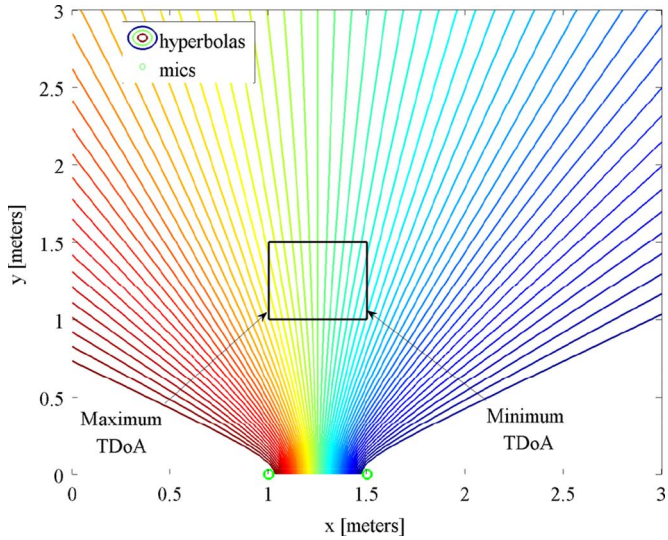

Fig. 2. Figure illustrating for 2-D space that the maximum and minimum TDoAs inside a square region will be located in its vertices.

Corollary 1: The minimum value $\tau^{\min}_{m_1,m_2}(\mathcal{V}) \in \mathbb{R}$ of the set $\{\tau_{m_1,m_2}(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{V}\}$, where $\mathcal{V}$ is a cuboid, can be approximated as

$$\tau^{\min}_{m_1,m_2}(\mathcal{V}) \approx \min_{[x\ \ y\ \ z]^T \in \mathcal{S}^{\text{v}}_{m_1,m_2}(\mathcal{V})} \{\tau_{m_1,m_2}(x,y,z)\}, \tag{20}$$

in which $\mathcal{S}^{\text{v}}_{m_1,m_2}(\mathcal{V})$ is the set containing 8 candidate solutions which are the {vertices} of $\mathcal{V}$.

Proof: As claimed in Appendix C, the sets $\mathcal{S}^{\text{f}}_{m_1,m_2}(\mathcal{V})$ and $\mathcal{S}^{\text{e}}_{m_1,m_2}(\mathcal{V})$ will be empty for most volumes $\mathcal{V} \in \mathbb{V}$, since the solutions are on a face or in an edge only if very specific symmetry conditions are satisfied (see Appendix 3 for further details). In practice, those symmetry conditions are not satisfied for most regions $\mathcal{V}$. ∎

In order to facilitate visualization, Fig. 2 depicts a 2-D representation illustrating how TDoAs and volumes are typically related for a given microphone pair. Actually, since it is a 2-D representation, cuboids and hyperboloids map to squares and hyperbolas, where each hyperbola should be understood as a collection of points that lead to the same TDoA value. The TDoA values gradually increase from right (dark-blue hyperbola) to left (brown hyperbola). Observe that the minimum TDoA $\tau^{\min}_{m_1,m_2}(\mathcal{V})$ and the maximum TDoA $\tau^{\max}_{m_1,m_2}(\mathcal{V})$ are, indeed, at the vertices of the square. Note that the same comments would still be valid if the square were in a different position.

Therefore, the TDoA bounds for a given volume and a given microphone pair can be approximated by the maximum and minimum TDoA values associated with the vertices of the related volume. It is worth mentioning that one can determine $\tau^{\max}_{m_1,m_2}(\mathcal{V})$ following the same steps previously described, but replacing the "min" operator with the "max" operator in the related expressions.

C. Hierarchical Search Initialization

In this subsection, a comparison between the computational complexity of a hierarchical search and an equivalent full (ex-

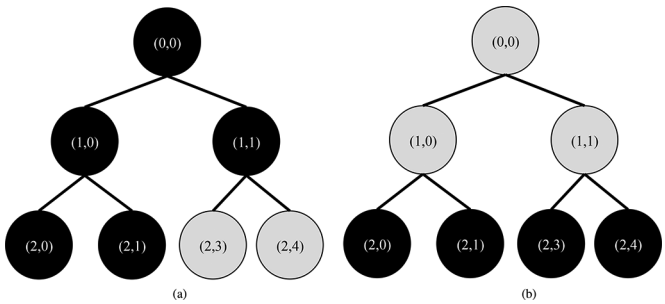

Fig. 3. Example denoting the minimum number of nodes that need to be visited (black circles) for the hierarchical search and the full-grid search, assuming that the node (2, 0) has the largest bounding value. (a) Hierarchical search. (b) Full-grid Search.

haustive) search is made. This comparison looks into the computational complexity of the hierarchical search, showing that, under certain circumstances, it might be more efficient to start the search from a subdivision of the search region.

In general, the computational complexity of the H-SRP method can be decomposed into two parts: one associated with the calculation of the cross-correlations between the signals of each microphone pair[9] and another arising from the calculation of the bounding function for each spatial region visited during the search. In the following discussion, focus will be put on the cost of computing the bounding function.

Assuming that the cross-correlations have been computed for all possible TDoAs of all microphone pairs, the cost of computing the bounding value for a given spatial region $\mathcal{V}$ is proportional to the number of terms in the summations of (13). In practice, this number will depend on the number of microphones and the number of discrete TDoAs in the set $\mathcal{K}_{m_1,m_2}(\mathcal{V})$, i.e., its cardinality.

The hierarchical search employed starts by computing the bounding value of the whole search space $\mathcal{F}$. Since the cardinality of $\mathcal{K}_{m_1,m_2}(\mathcal{F})$ cannot be smaller than that of any other spatial region $\mathcal{V} \subsetneq \mathcal{F}$ for the same microphone pair (see Theorem 2), the cost of computing the bounding value of $\mathcal{F}$ will be no less than that of any other spatial region. In fact, the computational complexity of computing the bounding value for a certain region $\mathcal{V}_i$ is no less than that of computing the bounding value for a region $\mathcal{V}_j$, if $\mathcal{V}_j \subset \mathcal{V}_i$. This comes from the fact that all delays, for the same microphone pair, present in $\mathcal{V}_j$ are also present in $\mathcal{V}_i$. Hence, the number of operations required to compute the bounding function grows with the region size.

