

Efficient Steered-Response Power Methods for Sound Source Localization Using Microphone Arrays

Markus V. S. Lima, Wallace A. Martins, Leonardo O. Nunes, Luiz W. P. Biscainho, Tadeu N. Ferreira,
Maurício V. M. Costa, Bowon Lee

Abstract—This paper proposes an efficient method based on the steered-response power (SRP) technique for sound source localization using microphone arrays: the volumetric SRP (V-SRP). As compared to the SRP, by deploying a sparser volumetric grid, the V-SRP achieves a significant reduction of the computational complexity without sacrificing the accuracy of the location estimates. By appending a fine search step to the V-SRP, its refined version (RV-SRP) improves on the compromise between complexity and accuracy. Experiments conducted in both simulated- and real-data scenarios demonstrate the benefits of the proposed approaches. Specifically, the RV-SRP is shown to outperform the SRP in accuracy at a computational cost of about ten times lower.

Index Terms—Sound source localization, steered-response power, microphone array, computational complexity.

I. INTRODUCTION

SOUND source localization (SSL) with microphone arrays is key to many applications such as 3-D audio capture, speech enhancement for hearing aids in medical applications, vehicle and gunshot localization for military use, automatic camera steering for event broadcasting or video conferencing, and video games [1]. SSL methods exploit spatial diversity by using multiple microphones to simultaneously acquire different versions of emitted source signals, which are then jointly processed. Knowing the location of a given source enables the enhancement of its associated acquired signals, e.g. beamforming [1], thus providing higher signal-to-noise ratio (SNR) than a single-microphone capture would achieve.

The SSL field has borrowed/extended many of the techniques proposed for source localization using antenna arrays, which has been an active research area for more than forty years [2]. In the antenna array framework, most classical algorithms [3]–[5] were developed under the assumption that transmitted signals are sufficiently narrowband to allow that

phase drifts between the impinging signals on the several receiving sensors can be attributed only to source positioning. Since this narrow-band hypothesis does not hold for speech signals, broadband algorithms are the best choice in the SSL scenario [6].

A common approach to solve the localization problem is first estimating the *time-differences-of-arrival* (TDOAs) between the acquired signals and then mapping them into a source position. When the far-field hypothesis is valid, i.e. the distance between source and array is greater than approximately ten times [1] the length of the array aperture, the algorithms for DoA (direction-of-arrival) estimation can be employed.

An intuitive way to estimate the TDOA related to a pair of microphones can be devised if the cross-correlation between their two acquired signals is known: the lag associated with the maximum measured correlation provides the TDOA estimate itself. This is the basis of the cross-correlation (CC) method for source localization [1]. The generalized cross-correlation (GCC) method [7] adds robustness to the CC method by including a weighting function in the cross-spectrum. Different choices for this function lead to different algorithms [1], [8], [9], among which the *phase-transform* (PHAT) GCC [7] is the preferred scheme.

A natural extension of the GCC technique is the *steered-response power* (SRP) [10]–[14] method, which from now on will be denominated *classical SRP* (C-SRP). Compared to the GCC, which first estimates the TDOAs between acquired signals, the C-SRP algorithm becomes more robust to reverberation and noise effects [1], [14], [15] by performing a global optimization using all available information. In general terms, the C-SRP method can be implemented in two steps: (i) compute the cross-correlation function between the signals acquired by each microphone pair; and (ii) search for the source location over a grid of spatial points. The second stage is usually the most computationally demanding one since high localization accuracy implies using dense grids, which can be a major problem especially when facing large search spaces.

In addition to the computational problem due to the use of dense grids, increasing the number of microphones within an array has also been used to increase accuracy when the target application allows, as in the huge microphone arrays presented in [16], [17]. The number of grid points as well as the number of microphones impact directly the computational burden of the C-SRP, which even under common practical situations may

Copyright ©2014 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

M. V. S. Lima, W. A. Martins, L. O. Nunes, L. W. P. Biscainho, and M. V. M. Costa are with the Signals, Multimedia and Telecommunications Laboratory, SMT-DEL/Poli & PEE/COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil (e-mail: markus.lima@ieee.org, {wallace.martins, leonardo.nunes, wagner, mauricio.costa}@smt.ufrj.br).

T. N. Ferreira is with the Telecommunications Engineering Department, Fluminense Federal University, Niterói, Brazil (e-mail: tadeu_ferreira@id.uff.br).

B. Lee is with the Department of Electronic Engineering, Inha University, Incheon, South Korea (e-mail: bowon.lee@inha.ac.kr). He performed the work while at Hewlett-Packard Laboratories.

not achieve the desired accuracy. Therefore, SSL solutions aiming at reducing the computational complexity of SRP-based methods are called for.

A. Main Contributions

This paper proposes an efficient SRP-based procedure to tackle the problem of sound source localization. In contrast to the point-wise C-SRP method, the proposed *volumetric SRP* (V-SRP) algorithm operates on pre-defined non-overlapping spatial regions (hereafter loosely called *volumes*), each one containing a set of two or more grid points. The fact that there are more grid points than volumes, which then define a *volumetric grid*, is explored to significantly reduce the computational cost of the method, as it will become clear in Section VI. The V-SRP method has proven to be effective even when using sparse volumetric grids, i.e. large volumes.

Additionally, a *refined V-SRP* (RV-SRP) method is devised to tackle those situations when high accuracy is extremely important. In fact, it looks for a compromise between computational complexity and localization accuracy. This alternative method departs from the V-SRP search result, over which it performs a second refining step. The overall search results are much more accurate at little additional cost over the V-SRP.

Besides, both V-SRP and RV-SRP focus on reducing the number of computations required by the second step inherent to most SRP-based techniques, namely the search stage. The idea of a volumetric SRP was originally proposed in [18] using a different formulation.

B. Related Works

Several other methods have also been proposed to reduce the computational complexity of the C-SRP. In [15], for example, an improved search method for the C-SRP was proposed, where Eq. (1) is employed, but the stochastic region contraction (SRC) algorithm is used to find the source position without having to evaluate the objective function for every grid point. This method was then further improved by the use of particle filters in [19]. In [20], a two-step approach is employed in order to reduce the computational complexity of the C-SRP method. In particular, only the TDoAs associated with high-energy cross-correlation values are considered in the search, thus reducing the computational complexity.

The previous methods tried to reduce the computational complexity by avoiding the computation of the objective function for every point in the search grid. The approach employed by the V-SRP proposed in this paper is to obtain an SRP-based method for volumetric regions, allowing for the use of sparser search grids without compromising its performance. If a more precise estimate of the source position is needed, then a second (low-cost) stage can be employed (RV-SRP).

As mentioned before, the original proposal of an SRP operating on volumetric regions is described in [18]. In addition to their searching process strategies, another key difference between the proposed V-SRP/RV-SRP and [18] lies in their objective functions. While the objective function of [18] performs an accumulation of the energy of each point inside

a volume, the proposed algorithm performs this accumulation over the TDoAs associated with the volume.

The algorithm proposed in [21] uses an objective function similar to the one proposed in this paper. In Section IV-A, the differences between the two approaches are detailed.

C. Organization

This paper is organized as follows. Section II reviews the C-SRP algorithm. In Section III the V-SRP and RV-SRP methods are proposed. Section IV discusses the algorithm proposed in [21], focusing on similarities and differences with respect to the proposed approach. Implementation aspects of the aforementioned techniques are addressed in Section V. In Section VI, assuming a given cost-reducing strategy is employed, the number of arithmetic operations required by each method is computed. Simulation results for simulated- and real-data scenarios demonstrating the good performance of the proposed methods are shown in Section VII. Conclusions are drawn in Section VIII.

D. Notation

The symbols \mathbb{R} , \mathbb{Z} , and \mathbb{N} denote the field of real numbers, the set of integer numbers, and the set of natural numbers, respectively. The set of non-negative real numbers is represented by \mathbb{R}_+ . In addition, vectors are denoted by lowercase boldface letters, $\|\cdot\|$ is the Euclidean norm, and the symbol $\lfloor \cdot \rfloor$ represents the highest integer number that is smaller than or equal to the argument (\cdot) (floor operator).

