Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 7972146, 13 pages
https://doi.org/10.1155/2017/7972146

Review Article

Acoustic Sensor Self-Localization: Models and Recent Results

Diego B. Haddad,[1] Markus V. S. Lima,[2] Wallace A. Martins,[2] Luiz W. P. Biscainho,[2] Leonardo O. Nunes,[3] and Bowon Lee[4]

[1]*Computer Engineering Department, Federal Center for Technological Education (CEFET/RJ), Petropolis, RJ, Brazil*
[2]*DEE-DEL/Poli & PEE/COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil*
[3]*Advanced Technology Labs, Microsoft, Rio de Janeiro, RJ, Brazil*
[4]*Department of Electronic Engineering, Inha University, Incheon, Republic of Korea*

Correspondence should be addressed to Bowon Lee; bowon.lee@inha.ac.kr

Received 19 May 2017; Revised 18 August 2017; Accepted 30 August 2017; Published 22 October 2017

Academic Editor: Nathalie Mitton

The wide availability of mobile devices with embedded microphones opens up opportunities for new applications based on acoustic sensor localization (ASL). Among them, this paper highlights mobile device self-localization relying exclusively on acoustic signals, but with previous knowledge of reference signals and source positions. The problem of finding the sensor position is stated as a function of estimated times-of-flight (TOFs) or time-differences-of-flight (TDOFs) from the sound sources to the target microphone, and the main practical issues involved in TOF estimation are discussed. Least-squares ASL solutions are introduced, followed by other strategies inspired by sound source localization solutions: steered-response power, which improves localization accuracy, and a new region-based search, which alleviates complexity. A set of complementary techniques for further improvement of TOF/TDOF estimates are reviewed: sliding windows, matching pursuit, and TOF selection. The paper proceeds with proposing a novel ASL method that combines most of the previous material, whose performance is assessed in a real-world example: in a typical lecture room, the method achieves accuracy better than 20 cm.

1. Introduction

The increasing number of smart mobile devices (such as laptops and smartphones) that can interact with their surrounding world through their many sensors enables one to envision several new and exciting applications. Examples of such applications include indoor navigation [1], location-based services [2], patient monitoring [3], and many others involving ad hoc microphone arrays [4–11]. These applications have one thing in common: they require the estimation of the mobile device position.

Based on the specific technology employed, it is possible to categorize methodologies for localization of mobile devices in three distinct groups [12]: (i) systems that utilize ancillary sensors, such as accelerometer and magnetometers; (ii) systems that rely on the received signal strength of radio signals; and (iii) systems that use time-of-flight (TOF) of acoustic signals to perform localization. This paper focuses on (iii), since acoustic sensor localization (ASL) systems enable centimeter-scale localization and are quite inexpensive as they only require acoustic sensors (i.e., microphones) in the mobile devices.

ASL techniques relying on TOF information may require or not prior knowledge of the loudspeakers' positions. For instance, many works have tackled the problem of microphone array calibration [13–16] assuming that sources' positions are known and aiming to estimate (or refine an initial estimate of) the acoustic sensor location. In such scenarios, synchronization problems are not a major concern, since all sources and sinks are usually under the control of whoever wants to calibrate the microphone array. On the other hand, if the sources' positions are unknown, then a joint source and sensor localization is commonly employed [17, 18]. In this case, the lack of synchronism may severely degrade the mobile position estimate. In [7], a solution that deals explicitly with synchronization was presented. In [8], a solution considering multiple sources and sensors per device was described.

When it comes to the performance of ASL techniques, the accuracy of TOF estimates is of paramount importance. Roughly speaking, in order to obtain an accurate TOF estimate, the sensor (microphone) should be able to determine the exact time a given acoustic signal takes to propagate from the source. Hence, the acoustic signals employed, herein called *probe signals*, should be carefully selected, taking into consideration practical issues like noise, interference, and mainly reverberation. Recently, several papers have addressed the problem of designing good probe signals and improving the TOF estimation. In [19], a probe signal design using pulse compression technique and hyperbolic frequency modulated signals was presented. The technique proposed in that work is able to localize an acoustic source and estimate its velocity and direction in case it is moving. In addition, a matching pursuit-based algorithm for TOF estimation was described in [20] and then refined in [12], in both papers with promising results. In [21], an iterative peak matching algorithm for the calibration of a wireless acoustic sensor network was described.