In order to study the computational complexity of the hierarchical search, its tree representation will be employed, where the jth node of the ith level of the tree will be denoted as $\mathcal{N}_{i,j}$. Considering the discussion in the previous paragraph, it can be seen that the number of operations required to compute the bound of a given node is always smaller than or equal to that of its parent node. One important question that must be answered, then, is if it is computationally efficient to select a node with highest bound through the hierarchical search algorithm or through a full search among all nodes at a given level of the tree. As an illustration, Fig. 3(a) shows the minimum number of nodes that would have to be visited during the hierarchical search in order to select node $\mathcal{N}_{2,0}$ as having the largest bound, whereas Fig. 3(b) shows which nodes would have to be considered in the case of a full search. Comparing both diagrams in this figure, one can easily see that the full search is more computationally efficient, since the hierarchical search computes nodes near the root node, which are likely to have more TDoAs and therefore a larger computational complexity. On the other hand, as the tree grows, more nodes could be ignored by the hierarchical algorithm, thus overcoming the cost of the first-level nodes.

In practical terms, the hierarchical search can be applied only from a level in which its best-case computational cost[10] is lower than that of the full search. This scheme allows one to avoid the computation of large-cost nodes. This initial level can be chosen offline, by calculating the number of operations required to compute the bounding value[11] of each node in the tree and the cost associated with the best-case hierarchical search and to the full search. Once the initial level is chosen, the hierarchical search can be applied. This way, the computation of the high-cost bounds of the first levels of the tree are avoided, while the hierarchical search can still avoid several branches of the tree, thus reducing the overall complexity. Algorithmically, the initialization can be performed by attributing the spatial regions and bounding values for all nodes of the initial level to the list $\mathbf{L}$ instead of the corresponding quantities for the whole search space $\mathcal{F}$ in the algorithm described in Section II. This way, all properties of the hierarchical search algorithm are preserved, including the convergence to a region containing the source position (in an anechoic environment).

D. Implementation Summary

In this subsection, the H-SRP algorithm is summarized. The first step of the algorithm consists of finding the level $I \in \mathbb{N}$ from which the hierarchical search will start, as explained in the last paragraph of Subsection IV.C. This can be determined as soon as the microphone positions and initial search space $\mathcal{F}$ are known. The algorithm then executes the following steps for each signal frame:

1) Compute the generalized cross-correlation function $\mathcal{R}_{m_i,m_j}[k]$ for each microphone pair using the PHAT filter;[12]
2) Subdivide the search region $\mathcal{F}$ into $8^I = 2^I \times 2^I \times 2^I$ parts (each dimension is subdivided 2^I times) as described in Section IV.A, without computing the bounding function. For each volume find its minimum and maximum delays for each microphone pair[13] using the method described in Section IV.B;
3) Compute the bounding function for each volumetric region in level I through the expression:

$$W_{\mathrm{PHAT}}(\mathcal{V}) = \sum_{m_1=0}^{M-1} \sum_{m_2=0}^{M-1} \sum_{k \in \mathcal{K}_{m_1,m_2}(\mathcal{V})} \mathcal{R}_{m_1,m_2}[k]; \quad (21)$$

4) Start the hierarchical search described in Section IV.A. For each new volume visited during the search algorithm, com-

[9]Usually performed in the frequency domain using a fast Fourier transform (FFT) algorithm.

[10]In the sense that the least number of nodes are evaluated.

[11]As mentioned previously, this number depends only on the initial search space $\mathcal{F}$ and the positions of each microphone.

[12]As mentioned before, the cross-correlation is efficiently computed in the frequency domain by the use of FFT algorithms.

[13]The bounds for each volume can also be pre-computed and stored in look-up tables to reduce computational complexity.

pute its maximum and minimum delays as described in Section IV.B and its bounding function through (21).

In each frame, at the end of the search algorithm, the region with "size" $V_{\min}$ that expectedly contains the source is selected; the source position is then chosen as the center of the winning volume.

So far, the term "size" has been intentionally left without a proper definition. Depending on the context, such term can have different meanings, such as the edges or the main diagonal of the cuboid. When presenting the results in Section VI, "size" means volume, i.e., $V_{\min}$ is the smallest volume that a cuboid can assume.

V. Comparison With a Previous Work

In [16], the authors describe a modification of the C-SRP algorithm in which several TDoAs associated with a certain region in space around a single point are taken into account. The objective of the method is to define a certain area of influence around each point in space, find a set of delays associated with this region, and then evaluate the modified SRP (hereafter called M-SRP) function for this point. Mathematically, this function is equal to

$$W_{\mathrm{MOD}}(\mathbf{x}) = \sum_{m_1=0}^{M-2} \sum_{m_2=m_1+1}^{M-1} \sum_{k=\hat{K}^{\min}_{\mathbf{x},m_1,m_2}}^{\hat{K}^{\max}_{\mathbf{x},m_1,m_2}} \mathcal{R}_{m_1,m_2}[k]. \quad (22)$$

In which the values $\hat{K}^{\min}_{\mathbf{x},m_1,m_2}$ and $\hat{K}^{\max}_{\mathbf{x},m_1,m_2}$ are found by first considering a cube centered on $\mathbf{x}$ and then performing a linear approximation to find the minimum and maximum delay values inside this cube, considering microphones m_1 and m_2.

