A Geometric Approach to Sound Source Localization from Time-Delay Estimates

Xavier Alameda-Pineda and Radu Horaud

***Abstract*—This paper addresses the problem of sound-source localization from time-delay estimates using arbitrarily-shaped non-coplanar microphone arrays. A novel geometric formulation is proposed, together with a thorough algebraic analysis and a global optimization solver. The proposed model is thoroughly described and evaluated. The geometric analysis, stemming from the direct acoustic propagation model, leads to necessary and sufficient conditions for a set of time delays to correspond to a unique position in the source space. Such sets of time delays are referred to as *feasible sets*. We formally prove that every feasible set corresponds to exactly one position in the source space, whose value can be recovered using a closed-form localization mapping. Therefore we seek for the optimal feasible set of time delays given, as input, the received microphone signals. This time delay estimation problem is naturally cast into a programming task, constrained by the feasibility conditions derived from the geometric analysis. A global branch-and-bound optimization technique is proposed to solve the problem at hand, hence estimating the best set of feasible time delays and, subsequently, localizing the sound source. Extensive experiments with both simulated and real data are reported; we compare our methodology to four state-of-the-art techniques. This comparison shows that the proposed method combined with the branch-and-bound algorithm outperforms existing methods. These in-depth geometric understanding, practical algorithms, and encouraging results, open several opportunities for future work.**

***Index Terms*—Geometric sound source localization, time delay estimates, constrained multivariate nonlinear programming.**

I. Introduction

FOR the past decades, source localization has been a fruitful research topic. Sound source localization (SSL) in particular, has become an important application, because many speech, voice and event recognition systems assume the knowledge of the sound source position. Time delay estimation (TDE) has proven to be a high-performance methodological framework for SSL, especially when it is combined with training [15], statistics [48] or geometry [8], [1]. We are interested in the development of a general-purpose TDE-based method for SSL, i.e., TDE-SSL, and we are particularly interested in indoor environments. This is extremely challenging for several reasons: (i) there may be several sound sources and their number varies over time, (ii) regular rooms are echoic, thus leading to reverberations, and (iii) the microphones are often embedded in devices (for example: robot heads and smart phones) generating high-level noise.

In this context, we focus on arbitrarily shaped non-coplanar microphone arrays, because of three main reasons. First, microphone arrays working on real (mobile) platforms may need to accommodate very restrictive design criteria, for which array geometries that have been traditionally studied, e.g., linear, circular, or spherical, are not well suited. We are particularly interested in embedding microphones into a robot head, such as the humanoid robot NAO[1] which possesses four microphones in a tetrahedron-like shape. There are robot design constraints that are not compatible with a particular type of microphone array. Moreover, solving for the most general non-coplanar microphone configuration opens the door to dynamically reconfigurable microphone arrays in arbitrary layouts. Such methods have been already studied in the specific case of spherical arrays [31]. Nevertheless, the most general case is worthwhile to be studied, since non-coplanar arrays include an extremely wide range of specific configurations.

This paper has the following original contributions:

- The *geometric analysis* of the microphone array. We are able to characterize those time delays that correspond to a position in the source space. Such time delays will be called *feasible* and the derived necessary and sufficient conditions will be called *feasibility conditions*.
- A *closed-form solution* for SSL. Indeed, we formally prove that every feasible set corresponds to exactly one position in the source space. Moreover, a localization mapping is built to recover, unambiguously, the sound source position from any set of feasible time delays.
- A *programming framework* in which the TDE-SSL problem is cast. More precisely, we propose a criterion for multichannel TDE designed to deal with microphone arrays in general configuration. The feasibility conditions derived from the geometric analysis constrain the optimization of the criterion, ensuring that the final TDEs correspond to a position in the source space.
- A branch-and-bound *global optimization method* solving the TDE-SSL task. Once the algorithm converges, the closed-form localization mapping is used to recover the sound source position. We state and prove that the sound source position is unique.

[1]http://www.aldebaran-robotics.com/

Manuscript received October 31, 2013; revised February 10, 2014; accepted April 08, 2014. Date of publication April 17, 2014; date of current version May 09, 2014. This work was supported by the EU project HUMAVIPS FP7-ICT-2009-247525. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mads Græsbøll Christensen.

The authors are with the Perception Team, INRIA Grenoble Rhône-Alpes, 38334 Montbonnot, France (e-mail: xavier.alameda-pineda@inria.fr; radu.horaud@inria.fr).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This includes the Matlab code, ISM filters, and real data. This material is 1.04 GB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2317989

- An *extensive set of experiments* benchmarking the proposed technique to the state-of-the-art. Our method is compared to four existing methods using both simulated and real data.

The remaining part of the paper is organized as follows. Section II describes the related work. Section III briefly summarizes the signal and propagation models. Section IV presents the full geometric analysis, together with the formal proofs. Section V casts the TDE-SSL task into a constrained optimization task. Section VI describes the branch-and-bound global optimization technique. The proposed SSL-TDE method is evaluated and compared to the state-of-the-art in Section VII. Finally, conclusions and a discussion for future work are provided in Section VIII. The associated MATLAB code and supplementary material are publicly available for download[2].

II. Related Work

The task of localizing a sound source from time delay estimates has received a lot of attention in the past; recent reviews can be found in [45], [39], [12].

One group of approaches (referred to as *bichannel SSL*) requires one pair of microphones. For example [33], [34], [53], [54] estimate the azimuth from the interaural time difference. These methods assume that the sound source is placed in front of the microphones and it lies in a horizontal plane. Consequently, they are intrinsically limited to one-dimensional localization. Other methods either guess both the azimuth and elevation [27], [14] or track them [24], [23]. These methods are based on estimating the impulse response function, which is a combination of the head related transfer function (HRTF) and the room impulse response (RIR). In order to guarantee the adaptability of the system, the intrinsic properties of the recording device encompassed in the HRTF must be estimated separately from the acoustic properties of the environment, modeled by the RIR. Furthermore, these methods lead to localization techniques which do not yield closed-form expressions, thus increasing the computational complexity. Moreover, the dependency on both HRTF and RIR of the source position is extremely complex, hence, it is difficult to model or to estimate this dependency. In conclusion, these methods suffer from two main drawbacks. On one side, large training sets and complex learning procedures are required. On the other side, the estimated parameters correspond to one particular combination of environment and microphone-pair position and orientation. Since estimating the parameters for all such possible combinations is unaffordable, these methods are hardly adaptable to unknown environments.

A second group of methods (referred to as *multilateration*) performs SSL from TDE with more than two microphones, by first estimating pairwise time delays, followed by localizing the source from these estimates. We note that the time delays are estimated independently of the location. In other words, the two steps are decoupled. Moreover, the TDEs do not incorporate the geometry of the array, that is, the estimates are computed regardless of the microphones' position. This is problematic because the existence of a source point consistent with all TDEs is not guaranteed. In order to illustrate this potential conflict, we consider a three-microphone linear array. Let t_m be the time-of-arrival at the mth microphone, and $t_{m,n} = t_n - t_m$ be the time delay associated to the microphone pair (m, n). In the particular set up of a three-microphone linear array, the case $t_{1,2} > 0$ and $t_{3,2} > 0$ is not physically possible. Indeed, this is equivalent to say that the acoustic wave reaches the first and the third microphones *before* reaching the middle one, which is inconsistent with the propagation path of the acoustic wave. In order to overcome this issue, multilateration is formulated either as maximum likelihood (ML) [10], [46], [48], [51], [49], [56], [55], as least squares (LS) [47], [4], [5], [17], [22], [8] or as global coherence fields (CFG) [37], [35], [6], [7]. Multilateration methods posses the advantage of being able to evaluate different TDE and SSL techniques. This allows for a better understanding of the interactions between TDE and SSL. Unfortunately, even if the ML/LS/GCF frameworks are able to discard TDE outliers, they can neither prevent nor reduce their occurrence. Consequently, the performance of these methods drops dramatically when used in highly reverberant environments.

A third group of methods (referred to as *multichannel SSL*) estimates all time delays at once, thus ensuring their mutual consistency. Multichannel SSL can be further split into two sub-groups. The first sub-group performs SSL using the TDEs extracted from the acoustic impulse responses [16], [42], [21], [32], [36]. These responses are directly estimated from the raw data, which is very challenging. As with bichannel SSL, large training sets and complex learning procedures are necessary. Moreover, the estimated impulse responses correspond to the acoustic signature of the environment associated with one particular microphone-array position and orientation. Therefore, such methods suffer from low adaptability to a changing environment. The second sub-group exploits the redundancy among the received signals. In [11] a multichannel criterion based on cross-correlation is proposed. Even if the method is based on pair-wise cross-correlation functions, the estimation of the time delays is performed at once. [11] has been extended using temporal prediction [20] and has also proven to be equivalent to two information-theoretic criteria [19], [2], under some statistical assumptions. However, all these methods were specifically designed for *linear* microphone arrays. Indeed, the line geometry is directly embedded in the proposed criterion and in the associated algorithms. Likewise, some methods were designed for other array geometries, such as *circular* [38] or *spherical* [43], [41], [40], [50] arrays. Again, the geometry is directly embedded in the methods in both cases. Hence, all these methods cannot be generalized to microphone arrays owing a general geometric configuration.

Recently, we addressed multichannel TDE-SSL in the case of arbitrary arrays, thus guaranteeing the system's adaptability [1]. TDE-SSL was modelled as a non-linear programming task, for which a gradient-based local optimization technique was proposed. However, this method has several drawbacks. First, the geometric analysis is incomplete. Indeed, the reported model is not valid for arrays with more than four microphones, thus limiting its generality. Second, the local optimization algorithm needed to be initialized on a grid. Consequently, the resulting procedure is prohibitively slow. Third, the evaluation was carried out in scenarios with almost no reverberations and only on simulated data. Last, no complexity analysis was performed.

Unlike most of the existing approaches on multichannel TDE, we did not embed the geometry of the array in the criterion. Instead, the geometry of the array is incorporated as two feasibility

[2]https://team.inria.fr/perception/research/geometric-sound-source-localization/.

constraints. Furthermore, our approach has several interesting features: (i) *generality*, since it is not designed for a particular array geometry and it may accommodate several microphones, (ii) *adaptability*, because the method neither constrains nor estimates the acoustic signature of the environment, (iii) *intuitiveness*, since the entire approach is built on a simple signal model and the geometry is derived from the direct-path propagation model, (iv) *soundness*, due to the thorough mathematical formalism underpinning the approach and (v) *robustness* and *reliability*, as shown by the extensive experiments and comparisons with state-of-the-art methods on both simulated and real data.

III. Signal and Propagation Models

In this Section we describe the sound propagation model and the signal model. While the first one is exploited to geometrically relate the time delays to the sound source position (see Section IV), the second one is used to derive a multichannel SSL criterion (see Section V). We introduce the following notations: the position of the sound source $\boldsymbol{S} \in \mathbb{R}^N$, the number of microphones M, as well as their positions, $\{\boldsymbol{M}_m\}_{m=1}^{m=M} \in \mathbb{R}^N$. Let $x(t)$ be the signal emitted by the source. The signal received at the mth microphone writes:

$$x_m(t) = x(t - t_m) + n_m(t), \tag{1}$$

where n_m is the noise associated with the mth microphone and t_m is the time-of-arrival from the source to that microphone. The microphones' noise signals are assumed to be zero-mean independent Gaussian random processes. Throughout the article, constant sound propagation speed, ν, and direct propagation path are assumed. Hence we write $t_m = \|\boldsymbol{S} - \boldsymbol{M}_m\|/\nu$. Using this model, the expression for the time delay between the mth and the nth microphones, $t_{m,n}$, is:

$$t_{m,n} = t_n - t_m = \frac{\|\boldsymbol{S} - \boldsymbol{M}_n\| - \|\boldsymbol{S} - \boldsymbol{M}_m\|}{\nu}. \tag{2}$$

IV. Geometric Sound Source Localization

We recall that the task is to localize the sound source from the TDE. In this Section we state the main theoretical results. Firstly, we describe under which conditions a set of time delays correspond to a sound source position–when a sound source can be localized. Such sets will be called *feasible* and the conditions, *feasibility constraints*. Secondly, we prove the uniqueness of the sound source positions for any feasible time delay set. Finally, we provide a closed-formula for sound source localization from any feasible set of time delays. Even if, in practice the problem is set in the source space, $\mathbb{R}^3$, the theory presented here is valid in $\mathbb{R}^N$, $N \geq 2$. In the following, Section IV-A describes the geometry of the problem for the two-microphone case, and Section IV-B delineates the geometry associated to the M-microphone case in general position.