The ASL methods described in this work assume that sources' positions and probe signals are known at the receiver (i.e., the mobile device containing a microphone), but not counting on synchronism between sensor and source nor cooperation with other mobile devices: all processing must take place on the sensor itself, allowing the sensor to self-localize indoors. Nevertheless, a robust ASL system with such characteristics needs to address several issues, such as reverberation, asynchrony between acoustic source and sensor, poor estimation of the speed of sound, and noise. Also, the proposed solutions need to be computationally inexpensive in order to be able to run on the mobile itself (which in general is a battery-operated device).

1.1. Objectives. The main objectives of this paper are

(i) to review some state-of-the-art techniques for ASL;

(ii) to describe in detail the challenges ASL algorithms face in practical environments;

(iii) to present in a unified context some recent methods proposed by the authors to tackle such practical challenges;

(iv) to propose a novel ASL algorithm that combines different state-of-the-art techniques with a new region-based search.

For beginners in the ASL field it can be a Herculean task to go through the myriad of techniques addressing ASL problems. The concise approach adopted here provides beginners with a glimpse of the main problems and solutions in the ASL field. Experienced researchers can also benefit from the unified approach covering recent ideas that are not standard in the area, as the ones described in [22–24]. In fact, many research opportunities may open up thanks to the unified way ASL problems and related solutions are described here. Finally, low power high-frequency probe signals inaudible for most people are proposed, along with a new region-based search that is shown to be robust when combined with matching pursuit TOF-selection strategies.

1.2. Organization. This paper is organized as follows. Section 2 provides basic definitions and formally states the ASL problem. Classical solutions are described in Section 3, whereas improved algorithms are described in Section 4. A novel ASL system is proposed in Section 5, and the respective experimental results are presented in Section 6. Concluding remarks are drawn in Section 7.

2. Background

2.1. Acoustic Networks. An acoustic network is formed when there exists an acoustic coupling among loudspeakers (herein also called sources) and microphones (herein also denominated as sensors). In a general configuration, an acoustic network is comprised of $S \in \mathbb{N}$ sources and $M \in \mathbb{N}$ microphones, which can be moving as long as they do not go further enough to break the acoustic coupling. In addition, these nodes (loudspeakers and microphones) may cooperate somehow in order to accomplish some task, like localizing a source or a sensor or enhancing a signal.

This paper addresses the problem of acoustic sensor localization (ASL) without sensor cooperation. In this case, each microphone individually uses the S acoustic signals emitted/acquired in order to passively estimate its own location. Thus, without loss of generality, $M = 1$ is assumed. In addition, the positions of the loudspeakers are fixed and known.

2.2. Basic Definitions. Let $\mathbf{m}, \mathbf{p}_s \in \mathbb{R}^3$ be the microphone position and the position of the sth loudspeaker, respectively. The *time-of-flight* (TOF), that is, the time that the acoustic signal takes to travel from the source to the sensor through a *direct path*, plays a central role in ASL. Indeed, TOF $t_s \in \mathbb{R}_+$ from the sth loudspeaker to the sensor is defined as

$$t_s \triangleq \frac{\|\mathbf{m} - \mathbf{p}_s\|}{v}, \tag{1}$$

where $v \in \mathbb{R}_+$ denotes the speed of sound. From (1), it is clear that $\mathbf{m}$ can be found if a sufficient number of TOFs are known. In practice, however, t_s cannot be exactly determined, but rather estimated by some procedure as $\widehat{t}_s \in \mathbb{R}_+$.