If one rewrites (21) as

$$W_{\mathrm{PHAT}}(\mathcal{V}) = 2 \sum_{m_1=0}^{M-2} \sum_{m_2=m_1+1}^{M-1} \sum_{k \in \mathcal{K}_{m_1,m_2}(\mathcal{V})} \mathcal{R}_{m_1,m_2}[k] + \sum_{m=0}^{M-1} \mathcal{R}_{m,m}[0], \quad (23)$$

then it is possible to notice the similarity between the M-SRP objective function and the proposed bounding function. Both methods take into account a set of delays in order to estimate a certain region. Nevertheless, the possible regions are different: cubes for the M-SRP and cuboids for the proposed H-SRP. Also, the maximum and minimum delays are only approximated in (22) whereas for (23) they are exact if some (mild) conditions hold (see Section IV.B). Overall, these differences allow the use of the bounding function in the hierarchical search, whereas (22) is more appropriate for use in exhaustive searches, as was its original purpose.[14]

VI. Experiments

In this section, two experiments are described. The first experiment uses simulated signals and shows how the parameters of the hierarchical search influence its performance under different acoustic conditions. The second experiment uses recorded signals and compares the performance of the hierarchical search with other methods found in the literature.

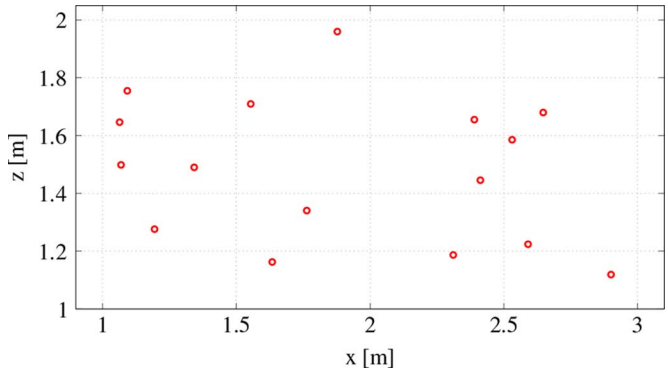

Fig. 4. Positioning of the microphones (circles) used in the experiment with simulated signals.

A. Evaluation Using Simulated Signals

The objective of this experiment using simulated signals is to understand how the choice of $V_{\min}$ affects the performance of the hierarchical search, especially the tradeoff between localization accuracy and number of operations performed. The hierarchical search is also tested under different reverberation conditions in order to show that, even though its theoretical formulation has been based on an anechoic scenario, it can also be employed under more realistic conditions.

1) Experimental Setup: The simulation considered three rooms sized 4 m × 6 m × 3 m with different acoustic characteristics. The first is an anechoic room, whereas the second and third have reverberation times [18] (T_{60}) of 250 and 500 ms, respectively.[15] As previously mentioned, these parameter choices allow the study of the algorithm in the idealized (anechoic) acoustic scenario used in its development as well as the impact of reverberation on its performance.

Similarly to the arrangement in [2], this experiment employed a planar[16] array with 16 microphones randomly placed in a 2 m × 1 m rectangular region, as illustrated in Fig. 4. In the simulations, the array was supposed to be placed on one of the room walls (which, according to the adopted convention, implies that all microphones have coordinate $y = 0$).

The source signal was a 1-s long female speech signal pre-processed by a voice-activity detector (VAD) to remove any silent intervals. This step guarantees that there is always a speech signal impinging on the microphones. In practice, the VAD step would have been executed after the acquisition of the signal by the microphones. The speech signal was sampled at 48 kHz with 24 bits of precision. The speech source was simulated at 5 randomly chosen positions, shown in Table I. Such positions were chosen so as to be representative of the region of interest regarding its discriminability index (DI) [20] range. The DI is a numerical measure of the capability of a

[14]While this paper was under review, a modified version of [16] which includes an iterative search appeared in [17]. The main feature of the present paper is the proposal of a hierarchical procedure with proved convergence under ideal conditions, which makes it reliable for use under not much demanding practical conditions. Since [17] follows a different path and also modifies the objective function of [16], the authors considered that the consequent content enhancement would not justify its inclusion among the experiments of this paper, at this point. Of course, further continuations of this work will include detailed comparison with [17].

[15]These acoustic characteristics were simulated using the image method [19] implemented in the Audio Systems Array Processing Toolbox available at http://www.engr.uky.edu/~donohue/audio/Arrays/MAToolbox.htm.

[16]Although it could be argued that a 3-D configuration would be better suited for SSL algorithms in general, we selected a planar array in this experiment because this simple arrangement is more suitable to practical systems.

TABLE I
COORDINATES OF THE 5 SIMULATED SOURCE POSITIONS

Pos.	Coordinates [m]
1	[2.92, 2.18, 1.64]
2	[2.09, 1.01, 1.88]
3	[1.28, 2.13, 1.88]
4	[1.30, 0.99, 1.69]
5	[1.52, 2.36, 1.78]

microphone array to distinguish a given point in space from its neighbors that is influenced mainly by the array geometry and the spatial region under consideration. Specifically, the 5 positions encompass the DI range of 30% to 90% found for the region from which they were picked up.

2) Algorithm Setup: The hierarchical search algorithm was used to estimate the source position in successive 4096-sample long frames with 50% of overlap. The PHAT pre-filter was employed in the estimation of the cross-correlation function for every microphone pair. As regards the configuration of the search algorithm itself, it was initialized with $I = 3$, and 6 different values for $V_{\min}$ ($V_0/2^{3i}$ for $0 \leq i \leq 5$, with $V_0 = 25.0 \times 37.5 \times 18.8$ cm) were tested in an attempt to quantify the compromise between the number of operations performed and the localization accuracy. The initial search space $\mathcal{F}$ was the whole room.

3) Figures of Merit: The localization error for each frame is measured as the Euclidean distance between actual and estimated source positions. In order to obtain an overall value for the localization performance, the median value of the error is calculated across all different frames and different positions. One median error value is computed for each chosen $V_{\min}$.

The computation of the number of operations performed is based on the number of evaluations of (21). Since the search algorithm is executed in a frame-by-frame basis, the number of operations can vary from frame to frame. Then, for each frame and source position, the total number of summations in each evaluation of (21) is stored, and an overall measure of complexity is obtained by averaging this number of summations over all frames and source positions.