A. The Two-Microphone Case

We start by characterizing the locus of sound-source locations corresponding to a particular time delay estimate $\hat{t}_{m,n}$, namely $\boldsymbol{S}$ satisfying $t_{m,n}(\boldsymbol{S}) = \hat{t}_{m,n}$. Since (2) defines a hyperboloid in $\mathbb{R}^N$, this equation embeds the *hyperbolic geometry* of the problem. For completeness, we state the following lemma:

Lemma 1: The space of sound-source locations $\boldsymbol{S} \in \mathbb{R}^N$ satisfying $t_{m,n}(\boldsymbol{S}) = \hat{t}_{m,n}$ is:

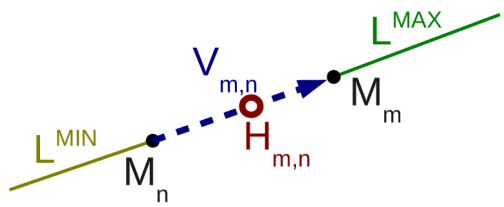


Fig. 1. The geometry associated with the very same, two-microphone case, located at $\boldsymbol{M}_m$ and $\boldsymbol{M}_n$ (see Lemma 1). $H_{m,n}$ is the mid-point of the microphones (in red) and $V_{m,n}$ designates the vector $\boldsymbol{M}_m - \boldsymbol{M}_n$ (in dashed-blue). $L_{m,n}^{\mathrm{MAX}}$ and $L_{m,n}^{\mathrm{MIN}}$ are the two half lines drawn in green and yellow respectively.

(i) the empty set if $|\hat{t}_{m,n}| > t_{m,n}^*$, where $t_{m,n}^* = \|\boldsymbol{M}_m - \boldsymbol{M}_n\|/\nu$;

(ii) the half line $\boldsymbol{L}_{m,n}^{\mathrm{MAX}}$ (or $\boldsymbol{L}_{m,n}^{\mathrm{MIN}}$), if $\hat{t}_{m,n} = t_{m,n}^*$ (or if $\hat{t}_{m,n} = -t_{m,n}^*$), where

$$\boldsymbol{L}_{m,n}^{\mathrm{MAX}} = \{\boldsymbol{H}_{m,n} + \mu \boldsymbol{V}_{m,n}\}_{\mu \geq 1/2},$$
$$\boldsymbol{L}_{m,n}^{\mathrm{MIN}} = \{\boldsymbol{H}_{m,n} - \mu \boldsymbol{V}_{m,n}\}_{\mu \geq 1/2},$$

$\boldsymbol{H}_{m,n} = (\boldsymbol{M}_m + \boldsymbol{M}_n)/2$ is the microphones' middle point and $\boldsymbol{V}_{m,n} = \boldsymbol{M}_m - \boldsymbol{M}_n$ is the microphones' vectorial baseline (see Fig. 1);

(iii) the hyperplane passing through $\boldsymbol{H}_{m,n}$ and perpendicular to $\boldsymbol{V}_{m,n}$, if $\hat{t}_{m,n} = 0$; or

(iv) one sheet of a two-sheet hyperboloid with foci $\boldsymbol{M}_m$ and $\boldsymbol{M}_n$ for other values of $\hat{t}_{m,n}$.

Proof: Using the triangular inequality, it is easy to see $-t_{m,n}^* \leq t_{m,n}(\boldsymbol{S}) \leq t_{m,n}^*, \forall \boldsymbol{S} \in \mathbb{R}^N$, which proves (i). (ii) is proven by rewriting $\boldsymbol{S} = \boldsymbol{H}_{m,n} + \mu_1 \boldsymbol{V}_{m,n} + \sum_{k=2}^N \mu_k \boldsymbol{W}_k$, where $(\boldsymbol{V}_{m,n}, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_N)$ is an orthogonal basis of $\mathbb{R}^N$, and then taking the derivatives with respect to the μ_i's. In order to prove (iii) and (iv) and without loss of generality, we can assume $\boldsymbol{M}_m = \boldsymbol{e}_1$, $\boldsymbol{M}_n = -\boldsymbol{e}_1$ and $\nu = 1$, where $\boldsymbol{e}_1$ is the first element of the canonical basis of $\mathbb{R}^N$. Hence, $t_{m,n}^* = 4$. Equation (2) rewrites:

$$(\hat{t}_{m,n})^2 + 4x_1 = -2\hat{t}_{m,n}\left((x_1+1)^2 + \sum_{k=2}^N x_k^2\right)^{\frac{1}{2}}, \tag{3}$$

where $(x_1, \ldots, x_N)^\top$ are the coordinates of $\boldsymbol{S}$. By squaring the previous equation we obtain:

$$a(4-a) + 4a\sum_{k=2}^N x_k^2 - 4(4-a)x_1^2 = 0, \tag{4}$$

where $a = (\hat{t}_{m,n})^2$. Notice that if $\hat{t}_{m,n} = 0$, (4) is equivalent to $x_1 = 0$, which corresponds to the statement in (iii). For the rest of values of a, that is $0 < a < (t_{m,n}^*)^2 = 4$, equation (4) represents a two-sheet hyperboloid because all the coefficients are strictly positive except the coefficient of x_1^2, which is strictly negative. In addition, we can rewrite (4) as:

$$x_1^2 = \frac{a(4-a) + 4a\sum_{k=2}^N x_k^2}{4(4-a)} \tag{5}$$

and notice that $x_1^2 > 0$. We observe that the solution space of (4) can be split into two subspaces $\mathcal{S}_{m,n}^+$ and $\mathcal{S}_{m,n}^-$ parametrized by $(x_2, \ldots, x_N)$, corresponding to the two solutions of (5). These two subspaces are the two sheets of the hyperboloid defined in (4). Moreover, one can easily verify that

$t_{m,n}(\mathcal{S}^+_{m,n}) = -t_{m,n}(\mathcal{S}^-_{m,n})$, so either $t_{m,n}(\mathcal{S}^+_{m,n}) = \hat{t}_{m,n}$ or $t_{m,n}(\mathcal{S}^-_{m,n}) = \hat{t}_{m,n}$, but both equalities cannot hold simultaneously. Hence the set of points $\boldsymbol{S}$ satisfying $t_{m,n}(\boldsymbol{S}) = \hat{t}_{m,n}$ is either $\mathcal{S}^+_{m,n}$ or $\mathcal{S}^-_{m,n}$: one sheet of a two-sheet hyperboloid. ■

We remark that the solutions of (4) are $\mathcal{S}^+_{m,n} \cup \mathcal{S}^-_{m,n}$. However, the solutions of (3) are either $\mathcal{S}^+_{m,n}$ or $\mathcal{S}^-_{m,n}$. This occurs because (4) depends only on $a = (\hat{t}_{m,n})^2$, and not on $\hat{t}_{m,n}$. Consequently, changing the sign of $\hat{t}_{m,n}$ does not modify the solutions of (4). In other words, the solutions of (4) contain not only the *genuine* solutions (those of (3)), but also a set of *artifact* solutions. More precisely, this set corresponds to the solutions of (3), replacing $\hat{t}_{m,n}$ by $-\hat{t}_{m,n}$. Geometrically, the solutions of (3) are one sheet of a two-sheet hyperboloid, and the solutions of (4) are the entire hyperboloid. Notice that, because $t_{m,n}(S^+_{m,n}) = -t_{m,n}(S^-_{m,n})$, we are always able to disambiguate the *genuine* solutions from the *artifact* ones.

B. The Case of M Microphones in General Position

We now consider the case of M microphones in *general position*, i.e., the microphones do not lie in a hyperplane of $\mathbb{R}^{\mathbb{N}}$. Firstly, we remark that, if a set of time delays $\hat{\boldsymbol{t}} = \{\hat{t}_{m,n}\}_{m=1,n=1}^{m=M,n=M} \subset \mathbb{R}^{M^2}$ satisfies (2) $\forall m, n$, then these time delays also satisfy the following constraints:

$$\begin{aligned} \hat{t}_{m,m} &= 0 \quad \forall m, \\ \hat{t}_{m,n} &= -\hat{t}_{n,m} \quad \forall m, n, \\ \hat{t}_{m,n} &= \hat{t}_{m,k} + \hat{t}_{k,n} \quad \forall m, n, k. \end{aligned}$$

As a consequence of these three equations we can rewrite any $\hat{t}_{m,n}$ in terms of $(\hat{t}_{1,2}, \ldots, \hat{t}_{1,M})$:

$$\hat{t}_{m,n} = -\hat{t}_{1,m} + \hat{t}_{1,n} \qquad \forall m, n. \tag{6}$$

This can be written as a vector $\hat{\boldsymbol{t}} = (\hat{t}_{1,2} \ldots \hat{t}_{1,M})^\top$ that lies in an $(M-1)$-dimensional vector subspace $\mathcal{W} \subset \mathbb{R}^{M^2}$. In other words, there are only $M-1$ linearly independent equations of the form (2). We remark that these $M-1$ linearly independent equations are still coupled by the sound source position $\boldsymbol{S}$. Geometrically, this is equivalent to seek the intersection of $M-1$ hyperboloids in $\mathbb{R}^N$ (see Fig. 2). Algebraically, this is equivalent to solve a system of $M-1$ non-linear equations in N unknowns. In general, this leads to finding the roots of a high-degree polynomial. However, in our case the hyperboloids share one focus, namely $\boldsymbol{M}_1$. As it will be shown below, in this case the problem reduces to solving a second-degree polynomial plus a linear system of equations. The $M-1$ equations (2) write:

$$\begin{cases} \nu\hat{t}_{1,2} &= \|\boldsymbol{S} - \boldsymbol{M}_2\| - \|\boldsymbol{S} - \boldsymbol{M}_1\| \\ &\vdots \\ \nu\hat{t}_{1,M} &= \|\boldsymbol{S} - \boldsymbol{M}_M\| - \|\boldsymbol{S} - \boldsymbol{M}_1\| \end{cases} \tag{7}$$

Because the M microphones are in general position (they do not lie in a hyperplane of $\mathbb{R}^N$), $M \geq N+1$ and the number of equations is greater or equal than the number of unknowns.

We now provide the conditions on $\hat{\boldsymbol{t}}$ under which (7) yields a real and unique solution for $\boldsymbol{S}$. More precisely, firstly, we provide a necessary condition on $\hat{\boldsymbol{t}}$ for (7) to have real solutions, secondly, we prove the uniqueness of the solution and build a mapping to recover the solution $\boldsymbol{S}$, and thirdly, we provide a necessary and sufficient condition on $\hat{\boldsymbol{t}}$ for (7) to have a real and unique solution.

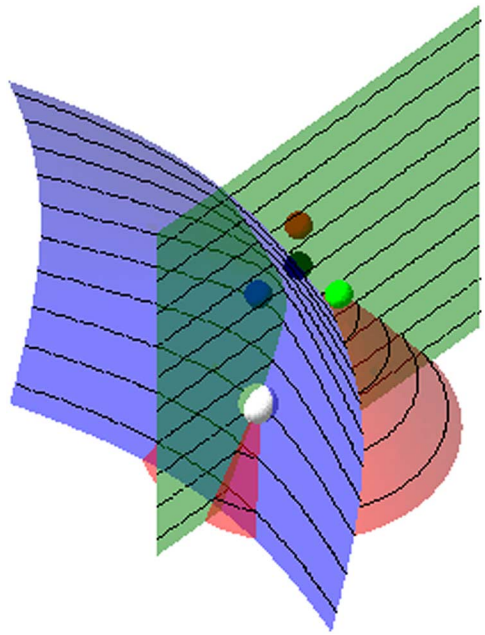


Fig. 2. Localization of the source using four microphones. Their position is shown in black ($\boldsymbol{M}_1$), blue ($\boldsymbol{M}_2$), red ($\boldsymbol{M}_3$) and green ($\boldsymbol{M}_4$). The blue hyperboloid corresponds to $\hat{t}_{1,2}$, the red to $\hat{t}_{1,3}$ and the green to $\hat{t}_{1,4}$. The intersection of the hyperboloids corresponds to the sound source position (white marker).