Instead of using the TOF directly, there are also some ASL techniques that rely on the *time-difference-of-flight* (TDOF) to estimate the sensor position. The TDOF related to loudspeakers $s_1, s_2 \in \mathcal{S}$ is given by

$$\tau_{s_1 s_2} \triangleq t_{s_1} - t_{s_2} = \frac{\|\mathbf{m} - \mathbf{p}_{s_1}\| - \|\mathbf{m} - \mathbf{p}_{s_2}\|}{v}, \tag{2}$$

which must also be estimated as $\widehat{\tau}_{s_1 s_2} \triangleq \widehat{t}_{s_1} - \widehat{t}_{s_2}$.

2.3. TOF Estimation. One of the simplest procedures to obtain a TOF estimate $\widehat{t}_s$ is to compute the *cross-correlation function* (CCF) $R_s[\cdot] : \mathbb{N} \to \mathbb{R}$ between the signal $y[n]$ acquired by the microphone and the signal $x_s[n]$ emitted by the sth loudspeaker, given by

$$R_s[k] \triangleq \sum_{n \in \mathbb{N}} y[n]\, x_s[n-k], \tag{3}$$

where it is assumed that $x_s[n]$ is known at the microphone node, and then find the delay associated with the highest peak of $R_s[k]$.

In order to fully understand the above procedure, consider an ideal setup in which there exists *line-of-sight* (LOS), that is, a direct path connecting loudspeaker and microphone, and the propagation medium does not introduce any distortion to the emitted signal. In this scenario, the received signal is a delayed version of the emitted signal, that is, $y[n] = x_s[n-k_0]$, where k_0 stands for the delay measured in samples. Thus, it is clear that the maximum of (3) is achieved when $k = k_0$, which is then converted from samples to seconds, as shown in (6), yielding the TOF estimate. In this ideal setup, the accuracy of $\widehat{t}_s$ is limited only by the sampling rate. However, in practical cases, there are many issues that may deteriorate the TOF estimates and, consequently, impair the accuracy of ASL techniques.

The CCF essentially measures the similarity between signals $y[n]$ and $x_s[n-k]$ for different values of k. In practice, however, the *generalized cross-correlation* (GCC) function [25], a filtered version of the CCF, is usually preferred. By denoting as $X(e^{j\omega})$ the discrete-time Fourier transform of a generic signal $x[n]$, the GCC function $R_s^{\mathrm{GCC}}[\cdot] : \mathbb{N} \to \mathbb{R}$ is

$$R_s^{\mathrm{GCC}}[k] \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} \Psi_s\left(e^{j\omega}\right) X_s^*\left(e^{j\omega}\right) Y\left(e^{j\omega}\right) e^{j\omega k} d\omega, \quad (4)$$

where $\Psi_s(e^{j\omega})$ is the frequency response of the filter employed to extract specific pieces of information from the CCF [26]. There are many different choices for this filter [26, 27], but the majority of applications involving localization through acoustic signals use the *phase transform* (PHAT) filter

$$\Psi_s^{\mathrm{PHAT}}\left(e^{j\omega}\right) \triangleq \frac{1}{\left|X_s^*\left(e^{j\omega}\right) Y\left(e^{j\omega}\right)\right|}. \quad (5)$$

Note that, in the ideal setup with $y[n] = x_s[n-k_0]$, $R_s^{\text{GCC-PHAT}}[k] = \delta[k-k_0]$, featuring only one peak that can be unambiguously associated with the TOF. This indicates that GCC-PHAT may facilitate the TOF-estimation task by eliminating spurious peaks that may appear in the standard CCF due to the emitted signals' self-similarities. This indeed occurs for wideband signals, but when narrowband signals are employed the division in (5) may not be well defined for all frequencies or may actually induce noise enhancement.

Practical scenarios usually face moderate to severe reverberation, which essentially means that the received signal $y[n]$ can be represented as the sum of delayed and attenuated versions of the signal emitted by the loudspeaker $x_s[n]$. Since the phase of a signal conveys the information related to delays and also because reverberation distorts the magnitude of the signal, the PHAT filter discards the magnitude response of the cross-spectrum density (Fourier transform of the CCF) and focuses solely on its phase response, which makes the GCC-PHAT more robust against reverberation, considering wideband signals [27].