II. CLASSICAL SRP METHOD

The main idea behind the C-SRP method is to steer the array directionality pattern to different regions, searching for the acoustic source position which is indicated by the maximum power of the array output signal. Mathematically, the goal of the C-SRP is to find a point $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$ that maximizes the objective function $W(\mathbf{x}) \in \mathbb{R}$ given by

$$W(\mathbf{x}) \triangleq \sum_{p=1}^P \phi_p[\zeta_p(\mathbf{x})], \quad (1)$$

in which $P \in \mathbb{N}$ denotes the number of distinct microphone pairs in the array, i.e.

$$P \triangleq \frac{M(M-1)}{2}, \quad (2)$$

where $M \in \mathbb{N}$ represents the number of microphones. In addition, $\zeta_p(\mathbf{x}) \in \mathbb{Z}$ is defined as

$$\zeta_p(\mathbf{x}) \triangleq \text{round} \left\{ \frac{\|\mathbf{m}_{p,2} - \mathbf{x}\| - \|\mathbf{m}_{p,1} - \mathbf{x}\|}{c} f_s \right\}, \quad (3)$$

that is, $\zeta_p(\mathbf{x})$ represents the TDoA (measured in samples) from point \mathbf{x} to the microphone locations $\mathbf{m}_{p,1}, \mathbf{m}_{p,2} \in \mathbb{R}^3$, in which $f_s, c \in \mathbb{R}_+$ denote the sampling rate and the propagation speed of sound, respectively. Denoting the signals acquired by the first and second microphones of the p th microphone pair by $s_{p,1}[n]$ and $s_{p,2}[n]$, the function $\phi_p[\zeta] \in \mathbb{R}$ is defined as

$$\phi_p[\zeta] \triangleq \sum_{n \in \mathbb{Z}} s_{p,1}[n] s_{p,2}[n - \zeta]. \quad (4)$$

Hence, $\phi_p[\zeta_p(\mathbf{x})]$ is the measured cross-correlation function between the signals acquired by the p th microphone pair for a given TDoA $\zeta_p(\mathbf{x})$.

Alternatively, $s_{p,1}[n]$ and $s_{p,2}[n]$ may also be filtered versions of the signals recorded by the two microphones of the p th pair. This is the case, for example, when PHAT filtering is used [1], [13], [14].

III. VOLUMETRIC SRP METHOD

As mentioned before, the capability of the C-SRP to localize sound sources relies on the assumption that the acoustic activity at the actual source position is larger than at other positions. Such acoustic activity is estimated by means of the objective function $W(\mathbf{x})$ in Eq. (1), which is computed based on the TDoAs between the point \mathbf{x} and each microphone pair. Thus, one can see $W(\mathbf{x})$ as a sum of $\phi_p[\zeta]$ across all microphone pairs, where for each index p the argument ζ assumes the value of the p th TDoA as if the source were at position \mathbf{x} . Such objective function, therefore, can be regarded as a “soft” way¹ of “counting” the number of hyperboloids² that pass through \mathbf{x} and that are coherent with the source and microphones’ positions. This TDoA-counting process elects the spatial point that maximizes $W(\mathbf{x})$ as the best choice for the source position estimate.

The idea of the proposed *volumetric SRP* (V-SRP) method is to consider spatial information from several points in order to obtain an estimate of the acoustic activity inside a spatial region. Thus, based on the aforementioned reasoning, the goal here is to elect the spatial region corresponding to the maximum value of a similar TDoA-counting process as the one most likely to contain the acoustic source. In this case, one should consider all grid points within a volume \mathcal{V} in order to compute the number of hyperboloids that pass through it. By taking into account the spatial information of these points, the V-SRP will be able to employ sparser volumetric grids, thus lowering the number of computations without impacting the performance. Following the same reasoning of the C-SRP, the spatial information of the points inside \mathcal{V} is contained in the TDoAs that they yield.

Mathematically, the V-SRP searches for the spatial region \mathcal{V} that maximizes the objective function $\overline{W}(\mathcal{V})$ given by

$$\overline{W}(\mathcal{V}) \triangleq \sum_{p=1}^P \sum_{\zeta=\zeta_{p,\mathcal{V}}^{\min}}^{\zeta_{p,\mathcal{V}}^{\max}} \chi_p[\zeta, \mathcal{V}] \phi_p[\zeta], \quad (5)$$

where

$$\chi_p[\zeta, \mathcal{V}] \triangleq \begin{cases} 1, & \text{if } \zeta = \zeta_p(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{V}, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Thus, for a given microphone pair with index p and for a pre-defined volume \mathcal{V} , $\chi_p[\zeta, \mathcal{V}] = 1$ implies that there exists at least one grid point $\mathbf{x} \in \mathcal{V}$ such that its associated TDoA $\zeta_p(\mathbf{x})$ is equal to ζ . In addition, $\phi_p[\zeta]$ is the measured

cross-correlation function given by Eq. (4), and $\zeta_{p,\mathcal{V}}^{\min}$ and $\zeta_{p,\mathcal{V}}^{\max} \in \mathbb{Z}$ are the minimum and maximum TDoA values considering both a specific volume \mathcal{V} and the p th microphone pair. Hence, one may regard $\chi_p[\zeta, \mathcal{V}]$ as a selector of lags $\zeta \in \{\zeta_{p,\mathcal{V}}^{\min}, \dots, \zeta_{p,\mathcal{V}}^{\max}\} \subset \mathbb{Z}$. This selector indicates which lags correspond to TDoAs for grid points inside the region \mathcal{V} being evaluated. After finding the volume \mathcal{V} that maximizes $\overline{W}(\mathcal{V})$, if one desires a point estimate for the source location, then the center of the volume may be chosen.³

The region \mathcal{V} is usually defined as a volume (e.g. a parallelepiped) in a search space contained in \mathbb{R}^3 , but if the search space is contained in \mathbb{R}^2 , then the region degenerates to a plane region (e.g. a rectangle). In order to allow a simple visualization, Fig. 1 shows a set of points inside a given spatial region \mathcal{V} contained in \mathbb{R}^2 , arbitrarily chosen as a square. It is important to highlight that only two edges of the square are closed, thus indicating that the points on them belong to \mathcal{V} , whereas their opposite edges (in dashed lines) are open.⁴ This construction guarantees that there is no overlap among adjacent volumes within a volumetric grid.

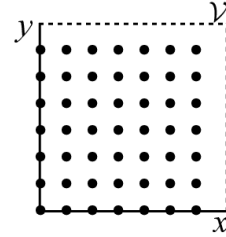


Fig. 1. Spatial region \mathcal{V} with associated points.

A. TDoA Smoothing and Its Spatial Effects

The V-SRP method in practice employs two grids: one for the spatial regions and another for the points inside each region. The volumetric grid is employed in the search itself and bounds the accuracy of the method, whereas the point grid can improve performance by adding more spatial information (through added TDoAs) about a given spatial region. On the other hand, by increasing the number of points inside a given spatial region, the computational complexity for computing the acoustic activity of a single spatial region is also increased (since more terms in χ_p in Eq. (6) are likely to be non-zero). It should be mentioned, however, that since more than one point can be associated with the same delay, the increase in complexity is not linear with the number of points. Moreover, the indicator function χ_p can be pre-computed for a given search grid (with associated points) and array geometry, allowing an efficient implementation of the algorithm by avoiding the multiplications, as will be explained in Section V.

¹Here, the word “soft” is employed due to the continuous nature of function ϕ_p , which is affected by noise and reverberation effects.

²The term hyperboloid is used in this paper to denote the geometric surface comprised of the points whose associated TDoAs are equal, i.e., the set $\{\mathbf{x} \in \mathbb{R}^3 : \zeta_p(\mathbf{x}) = \zeta\}$, where $\zeta \in \mathbb{Z}$ is a constant.

³This is an arbitrary choice that does not have to be made in this particular way.

⁴In this example, there are 7^2 points included in \mathcal{V} . If it were a cube rather than a square, \mathcal{V} would include 7^3 points, since there would be three concurrent faces closed with their opposite ones open.