Notice that each equation in (7) is equivalent to $(\nu\hat{t}_{1,m} + \|\boldsymbol{S} - \boldsymbol{M}_1\|)^2 = \|\boldsymbol{S} - \boldsymbol{M}_m\|^2$, from which we obtain $-2(\boldsymbol{M}_1 - \boldsymbol{M}_m)^\top \boldsymbol{S} + p_{1,m}\|\boldsymbol{S} - \boldsymbol{M}_1\| + q_{1,m} = 0$, where $p_{1,m} = 2\nu\hat{t}_{1,m}$ and $q_{1,m} = \nu^2(\hat{t}_{1,m})^2 + \|\boldsymbol{M}_1\|^2 - \|\boldsymbol{M}_m\|^2$. Hence, (7) can now be written in matrix form:

$$\mathbf{M}\boldsymbol{S} + \boldsymbol{P}\|\boldsymbol{S} - \boldsymbol{M}_1\| + \boldsymbol{Q} = 0, \tag{8}$$

where $\mathbf{M} \in \mathbb{R}^{(M-1)\times N}$ is a matrix with its mth row, $2 \leq m \leq M$, equal to $2(\boldsymbol{M}_m - \boldsymbol{M}_1)^\top$, $\boldsymbol{P} = (p_{1,2}, \ldots, p_{1,M})^\top$ and $\boldsymbol{Q} = (q_{1,2}, \ldots, q_{1,M})^\top$. Notice that $\boldsymbol{P}$ and $\boldsymbol{Q}$ depend on $\hat{\boldsymbol{t}}$.

Without loss of generality and because the points $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_M$ do not lie in the same hyperplane, we assume that $\mathbf{M}$ can be written as a concatenation of an invertible matrix $\mathbf{M}_L \in \mathbb{R}^{N\times N}$ and a matrix $\mathbf{M}_E \in \mathbb{R}^{(M-N-1)\times N}$ such that $\mathbf{M} = \begin{pmatrix} \mathbf{M}_L \\ \mathbf{M}_E \end{pmatrix}$. We can easily accomplish this by renumbering the microphones such that the first $N+1$ microphones do not lie in the same hyperplane. This implies that the first N rows of $\mathbf{M}$ are linearly independent, and therefore $\mathbf{M}_L$ is invertible. Similarly we have $\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_L \\ \boldsymbol{P}_E \end{pmatrix}$ and $\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_L \\ \boldsymbol{Q}_E \end{pmatrix}$. Thus, (8) rewrites:

$$\mathbf{M}_L\boldsymbol{S} + \boldsymbol{P}_L\|\boldsymbol{S} - \boldsymbol{M}_1\| + \boldsymbol{Q}_L = 0, \tag{9}$$

$$\mathbf{M}_E\boldsymbol{S} + \boldsymbol{P}_E\|\boldsymbol{S} - \boldsymbol{Q}_1\| + \boldsymbol{Q}_E = 0, \tag{10}$$

where $\boldsymbol{P}_L$, $\boldsymbol{Q}_L$ are vectors in $\mathbb{R}^N$ and $\boldsymbol{P}_E$, $\boldsymbol{Q}_E$ are vectors in $\mathbb{R}^{M-N-1}$. If we also decompose $\hat{\boldsymbol{t}}$ into $\hat{\boldsymbol{t}}_L$ and $\hat{\boldsymbol{t}}_E$, we observe that $\boldsymbol{P}_L$ and $\boldsymbol{Q}_L$ depend only on $\hat{\boldsymbol{t}}_L$ and that $\boldsymbol{P}_E$ and $\boldsymbol{Q}_E$ depend only on $\hat{\boldsymbol{t}}_E$. Notice that (7) is strictly equivalent to (9) - (10). In the following, (9) will be used for defining the necessary conditions on $\hat{\boldsymbol{t}}$ as well as localizing the sound source. The study of

(10) is reported further on. By introducing a scalar variable w, (9) can be written as:

$$\mathbf{M}_L \boldsymbol{S} + w\boldsymbol{P}_L + \boldsymbol{Q}_L = 0, \tag{11}$$

$$\|\boldsymbol{S} - \boldsymbol{M}_1\|^2 - w^2 = 0. \tag{12}$$

We remark that the system (11)–(12) is defined in the $(\boldsymbol{S}, w)$ space. Notice that (11) a straight line and (12) represents a two-sheet hyperboloid. Because two-sheet hyperboloids are not ruled surfaces, (12) cannot contain the straight line in (11). Hence (11) and (12) intersect in two (maybe complex) points.

In order to solve (11)–(12), we first rewrite (11) as

$$\boldsymbol{S} = \boldsymbol{A}w + \boldsymbol{B}, \tag{13}$$

where $\boldsymbol{A} = -\mathbf{M}_L^{-1}\boldsymbol{P}_L$ and $\boldsymbol{B} = -\mathbf{M}_L^{-1}\boldsymbol{Q}_L$, and then substitute $\boldsymbol{S}$ from (13) into (12) obtaining:

$$(\|\boldsymbol{A}\|^2 - 1)w^2 + 2\langle \boldsymbol{A}, \boldsymbol{B} - \boldsymbol{M}_1\rangle w + \|\boldsymbol{B} - \boldsymbol{M}_1\|^2 = 0. \tag{14}$$

We are interested in the real solutions, that is, $\boldsymbol{S} \in \mathbb{R}^N$. Because $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^N$, the solutions of (11) - (12) are real, if and only if, the solutions to (14) are real too. Equivalently, the discriminant of (14) has to be non-negative. Hence the solutions to (11) - (12) are real if and only if $\hat{\boldsymbol{t}}$ satisfies:

$$\Delta(\hat{\boldsymbol{t}}) := \langle \boldsymbol{A}, \boldsymbol{B} - \boldsymbol{M}_1\rangle^2 - \|\boldsymbol{B} - \boldsymbol{M}_1\|^2(\|\boldsymbol{A}\|^2 - 1) \geq 0. \tag{15}$$

The previous equation is a *necessary condition* for (11) - (12) to have real solutions. Albeit, we are interested in the solutions of (9). Obviously, if $\boldsymbol{S}$ is a solution of (9), then $(\boldsymbol{S}, \|\boldsymbol{S} - \boldsymbol{M}_1\|)$ is a solution of (11) - (12). However, the reciprocal is not true; these two systems are not equivalent. Indeed, since $\Delta(\hat{\boldsymbol{t}}) = \Delta(-\hat{\boldsymbol{t}})$, one of the solutions of (11) - (12) is the solution of (9) and the other is the solution of (9) replacing $\hat{\boldsymbol{t}}$ by $-\hat{\boldsymbol{t}}$. In other words, the two solutions of (11) - (12), namely $(\boldsymbol{S}^+, w^+)$ and $(\boldsymbol{S}^-, w^-)$, satisfy

$$\text{either} \begin{cases} \boldsymbol{t}(\boldsymbol{S}^+) = \hat{\boldsymbol{t}} \\ \boldsymbol{t}(\boldsymbol{S}^-) = -\hat{\boldsymbol{t}} \end{cases} \text{ or } \begin{cases} \boldsymbol{t}(\boldsymbol{S}^+) = -\hat{\boldsymbol{t}} \\ \boldsymbol{t}(\boldsymbol{S}^-) = \hat{\boldsymbol{t}} \end{cases}.$$

Notice that this situation has been already encountered on equations (3) and (4), where the same disambiguation reasoning has been used. To summarize, the solution to (9) is *unique*. Moreover, we can use (13) to define the following *localization mapping*, which retrieves the sound-source position from a feasible $\hat{\boldsymbol{t}}$:

$$L(\hat{\boldsymbol{t}}) := \begin{cases} \boldsymbol{S}^+ = \boldsymbol{A}w^+ + \boldsymbol{B} & \text{if } \boldsymbol{t}(\boldsymbol{S}^+) = \hat{\boldsymbol{t}} \\ \boldsymbol{S}^- = \boldsymbol{A}w^- + \boldsymbol{B} & \text{otherwise.} \end{cases} \tag{16}$$

Until now we provided the condition under which equation (9) yields real solutions, the uniqueness of the solution and a localization mapping. However, the original system includes also equation (10). In fact, (10) adds $M - N - 1$ constraints onto $\hat{\boldsymbol{t}}$. Indeed, if $L(\hat{\boldsymbol{t}})$ is a solution of (9), then in order to be a solution of (9) - (10), it has to satisfy:

$$\mathcal{E}(\hat{\boldsymbol{t}}) := \mathbf{M}_E L(\hat{\boldsymbol{t}}) + \boldsymbol{P}_E \|L(\hat{\boldsymbol{t}}) - \boldsymbol{M}_1\| + \boldsymbol{Q}_E = 0. \tag{17}$$

Moreover, the reciprocal is true. Summarizing, the system (9)–(10) has a unique real solution $L(\hat{\boldsymbol{t}})$ if and only if $\Delta(\hat{\boldsymbol{t}}) \geq 0$ and $\mathcal{E}(\hat{\boldsymbol{t}}) = 0$.

It is interesting to discuss these findings from three different perspectives: geometric, algebraic, and computational:

1) *The differential geometry point of view*. The set of feasible time delays,

$$\mathcal{T} = \left\{\hat{\boldsymbol{t}} \in \mathcal{W}, \Delta(\hat{\boldsymbol{t}}) \geq 0 \text{ and } \mathcal{E}(\hat{\boldsymbol{t}}) = 0\right\},$$

is a bounded N-dimensional manifold with boundary lying in a $(M-1)$-dimensional vector subspace of $\mathbb{R}^{M^2}$, Indeed, because $\mathcal{E}$ is a $(M - N - 1)$-dimensional vector-valued function, $\mathcal{T}$ has dimension N. The boundary of $\mathcal{T}$ is the set

$$\partial\mathcal{T} = \left\{\hat{\boldsymbol{t}} \in \mathcal{W} | \Delta(\hat{\boldsymbol{t}}) = 0 \text{ and } \mathcal{E}(\hat{\boldsymbol{t}}) = 0\right\}.$$

In this context, the localization mapping must be seen as a smooth bijection from $\mathcal{T}$ to $\mathbb{R}^N$, i.e., an isomorphism between the manifolds.

2) *The algebraic point of view.* Δ and $\mathcal{E}$ characterize the time delays corresponding to a position in the source space, $\mathbb{R}^N$. That is to say that Δ and $\mathcal{E}$ represent the feasibility constraints, the necessary and sufficient conditions for the existence of $\boldsymbol{S}$. Under this conditions $\boldsymbol{S}$ is unique and given by the closed-form localization mapping L.

3) *The computational point of view.* The mappings $\Delta(\hat{\boldsymbol{t}})$, $\mathcal{E}(\hat{\boldsymbol{t}})$ and $L(\hat{\boldsymbol{t}})$ which are computed from (15), (17), and (16) are expressed in closed-form and they only depend on the microphone locations. The most time-consuming part of these computations is the inversion of the *microphone matrix* $\mathbf{M}_L$, which can be performed off-line. Consequently, the use of these three mappings is intrinsically efficient.