The TOF estimate can eventually be computed as

$$\widehat{t}_s \triangleq \frac{1}{F}\left(\arg\max_{k\in\mathbb{N}}\left\{R_s^{\mathrm{GCC}}[k]\right\}\right), \quad (6)$$

where $F \in \mathbb{R}_+$ is the sampling frequency.

One can employ TOF- or TDOF-based ASL techniques depending on the kind of errors present in each $\widehat{t}_s$. Indeed, for each estimate $\widehat{t}_s$ there are random and systematic errors associated with it. If systematic errors are dominant, then it is better to rely on TDOF estimates, as similar systematic errors are canceled by the subtraction of the TOF estimates. On the other hand, if random errors are dominant, then it is better to directly use the TOF estimates, because TDOF estimates may amplify random errors. In the next subsection, practical issues that lead to some kinds of errors in $\widehat{t}_s$ are described.

2.4. Common Practical Issues. The accuracy of ASL techniques depends on how close the estimates $\widehat{t}_s$ are from the *actual* TOFs t_s. A problem might occur if the direct path connecting source and sensor does not exist or is temporarily obstructed, that is, if there is no LOS; in this case, $\widehat{t}_s$ in (6) may be the propagation time corresponding to some other path related to reflection. Similarly, the *directionality* of the loudspeakers is also important, as it is related to the coverage area of ASL systems. Both LOS and directionality issues should be taken into account when designing ASL systems, that is, when choosing and positioning the loudspeakers.

There are also some practical issues, like *reverberation* and *interference*, that corrupt the received signal $y[n]$. When such issues are present, the CCF function may have other peaks; as a consequence, the peak corresponding to the direct path (i.e., to the TOF) becomes less prominent, sometimes even smaller than surrounding peaks. Therefore, this kind of issue impairs the received signal and thus the CCF and can be circumvented by choosing an adequate acoustic signal $x_s[n]$ (herein also called *probe signal*) with desirable correlation properties and by exploiting the physics of the problem, as will be explained in Section 4.

In general, if source and/or sensor are moving, then the Doppler effect arises and may further corrupt the estimation of t_s. Nevertheless, in many ASL applications, the relative velocity between source and sensor is usually much smaller than the speed of sound v, which makes Doppler effect negligible.

The remainder of this section is devoted to some issues involving time measurements, namely, *asynchrony* and *frequency mismatch* between loudspeakers and microphones. When there is asynchrony between microphone and loudspeaker, the argument that maximizes $R_s^{\mathrm{GCC}}[k]$ can be written as $k_s^o + \Delta k_s \in \mathbb{N}$, where k_s^o is the actual discrete-time TOF and Δk_s is the discrete-time bias summarizing all time-lag impairments. In addition, if there is frequency mismatch, then the sensor's clock frequency can be written as $F + \Delta F \in \mathbb{R}_+$, where F is the source's clock frequency. Considering these two deviations, the TOF estimate obtained in (6) is related to the actual discrete-time TOF k_s^o by

$$\widehat{t}_s = \frac{k_s^o + \Delta k_s}{F + \Delta F} = \frac{k_s^o + \Delta k_s}{F}\left(\frac{1}{1 + \Delta F/F}\right). \quad (7)$$

By assuming that $|\Delta F| \ll F$, this expression can be simplified as

$$\widehat{t}_s \approx \frac{k_s^o + \Delta k_s}{F}\left(1 - \frac{\Delta F}{F}\right) \approx \alpha t_s + \beta_s, \quad (8)$$

where $\alpha \triangleq (F - \Delta F)/F$, $t_s \approx k_s^o/F$, and $\beta_s \triangleq \Delta k_s(F - \Delta F)/F^2$. Observe that if the discrete-time biases Δk_s, with $s \in \mathcal{S}$, are due to asynchrony and some constant playback-system delay only, that is, no acoustic impairment or LOS obstruction is present, then $\Delta k_s = \Delta k$ for all $s \in \mathcal{S}$, implying that $\beta_s = \beta$ for all $s \in \mathcal{S}$. Hence,

$$\|\mathbf{m} - \mathbf{p}_s\| = vt_s \approx \left(\frac{v}{\alpha}\right)\widehat{t}_s + \left(-\frac{v\beta}{\alpha}\right) = a\widehat{t}_s + b, \tag{9}$$

in which the speed of sound v is embedded into the unknown parameters a and b. In any context in which clock skew can be neglected (i.e., $\alpha = 1$), a actually models the speed of sound and b models a constant range-bias.