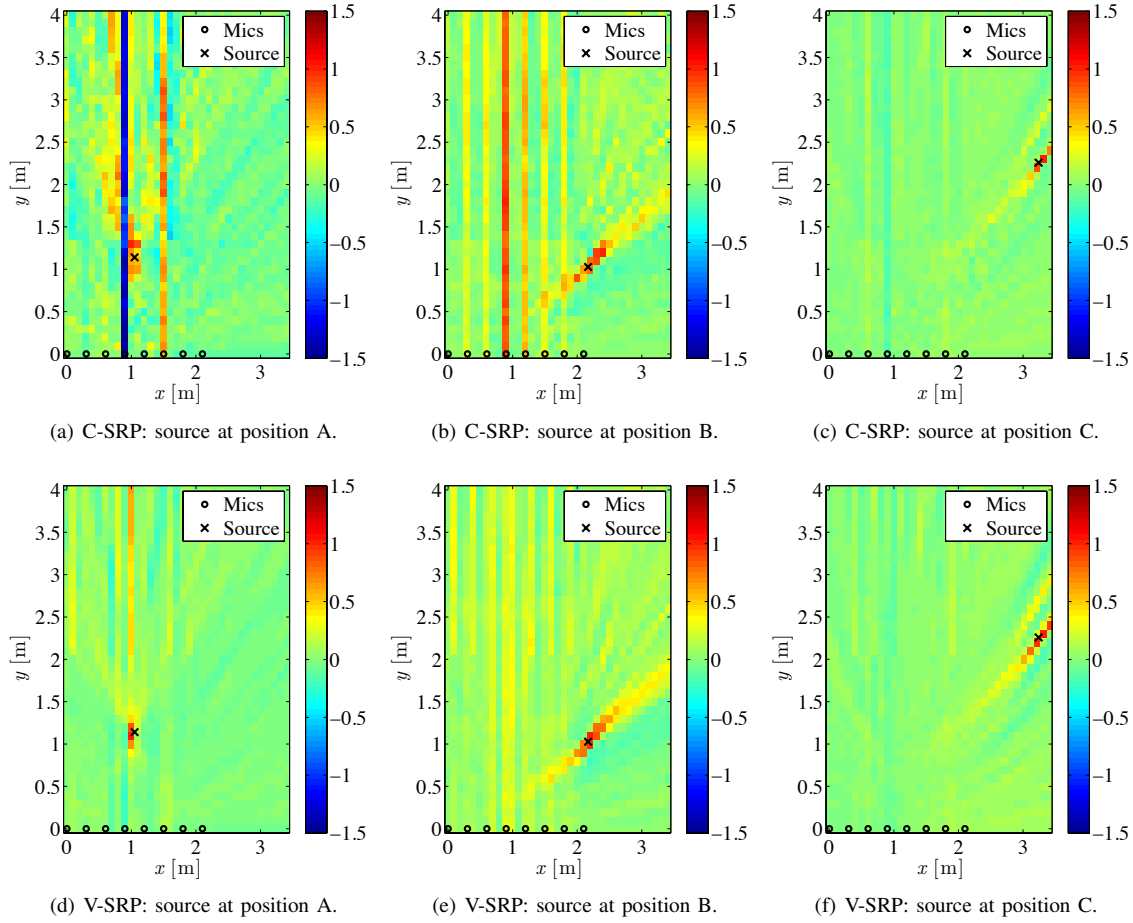


Fig. 2. Energy maps for C-SRP and V-SRP. Further details about this setup can be found in Subsection VII-A.

In addition to the sum across all microphone pairs, the presence of another sum operator over the lags in Eq. (5) indicates that the objective function $\bar{W}(\mathcal{V})$ performs a filtering process along the TDoAs, whose consequent smoothing can mitigate some interferences inherent to the practical TDoA “counting”. Indeed, as $\phi_p[\zeta]$ in Eq. (4) may be affected by reverberation and other acoustic interferences, if one takes into account many lags ζ instead of just one, then it is likely to achieve a more reliable estimate.

Fig. 2 illustrates the spatial effects of the TDoA smoothing considering recorded signals in a realistic setup described in detail in Subsection VII-A. This figure depicts energy maps related to C-SRP and V-SRP, which are a pictorial way of representing $W(\mathbf{x})$ and $\bar{W}(\mathcal{V})$, respectively. Noting that acoustic sources are indicated by an ‘x’ in the figures, by comparing Fig. 2(a) with Fig. 2(d) (source at position A) and Fig. 2(b) with Fig. 2(e) (source at position B), one can see the advantages of the TDoA smoothing as regards its capability of mitigating undesirable peaks that are not related to the actual source position. Figs. 2(c) and 2(f) (source at position C) indicate that, when there is no significant secondary peak, the TDoA smoothing does not work against the proposed method.

B. Refined Volumetric SRP Method

As previously mentioned, the C-SRP method requires dense grids in order to provide accurate estimates of the source position, especially in reverberant environments. However, the use of dense grids may be prohibitive in real-time applications and/or for microphone arrays comprised of a large amount of microphones.

In this context, the V-SRP emerges as a low-cost alternative that provides accurate estimates even when using sparse volumetric grids in reverberant environments. Nevertheless, if more accuracy is desired, a refinement stage can be implemented, leading to the *refined volumetric SRP* (RV-SRP) method, which is comprised of two steps:

- 1) First the entire search space is reduced to a volume $\hat{\mathcal{V}}$, chosen by the V-SRP method;
- 2) Considering that the new search space is the volume $\hat{\mathcal{V}}$, then the C-SRP method is applied inside $\hat{\mathcal{V}}$ with a dense grid.

The RV-SRP allows one to take advantage of the precise estimation provided by the V-SRP method even when coarse grids are employed, while obtaining precise point-estimates of the source by using a low-cost C-SRP due to the limited search region. The trade-offs as related to computational cost between these methods are detailed in Section V-B.

TABLE I
SUMMARY OF DIFFERENCES BETWEEN V-SRP AND M-SRP.

Differences	M-SRP	V-SRP
Grid	point grid	point and volumetric grids
Volume Shape	cubes	arbitrary
Lag Weights	1 (always)	1 or 0
TDoA Bounds	may span / extrapolate \mathcal{V}	span \mathcal{V}

IV. MODIFIED SRP

This section describes and compares against the V-SRP a recently proposed SRP-based method whose objective function looks similar to the one in Eq. (5): the modified SRP (M-SRP) method proposed in [21].

For a given grid point \mathbf{x} , the objective function associated with the M-SRP method depends not only on the TDoAs from \mathbf{x} to each pair p of microphones, but also on all other TDoAs related to a cubic volume surrounding \mathbf{x} . Mathematically, the M-SRP objective function can be written as

$$W_M(\mathbf{x}) \triangleq \sum_{p=1}^P \sum_{\zeta=\zeta_{p,\mathbf{x}}^{\min}}^{\zeta_{p,\mathbf{x}}^{\max}} \phi_p[\zeta], \quad (7)$$

where the limits of the summation are $\hat{\zeta}_{p,\mathbf{x}}^{\min} \triangleq \text{round}\{L_{p,1}(\mathbf{x})f_s\}$ and $\hat{\zeta}_{p,\mathbf{x}}^{\max} \triangleq \text{round}\{L_{p,2}(\mathbf{x})f_s\}$ with the following definitions:

$$L_{p,1}(\mathbf{x}) \triangleq \tau_p(\mathbf{x}) - \|\nabla\tau_p(\mathbf{x})\|d, \quad (8)$$

$$L_{p,2}(\mathbf{x}) \triangleq \tau_p(\mathbf{x}) + \|\nabla\tau_p(\mathbf{x})\|d, \quad (9)$$

$$\tau_p(\mathbf{x}) \triangleq \frac{\|\mathbf{m}_{p,2} - \mathbf{x}\| - \|\mathbf{m}_{p,1} - \mathbf{x}\|}{c}, \quad (10)$$

$$d \triangleq \frac{r}{2} \min\left(\frac{1}{|\sin\theta\cos\phi|}, \frac{1}{|\sin\theta\sin\phi|}, \frac{1}{|\cos\theta|}\right), \quad (11)$$

where r is the length of the cube's edge, and θ and ϕ are respectively the elevation and azimuth angles of the gradient $\nabla\tau_p(\mathbf{x})$ in cylindrical coordinates; see Eqs. (13), (14), and (9) in [21] for more details.⁵

A. V-SRP vs. M-SRP

In the following, a close comparison between the V-SRP and the M-SRP highlights their differences. Table I summarizes the topics discussed in this subsection.

Although Eqs. (5) and (7) are closely related, they differ from each other in fundamental aspects. They are:

- 1) Grid: The M-SRP uses a point grid, as does the C-SRP. On the other hand, V-SRP employs two grids, viz. a point grid and a volumetric grid. The volumetric grid determines the spatial regions. Each region \mathcal{V} is actually a set of points that belong to the point grid.
- 2) Volume shape: The spatial regions inherent to the V-SRP may follow arbitrary shapes and sizes, whereas the M-SRP assumes cubic regions surrounding each point of the grid.

- 3) Lag weights: V-SRP includes the weights $\chi_p[\zeta, \mathcal{V}]$, which allow one to skip all values of $\phi_p[\zeta]$ not associated with the adopted point grid and, as a consequence, providing a desirable control over the *trade-off between computational complexity and accuracy*.
- 4) TDoA bounds: In the V-SRP, the lags $\zeta_{p,\mathcal{V}}^{\min}$ and $\zeta_{p,\mathcal{V}}^{\max}$ are pre-computed by checking the TDoAs related to all grid points inside the volume \mathcal{V} . In the M-SRP, however, the TDoA bounds $\hat{\zeta}_{p,\mathbf{x}}^{\min}$ and $\hat{\zeta}_{p,\mathbf{x}}^{\max}$ are coarse estimates of the minimum and maximum TDoAs inside a cube surrounding \mathbf{x} for the p th microphone pair.