To conclude, we highlight that $\Delta(\hat{\boldsymbol{t}})$ and $\mathcal{E}(\hat{\boldsymbol{t}})$, i.e., (15) and (17), provide the conditions under which the $M - 1$ time delays correspond to a *valid point* in $\mathbb{R}^N$. *If these conditions are satisfied, the problem yields a unique location for the source* $\boldsymbol{S}$. Moreover, the mapping $L(\hat{\boldsymbol{t}})$ defined by (16) is a closed-form sound-source localization solution for any set of *feasible* time delays $\hat{\boldsymbol{t}}$:

$$\boldsymbol{S} = L(\hat{\boldsymbol{t}}). \tag{18}$$

V. Time Delay Estimation

In the previous section we described how to characterize the feasible sets of time delays and how to localize a sound source from them. We now address the problem of how to obtain an optimal set of time delays given the perceived acoustic signals. In the following, we delineate a criterion for multichannel time delay estimation (Section V-A), which will subsequently be used in Section V-B to cast the TDE-SSL problem into a non-linear multivariate constrained optimization task. Indeed, the multichannel TDE criterion is a non-linear cost function allowing to choose the best value for $\hat{\boldsymbol{t}}$. The feasibility constraints derived in the previous section are used to constrain the optimization problem, thus seeking the optimal feasible value for $\hat{\boldsymbol{t}}$.

A. A Criterion for Multichannel TDE

The criterion proposed in [11] is built from the theory of linear predictors and presented in the framework of *linear* microphone arrays. Following a similar approach, we propose to generalize this criterion to arrays owing a general microphone configuration. Given the M perceived signals $\{x_m(t)\}_{m=1}^{m=M}$, we would

like to estimate the time delays between them. As explained before (see (6) and after), only $M-1$ of the delays are linearly independent. Without loss of generality we choose, as above, the delays $t_{1,2},\dots,t_{1,m},\dots,t_{1,M}$. We select $x_1(t)$ as the reference signal and set the following prediction error:

$$e_{\boldsymbol{c},\boldsymbol{t}}(t) = x_1(t) - \sum_{m=2}^{M} c_{1,m} x_m(t+t_{1,m}), \tag{19}$$

where $\boldsymbol{c} = (c_{1,2},\dots,c_{1,m},\dots,c_{1,M})^\top$ is the vector of the prediction coefficients and $\boldsymbol{t} = (t_{1,2},\dots,t_{1,m},\dots,t_{1,M})^\top$ is the vector of the prediction time delays. Notice also that, when $\boldsymbol{t}$ takes the true value, the signals $x_m(t+t_{1,m})$ and $x_n(t+t_{1,n})$ are on phase. The criterion to minimize is the expected energy of the prediction error (19), leading to an unconstrained optimization problem:

$$(\boldsymbol{c}^*,\boldsymbol{t}^*) = \arg\min_{\boldsymbol{c},\boldsymbol{t}} \mathrm{E}\left\{e^2_{\boldsymbol{c},\boldsymbol{t}}(t)\right\}.$$

In addition, it can be shown (see [11]) that this problem is equivalent to:

$$\boldsymbol{t}^* = \arg\min_{\boldsymbol{t}} J(\boldsymbol{t}), \tag{20}$$

with

$$J(\boldsymbol{t}) = \det\left(\mathbf{R}(\boldsymbol{t})\right), \tag{21}$$

$\mathbf{R}(\boldsymbol{t}) \in \mathbb{R}^{M\times M}$ being the real matrix of normalized cross-correlation functions evaluated at $\boldsymbol{t}$. That is $\mathbf{R}(\boldsymbol{t}) = [\rho_{m,n}(t_{m,n})]_{m,n}$ with:

$$\rho_{m,n}(t_{m,n}) = \frac{\mathrm{E}\left\{x_m(t+t_{1,m})x_n(t+t_{1,n})\right\}}{\sqrt{E_m E_n}}, \tag{22}$$

where $E_m = R_{m,m}(0) = \mathrm{E}\{x_m^2(t)\}$ is the energy of the mth signal.

Importantly, the criterion J in (21) is designed to deal with microphones in a general configuration and not for a specific microphone-array geometry. Hence, this guarantees the generality of the proposed approach. Because the array's geometry is not embedded in J (as it is done in [11]), J is a multivariate function.

In the next Section, the feasibility constraints previously derived are combined with this cost function to set up a constrained multivariate optimization task.

B. The Constrained Optimization Formulation

So far we characterized the *feasible* values of $\boldsymbol{t}$, i.e., those corresponding to a sound source position (Section IV) and introduced a criterion to choose the best value for $\hat{\boldsymbol{t}}$ (Section V-A). The next step is to look for the best value among the feasible ones. This will be referred to as the *geometrically-constrained time delay estimation* problem, which naturally casts into the following *non-linear multivariate constrained optimization* problem:

$$\begin{cases} \min\limits_{\boldsymbol{t}} J(\boldsymbol{t}), \\ \text{s.t.} \quad \boldsymbol{t} \in \mathcal{W}\cap\mathcal{B}, \quad \Delta(\boldsymbol{t}) \geq 0, \quad \mathcal{E}(\boldsymbol{t}) = 0, \end{cases} \tag{23}$$

where $\mathcal{W}$, Δ and $\mathcal{E}$ are defined in Section IV and $\mathcal{B}$ is a compact set defined as:

$$\mathcal{B} = \left\{\boldsymbol{t}\in\mathbb{R}^{M^2} \,|\, |t_{m,n}| \leq t^*_{m,n}, \forall m,n = 1,\dots,M\right\}. \tag{24}$$

It is worth noticing that, in practice, the dimension of the optimization task is $M-1$. Indeed, since all time delays can be expressed as a function of $(t_{1,2},\dots,t_{1,M})$, the optimization is done with respect to these $M-1$ variables. We also remark that the optimization variables lay in a bounded space (as described in Section IV-A). The equality constraint is trivial when $M = N+1$. In other words, in a real world scenario, this constraint does not exist when the array consists on $4 = 3+1$ microphones. When using five or more microphones, the condition could be relaxed to $\|\mathcal{E}(\boldsymbol{t})\|^2 \leq \epsilon$, which is often more adapted to the existing optimization algorithms.

Notice that it is possible to use a redundant set of TDOAs, namely all the M^2 TDOAs available with M microphones. This implies that the dimension of the minimizer $\boldsymbol{t}$ in (20) is M^2, instead of $M-1$, and that the minimization (23) must be carried out under the presence of M^2-M+1 additional linear constraints (6).

We would like to highlight that all M microphone signals are used in the estimation procedure. In a sense, all received signals affect the estimation of all the time delays. This is why there is one $(M-1)$-dimensional optimization task and not several one-dimensional optimization tasks. The localization is carried out immediately after the time delay estimation thanks to a closed-form solution (18), thus with no other estimation procedure. The power of the proposed method relies on the intrinsic relation between signals and time delay estimates combined with the use of the geometric constraints given by the microphones position. By adding these constraints to the estimation procedure, we do not discard any infeasible sets, but we prevent them to be the outcome of our algorithm. In other words, the estimation procedure will always provide a set of time delays corresponding to a position in the sound source space. Next Section describes the branch & bound global optimization technique proposed to solve (23).

VI. Branch & Bound Optimization

Global optimization is, in most cases, an extremely challenging task. Nevertheless, the optimization of (23) is well suited for a global optimizer. Indeed, J is continuously differentiable on $\mathcal{B}$, therefore ∇J is continuous. This implies that ∇J is bounded on any compact set, in particular on $\mathcal{B}$. Hence, by means of theorem 9.5.1 in [44], J is Lipschitz on $\mathcal{B}$. Subsequently, a branch & bound (B&B) type of algorithm is well suited.

Such optimization techniques were initially proposed for linear mixed-integer programming [13] and extended later on to the non-linear case [30]. They alternate between the branch and bound procedures in order to recursively seek the potential regions where the global minimum is. While the branch step splits the potential regions into smaller pieces, the bound step estimates the lower and upper bounds of each potential region. After the bounding, the *discarding threshold* is set to the minimum of the upper bounds. Then, all regions whose lower

bound is bigger than the discarding threshold are discarded (since they cannot contain the global minimum).

The B&B algorithm that we propose maintains two lists of regions: $\mathcal{P}$ containing the potential regions and $\mathcal{D}$ containing the discarded regions (see Algorithm 1). The B&B inputs are the initial list of potential regions and the Lipschitz constant L. The outputs are the two maintained lists, $\mathcal{P}$ and $\mathcal{D}$ after convergence. Each of the regions in $\mathcal{P}$ and in $\mathcal{D}$ represents a $(M-1)$-dimensional cube $(\boldsymbol{t}, s)$, where $\boldsymbol{t}$ is the cube's centre and s is its side's length. The Branch routine splits each of the cubes in $\mathcal{P}$ into 2^{M-1} smaller cubes of size $s/2$. Next, the bound routine (see Algorithm 2) estimates the upper (u) and the lower (l) bounds of all regions in $\mathcal{P}$. The discarding threshold τ is then set to the minimum of all upper bounds. All sets in $\mathcal{P}$ with lower bound higher than τ are moved to the discarded list $\mathcal{D}$. One prominent feature of the optimization task in (23) us that we seek for the minimum on the set $\mathcal{B}$. The set of potential sets, $\mathcal{P}$ is naturally initialized to the set $\mathcal{B}$. Consequently, the B&B procedure does not require a grid-based optimization.

Algorithm 1 Branch and Bound

Input: The Lipschitz constant L and the initial list of potential regions $\mathcal{P}$.

Output: The list of potential solutions $\mathcal{P}$ and the list of discarded regions $\mathcal{D}$.

repeat

  (a) $\mathcal{P} = \text{Branch}(\mathcal{P})$

  (b) $[\mathcal{P}, \mathcal{RD}] = \text{Bound}(\mathcal{P}, L)$

  (c) $\mathcal{D} = \mathcal{D} \cup \mathcal{RD}$

until Convergence

Algorithm 2 Bound routine of Algorithm 1

Input: The Lipschitz constant L and the list of potential regions $\mathcal{P}$.

Output: The list of potential regions updated $\mathcal{P}$ and the list of recently discarded regions $\mathcal{RD}$.

for $i = 1, \ldots, |\mathcal{P}|$ **do**

  (a) $l^{(i)} = J(\boldsymbol{t}^{(i)}) - s^{(i)}L$.

  (b) $u^{(i)} = J(\boldsymbol{t}^{(i)}) + s^{(i)}L$.

  (c) $\tau = \min_{i=1,\ldots,|\mathcal{P}|} u^{(i)}$.

end for

for $i = 1, \ldots, |\mathcal{P}|$ **do**

  if $l^{(i)} > \tau$ **then**

    Move $(\boldsymbol{t}^{(i)}, s^{(i)})$ from $\mathcal{P}$ to $\mathcal{RD}$.

  end if

end for

The branch and bound routines are alternated until convergence. Many criteria could be used to stop the algorithm. In order to guarantee an accurate solution, we may force a maximum size for the potential regions. Also, in case we would like to guarantee the stability of the solution, we may track the variation of the smallest of the lower bounds. Either way, once the algorithm has converged, we select the best region in $\mathcal{P}$ among those satisfying the constraints Δ and $\mathcal{E}$. If there is no such region in $\mathcal{P}$[3], the B&B algorithm is run again providing $\mathcal{D}$ as the initial list of potential regions.

VII. Experimental Validation

A. Experimental Setup

In order to validate the proposed model and the associated estimation technique, we used an evaluation protocol with simulated and real data. In both cases, the environment was a room of approximately $4 \times 4 \times 4$ meters ($N = 3$), with an array of $M = 4$ microphones placed at (in meters) $\boldsymbol{M}_1 = (2.0, 2.1, 1.83)^\top$, $\boldsymbol{M}_2 = (1.8, 2.1, 1.83)^\top$, $\boldsymbol{M}_3 = (1.9, 2.2, 1.97)^\top$ and $\boldsymbol{M}_4 = (1.9, 2.0, 1.97)^\top$. *The microphones are the vertices of a tetrahedron, resulting in a non-coplanar configuration*. The sound source was placed on a sphere of 1.7 m radius centred at the microphone array. More precisely, the source was placed at 21 different azimuth values, between $-160°$ and $160°$, and at 9 different elevation values between $-60°$ and $60°$, hence at 189 different directions. The speech fragments emitted by the source were randomly chosen from a publicly available data set [18]. One hundred millisecond cuts of these sounds were used as input of the evaluated methods.