Once the main definitions and common practical issues have been described, one now has all basic tools to address the sensor localization problem itself.

3. ASL Algorithms

As mentioned in Section 1, there are many different solutions to the problem of localizing an acoustic sensor. This section describes three state-of-the-art families of algorithms that span the main ASL approaches. Section 3.1 details TOF- and TDOF-based least-squares (LS) solutions [23] built on the affine model in (9). Section 3.2 indicates how the steered-response power (SRP) technique, originally targeted to the problem of sound source localization (SSL) using microphone array [27], can be adapted to work in the ASL context. Section 3.3 describes a new localization approach that relies on searching for regions (cuboids) that are likely to contain the sensor.

3.1. Least-Squares. In the LS procedure, many TOF or TDOF measurements are collected and used to estimate the microphone position. In what follows, the TOF-based LS method known as T*ab* and the TDOF-based LS method known as D*a* proposed in [22, 23] are described.

3.1.1. TOF-Based Problem (Tab). Motivated by the relation in (9), let the error e_s corresponding to the sth loudspeaker be defined as $e_s \triangleq (a\widehat{t}_s + b)^2 - \|\mathbf{m} - \mathbf{p}_s\|^2$, for all $s \in \mathcal{S}$, where the unknowns are $\mathbf{m}, a, b$. Define the error vector $\mathbf{e}_{(\mathrm{T}ab)} \triangleq \begin{bmatrix} e_1 & e_2 & \cdots & e_S \end{bmatrix}^T$, in which the subscript T*ab* stands for TOF-based solutions with unknown parameters a, b, besides $\mathbf{m}$ itself. Then, the T*ab* problem is formulated as

$$\min_{\mathbf{m},a,b} \quad \left\|\mathbf{e}_{(\mathrm{T}ab)}\right\|^2, \tag{10}$$

which is nonlinear and has no closed-form solution.

3.1.2. TDOF-Based Problem (Da). Without loss of generality, assume that $s = 1$ corresponds to the reference loudspeaker, whose associated TOF and position are $\widehat{t}_1$ and $\mathbf{p}_1 = \mathbf{0}$, respectively. In order to work with TDOFs, one can simply apply the following relation to (9): $a\widehat{t}_s + b = (a\widehat{t}_s - a\widehat{t}_1) + (a\widehat{t}_1 + b) \approx a\widehat{\tau}_{s1} + \|\mathbf{m}\|$, where $\widehat{\tau}_{s1}$ is the TDOF related to the sth and the reference loudspeakers. In this way, by defining $\widetilde{e}_s \triangleq (a\widehat{\tau}_{s1} + \|\mathbf{m}\|)^2 - \|\mathbf{m} - \mathbf{p}_s\|^2$, $s \in \mathcal{S} \setminus \{1\}$, and forming the error vector $\mathbf{e}_{(\mathrm{D}a)} \triangleq \begin{bmatrix} \widetilde{e}_2 & \cdots & \widetilde{e}_S \end{bmatrix}^T$, the D*a* problem can be written as

$$\min_{\mathbf{m},\mathbf{a}} \quad \left\|\mathbf{e}_{(\mathrm{D}a)}\right\|^2, \tag{11}$$

which is also nonlinear. Observe that, by working with TDOFs, the D*a* approach eliminates the systematic error b, as stated in Section 2.2. The acronym D*a* points out that this solution uses TDOFs and estimates a, besides $\mathbf{m}$ itself.