4) Results: Fig. 5 shows three curves depicting the different results of median error and number of operations found for the 6 chosen values of $V_{\min}$ for the three different rooms. As can be gathered from the figure, for a given acoustic configuration, as the size of $V_{\min}$ is reduced, the error decreases with an associated increase in the number of operations performed. Increasing the reverberation time, on the other hand, increases the error for a given number of operations. Hence, it appears that the more challenging the acoustic scenario is, the smaller $V_{\min}$ should be in order to attain a given performance, at the cost of an increase in the necessary number of operations. Nevertheless, the fact that the error decreases with $V_{\min}$, even under reverberant conditions, validates the method. The minimum median error achieved by the hierarchical search was 1.5 cm for both the anechoic room and the room with $T_{60} = 250$ ms, and 2.5 cm for the room with $T_{60} = 500$ ms.

B. Evaluation Using Recorded Signals

In this experiment, a recording was made in order to obtain the signals at each microphone. In this more realistic scenario,

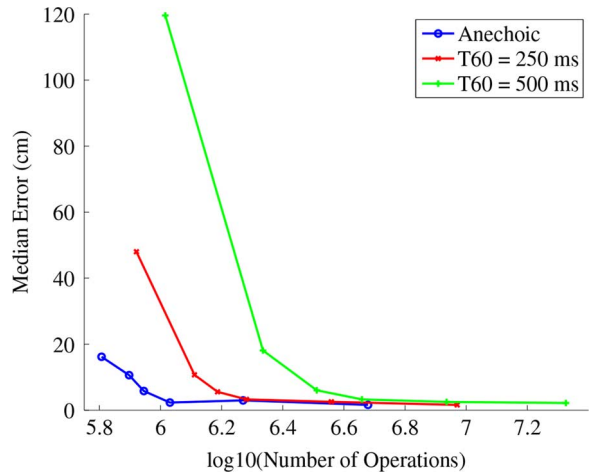

Fig. 5. Results obtained for the experiment with the simulated signals. All results refer to the hierarchical search method with varying values of $V_{\min}$ (the leftmost point corresponds to V_0 in all plots).

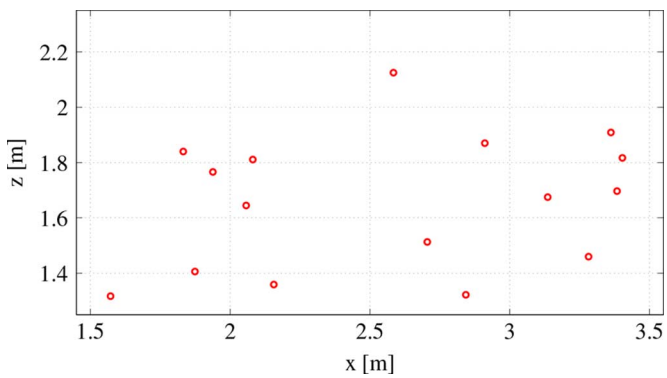

Fig. 6. Positioning of the microphones (circles) used in the experiment with simulated signals.

the objective is to compare the localization accuracy of the hierarchical search with that of other methods found in the literature. The methods chosen were the classical SRP (employing a full-grid search), the modified SRP (M-SRP) described in [16] (which also employs a full-grid search), and the C-SRP with stochastic region contraction (SRC) [10].

1) Experimental Setup: The recording took place in an acoustically-treated presentation room located at the Federal University of Rio de Janeiro. The room dimensions are 5.2 m $\times$ 7.5 m $\times$ 2.6 m and its measured reverberation time is approximately 500 ms.

The array employed omnidirectional high-sensitivity microphones.[17] The array geometry is planar as in the previous experiment, this time with the microphones placed as shown in Fig. 6. The MA was placed on one of the walls of the room, i.e., $y = 0$ for all microphones.

The source signal was a 5-s long silence-free female speech signal sampled at 48 kHz with 24 bits of precision. Loudspeakers with a diameter of 6 cm[18] were employed as sound sources. Fig. 7 shows a top-down view of the 9 different source

[17]Model SMK 4060 by DPA.

[18]By Orb Audio.

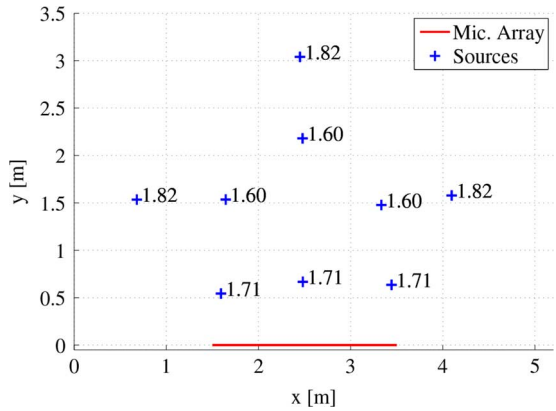

Fig. 7. Top-down view of the sources in the experiment with recorded signals. The number next to each source position represents the height of the source in meters.

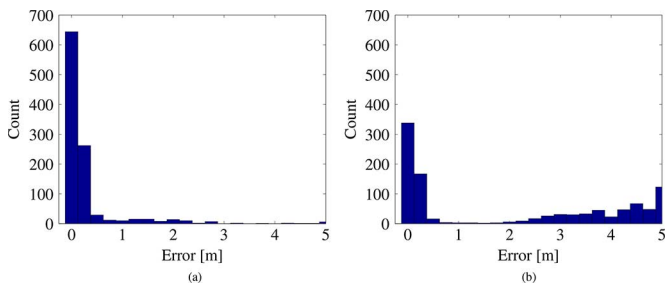

Fig. 8. Histograms of the errors committed for each frame and each source position by both the hierarchical search and SRC algorithms. (a) Hierarchical search. (b) SRC.

positions recorded. Regardless of the source position, the loudspeaker was always facing the microphone array.

2) Algorithm Setup: All of the evaluated algorithms estimate the source position in successive 4096-sample long frames with 50% of overlap, and employed the PHAT pre-filter when computing cross-correlations.