In the simulated case, we controlled two parameters. Firstly, the value of T_{60}, which is a parameter of the image-source model [29] (available at [28]), controlling the amount of reverberations. More precisely, T_{60} measures the time needed for the emitted signal to decay 60 dB. The higher the T_{60}, the larger the amount of reverberations and their energy. In our simulations, T_{60} took the following values (in seconds): 0, 0.1, 0.2, 0.4 and 0.6. Secondly, we controlled the amount of white noise added to the received signals by setting the signal-to-noise ratio (SNR) to -10, -5, or 0 dB.

In the real case, we used a slightly modified version of the acquisition protocol defined in [15]. This protocol was designed to automatically gather sound signals coming from different directions, using the motor system of a robotic platform placed in a regular indoor room. Such realistic environment inherently limits the quality of the recordings. First of all, the noise of the acquisition device (the computer's fans) is also recorded. Second, ambient noise associated with the room's location (between a corridor and a server room) has also a negative, but very realistic effect on the data. We roughly estimated the acoustic characteristics of the real recordings, namely $T_{60} \approx 0.5$ s and $SNR \approx 0$ dB. In our case, we replaced the dummy head used in [15], by a tetrahedron-shaped microphone array. The motor platform has two degrees of freedom (pan and tilt), and was designed to guarantee the repeatability of the movements. A loud-speaker was placed 1.7 m away from the array to emulate the sound source. We recorded sound waves coming from 189 different directions. Consequently, the performed tests cover an extensive range of realistic situations.

B. Implementation

We implemented several methods and compared them within the previously described set up. *b&b* stands for the branch

[3]In order to decide whether a region satisfies the constraints or not, we test its centre. This approximation is justified by the fact that, at this stage of the algorithm, the regions are extremely small (since we force a maximum region size).

& bound method proposed in this paper. *unc*, *d-lb* and *s-lb*, stemmed from [1], represent the state-of-the-art in multichannel TDE-SSL for arbitrarily shaped microphone arrays. *dm* stands for "direct multichannel", which is the generalization of [11] to arrays with arbitrary configuration. *n-mult*, *t-mult*, and *f-mult* are variants of [8], which represent the state-of-the-art in multilateration methods. *pi* is a straightforward multilateration algorithm. We have chosen to implement all nine methods to (i) compare the proposed algorithm to the state-of-the-art and (ii) push the limits of existing TDE-SSL algorithms of both multilateration and multichannel SSL disciplines.

- *b&b* corresponds to the branch & bound algorithm described in Section VI. The list of potential regions is naturally initialized as $\mathcal{P} = \{\mathcal{B}\}$. The Lipschitz constant L is estimated by computing the maximum slope among one thousand point pairs randomly drawn inside the feasibility domain.
- *unc*, *d-lb* and *s-lb* are directly derived from the procedure described in [1]. All of them perform multichannel TDE to further localize the sound source. *unc* and *s-lb* are proposed here to push the limits of the base method, *d-lb*. These local optimization techniques are initialized on the unconstrained ($\mathcal{G}_u$), constrained ($\mathcal{G}_c$) and sparse ($\mathcal{G}_s$) grids respectively. The details are given in Section VII-B1.
- *dm* is the straightforward generalization of [11] to arbitrarily-shaped microphone arrays. J, defined in (21) is evaluated on $\mathcal{G}_c$, and the minimum over the grid is selected. The difference between *dm* and *d-lb* is that in the former no local minimization is carried out.
- *n-mult*, *t-mult* and *f-mult* are implementations of the method described in [8]. In this case the time delay estimates, $\hat{\boldsymbol{t}}$ are computed independently (using [26]), and the sound source position, $\boldsymbol{S}$, is chosen to be as close as possible to the hyperboloids associated with $\hat{\boldsymbol{t}}$. Because the algorithm was designed for distributed sensor networks and not for egocentric arrays, we had to modify it. Further explanations are given in Section VII-B2.
- *pi* corresponds to pair-wise independent time delay estimation based on cross-correlation [26]. That is, $t_{1,j}$ is the maximum of the function $\rho_{1,j}(\tau)$; This is the simplest multilateration algorithm one can think of.

Except for *n-mult*, *t-mult*, and *f-mult*, which provide $\boldsymbol{S}$ directly, all other algorithms provide a time delay estimate. If this estimate is feasible, $\boldsymbol{S}$ is recovered using (18).

1) Methods unc, d-lb and s-lb: In [1], the constrained problem is converted into an unconstrained problem with a different cost function. The intuition is that the cost function is modified to penalize those points that are closer to the feasibility border. In practice, the inequality constraint is added to the cost by means of a log-barrier function

$$\begin{cases} \min\limits_{\boldsymbol{t}} J(\boldsymbol{t}) - \mu \log(\Delta(\boldsymbol{t})), \\ \text{s.t.} \quad \boldsymbol{t} \in \mathcal{W} \cap \mathcal{B}, \quad \mathcal{E}(\boldsymbol{t}) = 0, \end{cases} \tag{25}$$

where $\mu \geq 0$ is a regularizing parameter.

Consequently, the original task (23) is converted into a sequence of tasks indexed by μ. Each of the problems has an optimal solution $\hat{\boldsymbol{t}}_\mu$. It can be proven (see [3]) that $\hat{\boldsymbol{t}}_\mu \to \hat{\boldsymbol{t}}$ when $\mu \to 0$. Log-barrier methods are gradient-based techniques, which decrease the value of μ with the iterations, thus converging to the closest feasible local minimum of J. Therefore, it is recommended to provide the analytic derivatives in order to increase both the convergence speed and the accuracy (see Appendices A and B for the expressions of the gradients and Hessians of the cost function and the constraints, respectively).

Unfortunately, log-barrier methods are designed for convex problems. In other words, these methods find the local minimum closest to the initialization point. Hence, in order to find the global minimum, the algorithm must be multiply initialized from points lying on a grid $\mathcal{G}$. After convergence, the minimum among all the local minima found is assumed to be the global solution of the problem. *d-lb* (dense-log-barrier) corresponds to the method in [1], hence solving for (23), initialized on a grid $\mathcal{G}_c$ of 352 feasible points. *unc* solves for the unconstrained problem, i.e., (20), and it is initialized on a grid $\mathcal{G}_u$. The difference between the two grids is that while $\mathcal{G}_c$ contains just feasible points, $\mathcal{G}_u$ contains unfeasible points as well. In practice, $\mathcal{G}_u$ contains 456 points. The rationale of implementing *unc* is to better assess and quantify the role played by the feasibility constraints, Δ and $\mathcal{E}$. *s-lb* (sparse-log-barrier) corresponds to the same log-barrier method initialized on a sparse grid $\mathcal{G}_s$. We conjecture that the global minimum of J corresponds to one of the local maxima of $\rho_{1,m}$ in (22) for $m = 2, \ldots, M$. For each microphone pair $(1, m)$ we extract $K = 3$ local maxima of $\rho_{1,m}$. $\mathcal{G}_s$ consists of all possible combinations of these values, thus containing $K^{M-1} = 27$ points (in the case of $M = 4$ microphones). *s-lb* is implemented to assess the robustness towards initialization of the local optimization technique. Both, *d-lb* and *s-lb* are reimplementations of the publicly available MATLAB log-barrier dual interior-point method [9].

2) Methods n-mult, t-mult and f-mult: As already mentioned, we implemented [8] with some modifications. Indeed, the method was designed for *distributed* microphone arrays. With such a setup, the sound source position lies inside the volume defined by the microphone positions in the room. In the case of an egocentric array the sound source is necessarily located outside the volume delimited by the microphone array. The method described in [8] seeks the locations the closest to the hyperboloids given by the independently estimated time delays $\hat{\boldsymbol{t}}$. More precisely, the following criterion is minimized:

$$H(\boldsymbol{S}) = \sum_{1 \leq m < n \leq M} \left(h_{m,n}(\boldsymbol{S})\right)^2, \tag{26}$$

where $h_{m,n}$ is the equation of a two-sheet hyperboloid (i.e., equation (4)) with foci $\boldsymbol{M}_m$ and $\boldsymbol{M}_n$ and differential value $\hat{t}_{m,n}$. We will call H the *multilateration cost function*, to distinguish is from the cost function J. If the estimated time delays are feasible, the minimization of (26) is equivalent to solve the system $\{h_{m,n}(\boldsymbol{S}) = 0\}_{1 \leq m < n \leq M}$ or to compute $L(\hat{\boldsymbol{t}})$, otherwise the methods seeks the value of $\boldsymbol{S}$ that best explains the TDE. We have experimentally observed that, in most of the cases, one solution is "inside" the microphone array and the other one is "outside" the array. However, the cost function behaves differently around these two solutions. Indeed, H is much sharper around the solution inside the microphone array. Hence, this is usually the one found by the optimization procedure. That is why we had to modify the cost function in order to bias the optimization:

$$\tilde{H}(\boldsymbol{S}) = H(\boldsymbol{S}) + \lambda\left(\|\boldsymbol{S}\|^2 - r\right)^2, \tag{27}$$

where r is the desired radius of the solution and λ is a regularization parameter. This way of constraining the optimization

TABLE I

RESULTS OBTAINED WITH BOTH SIMULATED DATA AND REAL DATA (LAST COLUMN). THE FIRST ROW SHOWS THE SNR RATIO IN DB. THE SECOND ROW SHOWS THE VALUES OF T_{60} IN SECONDS. THE REMAINING ROWS SHOW THE RESULTS WITH THE METHODS OUTLINED IN SECTION VII-B. FOR EACH SNR–T_{60} COMBINATION AND FOR EACH METHOD, WE DISPLAY THREE VALUES: (I) THE PROPORTION OF INLIERS (THE ANGULAR ERROR IS LESS THAN $30°$), (II) THE MEAN ANGULAR ERROR OF INLIERS, AND (III) THEIR STANDARD DEVIATION. COLUMN WISE, THE BEST RESULTS ARE SHOWN IN **BOLD**. NOTICE THAT *X-MULT* METHODS OFTEN YIELD THE BEST MEAN ANGULAR ERROR OF INLIERS. HOWEVER, *X-MULT* REQUIRES AN ESTIMATE OF THE ARRAY-TO-SOURCE DISTANCE