3.1.3. LS Solutions. For both LS approaches, the error vector can be written in a general form as $\mathbf{e} = \mathbf{H}\boldsymbol{\Theta} - \mathbf{g}$, where vector $\boldsymbol{\Theta}$ is comprised of the unknowns to be estimated, whereas $\mathbf{H}$ and $\mathbf{g}$ are known matrix and vector, respectively. The structures of the involved vectors and matrix are given in the following equations for both T*ab* and D*a* [23]:

$$\begin{aligned} \mathbf{e}_{(\mathrm{T}ab)} &= \begin{bmatrix} 1 & 2\mathbf{p}_1^T & \widehat{t}_1^2 & 2\widehat{t}_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 2\mathbf{p}_S^T & \widehat{t}_S^2 & 2\widehat{t}_s \end{bmatrix} \begin{bmatrix} b^2 - \|\mathbf{m}\|^2 \\ \mathbf{m} \\ a^2 \\ ab \end{bmatrix} - \begin{bmatrix} \|\mathbf{p}_1\|^2 \\ \vdots \\ \|\mathbf{p}_S\|^2 \end{bmatrix}, \\ \mathbf{e}_{(\mathrm{D}a)} &= \begin{bmatrix} 2\widehat{\tau}_{21} & 2\mathbf{p}_2^T & \widehat{\tau}_{21}^2 \\ \vdots & \vdots & \vdots \\ 2\widehat{\tau}_{S1} & 2\mathbf{p}_S^T & \widehat{\tau}_{S1}^2 \end{bmatrix} \begin{bmatrix} \|\mathbf{m}\|\, a \\ \mathbf{m} \\ a^2 \end{bmatrix} - \begin{bmatrix} \|\mathbf{p}_2\|^2 \\ \vdots \\ \|\mathbf{p}_S\|^2 \end{bmatrix}. \end{aligned} \tag{12}$$

Therefore, the entries and dimension of these variables depend upon whether a TOF- or a TDOF-based solution is searched.

Although the entries of $\boldsymbol{\Theta}$ are not independent from each other, a better tradeoff between computational burden and estimation accuracy is achieved when such a constraint is neglected [23, 28]. By doing so, the nonlinear problems in (10) and (11) can be written as the following unconstrained LS problem:

$$\min_{\boldsymbol{\Theta}} \quad \|\mathbf{e}\|^2, \tag{13}$$

whose closed-form solution is

$$\widehat{\boldsymbol{\Theta}} \triangleq \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{g}. \tag{14}$$

In (14), it is assumed that $\mathbf{H}$ is a full-rank matrix, which means that $S \geq 6$ for both the T*ab* and D*a* techniques.

When one can assume that $a = v$, that is, there are no sampling frequency mismatches, and the speed of sound is known a priori, then the TDOF-based model can be written as

$$\mathbf{e}_{(\mathrm{D})} = \begin{bmatrix} 2v\widehat{\tau}_{21} & 2\mathbf{p}_2^T \\ \vdots & \vdots \\ 2v\widehat{\tau}_{S1} & 2\mathbf{p}_S^T \end{bmatrix} \begin{bmatrix} \|\mathbf{m}\| \\ \mathbf{m} \end{bmatrix} - \begin{bmatrix} \|\mathbf{p}_2\|^2 - v^2\widehat{\tau}_{21}^2 \\ \vdots \\ \|\mathbf{p}_S\|^2 - v^2\widehat{\tau}_{S1}^2 \end{bmatrix}. \tag{15}$$

Its corresponding LS solution, denoted by subscript D (TDOF-based with no auxiliary parameter), is related to the one presented in [28, 29] for the SSL problem.

It is worth pointing out the strong dependence of the LS solutions on the estimated TOFs or TDOFs, whose errors may have a harmful effect on the localization accuracy. This fact motivates the use of a different approach that tries to estimate the sensor position directly, without resorting to prior TOF or TDOF estimation. SRP-inspired solutions allow one to do that, as described in the next section.

3.2. Steered-Response Power. The SRP technique was originally proposed for SSL problems using microphone array, but it can be adjusted to ASL as follows.