The individual parameters of each search method were chosen so as to yield a similar number of operations. The C-SRP algorithm was set to operate over a grid with a spacing between adjacent points equal to 10 cm. The M-SRP method employed a grid with a spacing of 20 cm. The SRC method used 9000 points per iteration, selecting the 300 best points. The hierarchical search used a minimum volume of size 1.0 cm $\times$ 1.5 cm $\times$ 0.5 cm and was initialized with $I = 3$. All methods searched for the source position in the whole room.

3) Figures of Merit: The localization performance of each method was assessed by the median of the Euclidean distances between actual and estimated source positions across all frames and positions.

The number of operations, on the other hand, considered the number of summations performed by each evaluation of the objective function associated with each method. For all methods, the computational cost of the cross-correlation function for each microphone pair was not considered. Also, computational cost of the TDoA associated with each position in space (for the C-SRP and SRC algorithms) or the TDoA bounds (for the hierarchical method and M-SRP) were not considered.[19] For the C-SRP and M-SRP, which employ a full-grid search, the number of summations per frame is used since it does not vary from frame to frame. For the SRC and hierarchical search, the number of operations per frame varies, hence the average number of operations per frame is employed.

TABLE II
EXPERIMENTAL RESULTS

Algorithm	Median Error [cm]	$\log_{10}$(Number of operations)
H-SRP	7.2	7.1
C-SRP	15.5	7.1
M-SRP	18.9	7.0
SRC	22.3	7.1

4) Results: Table II shows the median error and number of operations[20] obtained for each method. By design, the number of operations for each method is similar; under these conditions, the median error varies significantly.[21] Overall, the best performance was achieved by the proposed hierarchical search method, which attained an error equal to less than half of the error of the next best method. It should be noted that the loudspeaker membrane itself has a diameter of approximately 6 cm, which induces an uncertainty on the exact position of the source signals.

When one compares the two methods that do not perform full-grid searches, the hierarchical search and the SRC, one can see the advantages of having a method that deterministically searches for the source position. Fig. 8 shows the histograms depicting the error distributions for the hierarchical search and the SRC methods for all frames and all source positions. As can be seen, the main difference between both methods is the small number of erroneous estimates found in the hierarchical search, indicating that it usually selects the global maximum associated with the source position. On the other hand, the SRC seems to quite often elect positions that are far away from the actual source position.

VII. CONCLUDING REMARKS

In this paper, a novel sound source localization method, called H-SRP, is proposed. The method relies on measurements corresponding to the acoustic activity inside a given volumetric region and employs a hierarchical search method inspired by the branch-and-bound paradigm. It is proved that the proposed method finds a volume with a given minimum size (stop criterion) that contains the global maximum targeted by the C-SRP method in an anechoic and single-source scenario. Tests in simulated and practical reverberant rooms have shown that the proposed method can be successfully applied in more realistic scenarios. Moreover, the results indicate that the H-SRP can pro-

[19]In practice, these values can be efficiently pre-computed [21], since the microphone array geometry and search space are usually fixed.

[20]For example, the C-SRP performs $(\frac{5.2+0.1}{0.1}) \times (\frac{7.5+0.1}{0.1}) \times (\frac{2.6+0.1}{0.1}) = 108,756$ functional evaluations requiring $\binom{16}{2} - 1 = 119$ operations (additions) each, thus yielding a total of $12,941,964 \approx 10^{7.1}$ numerical operations.

[21]The equivalence by the number of operations was chosen because one usually wants to know what is the best performance achievable given a computational "budget". Moreover, in order to equate the performances of the methods, a prohibitive number of operations would be required by the C-SRP.

vide an accurate position estimate with fewer numerical operations than competing localization methods. Finding a sufficient condition for guaranteeing the convergence to a region containing the source position in reverberant environments, and performing extensive tests with different microphone array geometries and speech signals are interesting topics for future research.

Appendix A
Proof of Theorem 1

Assume that there exists $\mathcal{V}$ such that $W_{\mathrm{PHAT}}(\mathcal{V}) > W_{\mathrm{PHAT}}(\bar{\mathcal{V}})$. This means that

$$\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\sum_{k\in\mathcal{K}_{m_1,m_2}(\mathcal{V})}\delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})] > \sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\sum_{k\in\mathcal{K}_{m_1,m_2}(\bar{\mathcal{V}})}\delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})]. \quad (24)$$

Now, for each fixed pair of microphones (m_1, m_2), one has that $\delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})]$ can only be nonzero for *at most one* k within $\mathcal{K}_{m_1,m_2}(\mathcal{V})$. This implies that

$$W_{\mathrm{PHAT}}(\mathcal{V}) = \sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\underbrace{\sum_{k\in\mathcal{K}_{m_1,m_2}(\mathcal{V})}\delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})]}_{\leq 1} \leq M^2. \quad (25)$$

On the other hand, as $\bar{\mathbf{x}} \in \bar{\mathcal{V}}$, then $\delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})]$ will be nonzero for *exactly one* k within $\mathcal{K}_{m_1,m_2}(\bar{\mathcal{V}})$. This fact yields

$$W_{\mathrm{PHAT}}(\bar{\mathcal{V}}) = \sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\sum_{k\in\mathcal{K}_{m_1,m_2}(\bar{\mathcal{V}})}\delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})] = M^2, \quad (26)$$

thus arriving at a contradiction with (24).

Appendix B
Proof of Theorem 2

Theorem 2 holds since $\mathcal{K}_{m_1,m_2}(\mathcal{V}_1) \subset \mathcal{K}_{m_1,m_2}(\mathcal{V}_2)$, i.e., all discrete TDoAs that are associated with region $\mathcal{V}_1$ are also associated with region $\mathcal{V}_2$, thus implying that

$$\begin{aligned}\sum_{k\in\mathcal{K}_{m_1,m_2}(\mathcal{V}_2)}&\delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})]\\ &= \sum_{k\in\mathcal{K}_{m_1,m_2}(\mathcal{V}_1)}\delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})]\\ &\quad+\underbrace{\sum_{k\in\mathcal{K}_{m_1,m_2}(\mathcal{V}_2\setminus\mathcal{V}_1)}\delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})]}_{\geq 0}\\ &\geq \sum_{k\in\mathcal{K}_{m_1,m_2}(\mathcal{V}_1)}\delta[k-k_{m_1,m_2}(\bar{\mathbf{x}})], \end{aligned} \quad (27)$$

which, after summing over all microphone pairs, leads to $W_{\mathrm{PHAT}}(\mathcal{V}_2) \geq W_{\mathrm{PHAT}}(\mathcal{V}_1)$.