In order to numerically exemplify the difference between V-SRP and M-SRP, consider item 4 above. Due to the approximation adopted by M-SRP, it is easy to find examples in which the set of lags $\{\hat{\zeta}_{p,\mathbf{x}}^{\min}, \dots, \hat{\zeta}_{p,\mathbf{x}}^{\max}\}$ either includes TDoAs not found or does not include all TDoAs found within the volume, as in the next example. Let the sampling rate be $f_s = 48$ kHz, the speed of sound be 340 m/s, the two microphones of a given p th microphone pair be located at $\mathbf{m}_{p,1} = [-2 \ 0 \ 0]^T$ m and $\mathbf{m}_{p,2} = [2 \ 0 \ 0]^T$ m, and the center of the cube with $r = 1$ -m length edges be located at $[0 \ 2 \ 0]^T$ m. Following the equations in Section IV, one arrives at $\hat{\zeta}_{p,\mathbf{x}}^{\min} = -100$ and $\hat{\zeta}_{p,\mathbf{x}}^{\max} = 100$. However, by directly checking the TDoAs corresponding to the vertices of the cube, the minimum and maximum lags are found as respectively -110 and 110 . Clearly, in this example several TDoAs associated with points within the cube would be left out. Such issue might prevent the M-SRP from localizing sources placed close to the borders of the cube.

V. REMARKS ON IMPLEMENTATION

In order to maximally reduce the overall number of arithmetic operations required by each method in real time, and to enable a fair comparison of their computational costs, the strategy of *pre-computing whatever is possible* is assumed in this paper. Such strategy yields a significant reduction of the computational burden in the search stage at the expense of requiring a larger memory to store look-up tables.

Observe that SRP-based methods basically compute three quantities: (i) TDoAs, (ii) cross-correlations, and (iii) objective function values for each grid element. The implementation of each of these computations is described in the next subsections.

A. Computing the TDoAs

Prior to any processing and as soon as the spatial grid is defined, the TDoAs can be pre-computed and stored in look-up tables. For the C-SRP method, given the position of the microphones and the points of the grid, all TDoAs $\zeta_p(\mathbf{x})$ can be computed for all microphone pairs. Thus, P look-up tables are constructed, in which each entry is indexed by a point of the grid and stores its corresponding TDoA.

As for the V-SRP method, a table whose entries are indexed by volumes can be constructed as well. In this case, each entry stores the set $\mathcal{Z}_{p,\mathcal{V}}$, which corresponds to a list of lags ζ where $\chi_p[\zeta, \mathcal{V}] = 1$ (see Eq. (6)), i.e.

$$\mathcal{Z}_{p,\mathcal{V}} \triangleq \{\zeta \in \mathbb{Z} \mid \chi_p[\zeta, \mathcal{V}] = 1\}. \quad (12)$$

⁵ While this paper was under review, a new version of the M-SRP was published in [22]. Such new version employs an iterative search.

By doing so, the expression actually implemented is

$$\bar{W}(\mathcal{V}) = \sum_{p=1}^P \sum_{\zeta \in \mathcal{Z}_{p,\mathcal{V}}} \phi_p[\zeta], \quad (13)$$

whose main advantage over Eq. (5) is skipping trivial multiplications by 1 and 0.

In this subsection we presented an initialization procedure applicable to all SRP-based methods. Since the TDoAs do not vary with time, for a fixed grid and array, they can be computed just once (during initialization) and stored in look-up tables. This strategy can significantly reduce the number of arithmetic operations performed in the long run, especially when one is dealing with large rooms and/or using dense grids.

B. Computing the Cross-Correlation Function

The cross-correlation function (CCF) ϕ_p defined in Eq. (4) can be pre-computed as soon as each signal frame reaches the microphones. Aiming at real-time applications, it is preferable to compute ϕ_p in the frequency domain, using a fast Fourier transform (FFT) algorithm. Besides reducing the number of arithmetic operations, working on the frequency domain is amenable to the use of PHAT.

It should be noticed that all SRP-based methods considered in this paper require the computation of P CCFs and, therefore, the number of arithmetic operations required at this step is the same for these methods.

C. Objective Function Evaluation

For the C-SRP method, since the values of $\zeta_p(\mathbf{x})$ are already stored in look-up tables, the evaluation of the objective function given by Eq. (1) involves only $P - 1$ additions per point of the grid. As for the V-SRP method, provided that $\chi_p[\zeta, \mathcal{V}]$ is efficiently stored, as explained in Subsection V-A, the evaluation of the objective function given by Eq. (13) for a given volume \mathcal{V} requires only $\left(\sum_{p=1}^P |\mathcal{Z}_{p,\mathcal{V}}| \right) - 1$ additions, where $|\mathcal{Z}_{p,\mathcal{V}}|$ is the cardinality of the set $\mathcal{Z}_{p,\mathcal{V}}$ defined in Eq. (12). In the next section, the number of computations performed in this step is detailed for both the C-SRP and the V-SRP.

VI. NUMBER OF ARITHMETIC OPERATIONS PER FRAME

It is known that the number of arithmetic operations required by objective function evaluations tend to be dominant in the long run [20], and this section addresses this issue. Observe that, as explained in Subsection V-C, only one type of arithmetic operation is actually required: summation. Specifically, given the size of the search space and the definition of a grid, approximations for the *number of arithmetic operations per frame due to objective function evaluations* are provided for the C-SRP and V-SRP methods.

Assume that the search space has the shape of a rectangular parallelepiped with length L , width W , and height H , all of them being positive real numbers. In addition, let $g_{\mathbf{x}} \in \mathbb{R}_+$ denote the smallest distance between adjacent points of the

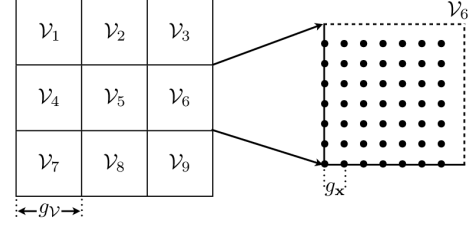


Fig. 3. Example of a 2-D volumetric grid: the volumes degenerate to planar regions.

grid and $g_{\mathbf{y}} \in \mathbb{R}_+$ denote the smallest distance between adjacent volumes of the grid, i.e. each volume is actually a set corresponding to the points of the grid within a cube of edge $g_{\mathbf{y}}$. An illustration of a 2-D volumetric grid where the quantities $g_{\mathbf{x}}$ and $g_{\mathbf{y}}$ appear is shown in Fig. 3. Observe that these quantities represent a uniform sampling of the search space in terms of points and volumes.

The number of points within the grid, $N_{\mathbf{g}} \in \mathbb{N}$, is given by

$$N_{\mathbf{g}} = \left\lfloor \left(\frac{L + g_{\mathbf{x}}}{g_{\mathbf{x}}} \right) \left(\frac{W + g_{\mathbf{x}}}{g_{\mathbf{x}}} \right) \left(\frac{H + g_{\mathbf{x}}}{g_{\mathbf{x}}} \right) \right\rfloor \quad (14)$$

In addition, the number of volumes within the volumetric grid, $N_{\mathbf{v}} \in \mathbb{N}$, is

$$N_{\mathbf{v}} = \left\lfloor \frac{LWH}{g_{\mathbf{y}}^3} \right\rfloor \quad (15)$$

The number of points and volumes in a grid determines how many times the objective function is evaluated. Therefore, the total number of arithmetic operations for the C-SRP method, $N_{\text{op}}^{\text{C-SRP}} \in \mathbb{N}$, can be written as

$$N_{\text{op}}^{\text{C-SRP}} = N_{\mathbf{g}} (P - 1). \quad (16)$$

On the other hand, for the V-SRP method, the total number of arithmetic operations, $N_{\text{op}}^{\text{V-SRP}} \in \mathbb{N}$, is given by

$$N_{\text{op}}^{\text{V-SRP}} = \sum_{\mathcal{V} \in \Gamma} \left[\left(\sum_{p=1}^P |\mathcal{Z}_{p,\mathcal{V}}| \right) - 1 \right], \quad (17)$$

where Γ is a set containing all volumes. Note that, by defining the average cardinality of the set $\mathcal{Z}_{p,\mathcal{V}}$ as

$$\langle |\mathcal{Z}| \rangle = \frac{1}{N_{\mathbf{v}} P} \sum_{\mathcal{V} \in \Gamma} \sum_{p=1}^P |\mathcal{Z}_{p,\mathcal{V}}|, \quad (18)$$

then Eq. (17) can be rewritten as

$$N_{\text{op}}^{\text{V-SRP}} = N_{\mathbf{v}} (P \langle |\mathcal{Z}| \rangle - 1) \quad (19)$$

Equations (16) and (19) represent the number of arithmetic operations per frame due to objective function evaluations for the C-SRP and V-SRP, respectively. Observe that we are not taking into account the computations of the TDoAs and CCFs due to the reasons explained in Section V. In order to fully understand these expressions, consider that the edge of each cube and the distance between adjacent points of the grid are

related by $g_V = \alpha g_x$, where $1 < \alpha \in \mathbb{N}$. In this case, $N_{\text{op}}^{\text{V-SRP}}$ and $N_{\text{op}}^{\text{C-SRP}}$ are related by

$$\begin{aligned} N_{\text{op}}^{\text{V-SRP}} &= \left\lfloor \frac{LWH}{g_V^3} \right\rfloor (P\langle|\mathcal{Z}|\rangle - 1) \\ &< \frac{LWH}{\alpha^3 g_x^3} P\langle|\mathcal{Z}|\rangle \\ &< \frac{(L + g_x)(W + g_x)(H + g_x)}{g_x^3} P \frac{\langle|\mathcal{Z}|\rangle}{\alpha^3} \approx N_{\text{op}}^{\text{C-SRP}} \frac{\langle|\mathcal{Z}|\rangle}{\alpha^3}, \end{aligned} \quad (20)$$

where the approximation is valid as long as the number of microphone pairs $P \gg 1$, which usually is the case. Additionally, one may regard L, W , and H as multiples of g_x in order to disregard the floor operator.