In this context, the SRP divides the search space into a grid of points representing possible candidates for the microphone position and measures the overall similarity among emitted and received signals related to each of these points. The point with highest similarity is selected as the estimate of the microphone position. Thus, the SRP in the ASL context searches for the grid point $\mathbf{x} \in \mathbb{R}^3$ that maximizes the following objective function:

$$W(\mathbf{x}) \triangleq \sum_{s=1}^{S} R_s^{\mathrm{GCC}}\left[k_s(\mathbf{x})\right], \tag{16}$$

where $k_s(\mathbf{x})$ is the discrete-time TOF assuming that the microphone position is $\mathbf{m} = \mathbf{x}$. When $R_s^{\mathrm{GCC}}[k]$ is the result of applying the PHAT filter to the CCF, the technique is known as SRP-PHAT.

When using a dense grid of points covering the entire search space, the SRP technique achieves accurate position estimates even in reverberant scenarios, as all TOFs are considered simultaneously in a joint optimization procedure. However, the SRP drawback is its high computational burden due to the exhaustive search throughout the many points of the grid.

It is important to highlight here that SRP-inspired solutions have not been thoroughly explored in the ASL context. For example, the impact of the affine model in (9) on the performance of SRP-inspired solutions should be studied, especially with regard to the bias b (this problem does not appear in SSL techniques, for they use TDOFs). This seems to be a reasonable research direction, which could also benefit from modifications on the SRP proposed to circumvent the high computational burden when solving SSL problems [24, 30–33]. Some of these strategies avoid the exhaustive search throughout the entire grid, like in [30], and therefore cannot guarantee accurate position estimates, as parts of the search space are not explored. More interesting results were obtained by the strategies that perform the exhaustive search throughout a grid of regions covering the entire search space, such as in [24, 32, 33]. The advantage over the standard SRP comes from the fact that the grid of regions can be made much coarser than the grid of points and still reach accurate estimates.

The next section describes a new technique that puts together a prior step of TDOF estimation, which usually yields computationally simple solutions, and the SRP-inspired idea of an exhaustive search over a grid of regions, which tends to improve the robustness of the method.

3.3. Region-Based Search. Consider a division of the whole search region $\mathbb{V} \subset \mathbb{R}^3$ in a number $C \in \mathbb{N}$ of compact, disjoint, and connected regions (e.g., cuboids) $\mathscr{V}_c \subset \mathbb{V}$, with $c \in \{1, 2, \ldots, C\}$, over which the search will be performed (hereinafter, it is assumed that the regions are cuboids). Each TDOF gives rise to a hyperboloid that is the locus of all candidate spatial points that are consistent with that TDOF. If the hyperboloid related to TDOF $\tau_{s_1 s_2}$ intersects cuboid $\mathscr{V}_c$, then TDOF $\tau_{s_1 s_2}$ is coherent with that cuboid [24]. It can be proven using the Karush-Kuhn-Tucker (KKT) conditions [34] that the coherence test can be implemented in a very efficient way—see Theorem 3 and Corollary 1 of [24]. This result shows that the TDOF $\tau_{sr}(\mathbf{x})$ of any point $\mathbf{x}$ inside $\mathscr{V}_c$ is within the interval (observe that the TDOF notation $\tau_{s_1 s_2}$ was replaced by τ_{sr}, with $s, r \in \mathcal{S}$, to emphasize that the second loudspeaker acts as a reference speaker):

$$\mathscr{I}_c^{sr} = \left[\min_j \tau_{sr}\left(\mathbf{v}_c^j\right), \max_j \tau_{sr}\left(\mathbf{v}_c^j\right)\right] \subset \mathbb{R}, \tag{17}$$

where $\mathbf{v}_c^j \in \mathbb{R}^3$ ($j \in \{1, 2, \ldots, 8\}$) is the position of the jth vertex of cuboid $\mathscr{V}_c$. Note that (17) implies that comparisons are the only operations required to perform the coherence test, since the TDOF bounds for each cuboid can be precomputed just once and stored in a lookup table.