Appendix C
Proof of Theorem 3

From the discussions on the theorem stated in [16], at least one of the inequality constraints must be active in the optimization problem that defines $\tau_{m_1,m_2}(\mathbf{x})$ (see expressions (14), (15), (16), and (17)), since the minimum/maximum values of $\tau_{m_1,m_2}(\mathbf{x})$ are achieved at the boundary surface of $\mathcal{V}$. Therefore, one must have $x = \text{constant}$ and/or $y = \text{constant}$ and/or $z = \text{constant}$.

Let $\mathcal{V} \in \mathbb{V}$ be a fixed region. For the sake of notational simplicity, the dependency on $\mathcal{V}$ will be omitted in the forthcoming derivations. Due to the symmetry of the coordinate variables that appear in the definition of the TDoA $\tau_{m_1,m_2}(x,y,z)$, one can assume that $z = \delta_z \in \{\delta_z^{\min}, \delta_z^{\max}\}$ is a fixed real number, without loss of generality. This means that

$$\begin{aligned} c\tau_{m_1,m_2}(x,y,\delta_z) &= \sqrt{(x-x_{m_1})^2+(y-y_{m_1})^2+(\delta_z-z_{m_1})^2}\\ &\quad-\sqrt{(x-x_{m_2})^2+(y-y_{m_2})^2+(\delta_z-z_{m_2})^2}. \end{aligned} \quad (28)$$

Now, assume that there is only one active constraint (in this case, $z = \delta_z$), i.e., the candidate solution is on a face. Thus, by defining $f(x,y) = c\tau_{m_1,m_2}(x,y,\delta_z)$, then the application of the Karush-Kuhn-Tucker (KKT) conditions [22] leads to

$$\nabla f(x_0, y_0) = 0, \quad (29)$$

where $\delta_x^{\min} < x_0 < \delta_x^{\max}$ and $\delta_y^{\min} < y_0 < \delta_y^{\max}$.

Since

$$\nabla f(x,y) = \left(\frac{x-x_{m_1}}{\|\mathbf{x}^z-\mathbf{x}_{m_1}\|}-\frac{x-x_{m_2}}{\|\mathbf{x}^z-\mathbf{x}_{m_2}\|}, \frac{y-y_{m_1}}{\|\mathbf{x}^z-\mathbf{x}_{m_1}\|}-\frac{y-y_{m_2}}{\|\mathbf{x}^z-\mathbf{x}_{m_2}\|}\right), \quad (30)$$

with $\mathbf{x}^z \triangleq [x \;\; y \;\; \delta_z]^T$, then the optimal point (x_0, y_0) must satisfy[22]

$$\frac{x_0-x_{m_2}}{x_0-x_{m_1}} = \frac{y_0-y_{m_2}}{y_0-y_{m_1}} = \frac{\|\mathbf{x}^z-\mathbf{x}_{m_2}\|}{\|\mathbf{x}^z-\mathbf{x}_{m_1}\|} > 0. \quad (31)$$

The former equation implies that

$$\begin{aligned}\frac{(x_0-x_{m_2})^2}{(x_0-x_{m_1})^2} &= \frac{(y_0-y_{m_2})^2}{(y_0-y_{m_1})^2}\\ &= \frac{(x_0-x_{m_2})^2+(y_0-y_{m_2})^2+(\delta_z-z_{m_2})^2}{(x_0-x_{m_1})^2+(y_0-y_{m_1})^2+(\delta_z-z_{m_1})^2}\\ &= \frac{(\delta_z-z_{m_2})^2}{(\delta_z-z_{m_1})^2}, \end{aligned} \quad (32)$$

yielding

$$\frac{x_0-x_{m_2}}{x_0-x_{m_1}} = \frac{y_0-y_{m_2}}{y_0-y_{m_1}} = \left|\frac{\delta_z-z_{m_2}}{\delta_z-z_{m_1}}\right| \triangleq w^z, \quad (33)$$

[22]Assuming that $x_0 \neq x_{m_1}$, $y_0 \neq y_{m_1}$, and $\mathbf{x}_{m_2} \neq \mathbf{x}^z \neq \mathbf{x}_{m_1}$.

which is a known (δ_z, z_{m_2}, and z_{m_1} are given), fixed, and nonnegative real number. Therefore, assuming $w^z \neq 1$, one has

$$(x_0, y_0, z_0) = \left(\frac{x_{m_2} - w^z x_{m_1}}{1 - w^z}, \frac{y_{m_2} - w^z y_{m_1}}{1 - w^z}, \delta_z \right), \quad (34)$$

in which one must necessarily check if $\delta_x^{\min} < x_0 < \delta_x^{\max}$ and $\delta_y^{\min} < y_0 < \delta_y^{\max}$. In this case, one has $(x_0, y_0, z_0) \in \mathcal{S}^{\mathrm{f}}_{m_1,m_2}$. However, if the solutions x_0 and y_0 do not satisfy the former inequality constraints, then $(x_0, y_0, z_0) \notin \mathcal{S}^{\mathrm{f}}_{m_1,m_2}$. It is worth highlighting that, if $\mathrm{sign}\{\delta_z - z_{m_2}\} = \mathrm{sign}\{\delta_z - z_{m_1}\}$, then the points $(x_0, y_0, z_0), (x_{m_1}, y_{m_1}, z_{m_1})$, and $(x_{m_2}, y_{m_2}, z_{m_2})$ are co-linear, which gives us a hint that the solutions on a face seem to be rare, only occurring for a few volumes $\mathcal{V} \in \mathbb{V}$.