In addition, note that the number of points within a cube of edge $g_V = \alpha g_x$ is α^3 . Therefore, in the *worst case scenario*, each point of the cube leads to a different TDoA implying that the maximum number of different TDoAs in a volume is α^3 (for a given microphone pair). Thus, one has $\langle|\mathcal{Z}|\rangle \leq \alpha^3$, in which the equality is achieved only when all volumes fall into the *worst case scenario* for all microphone pairs, a phenomenon that has not been observed with the data we tested. Therefore, the computational cost of the V-SRP method is, in the worst case, equivalent to the one of the C-SRP. However, it was observed that many points within a volume lead to the same TDoA and, consequently, the number of arithmetic operations per frame of the V-SRP is usually much lower than the one of the C-SRP method, since it is rather common to have $\langle|\mathcal{Z}|\rangle \ll \alpha^3$, especially for volumes relatively far from the array, as it was illustrated by the discriminability index results presented in [23]. Moreover, the minimum value for $\langle|\mathcal{Z}|\rangle$ is 1, at least theoretically. In such case, the V-SRP would perform about α^3 times fewer arithmetic operations, as compared to the C-SRP.

For the RV-SRP method, the number of arithmetic operations per frame can be determined by summing the following two terms: (i) number of arithmetic operations for the V-SRP considering the entire search space and (ii) number of arithmetic operations required by the C-SRP considering that the size of the search space is reduced to the size of the winning volume. If one compares the costs of the C-SRP and V-SRP, it is possible to see the motivation behind the development of the RV-SRP. By using a coarse volumetric grid, it is possible to reduce the computational complexity of the V-SRP. On the other hand, by applying the C-SRP to a smaller search region, its number of arithmetic operations are drastically reduced (smaller L, W , and H values), allowing the use of a denser internal grid in the second stage.⁶ As will be shown in the next section, by choosing the grids appropriately, one can achieve a low estimation error with low computational complexity when using the RV-SRP method. Of course, if the volumetric grid becomes too coarse, the overall complexity of the RV-SRP method tends to increase again, since the high

cost of the C-SRP refining stage applied to large volumes dominates.

VII. RESULTS

The performance of the proposed methods have been assessed through experiments with simulated and recorded signals, as explained in this section. The target here is to point out some attractive features of the proposed algorithms, as well as to highlight some trade-offs between computational complexity and localization performance. The C-SRP and M-SRP methods are used as benchmarks for comparisons.⁷ Two experiments are described: one using acquired signals (Subsection VII-A) and the other using simulated signals (Subsection VII-B). After having investigated the performance of the proposed methods, a brief discussion on how to set the grids is provided (Subsection VII-C).

A. Data from a Real Scenario

In this subsection, the performances of both V-SRP and RV-SRP methods are assessed when applied to signals acquired by a uniform linear array (ULA) in a reverberant room. First the experimental setup is described, then the results are presented along with a brief discussion of the trade-offs between computational complexity and accuracy.

1) *Experimental Setup*: The experiments are conducted in a 5.2 m \times 7.5 m \times 2.6 m room whose measured T60 is approximately 500 ms. The microphone array is a ULA composed of 8 microphones and with aperture of 2.1 m. A small-size loudspeaker (10-cm diameter)⁸ plays the role of the single acoustic source. The source signal consists of 3 sentences emitted by a female speaker and has a total duration of 4.5 s. The sentences were recorded in a professional studio and PCM-coded with a sampling rate of 48 kHz and 24-bit precision. A voice-activity detector (VAD) is employed before playing back the signals in order to discard speech-free segments of the original source signal.

The loudspeaker is placed at 10 different positions chosen at random, as illustrated in Fig. 4. Both microphones and source are always at the height of 72.5 cm. The sound source is amplified at the loudspeaker output in such a way to maximize the signal-to-interference-plus-noise ratio (SINR) at the microphones without saturating any of the amplifiers on the signal path.

2) *Setup of the Localization Methods*: Due to the inherent limitations associated with the ULA geometry to localize sources in the 3-D Euclidean space, and as the sources and microphones are always at the same height, the source location is estimated over the xy -plane whose height is 72.5 cm. The search region is the square with opposite vertices (0, 0, 0.725) m and (3.5, 4.0, 0.725) m, as illustrated in Fig. 4.

All the source localization methods were applied in successive 4096-sample long frames (85 ms at 48-kHz sampling

⁶Under such conditions, previous computation of the TDoAs required by the refining stage could be avoided, since its contribution to the overall complexity of the method is marginal.

⁷The PHAT pre-filtering is employed when computing the related cross-correlations in all methods.

⁸Since the loudspeaker is not a point source, there is an inherent uncertainty relative to the source position which limits the minimum error that can be achieved by the methods.

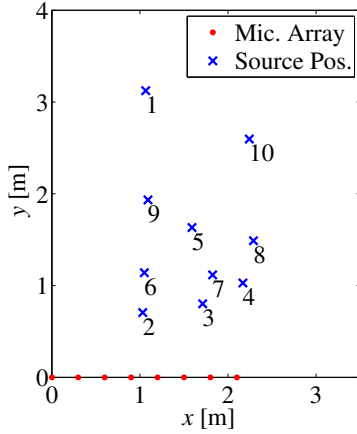


Fig. 4. Positions of microphones and sound sources (Subsection VII-A).

rate) with 50% of overlap. Considering that the signals have a duration of 4.5 s and that there are 10 different source positions, then a total of 1040 positions were estimated by each method. The error of each estimate was calculated as the Euclidean distance between the actual and estimated source position considering only x and y coordinates.

The C-SRP method was run with two different search grids, with 1-cm and 10-cm distance between adjacent points respectively. These two values will be used to illustrate the direct compromise between grid resolution and position estimation error. The V-SRP method was run over a grid of squares of 10-cm edges, each one enclosing 16 grid points (refer to Fig. 3). Regarding the RV-SRP method, a refinement stage using a grid of 1-cm is employed on the winning volume of the V-SRP. The performance of the M-SRP method with points spaced 10-cm apart was also evaluated.

3) *Results and Discussion:* Figs. 5, 6, and 7 show histograms of the estimation errors of the C-SRP, the (R)V-SRP, and the M-SRP, respectively. Those histograms consider all 1040 estimates for each method and for their different configurations. Table II displays the mean and median errors for each method along with the associated number of arithmetic operations performed during the search stage.

One can verify that the majority of the location estimation errors of the C-SRP method with a 1-cm resolution are between 0 and 5 cm (Fig. 5(a)), with a mean estimation error of 19.62 cm and a median estimation error of 3.15 cm (Table II). In addition, it is also possible to see that there are very few frames whose associated C-SRP outputs give completely wrong source position estimates, a very desirable characteristic. But from a practical perspective, the method suffers from a striking drawback: the computational complexity associated with the search stage may hinder the application of such a high-density grid. In this experiment, for example, the number of arithmetic operations due to the C-SRP objective function evaluation was around 38.0×10^5 per 4096-sample frame to perform a 2-D search. This is really an issue that the designer of such a system must face. The computational burden can dramatically increase as more microphones are added (quadratic dependence, see Eq. (2)) and/or the number of grid points grows (as in the case of 3-D regions with dense

grids).

A possible solution to the high computational demands required by source-localization algorithms is to decrease the number of points within the grid search. When the 10 cm resolution C-SRP is employed (see Fig. 5(b)) to the same problem, the number of arithmetic operations associated with the functional evaluations of the C-SRP objective function goes down to approximately 3.98×10^4 per frame, thus decreasing around two orders of magnitude as compared to the 1-cm resolution grid. Nonetheless, the performance is dramatically sacrificed, since the number of anomalous estimates increases too much. Indeed, the mean value of the estimation error in this case is around 52.09 cm, whereas the corresponding median error is around 11.17 cm.