SNR	0					−5					−10					Real data
T_{60}	0	0.1	0.2	0.4	0.6	0	0.1	0.2	0.4	0.6	0	0.1	0.2	0.4	0.6	0.5
b&b	**82.1%** 9.59 **3.66**	**82.8%** 10.49 **4.47**	**73.8%** 12.65 **6.14**	**48.3%** 14.99 **7.09**	**35.7%** 16.10 7.30	**84.1%** 10.46 **4.64**	**82.7%** 11.58 **5.45**	**68.6%** 13.91 **6.75**	**41.1%** 16.07 7.44	**29.8%** 16.97 7.35	**77.5%** 13.45 **6.56**	**66.6%** 14.35 **6.85**	**44.5%** 16.53 7.36	**24.8%** 18.29 7.29	**19.0%** 18.63 7.26	**27.5%** 16.04 7.55
unc	38.3% 15.89 7.47	36.2% 16.15 7.30	33.3% 17.01 7.46	23.3% 17.94 7.30	19.2% 18.60 7.69	37.5% 16.76 7.38	37.0% 16.92 7.36	31.5% 17.71 7.41	21.4% 18.48 7.30	16.7% 18.74 **7.26**	33.4% 17.28 7.51	29.2% 17.75 7.62	21.6% 18.67 7.34	14.3% 18.95 **7.21**	11.6% 19.54 7.03	14.1% 18.93 7.13
d-lb	75.3% 10.54 4.57	75.4% 11.55 5.26	67.5% 13.54 6.51	44.6% 15.53 7.21	33.4% 16.25 7.22	80.4% 11.74 5.52	77.9% 12.99 6.35	61.9% 14.74 6.91	36.8% 16.54 7.63	28.0% 17.11 7.38	66.6% 14.69 6.90	56.5% 16.01 7.19	36.3% 17.27 7.37	20.6% 18.28 7.30	15.8% 18.61 7.20	22.3% 17.51 7.53
s-lb	46.9% 11.63 5.54	46.7% 12.58 6.17	40.9% 14.79 6.98	27.9% 17.05 7.33	22.2% 17.67 7.13	39.3% 13.41 6.41	40.8% 14.58 6.74	34.4% 16.60 7.21	23.6% 17.76 7.33	18.7% 18.19 7.28	31.0% 17.13 7.36	29.7% 17.85 7.31	22.1% 18.46 **7.00**	14.5% 19.17 7.41	13.2% 19.65 7.26	13.2% 18.80 7.09
dm	80.3% 15.75 7.11	77.4% 15.94 7.18	62.4% 16.49 7.38	41.2% 17.04 7.50	30.2% 17.76 7.48	78.3% 15.80 7.12	74.0% 16.09 7.15	57.7% 16.48 7.35	34.8% 17.53 7.55	26.9% 17.82 7.45	60.4% 16.65 7.30	51.3% 16.90 7.31	35.6% 17.86 7.33	21.4% 18.76 7.29	16.3% 19.03 7.34	16.7% 19.34 7.00
n-mult	60.8% 7.99 5.45	54.6% 9.00 6.23	40.5% 11.18 7.43	22.6% **14.11** 8.11	16.0% 15.46 7.97	49.7% **9.23** 6.30	46.3% **10.56** 6.98	35.1% 13.18 7.51	18.6% **15.82** 7.93	13.7% 16.79 7.51	41.9% **13.11** 7.32	36.0% **14.28** 7.46	23.0% 16.26 7.68	12.8% **17.84** 7.40	10.1% 18.23 **7.01**	13.2% 17.54 7.27
t-mult	61.0% **7.81** 5.09	53.1% **8.75** 6.03	39.4% **11.06** 7.22	22.0% 14.14 8.10	15.5% **15.41** 7.74	50.0% 9.42 6.14	45.5% 10.72 7.00	34.1% **13.09** 7.43	18.3% 16.02 7.88	13.6% **16.52** 7.36	41.3% 13.65 7.34	35.1% 14.50 7.38	22.4% **16.22** 7.39	12.3% 18.15 7.45	9.9% **18.17** 7.22	17.38% 16.1 7.80
f-mult	59.6% 8.38 5.35	52.5% 9.31 6.16	38.6% 11.51 7.21	21.7% 14.27 7.90	15.0% 15.69 7.66	48.8% 9.92 6.24	44.5% 11.22 7.03	34.0% 13.54 7.33	17.9% 16.21 7.70	13.5% 16.92 7.37	40.6% 13.87 7.27	34.7% 14.75 7.41	21.9% 16.34 7.36	12.0% 17.87 7.41	9.8% 18.31 7.21	17.91% **15.3** 7.78
pi	53.7% 11.31 5.55	53.3% 12.47 6.17	44.3% 14.60 6.92	30.3% 16.81 7.33	23.6% 17.67 7.60	41.4% 13.24 6.11	41.5% 14.23 6.71	32.7% 16.50 7.09	21.9% 18.12 **7.15**	16.9% 18.08 7.43	29.6% 17.04 7.19	28.9% 17.76 7.17	20.8% 18.92 7.49	14.7% 18.98 7.58	12.5% 19.25 7.22	12.9% 19.03 **6.88**
random	7.04% 19.63 7.1491															

problem is justified by the well-known fact that the distance to the source is very difficult to estimate with egocentric arrays. We tested three different values for r: *n-mult* (near-multilateration) corresponds to $r = 0.9$, *t-mult* (true-multilateration) corresponds to $r = 1.7$ (which is the actual distance from the array to the source), and *f-mult* (far-multilateration) corresponds to $r = 2.5$ (all measures are in meters). In all cases the optimization procedure was initialized on a grid of 200 directions, $\mathcal{G}_d$, and, in order to increase the accuracy and convergence speed, the analytic derivatives were computed (see Appendix C).

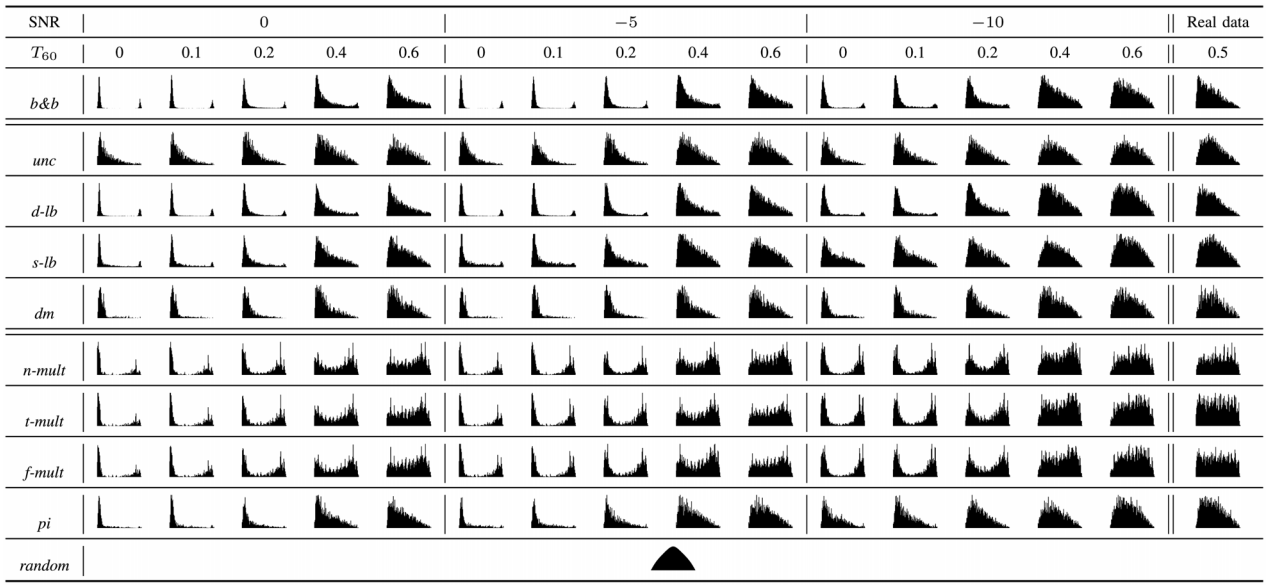
C. Results and Discussion

Table I summarizes the results obtained with the evaluated methods on simulated as well as on real data. While columns 2 to 16 correspond to simulated data, the last column corresponds to real data. In the simulated case, the first row displays the SNR in dB and the second row displays T_{60} in seconds. The next rows display the performance of the evaluated methods, and are split into three groups by double lines: the proposed *b&b* method (top), state-of-the-art multichannel TDE methods (middle), and state-of-the-art multilateration methods (bottom). For each SNR–T_{60} combination and for each method we provide three numbers: (i) the percentage of inliers (angular error $< 30°$), (ii) the average and (iii) the standard deviation of the angular error over the inliers. These quantities are computed on 100 ms long signals received from the 189 different sound source positions, and the best values are shown in bold. On an average, each entry of the table roughly corresponds to 3,000 localization trials. Overall, we performed more than 400,000 localizations. The error of the random localizer follows an unimodal distribution symmetric with respect to its mean, $90°$, with a standard deviation of approximately $40°$. Last row of Table I (*random*) characterizes the random localizer in terms of inliers, inlier mean and standard deviation.

Roughly speaking, multichannel algorithms yield better results than multilateration. Among the algorithms belonging to the latter class, *n-mult*, *n-mult*, and *n-mult* (jointly referred to as *x-mult*) perform better than *pi* in low-reverberant environments, independently of the level of noise. However, in high reverberant environments, *x-mult* performs slightly worse than *pi*. It is worth noticing that, the value of the parameter r (in the constrained optimization formulation (27)) has a small impact onto the method's performance. Indeed, the variation of the inlier percentage is not greater than 2% and the variation of the mean and standard deviation is less than $1°$.

All the multichannel TDE algorithms perform as expected with respect to the environmental parameters: The performance decreases as T_{60} is increased. However, the SNR and T_{60} have different effects on the objective function, J. On one side, the sensor noise decorrelates the microphone signals leading to many randomly spread local minima and increasing the value of the true minimum. If this effect is extreme, the hope for a good estimate decreases fast. On the other side, the reverberations produce only a few strong local minima. This perturbation is systematic given the source position in the room. Hence, there is hope to learn the effect of such reverberations in order to

TABLE II
THIS TABLE DISPLAYS THE HISTOGRAMS ASSOCIATED WITH THE LOCALIZATION ERROR, ORGANIZED IN THE SAME WAY AS TABLE I. THE HISTOGRAM ABSCISSAE START AT 0° (NO ERROR) AND SPAN TO 180° (MAXIMUM ERROR)



improve the quality of the estimates. Clearly, these perturbation types (noise and reverberations) have different effects on the results.

Concerning the methods themselves, we noticed that *unc* achieves a very low percentage of inliers. This fully justifies the need of the geometric constraint introduced in this paper. In other words, the cost function (21) suffers from a lot of local minima outside the feasible domain. Thus, the naive idea of estimating the time delays without adding information about the geometry of the microphone array, does not really work. We also remark that, except for the two first cases (the easiest ones), the *d-lb* method outperforms the *dm* method, since the former carries out a local minimization. Regarding the two easiest cases, it is not clear which of the two methods shows a better performance. On one side, *dm* captures more inliers. On the other side, both the mean angular error and standard deviation over the set of inliers are significantly lower with *d-lb* than with *dm*. The sparse initialization strategy does not show a remarkable performance. Indeed, the localization quality is comparable to *d-lb*, but the percentage of inliers is much lower. Thus, a method able to deal with large amounts of outliers should be added in order to clean up the localization results provided by *s-lb*. More importantly, we highlight the performance of *b&b*. This method yields the highest percentage of inliers in all the tests. Moreover, the quality of the localization is comparable, if not better, with *d-lb* or with *x-mult*. Regarding the percentage of outliers and the standard deviation, the *b&b* methods proves to obtain the best results. Notice that multilateration methods often yield the best mean angular error of inliers. However, these methods must be provided an estimate of the distance to the source: *t-mult* was provided with the true array-to-source distance. We conclude that *b&b* is the method of choice in the presence of noise and outliers.

The behavior of the methods described above on simulated data is similar on real data. The proposed multichannel method outperforms the state-of-the-art on both multichannel TDE and multilateration. Among the multichannel algorithms, *unc* and *s-lb* show very bad performance. Even if *dm*, and specially *d-lb* prove to work to some extent, the best method is *b&b* since it has the highest percentage of inliers and the localization quality is comparable to the one of *d-lb* and of *x-mult*. Finally, we noticed that the results on real data roughly correspond to the simulated case with $T_{60} = 0.6$ s and $SNR = -5$ dB, which is a very challenging scenario.

The results of Table I are complemented by error histograms displayed in Table II. This table has a row-column structure that is strictly identical with Table I. The histograms count all localization trials, contrary to Table I where only the inliers were used to compute the mean and the standard deviation. Table II is meant to provide a qualitative evaluation of the error distribution. An ideal localization situation would show a histogram with a full leftmost bin (0° error) and all the other bins being empty. One can see that the results of both tables are correlated. On one hand, good localization results correspond to histograms whose mass is mainly concentrated to the left. On the other hand, bad localization results correspond to histograms whose mass is evenly distributed over the histogram bins. However, we observe three different histogram patterns among the methods reporting low performance. First, in some cases, a very rough estimate of the position could be extracted, namely the mass is concentrated on the left half of the histogram (but not close to 0°), e.g., (*b&b*, −10, 0.2), (*dm*, −5, 0.4) or (*pi*, 0, 0.6). Second, in some cases a fairly accurate estimate is obtained, but only in 50% of the trials. This translates into a histogram with two large peaks, one of them close to 0° and the other far apart. Consequently, some rough estimate of the source position would allow to discard most of the outliers. Examples of this are the *x-mult* methods. Third, the worst case scenario, in which the error is uniformly distributed, do not provide any meaningful result, namely for $SNR = -10$ dB and $T_{60} = 0.6$ s.