Note that, if $w^z = 1$, then $\delta_z = \frac{z_{m_2}+z_{m_1}}{2}$ and $\frac{x_0 - x_{m_2}}{x_0 - x_{m_1}} = \frac{y_0 - y_{m_2}}{y_0 - y_{m_1}} = 1$, which implies that $x_{m_2} = x_{m_1}$ and $y_{m_2} = y_{m_1}$.[23] In this case (which may occur, for instance, in uniform linear arrays aligned with the z-axis), any $\delta_x^{\min} < x_0 < \delta_x^{\max}$ and $\delta_y^{\min} < y_0 < \delta_y^{\max}$ can be taken as solutions. Nevertheless, this possibility may be regarded as only existing in theory, since just few array geometry could satisfy it and, besides that, there is always an uncertainty about the true position of the microphones (they are not ideal points), so that having $x_{m_2} = x_{m_1}$ and $y_{m_2} = y_{m_1}$ is virtually impossible to occur in practice. Moreover, one can always avoid these conditions when building the microphone array and choosing the search region.

Everything performed considering $z = \delta_z$ must also be done for $x = \delta_x$ and $y = \delta_y$. The resulting set $\mathcal{S}^{\mathrm{f}}_{m_1,m_2}$ will therefore contain at most 6 spatial points whose related TDoAs must be evaluated. If the candidate solutions previously found do not satisfy the related inequality constraints ($\mathcal{S}^{\mathrm{f}}_{m_1,m_2} = \emptyset$), then there is no solution with only one active constraint.

Consider now the case in which only two inequality constraints are active, which means that the optimal point is on the edge of the rectangular parallelepiped. Without loss of generality, assume that one has $y = \delta_y \in \{\delta_y^{\min}, \delta_y^{\max}\}$ and $z = \delta_z \in \{\delta_z^{\min}, \delta_z^{\max}\}$. Then, by applying the KKT conditions to $f(x) = c\tau_{m_1,m_2}(x, \delta_y, \delta_z)$ so that $\nabla f(x_0) = 0$, one arrives at

$$\frac{x_0 - x_{m_2}}{x_0 - x_{m_1}} = \sqrt{\frac{(\delta_y - y_{m_2})^2 + (\delta_z - z_{m_2})^2}{(\delta_y - y_{m_1})^2 + (\delta_z - z_{m_1})^2}} \triangleq w^{y,z}, \quad (35)$$

which is a known, fixed, and nonnegative real number. If $w^{y,z} \neq 1$, one has

$$(x_0, y_0, z_0) = \left(\frac{x_{m_2} - w^{y,z} x_{m_1}}{1 - w^{y,z}}, \delta_y, \delta_z \right), \quad (36)$$

in which one must check if $\delta_x^{\min} < x_0 < \delta_x^{\max}$. In this case, one has $(x_0, y_0, z_0) \in \mathcal{S}^{\mathrm{e}}_{m_1,m_2}$. When this inequality condition is not satisfied, then $(x_0, y_0, z_0) \notin \mathcal{S}^{\mathrm{e}}_{m_1,m_2}$ and the optimal solution is one of the vertices of the rectangular parallelepiped.

If $w^{y,z} = 1$, then $x_{m_2} = x_{m_1}$. This condition occurs when, for example, $\delta_y = \frac{y_{m_1}+y_{m_2}}{2}$ and $\delta_z = \frac{z_{m_1}+z_{m_2}}{2}$. Another possibility is having $z_{m_2} = z_{m_1}$, thus implying that $\delta_y = \frac{y_{m_1}+y_{m_2}}{2}$. As in the previous case of one active inequality constraint, some ULAs may theoretically satisfy these last conditions, but having $w^{y,z} = 1$ is virtually impossible to occur in practice.

[23]Here, it is taken into account the fact that $\mathbf{x}_{m_2} \neq \mathbf{x}_{m_1}$ whenever $m_2 \neq m_1$.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer, 2010.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, ch. 8, pp. 157–180.

[3] H. Gamper, S. Tervo, and T. Lokki, "Speaker tracking for teleconferencing via binaural headset microphones," presented at the Int. Workshop Acoust. Signal Enhancement, Aachen, Germany, Sep. 2012.

[4] H. Puder, E. Fischer, and J. Hain, "Optimized directional processing in hearing aids with integrated spatial noise reduction," presented at the Int. Workshop Acoust. Signal Enhancement, Aachen, Germany, Sep. 2012.

[5] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[6] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proc. IEEE*, vol. 61, no. 10, pp. 1497–1498, Oct. 1973.

[7] M. S. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," presented at the IEEE ASSP Workshop Appl. Signal Process. Audio Acoust., New Paltz, NY, USA, Oct. 1997.

[8] J. DiBiase, "High-accuracy, low-latency technique for talker localization in reverberant environments," Ph.D. dissertation, Brown Univ., Providence, RI, USA, May 2000.

[9] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[10] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, USA, Apr. 2007, vol. 1, pp. 121–124.

[11] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New York, NY, USA, Oct. 2007, pp. 295–298.

[12] H. Do and H. F. Silverman, "Stochastic particle filtering: A fast SRP-PHAT single source localization algorithm," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2009, pp. 213–216.

[13] A. Said, B. Lee, and T. Kalker, "Fast steered response power computation in 3D spatial regions," HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2013-40, Apr. 2013.

[14] J. Clausen, Branch and Bound Algorithms: Principles and Examples, Mar. 1999 [Online]. Available: http://www.diku.dk/OLD/undervisning/2003e/datVoptimer/JensClausenNoter.pdf

[15] B. Gendron and T. G. Cranic, "Parallel branch-and-bound algorithms: Survey and synthesis," *Oper. Res.*, vol. 42, no. 6, pp. 1042–1066, Jun. 1994.

[16] 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 Process. Lett.*, vol. 18, no. 1, pp. 71–74, Jan. 2011.

[17] A. Marti, M. Cobos, J. J. Lopez, and J. Escolaro, "A steered response power iterative method for high-accuracy acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 134, no. 4, pp. 2627–2630, Oct. 2013.

[18] M. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, no. 409–412, Mar. 1965.