The V-SRP method (see Fig. 6(a)) yielded mean and median estimation errors around 18.29 cm and 6.21 cm, respectively, outperforming the C-SRP with grid resolution of 10 cm. This mean value is even a bit smaller than the mean estimation error value for the C-SRP with 1-cm resolution, but when one compares Fig. 5(a) with Fig. 6(a) it is straightforward to verify that most of the results of Fig. 5(a) are better than the ones in Fig. 6(a), a fact that is reflected in the median errors of those methods. The key advantage of the V-SRP method comes when one compares the computational complexity of the C-SRP with 1-cm resolution to the V-SRP with 10-cm edges. Indeed, the number of arithmetic operations associated with the functional evaluations in the search stage is around 2.08×10^5 per frame, which yields a significant reduction of the computational burden, without sacrificing performance significantly.

By employing the RV-SRP strategy, the total number of arithmetic operations is slightly increased to approximately 2.11×10^5 arithmetic operations per frame, while the mean and median estimation errors are reduced to 15.98 cm and 3.77 cm, respectively. Such results let clear the inherent capability of the proposed RV-SRP method to trade off performance and computational burden, thus providing an additional degree of freedom to the design of sound source localization systems.

The M-SRP method proposed in [21] is outperformed (see Fig. 7 and Table II) in this particular experimental setup by both V-SRP and RV-SRP as regards estimation error and number of arithmetic operations.

Finally, Table III contains the actual number of arithmetic operations per frame required to evaluate the objective functions over the entire search region for five different search grids. In the cases of the V-SRP and RV-SRP, the resolution indicates the size of the edges of the square spatial regions, each one enclosing 16 grid points. In addition, the refinement of the RV-SRP is always implemented by employing a C-SRP with a 1-cm grid resolution within the selected volume.

B. Simulated Scenario

In this subsection, the performance of the proposed algorithms is evaluated using simulated signals. The objective of the simulation is to search for a very high localization accuracy in the 3-D space, yet focusing on saving computational resources, especially as regards to decreasing the amount of

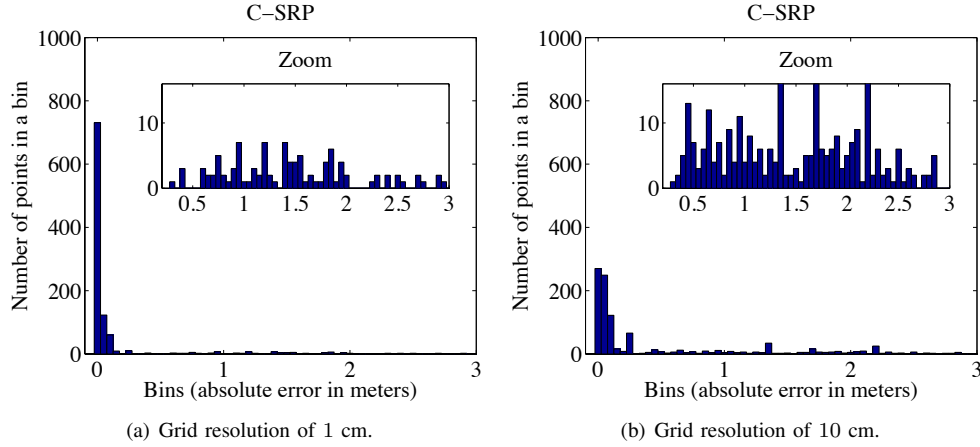


Fig. 5. Histograms of location estimation errors for the C-SRP (bin width is 5 cm). The inside histograms (Zoom) show the number of estimation errors larger than 30 cm.

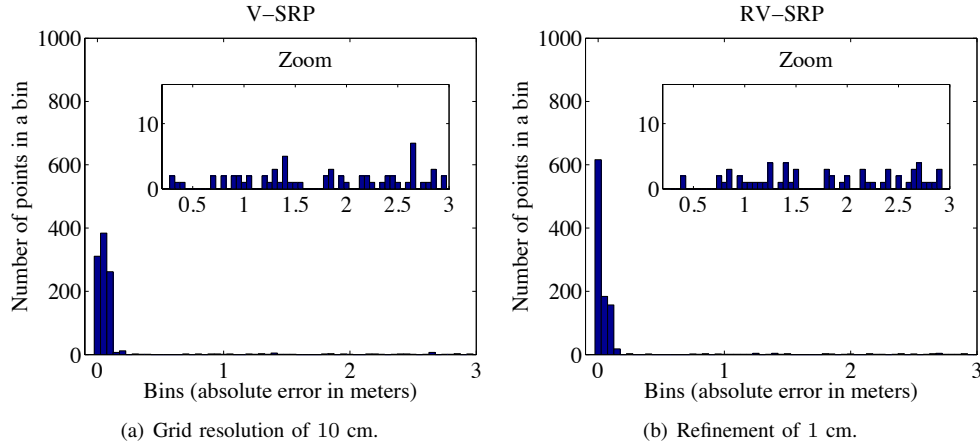


Fig. 6. Histograms of location estimation errors for the V-SRP and the RV-SRP (bin width is 5 cm). The inside histograms (Zoom) show the number of estimation errors larger than 30 cm.

TABLE II
SUMMARY OF THE RESULTS FOR THE REAL-DATA SCENARIO.

Method [grid resolution]	Mean error [cm]	Median error [cm]	Approx. number of op. per frame ($\times 10^5$)
C-SRP [1 cm]	19.62	3.15	38.0
C-SRP [10 cm]	52.09	11.17	0.398
V-SRP [10 cm, 16 pt]	18.29	6.21	2.08
RV-SRP [10 cm, 16 pt / ref. 1 cm]	15.98	3.77	2.11
M-SRP [10 cm]	19.67	6.76	2.71

TABLE III
NUMBER OF ARITHMETIC OPERATIONS PER FRAME DUE TO FUNCTIONAL EVALUATIONS. (* MEANS THAT THE RV-SRP COINCIDES WITH THE V-SRP FOR THIS RESOLUTION, I.E., NO REFINEMENT IS ACTUALLY PERFORMED)

Resolution	C-SRP	V-SRP	RV-SRP	M-SRP
1 cm	3,800,277	5,821,039	5,821,039*	6,154,534
10 cm	39,852	208,378	211,078	270,682
20 cm	10,206	79,419	90,219	128,072
50 cm	1,944	21,439	88,939	53,932

arithmetic operations related to functional evaluations. In order to achieve such high accuracy results, one needs to properly choose the array geometry, the grid resolution, and the specific

sound source localization method. The array geometry that allows for high 3-D localization accuracy must have both relatively large aperture and great amount of microphone pairs, so that the resulting spatial resolution is substantially increased. Along with these choices, the spatial grid must also be dense enough. In the following, the simulation setup and the chosen localization methods are described along with their associated results.

1) *Simulation Setup:* The environmental setup simulated for this example consists of a 4.0 m \times 6.0 m \times 3.0 m reverberant room whose T60 can be either 250 ms or 500 ms. Such reverberant environments are simulated using the image model method [20], [24]. The speech signal employed is a 1-s segment from the same source signal used in Subsection VII-A.

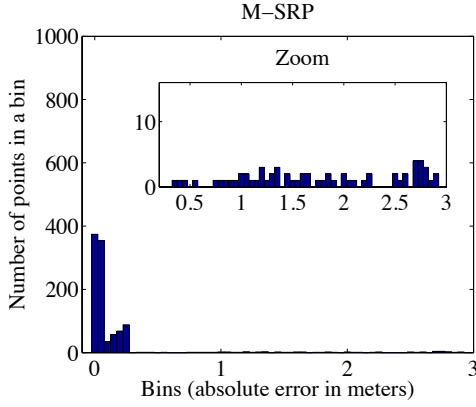


Fig. 7. Histograms of location estimation errors for the M-SRP (bin width is 5 cm). The inside histogram (Zoom) shows the number of estimation errors larger than 30 cm.

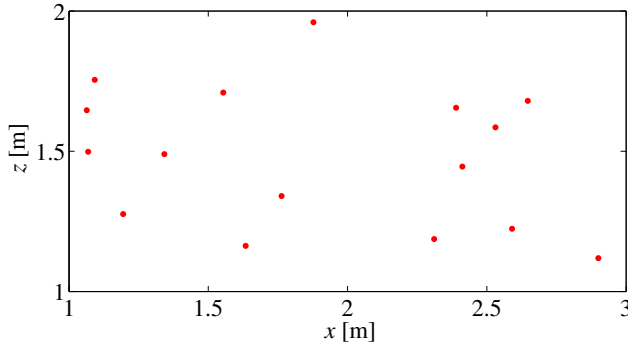


Fig. 8. Locations of each microphone in the planar array (Subsection VII-B).

TABLE IV
SOURCE POSITIONS OF THE ARTIFICIAL SCENARIO.

#	Source Position
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]

The array has 16 microphones distributed as depicted in Fig. 8 on a $2.0 \text{ m} \times 1.0 \text{ m}$ region perpendicular to the floor plane (i.e. to the xy -plane), mounted on one of the walls of the room (at $y = 0$). The source signal was artificially located at 5 different positions, which are shown in Table IV.