Finally, we performed a statistical analysis of the execution times. Table III shows the average and standard deviation of the execution times for each one of the tested methods and on 400 trials. All the methods were implemented in MATLAB and

TABLE III
AVERAGE EXECUTION TIMES AND STANDARD DEVIATIONS FOR EACH METHOD ON 400 TRIALS. ALL QUANTITIES ARE EXPRESSED IN SECONDS

Method	*b&b*	*unc*	*d-lb*	*s-lb*	*dm*	*x-mult*	*pi*
Mean	9.17	25.13	10.34	0.958	0.103	2.02	0.014
Std	1.082	2.759	1.316	0.810	0.004	0.265	0.0003

the code was run on the very same computer. The methods *n-mult*, *t-mult* and *f-mult* were not evaluated separately because the parameter r does not have any effect on the execution time. We first observe that methods with low computational complexity correspond to methods that are not robust (*pi*, *x-mult*, *dm* and *s-lb*). There are some methods that are neither robust nor fast (*unc*). A couple of methods present high robustness but high complexity (*d-lb* and *b&b*). However, optimization techniques and smart approximations will lead to *b&b*-based localization algorithms that are both efficient and robust. Indeed, platform-dedicated algorithm optimization will reduce the computational time of the proposed sound source localization procedure. Moreover, the accuracy of the localization results may be adjusted to the desired application/environment. For example, we could use the proposed framework to obtain a rough estimate of the sound source position. The semantic context of the ongoing social interplay will help us select the location of interest among the rough estimates. This coarse location of interest could be then refined with the very same algorithm, but with a much smaller search space.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper we addressed the problem of sound source localization from time delay estimates using non-coplanar microphone arrays. Starting from the direct path propagation model, we derived the full geometric analysis associated with an arbitrarily shaped non-coplanar microphone array. The necessary and sufficient conditions for the time delays to correspond to a position in the source space are expressed by means of two feasibility conditions. If they are satisfied, the position of the sound source can be recovered in closed-form, from the TDEs. Remarkably, the only knowledge required to build the feasibility conditions and the localization mapping is the microphones' position. A multichannel criterion for TDE allows us to cast the problem into an optimization task, which is constrained by the feasibility conditions. A branch-and-bound global optimization technique is proposed to solve the programming task, and hence to estimate the time delays and to localize the sound source. An extensive set of experiments is performed on simulated and real data. The experiments clearly show that the global optimization technique that we proposed outperforms existing methods in both the multilateration and the multichannel SSL literatures.

This work could be extended in several ways. First of all, considering the multiple source case. This could be achieved using a frequency filter bank, that would also discard empty frequency bands as in [52]. Second, a different set of experiments could be performed on distributed microphone arrays, to evaluate the behavior of the proposed methods in such settings. Other TDOA estimators, possibly more accurate than [26], could be used in our benchmark, such that *f-mult*, *t-mult*, and *n-mult* yield better results. Third, the method could also be used in calibration applications. Indeed, the positions of the microphones could be estimated if they were free parameters in our current formulation. In that case, measures from many different source positions would certainly be required, e.g., [25]. Fourth, by testing the proposed model and algorithms in the case of dynamic sources, and subsequently extending the framework to perform tracking. Finally, experiments with higher number of microphones should be performed, and the influence of the microphones' positions should be evaluated.

APPENDIX A
THE DERIVATIVES OF THE COST FUNCTION

The log-barrier algorithm relies on the use of the gradient and the Hessian of both, the objective function and the constraint(s). Providing the analytic expression for them would lead to a much more efficient and precise algorithm than estimating them using finite differences. Hence, this Section is devoted to the derivation of both the gradient and the Hessian of J, the cost function of (23).

We will use three formulas from matrix calculus. Let $\mathbf{Y} : \mathbb{R} \rightarrow \mathbb{R}^{M\times M}$, be a matrix function depending on y, the following formulas hold:

$$\frac{\partial \det(\mathbf{Y})}{\partial y} = \det(\mathbf{Y})\mathrm{trace}\left(\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial y}\right)$$
$$\frac{\partial \mathrm{trace}(\mathbf{Y})}{\partial y} = \mathrm{trace}\left(\frac{\partial \mathbf{Y}}{\partial y}\right)$$
$$\frac{\partial \mathbf{Y}^{-1}}{\partial y} = -\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial y}\mathbf{Y}^{-1}$$

Recall that the function we want to derivative is $J = \det(\mathbf{R})$. From the rules of matrix calculus we have:

$$\frac{\partial J}{\partial t_{1,m}} = \frac{\partial \det(\mathbf{R})}{\partial t_{1,m}} = \det(\mathbf{R})\,\mathrm{trace}\left(\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,m}}\right). \quad (28)$$

In addition we can compute the second derivative:

$$\frac{\partial^2 J}{\partial t_{1,n}\partial t_{1,m}} = \frac{\partial}{\partial t_{1,n}}\left[\det(\mathbf{R})\,\mathrm{trace}\left(\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,m}}\right)\right], \quad (29)$$

$$\frac{\partial^2 \tilde{J}}{\partial t_{1,j}\partial t_{1,k}} = \det(\mathbf{R})\left[\mathrm{trace}\left(\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,j}}\right)\mathrm{trace}\left(\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,k}}\right) + \mathrm{trace}\left(-\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,j}}\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial t_{1,k}} + \mathbf{R}^{-1}\frac{\partial^2 \mathbf{R}}{\partial t_{1,j}\partial t_{1,k}}\right)\right]. \quad (30)$$

whose full expression can be found in (30). Hence, in order to compute the first and second derivatives of the criterion J, we need the first and second derivatives of the matrix $\mathbf{R}$. We recall its expression:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho_{1,2}(t_{1,2}) & \rho_{1,3}(t_{1,3}) & \cdots \\ \rho_{1,2}(t_{1,2}) & 1 & \rho_{2,3}(-t_{1,2}+t_{1,3}) & \cdots \\ \rho_{1,3}(t_{1,3}) & \rho_{2,3}(-t_{1,2}+t_{1,3}) & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

We notice that only the $m-1$th row and column depend on $t_{1,m}$. Since $\mathbf{R}$ is symmetric, we do not need to take derivative

of the $m-1$th row and column separately, but compute only the derivative of:

$$\boldsymbol{r}_m = \begin{pmatrix} \rho_{1,m}(t_{1,m}) \\ \rho_{2,m}(-t_{1,2}+t_{1,m}) \\ \vdots \\ \rho_{m-1,m}(-t_{1,m-1}+t_{1,m}) \\ 1 \\ \rho_{m,m+1}(-t_{1,m}+t_{1,m+1}) \\ \vdots \\ \rho_{m,M}(-t_{1,m}+t_{1,M}) \end{pmatrix},$$

which is

$$\frac{\partial \boldsymbol{r}_m}{\partial t_{1,m}} = \begin{pmatrix} \rho'_{1,m}(t_{1,m}) \\ \rho'_{2,m}(-t_{1,2}+t_{1,m}) \\ \vdots \\ \rho'_{m-1,m}(-t_{1,m-1}+t_{1,m}) \\ 0 \\ -\rho'_{m,m+1}(-t_{1,m}+t_{1,m+1}) \\ \vdots \\ -\rho'_{m,M}(-t_{1,m}+t_{1,M}) \end{pmatrix}.$$

When computing the second derivative of $\mathbf{R}$ with respect to $t_{1,m}$ and $t_{1,n}$ we need to differentiate two cases: $\boxed{m=n}$ This fills the diagonal of the Hessian. Notice that:

$$\frac{\partial^2 \boldsymbol{r}_m}{\partial t_{1,m}^2} = \begin{pmatrix} \rho''_{1,m}(t_{1,m}) \\ \rho''_{2,m}(-t_{1,2}+t_{1,m}) \\ \vdots \\ \rho''_{m-1,m}(-t_{1,m-1}+t_{1,m}) \\ 0 \\ \rho''_{m,m+1}(-t_{1,m}+t_{1,m+1}) \\ \vdots \\ \rho''_{m,M}(-t_{1,m}+t_{1,M}) \end{pmatrix}.$$

$\boxed{m>n}$ This fills the lower triangular matrix of the Hessian (and the upper triangular part since the Hessian is symmetric, i.e., that $\tilde{J}$ is twice continuously differentiable). Only the $n-1$th position of the vector $\frac{\partial^2 \boldsymbol{r}_m}{\partial t_{1,n}\partial t_{1,m}}$ is not null, taking the value: $-\rho''_{m,n}(t_{1,m}-t_{1,n})$.

Appendix B
The Derivatives of the Constraints

In this Section we compute the formulae for the first and the second derivatives of the non-linear constraint Δ. Recall the expression from (15):

$$\Delta = \langle \boldsymbol{A}, \boldsymbol{B}-\boldsymbol{M}_1\rangle^2 - \|\boldsymbol{B}-\boldsymbol{M}_1\|^2\left(\|\boldsymbol{A}\|^2-1\right),$$

where $\boldsymbol{A} = -\mathbf{M}_L^{-1}\boldsymbol{P}_L$ and $\boldsymbol{B} = -\mathbf{M}_L^{-1}\boldsymbol{Q}_L$. It is easy to show that:

$$\nabla\Delta = 2\left(\langle \boldsymbol{A}, \boldsymbol{B}-\boldsymbol{M}_1\rangle\left(\mathbf{J}^\top{}_{\boldsymbol{A}}(\boldsymbol{B}-\boldsymbol{M}_1)+\mathbf{J}^\top{}_{\boldsymbol{B}}\boldsymbol{A}\right) - \right. \\ \left. - (\|\boldsymbol{A}\|^2-1)\mathbf{J}^\top{}_{\boldsymbol{B}}(\boldsymbol{B}-\boldsymbol{M}_1) - \|\boldsymbol{B}-\boldsymbol{M}_1\|^2\mathbf{J}^\top{}_{\boldsymbol{A}}\boldsymbol{A}\right)$$

where $\mathbf{J}_{\boldsymbol{A}} = -2\nu\mathbf{M}_L^{-1}$ and $\mathbf{J}_{\boldsymbol{B}} = -2\nu^2\mathbf{M}_L^{-1}\mathrm{diag}(\hat{\boldsymbol{t}})$. We can also compute the Hessian of Δ:

$$\mathbf{H}\Delta = 2\Big(\Big(\mathbf{J}_{\boldsymbol{A}}^\top(\boldsymbol{B}-\boldsymbol{M}_1)+\mathbf{J}_{\boldsymbol{B}}^\top\boldsymbol{A}\Big)\Big(\mathbf{J}_{\boldsymbol{A}}^\top(\boldsymbol{B}-\boldsymbol{M}_1)+\mathbf{J}_{\boldsymbol{B}}^\top\boldsymbol{A}\Big)^\top + \\ + \langle \boldsymbol{A}, \boldsymbol{B}-\boldsymbol{M}_1\rangle\Big(\mathbf{J}_{\boldsymbol{A}}^\top\mathbf{J}_{\boldsymbol{B}}+\mathbf{D}+\mathbf{J}_{\boldsymbol{B}}^\top\mathbf{J}_{\boldsymbol{A}}\Big) - \\ -\Big[2(\mathbf{J}_{\boldsymbol{B}}^\top(\boldsymbol{B}-\boldsymbol{M}_1))(\mathbf{J}_{\boldsymbol{A}}^\top\boldsymbol{A})^\top+(\|\boldsymbol{A}\|^2-1)(\mathbf{E}+\mathbf{J}_{\boldsymbol{B}}^\top\mathbf{J}_{\boldsymbol{B}}) + \\ + 2(\mathbf{J}_{\boldsymbol{A}}^\top\boldsymbol{A})(\mathbf{J}_{\boldsymbol{B}}^\top(\boldsymbol{B}-\boldsymbol{M}_1))^\top+\|\boldsymbol{B}-\boldsymbol{M}_1\|^2\mathbf{J}_{\boldsymbol{A}}^\top\mathbf{J}_{\boldsymbol{A}}\Big]\Big)$$

where $\mathbf{D} = -2\nu^2\mathrm{diag}(\mathbf{M}_L^{-1}\boldsymbol{A})$ and $\mathbf{E} = -2\nu^2\mathrm{diag}(\mathbf{M}_L^{-1}(\boldsymbol{B}-\boldsymbol{M}_1))$.