[19] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 3, pp. 943–950, Apr. 1979.

[20] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, B. Lee, A. Said, and R. W. Schafer, "Discriminability index for microphone array source localization," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Aachen, Germany, Sep. 2012, vol. 1, pp. 1–4.

[21] B. Lee and T. Kalker, "A vectorized method for computationally efficient SRP-PHAT sound source localization," presented at the 12th Int. Workshop Acoust., Echo, Noise Control, Tel Aviv, Israel, Aug. 2010.

[22] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. New York, NY, USA: Springer, 2008.

Leonardo O. Nunes (S'06) received his B.Sc. degree in electronics and computer engineering and his M.Sc. degree in electrical engineering, both from the Federal University of Rio de Janeiro (UFRJ) in 2007 and 2009, respectively. He is now a researcher working at the Applied Photonics Center & Halliburton Brazil Technology Center on distributed acoustic sensing technology and acoustic information retrieval with application to the oil & gas industry. He is also a pursuing a D.Sc. title from COPPE/UFRJ, in electrical engineering. His main interests are in: digital signal processing, adaptive signal processing, array signal processing, and machine learning. He is a student member of the IEEE (Institute of Electrical and Electronics Engineers).

Wallace A. Martins (M'12) was born in Brazil in 1983. He received the Electronics Engineer degree from the Federal University of Rio de Janeiro (UFRJ) in 2007, the M.Sc. and D.Sc. degrees in electrical engineering from COPPE/UFRJ in 2009 and 2011, respectively. In 2008, he was a research visitor at the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN. Since 2013 he has been with the Department of Electronics and Computer Engineering (DEL/Poli) and Electrical Engineering Program (PEE/COPPE), UFRJ, where he is presently an Associate Professor. His research interests are in the fields of digital communications, underwater communications, microphone/sensor array processing, adaptive signal processing, and machine learning techniques. Dr. Martins received the Best Student Paper Award from EURASIP at EUSIPCO-2009, Glasgow, Scotland, and the 2011 Best Brazilian D.Sc. Dissertation Award from Capes.

Markus V. S. Lima (S'08–M'14) was born in Rio de Janeiro, Brazil, in 1984. He received the Electronics Engineer degree from Universidade Federal do Rio de Janeiro (UFRJ) in 2008, and the M.Sc. and D.Sc. degrees in Electrical Engineering from COPPE/UFRJ in 2009 and 2013, respectively. Since 2014, he has been with the Department of Electrical Engineering, DEE/Poli/UFRJ, where he is presently an Associate Professor. His research interests include digital communications, microphone/sensor array processing, and adaptive filtering.

Luiz W. P. Biscainho (S'95–M'01) was born in Rio de Janeiro, Brazil, in 1962. He received the Electronics Engineering degree (magna cum laude) from the EE (now Poli) at Universidade Federal do Rio de Janeiro (UFRJ), Brazil, in 1985, and the M.Sc. and D.Sc. degrees in electrical engineering from the COPPE at UFRJ in 1990 and 2000, respectively. Having worked in the telecommunication industry between 1985 and 1993, Dr. Biscainho is now Associate Professor at the Department of Electronics and Computer Engineering (DEL) of Poli and the Electrical Engineering Program (PEE) of COPPE at UFRJ. His research area is digital audio processing. He is currently a member of the IEEE (Institute of Electrical and Electronics Engineers), the AES (Audio Engineering Society), the SBrT (Brazilian Telecommunications Society), and the SBC (Brazilian Computer Society).

Maurício V. M. Costa was born in Rio de Janeiro, Brazil, in 1989. He received the Electronics Engineer degree from Universidade Federal do Rio de Janeiro (UFRJ) in 2013 and is currently a graduate student in signal processing at Electrical Engineering Program, COPPE/UFRJ. His research area include microphone/sensor array processing and Query-by-Humming systems.

Felipe M. Gonçalves was born in Rio de Janeiro, Brazil, in 1990. He received the Electronics Engineer degree from Universidade Federal do Rio de Janeiro (UFRJ) in 2013 and is currently a graduate student in Finance and Managerial Control at COPPEAD Graduate School of Business—UFRJ. His research area includes Venture Capital and support services for New Technology-Based Firms.

Amir Said (S'90–M'95–SM'06–F'14) received the B.S. and M.S. degrees in electrical engineering from University of Campinas, Brazil, and the Ph.D. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY. After working at IBM, University of Campinas, and HP Labs, in 2013 he joined LG Electronics, where he is now principal researcher and program manager. His current research interests are in the areas of multimedia signal processing, compression, and 3D visualization, and their efficient implementation in new processing architectures. He has more than 100 technical publications among book chapters, conference and journal papers, more than 30 US patents and applications. Dr. Said received several awards including Best Paper Award from the IEEE Circuits and Systems Society, the IEEE Signal Processing Society Best Paper Award for his work on multi-dimensional signal processing, and the Most Innovative Paper Award at the 2006 IEEE International Conference on Image Processing. Among his technical activities, he was Associate Editor for the SPIE/IS&T *Journal of Electronic Imaging*, and IEEE TRANSACTIONS ON IMAGE PROCESSING; a member of the IEEE SPS Multimedia Signal Processing, and the Image, Video, and Multidimensional Signal Processing Technical Committees, was technical co-chair of the 2009 IEEE Workshop on Multimedia Signal Processing, the 2013 Picture Coding Symposium, and has co-chaired conferences at the SPIE/IS&T Electronic Imaging since 2006.

Bowon Lee (SM'10) received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea in 2000, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 2003 and 2006 respectively. From 2007 to 2014, he worked as a research scientist at Hewlett-Packard Laboratories in Palo Alto, California until he joined the faculty of the Department of Electronic Engineering at Inha University in March 2014. His research interests include statistical signal processing on audio and speech, microphone array signal processing, acoustic event detection and localization, and multimodal signal processing. He received two top 10% awards from the IEEE Workshop on Multimedia Signal Processing in 2009 and has served as the technical program committee of numerous IEEE conferences and workshops. He is a member of the Audio Engineering Society.