2) *Setup of the Localization Methods:* In this scenario, the search space is the entire room (a $4.0 \text{ m} \times 6.0 \text{ m} \times 3.0 \text{ m}$) region, and the error is the 3-D Euclidean distance between estimated and known source positions. The C-SRP method employed a grid resolution of 3 cm.⁹ Since the objective of this experiment is to attain high accuracy, the RV-SRP was employed. The V-SRP step was run over cubic spatial regions with 10 cm edges enclosing 64 points each. The refinement stage considered a search grid with 1-cm resolution. As a benchmark, the M-SRP with a resolution of 10 cm was also

⁹As will be mentioned, the computational cost of the C-SRP using a 1-cm resolution over the whole 3-D space is too high for this experiment, hence a coarser resolution is used.

evaluated. As in the previous experiment, source position estimates were calculated for successive 4096-sample long frames with 50% of overlap.

3) *Results and Discussion:* Fig. 9 shows the histogram of the estimation error for the three localization methods employed. Table V summarizes the results for this scenario, including the mean and median estimation errors for each method and its approximate number of arithmetic operations. The C-SRP with a 1-cm resolution was included in order to illustrate its demanding computational complexity, even though this resolution was not used in the evaluation.

By observing Figs. 9(a), 9(b), and 9(c) one can verify that all algorithms are able to localize the acoustic source at the positions tested with a relatively high accuracy when the reverberation time is moderately low ($T60 = 250 \text{ ms}$). Indeed, the zoom plots show that those methods do not yield any anomalous estimate, which is reflected in the close mean and median estimation error values in the respective column of Table V. Moreover, when computational complexity of the methods is also taken into account, one can observe from these results that the proposed RV-SRP algorithm achieves the best trade-off between performance and computational complexity in this particular environment.

Regarding the environment with $T60 = 500 \text{ ms}$, in the case of the C-SRP method with grid resolution of 3 cm, its mean estimation error was approximately 63.15 cm, even though the majority of the absolute errors were smaller than 15 cm (see Fig. 9(d)). If the designer of the system is somehow able to discard most of the anomalous estimates, then the average value of the estimation error would be obviously smaller. For instance, the median estimation error of this example is around 6.23 cm. This result is achieved by performing about 32.4×10^7 arithmetic operations per frame. Even after this significant reduction as compared to the 1-cm resolution case, the computational complexity might be still too high for the envisaged application.

When using the RV-SRP, the estimation error value is drastically reduced to approximately 9.76 cm, while its median value is around 2.86 cm. Note that the zoom plot of the RV-SRP in Fig. 9(e) shows that the proposed method yields very few anomalous estimates in this particular setup. In addition to this significant performance enhancement, the total number of arithmetic operations per frame required by the functional evaluations of the proposed method is around 4.59×10^7 , about one order of magnitude smaller than the C-SRP algorithm with 3-cm grid resolution. It is worth pointing out that such improvements obtained by the proposed algorithm come at the price of spending more memory resources.

The M-SRP obtained mean and median estimation errors of approximately 24.30 cm and 9.41 cm, respectively, both higher than those obtained by the RV-SRP method. Besides, the number of arithmetic operations required by the functional evaluations of the M-SRP is around 4.98×10^7 per frame, higher than the computational cost of the method proposed in this paper, for this particular scenario.

C. More on the Relation Between Point and Volumetric Grids

In the results shown in the previous subsections, volumes

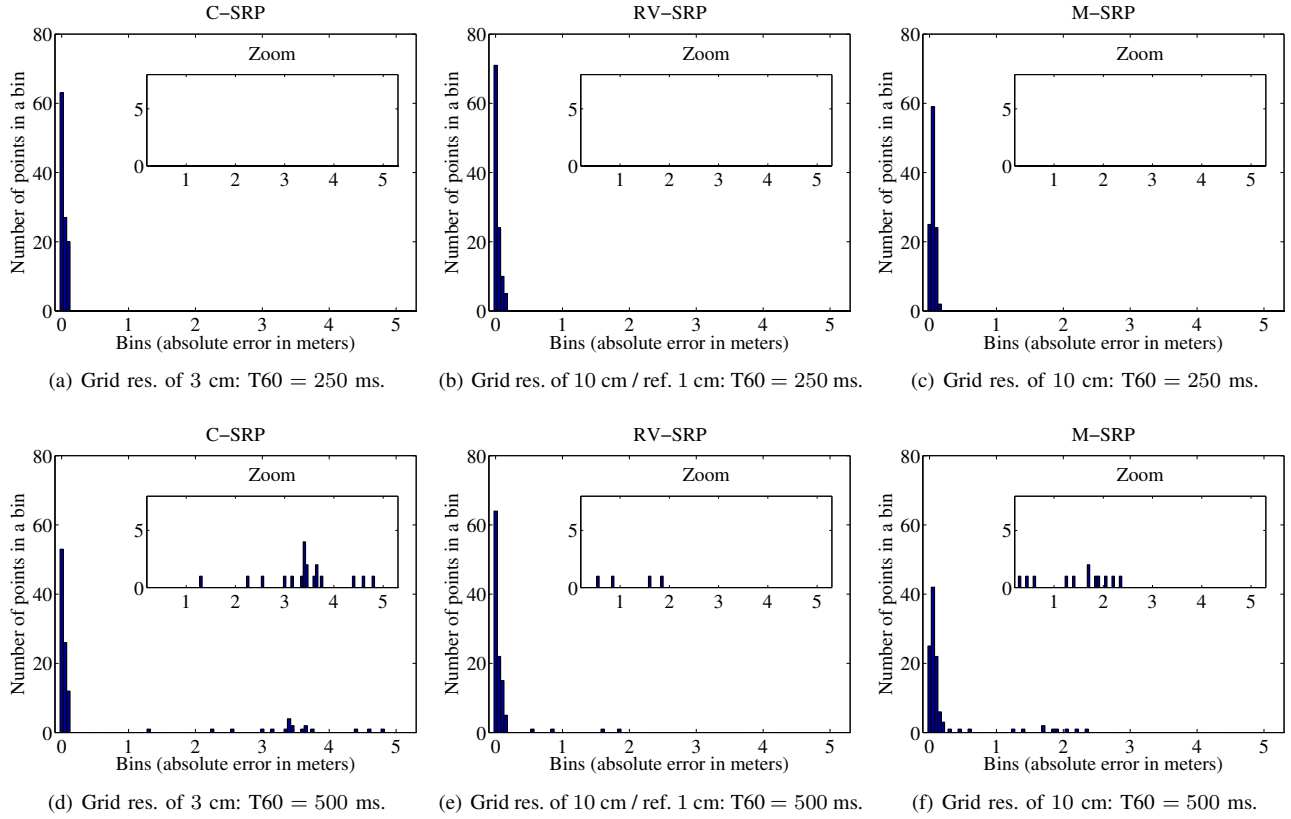


Fig. 9. Histograms of location estimation errors for the C-SRP, RV-SRP, and M-SRP (bin width is 5 cm). The inside histograms (Zoom) show the number of estimation errors larger than 30 cm.

TABLE V
RESULTS FOR THE SIMULATED SCENARIO.

Method [grid resolution]	Error [cm] for T60 = 250 ms		Error [cm] for T60 = 500 ms		Approx. number of op. per frame ($\times 10^7$)
	Mean	Median	Mean	Median	
C-SRP [1 cm]	—	—	—	—	863
C-SRP [3 cm]	4.87	3.56	63.15	6.23	32.4
RV-SRP [10 cm, 64 pt / ref. 1 cm]	5.04	2.33	9.76	2.86	4.59
M-SRP [10 cm]	7.91	7.85	24.30	9.41	4.98

with 10 cm of edge were used. Inside each volume there were 16 points for the real-data (2-D) scenario and 64 points for the simulated (3-D) scenario, thus implying that in both cases the number of points per edge was 4 (refer to Figs. 1 and 3). For these two scenarios, an increase in the number of points per edge, and thus per volume, did not lead to a significant performance gain that would justify an increase in the computational complexity. Hence, 4 points per edge proved to be a good choice when using volumes with 10 cm of edge.

If one intends to use the proposed methods with coarser volumetric grids, i.e. using volumes with larger edges, then one should be aware that the complexity of the RV-SRP is not a monotonic decreasing function of the size of the volume. For instance, the number of arithmetic operations required by the RV-SRP with 100 cm of grid resolution (edge) is higher than the one with 10 cm; besides the results are much worse due to the coarser volumetric grid. Experimental observations have pointed out that even when using volumes with large edges, such as 50 and 100 cm, 8 points per edge were enough. This choice depends on several parameters, among them the

sampling frequency (in this section, 48 kHz).