Similarly, we can compute the derivatives of the equality constraint $\mathcal{E}$.

Appendix C
The Derivatives of the Multilateration Cost Function

In this Section we provide the first and second derivatives of the cost function used by the methods *n-mult*, *t-mult* and *f-mult*. Denoted by H, the cost function has the following expression:

$$H(\boldsymbol{S}) = \sum_{1\le m<n\le M}(h_{m,n}(\boldsymbol{S}))^2, \tag{31}$$

where $h_{m,n}$ are the equivalent of (4) with foci $\boldsymbol{M}_m$ and $\boldsymbol{M}_n$ and differential value $\hat{t}_{m,n}$. In all, $h_{m,n}$ takes the following expression:

$$h_{m,n}(\boldsymbol{S}) = q_{m,n}^2 + 4\langle \boldsymbol{S}, \boldsymbol{M}_n-\boldsymbol{M}_m\rangle^2 - \\ - 4q_{m,n}\langle \boldsymbol{S}, \boldsymbol{M}_n-\boldsymbol{M}_m\rangle - p_{m,n}^2\|\boldsymbol{S}-\boldsymbol{M}_n\|^2.$$

The gradient of H writes:

$$\nabla H = 2\sum_{1\le m<n\le M} h_{m,n}\nabla h_{m,n},$$

where

$$\nabla h_{m,n} = 8\langle \boldsymbol{S}, \boldsymbol{M}_n-\boldsymbol{M}_m\rangle(\boldsymbol{M}_n-\boldsymbol{M}_m) - \\ - 4q_{m,n}(\boldsymbol{M}_n-\boldsymbol{M}_m) - 2p_{m,n}^2(\boldsymbol{S}-\boldsymbol{M}_n).$$

Similarly, the Hessian of H can be computed as:

$$\mathbf{H}H = 2\sum_{1\le m<n\le M}\nabla h_{m,n}(\nabla h_{m,n})^\top + h_{m,n}\mathbf{H}h_{m,n},$$

where the Hessian of $h_{m,n}$ has the following expression:

$$\mathbf{H}h_{m,n} = 8(\boldsymbol{M}_n-\boldsymbol{M}_m)(\boldsymbol{M}_n-\boldsymbol{M}_m)\top - 2p_{m,n}^2\mathbf{I}_N.$$

References

[1] X. Alameda-Pineda and R. Horaud, "Geometrically-constrained robust time delay estimation using non-coplanar microphone arrays," in *Proc. EUSIPCO*, Bucharest, Romania, Aug. 2012, pp. 1309–1313.

[2] J. Benesty, Y. Huang, and J. Chen, "Time delay estimation via minimum entropy," *IEEE Signal Process. Lett.*, vol. 14, no. 3, pp. 157–160, Mar. 2007.

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[4] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 45–50, Jan. 1997.

[5] M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput. Speech Lang.*, vol. 11, no. 2, pp. 91–126, 1997.

[6] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *Proc. Interspeech*, 2005.

[7] A. Brutti, M. Omologo, and P. Svaizer, "Speaker localization based on oriented global coherence field," in *Proc. Interspeech*, 2006, vol. 7, p. 8.

[8] A. Canclini, E. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 439–443, Feb. 2013.

[9] P. Carbonetto, MATLAB primal-dual interior-point solver for convex programs with constraints, 2008, [Online]. Available: http://www.cs.ubc.ca/ pcarbo/convexprog.html

[10] J. Chen, K. Yao, and R. Hudson, "Acoustic source localization and beamforming: Theory and practice," *EURASIP J. Appl. Signal Process.*, pp. 359–370, 2003.

[11] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, Nov. 2003.

[12] J. Chen, J. Benesty, and Y. Huang, "Time Delay Estimation in Room Acoustic Environments: An Overview," *EURASIP J. Adv. Signal Process.*, vol. 2006, no. i, pp. 1–20, 2006.

[13] R. J. Dakin, "A tree-search algorithm for mixed integer programming problems," *Comput. J.*, vol. 8, no. 3, pp. 250–255, 1965.

[14] A. Deleforge and R. Horaud, "2d sound-source localization on the binaural manifold," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, 2012, pp. 1–6.

[15] A. Deleforge and R. Horaud, "The cocktail party robot: Sound source separation, and localisation with an active binaural head," in *Proc. IEEE/ACM Int. Conf. Human Robot Interact.*, Boston, MA, USA, Mar. 2012.

[16] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 11, pp. 1110–1124, 2003.

[17] B. Friedlander, "A passive localization algorithm and its accuracy analysis," *IEEE J. Ocean. Eng.*, vol. 12, no. 1, pp. 234–245, 1987.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *Timit acoustic-phonetic continuous speech corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[19] H. He, J. Lu, L. Wu, and X. Qiu, "Time delay estimation via non-mutual information among multiple microphones," *Appl. Acoust.*, vol. 74, no. 8, pp. 1033–1036, 2013.

[20] H. He, L. Wu, J. Lu, X. Qiu, and J. Chen, "Time difference of arrival estimation exploiting multichannel spatio-temporal prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 463–475, Mar. 2013.

[21] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[22] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.

[23] F. Keyrouz, K. Diepold, and S. Keyrouz, "Humanoid binaural sound tracking using Kalman Filtering and HRTFs," *Robot Motion and Control*, pp. 329–340, 2007.

[24] F. Keyrouz and K. Diepold, "An enhanced binaural 3D sound localization algorithm," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, Aug. 2006, pp. 662–665.

[25] V. Khalidov, F. Forbes, and R. Horaud, "Alignment of binocular-binaural data using a moving audio-visual target," in *Proc. IEEE Workshop Multimedia Signal Process. (MMSP)*, Pula, Sardinia, Italy, Sep.–Oct. 2013.

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

[27] A. R. Kullaib, M. Al-Mualla, and D. Vernon, "2D binaural sound localization: For urban search, and rescue robotics," in *Proc. Mobile Robot.*, Istanbul, Turkey, Sep. 2009, pp. 423–435.

[28] E. A. Lehmann, Matlab code for image-source model in room acoustics, [Online]. Available: http://www.eric-lehmann.com/ism_code.html 2012, accessed Nov. 2011

[29] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, 2008.

[30] S. Leyffer, "Integrating sqp and branch-and-bound for mixed integer nonlinear programming," *Comput. Optimizat. Applicat.*, vol. 18, no. 3, pp. 295–309, 2001.

[31] Z. Li and R. Duraiswami, "A robust, and self-reconfigurable design of spherical microphone array for multi-resolution beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '05)*, 2005, vol. 4, pp. iv/1137–iv/1140.

[32] J.-S. Lim and H.-S. Pang, "Time delay estimation method based on canonical correlation analysis," *Circuits, Syst., Signal Process.*, vol. 32, no. 5, pp. 2527–2538, Mar. 2013.

[33] R. Liu and Y. Wang, "Azimuthal source localization using interaural coherence in a robotic dog: Modeling and application," *Robotica*, vol. 28, no. 7, pp. 1013–1020, 2010.

[34] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Proc. NIPS*, Cambridge, MA, USA, 2007, pp. 953–960.

[35] R. D. Mori, *Spoken Dialogues with Computers*. Orlando, FL, USA: Academic., 1997.

[36] K. Nakamura, K. Nakadai, F. Asano, and G. Ince, "Intelligent sound source localization, and its application to multimodal human tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2011, pp. 143–148.

[37] M. Omologo, P. Svaizer, A. Brutti, and L. Cristoforetti, "Speaker localization in chil lectures: Evaluation criteria, and results," in *Machine Learning for Multimodal Interaction, volume 3869 of Lecture Notes in Computer Science*, S. Renals and S. Bengio, Eds. Berlin/Heidelberg, Germany: Springer, 2006, pp. 476–487.

[38] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, Oct. 2013.

[39] P. Pertilä, Acoustic source localization in a room environment and at moderate distances Tampereen teknillinen yliopisto. Julkaisu-Tampere Univ. of Technol., Publication; 794, 2009.

[40] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[41] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication, volume 3 of Springer Topics in Signal Processing*, I. Cohen, J. Benesty, and S. Gannot, Eds. Berlin/Heidelberg, Germany: Springer, 2010, pp. 281–305.

[42] D. Salvati and S. Canazza, "Adaptive time delay estimation using filter length constraints for source localization in reverberant acoustic environments," *IEEE Signal Process. Lett.*, vol. 20, no. 5, pp. 507–510, May 2013.

[43] Y. Sasaki, M. Kabasawa, S. Thompson, S. Kagami, and K. Oro, "Spherical microphone array for spatial sound localization for a mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2012, pp. 713–718.

[44] M. Ó. Searcóid, *Metric spaces*. New York, NY, USA: Springer, 2006.

[45] F. Seco, A. Jiménez, C. Prieto, J. Roa, and K. Koutsou, "A survey of mathematical methods for indoor localization," in *Proc. IEEE Int. Symp. Intell. Signal Process. (WISP'09)*, 2009, pp. 9–14.

[46] X. Sheng and Y.-h. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 44–53, Jan. 2005.

[47] J. Smith and J. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 12, pp. 1661–1669, Dec. 1987.

[48] H. So, Y. Chan, and F. Chan, "Closed-form formulae for time-difference-of-arrival estimation," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2614–2620, Jun. 2008.

[49] N. Strobel and R. Rabenstein, "Classification of time delay estimates for robust speaker localization," in *Proc., IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'99)*, 1999, vol. 6, pp. 3081–3084.

[50] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using eb-esprit with spherical microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'11)*, 2011, pp. 117–120.

[51] A. Urruela and J. Riba, "Novel closed-form ML position estimator for hyperbolic location," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'04)*, 2004, vol. 2, pp. ii–149.
[52] J. Valin and F. Michaud, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *Proc. IEEE Trans. Acoust., Speech, Signal Process.*, vol. 2, no. 1, pp. 841–844.
[53] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," *Proc. DAFx*, pp. 209–213, 2003.
[54] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012.
[55] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 538–548, Apr. 2008.
[56] C. Zhang, Z. Zhang, and D. Florencio, "Maximum Likelihood Sound Source Localization for Multiple Directional Microphones," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '07)*, 2007, no. 6, pp. 125–128.

Xavier Alameda-Pineda was born in Barcelona, Catalunya. He received the M.Sc. degree in mathematics and telecommunications engineering from the Universitat Politècnica de Catalunya–BarcelonaTech in 2008 and 2009 respectively, the M.Sc. degree in computer science from the Université Joseph Fourier and Grenoble INP in 2010, and the Ph.D. degree in mathematics/computer science from the Université Joseph Fourier in 2013. He worked towards his Ph.D. degree in the Perception Team, at INRIA Grenoble Rhône-Alpes, where he currently holds a postdoctoral. His research interests are machine learning and signal processing for scene understanding, speaker diaritzation and tracking and sound source separation.

Radu Horaud received the B.Sc. degree in electrical engineering, the M.Sc. degree in control engineering, and the Ph.D. degree in computer science from the Institut National Polytechnique de Grenoble, Grenoble, France. Currently he holds a position of director of research with the Institut National de Recherche en Informatique et Automatique (INRIA), Grenoble Rhône-Alpes, Montbonnot, France, where he is the founder and head of the PERCEPTION team. His research interests include computer vision, machine learning, audio signal processing, audiovisual analysis, and robotics. He is an area editor of the *Elsevier Computer Vision and Image Understanding*, a member of the advisory board of the *Sage International Journal of Robotics Research*, and an associate editor of the *Kluwer International Journal of Computer Vision*. He was Program Cochair of the Eighth IEEE International Conference on Computer Vision (ICCV 2001). In 2013, Radu Horaud was awarded a five year ERC Advanced Grant for his project *Vision and Hearing in Action* (VHIA).