VIII. CONCLUDING REMARKS

This paper introduced a novel approach to the application of the SRP method to the problem of sound source localization using microphone arrays. In order to tackle high resolution requirements without resorting to a superfine grid, which would lead to an exceedingly complex procedure, the proposed V-SRP performs the search over a sparse volumetric grid; the volume with the highest objective function value is expected to contain the sound source. Its variant, the RV-SRP, further refines the search by applying the classical SRP method inside the winning volume. Complexity analysis as well as two sets of experiments (2-dimensional search using uniform linear array and natural signals, and 3-dimensional search using planar array and simulated signals) demonstrate that the V-SRP and the RV-SRP outperform the classical SRP and another recent competing method (the M-SRP), achieving a comparable accuracy with much reduced complexity.

The proposed approach provides some degrees of freedom that can be customized for a given application. For instance, one can use other volumes different from cubes, with variable sizes—possibly chosen with the aid of a discriminability measure [23]—or perform the refinement step using a source localization method other than the classical SRP. In addition, when accuracy is of paramount importance and computational resources are abundant (e.g. in cloud computing cases), the RV-SRP method can be adjusted to retain not just a single winning volume, but the N volumes that lead to the highest objective function values; their respective refinement steps can then be performed in a distributed fashion.

ACKNOWLEDGMENT

This R&D project resulted from a cooperation between Hewlett-Packard Brasil Ltda. and COPPE/UFRJ, being supported with resources of Informatics Law (no. 8.248, from 1991). L. W. P. Biscainho, T. N. Ferreira, M. V. S. Lima, W. A. Martins, and L. O. Nunes would like to thank also CAPES, CNPq, and FAPERJ agencies for funding their research work.

NOTE

An almost identical version of this manuscript was submitted to the IEEE Transactions on Audio, Speech, and Language Processing (TASLP) in June 2013, and eventually rejected in May 2014. Since one of the reviewers' concerns was the paper length, the authors decided to reshape the work as a letter to be submitted to the IEEE Signal Processing Letters, while at the same time making the longer version, which discusses e.g. computational requirements, available.

REFERENCES

- [1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Heidelberg: Springer, 2010.
- [2] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Upper Saddle River: Prentice Hall, 1993.
- [3] M. Haardt and J. A. Nosske, "Unitary ESPRIT: how to obtain increased estimation accuracy with a reduced computational burden," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1232–1242, May 1995.
- [4] Y. Hua and T. K. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 5, pp. 814–824, May 1990.
- [5] B. D. Rao and K. V. S. Hari, "Performance analysis of root-MUSIC," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 12, pp. 1939–1949, December 1989.
- [6] Y. Huang, J. Benesty, and J. Chen, "Time delay estimation and source localization," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Heidelberg: Springer, 2008, ch. 51, pp. 1043–1062.
- [7] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.
- [8] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proceedings of the IEEE*, vol. 61, no. 10, pp. 1497–1498, October 1973.
- [9] M. S. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, October 1997.
- [10] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, Adelaide, Australia, April 1994, pp. 273–276.
- [11] —, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, Atlanta, USA, May 1996, pp. 921–924.
- [12] M. Omologo, P. Svaizer, and R. D. Mori, *Spoken Dialogues with Computers*. Orlando: Academic Press, 1997, ch. Acoustic Transduction.
- [13] J. H. DiBiase, "A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments," Ph.D. dissertation, Brown University, Providence, May 2000.
- [14] 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. Heidelberg: Springer, 2001, ch. 8, pp. 157–180.
- [15] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, Honolulu, USA, April 2007, pp. 121–124.
- [16] H. F. Silverman, W. R. P. III, and J. L. Flanagan, "The huge microphone array," *Concurrency IEEE*, vol. 6, no. 4, pp. 36–46, October–December 1998.
- [17] Y. Tamai, S. Kagami, H. Mizoguchi, Y. Amemiya, K. Nagashima, and T. Takano, "Real-time 2 dimensional sound source localization by 128-channel huge microphone array," in *IEEE International Workshop on Robot and Human Interactive Communication*, Kurashiki, Japan, September 2004, pp. 65–70.
- [18] A. Said, B. Lee, and T. Kalker, "Fast steered response power computation in 3D spatial regions," HP Labs, Palo Alto, USA, Tech. Rep. HPL-2013-40, April 2013.
- [19] H. Do and H. F. Silverman, "Stochastic particle filtering: a fast SRP-PHAT single source localization algorithm," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, October 2009, pp. 213–216.
- [20] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2510–2526, November 2007.
- [21] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, January 2011.
- [22] A. Marti, M. Cobos, J. J. Lopez, and J. Escolaro, "A steered response power iterative method for high-accuracy acoustic source localization," *Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 2627–2630, October 2013.
- [23] 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 *International Workshop on Acoustic Signal Enhancement*, Aachen, Germany, September 2012, pp. 1–4.
- [24] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, April 1979.
- [25] M. V. S. Lima, W. A. Martins, L. O. Nunes, L. W. P. Biscainho, T. N. Ferreira, M. V. M. Costa, and B. Lee, "A volumetric SRP with refinement step for sound source localization," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1098–1102, Aug. 2015.

SUPPLEMENTARY MATERIAL

Abstract—The real-data and simulated scenarios presented in Section VII are revisited and results using other competing algorithms are included. These new algorithms are the stochastic region contraction (SRC) [15] and the Iterative SRP-based method (I-SRP) [22].

It is important to highlight that this part of the text contains new results, which were not part of the original manuscript submitted to the IEEE TALSP, that supplement the paper we submitted to IEEE Signal Processing Letters entitled “A Volumetric SRP with Refinement Step for Sound Source Localization”.

IX. ADDITIONAL RESULTS

In this section, we summarize the additional results for the real-data scenario (refer to Section VII-A) and for the simulated scenario (refer to Section VII-B). The new competing algorithms are the SRC [15] and the I-SRP [22]. The SRC evaluated 3000 points per volume among which 100 were chosen to define the new volume employed on the next iteration; this process was repeated until the volume’s edge achieved 10 cm. As for the I-SRP, three configurations were used:

- 1) I-SRP with a single iteration and grid resolution of 10 cm: this allows a fair comparison between the M-SRP and the I-SRP, which were proposed by the same group;
- 2) I-SRP with two iterations (first iteration with grid resolution of 10 cm and second iteration with grid resolution of 1 cm): this allows a fair comparison between the RV-SRP and the I-SRP;
- 3) I-SRP with three iterations (first iteration with grid resolution of 50 cm and last iteration with grid resolution of 1 cm): this is the configuration used in [22].

The results for the real-data and simulated scenarios are summarized in Tables VI and VII, respectively.

In summary, both the SRC and the I-SRP yielded inferior results, especially when facing large rooms and/or higher reverberation. In these experiments, the proposed V-SRP and RV-SRP were the most robust methods with respect to the different array geometries, reverberation time, and room dimension.

TABLE VI
RESULTS FOR THE REAL-DATA SCENARIO.

Method [grid resolution]	Mean error [cm]	Median error [cm]	Approx. number of op. per frame ($\times 10^5$)
C-SRP [1 cm]	19.62	3.15	38.0
C-SRP [10 cm]	52.09	11.17	0.398
M-SRP [10 cm]	19.67	6.76	2.71
I-SRP [10 cm (1 iteration)]	38.69	7.87	3.54
I-SRP [10 cm, 1 cm (2 iterations)]	36.24	6.05	3.65
I-SRP [50 cm, 10 cm, 1 cm (3 iterations)]	130.78	85.74	0.77
SRC [10 cm]	47.99	8.99	7.00
V-SRP [10 cm, 16 pt]	18.29	6.21	2.08
RV-SRP [10 cm, 16 pt / ref. 1 cm]	15.98	3.77	2.11

TABLE VII
RESULTS FOR THE SIMULATED SCENARIO.

Method [grid resolution]	Error [cm] for T60 = 250 ms		Error [cm] for T60 = 500 ms		Approx. number of op. per frame ($\times 10^7$)
	Mean	Median	Mean	Median	
C-SRP [1 cm]	–	–	–	–	863
C-SRP [3 cm]	4.87	3.56	63.15	6.23	32.4
M-SRP [10 cm]	7.91	7.85	24.30	9.41	4.98
I-SRP [10 cm (1 iteration)]	68.96	14.32	166.40	18.94	6.84
I-SRP [10 cm, 1 cm (2 iterations)]	65.01	9.41	164.16	12.91	6.98
I-SRP [50 cm, 10 cm, 1 cm (3 iterations)]	285.07	320.13	337.99	374.26	–
SRC [10 cm]	86.42	3.98	192.89	198.88	0.5
V-SRP [10 cm, 64 pt]	9.88	7.55	14.41	9.95	< 4.59
RV-SRP [10 cm, 64 pt / ref. 1 cm]	5.04	2.33	9.76	2.86	4.59