Each TDOF information is usually consistent with several cuboids and thus does not suffice to determine the cuboid that indeed contains the sensor; in fact, one has to aggregate information from all loudspeakers in order to choose the cuboid that most likely contains the sensor. Mathematically, by defining the coherence test through the indicator function

$$\mathbb{I}_c(\tau_{sr}) = \begin{cases} 1, & \text{if } \tau_{sr} \in \mathscr{I}_c^{sr} \\ 0, & \text{otherwise,} \end{cases} \tag{18}$$

a robust objective function that implicitly assesses the likelihood of the cth cuboid containing the sensor is defined as

$$\mathscr{L}_c = \sum_{r=1}^{S-1} \sum_{s=r+1}^{S} \mathbb{I}_c(\tau_{sr}). \tag{19}$$

Assuming the TDOFs associated with every microphone pair are known (in practice, they are estimated), $\mathscr{L}_c$ essentially counts the number of those TDOFs which are coherent with the cth cuboid. The cuboid with highest $\mathscr{L}_c$ is the one that most likely contains the sensor. The reader should observe that the computational complexity of evaluating (19) is very low, as it requires only counting and comparisons. It should be highlighted that there are alternative objective functions in the SSL literature whose definition and motivation are very similar to that of (19)—see [30, 32, 33].

Alternatively, the following objective function could also be employed:

$$\mathscr{L}_c' = \sum_{s=1, s\neq r}^{S} \mathbb{I}_c(\tau_{sr}). \tag{20}$$

In comparison to $\mathscr{L}_c$, $\mathscr{L}_c'$ uses a single loudspeaker (the rth one) as reference and, therefore, it is not as robust as

$\mathscr{L}_c$ for it relies on an accurate estimate of the rth TOF. Equation (19), on the other hand, employs all pairs of TDOFs, which is in some sense related to the "principle of least commitment" [35] to which some researchers attribute the excellent performance of SRP techniques [36].

The search procedure usually begins with a coarse volumetric grid covering the entire search space [32]. For instance, one can divide each spatial dimension in $L \in \mathbb{N}$ parts of equal length, generating $C = L^3$ cuboids, each with dimensions $(l_x/L) \times (l_y/L) \times (l_z/L)$, where the entire search space $\mathbb{V}$ is supposed to be a cuboid with edges of sizes l_x, l_y, and l_z. Then, a hierarchical search takes place where the candidate cuboids are tested in order to verify their coherence with the available TOF/TDOF information [31]. This search can sometimes be posed within the branch-and-bound paradigm [24] originally devised for combinatorial and discrete optimization problems [37]. Given the current candidate cuboids, the generation of the next set of candidate regions starts by the selection of those (winner) cuboids that maximize some objective function. Then, each winner cuboid is recursively partitioned into smaller cuboids, each of them carrying more specific information about its portion of space [31]. The iterative process advances until some stop criterion is met—for instance, the maximum number of iterations is achieved.

At this point, it might be clear that the practical issues mentioned in Section 2.4 impair the accuracy of the TOF or TDOF estimates. Since the new region-based approach of this section and the LS approaches in Section 3.1 strongly depend on these estimates, it becomes crucial to obtain accurate TOF/TDOF estimates which are robust to these practical issues. This is why the techniques presented in Section 4 focus on improving TOF estimates.

4. Improving TOF/TDOF Estimates

This section describes three techniques from the literature that improve the accuracy of TOF/TDOF estimates. Section 4.1 details the sliding windows approach, which resorts to physical constraints to define time-windows likely to contain the actual TOFs. Section 4.2 describes a method that uses matching pursuit (MP) algorithms to select candidate TOF estimates within the previously defined time-windows, while cleaning spurious components that appear in CCFs or GCC functions. Section 4.3 shows how one can change the selected TOFs in order to improve the final localization accuracy.

4.1. Sliding Windows. Each CCF (or GCC) may present many peaks, hampering the task of detecting the peak associated with the direct path. In order to overcome this difficulty, a set of physical constraints the actual TOFs must satisfy can be imposed on the search for the direct-path CCF peak. These constraints stem from room geometry and probe signals' inherent structure, such as duration and cyclical nature—ubiquitous in asynchronous passive ASL systems. The sliding windows (SW) approach proposed in [23] is a robust and computationally efficient technique employed to search for CCF peaks while taking into account the aforementioned constraints.

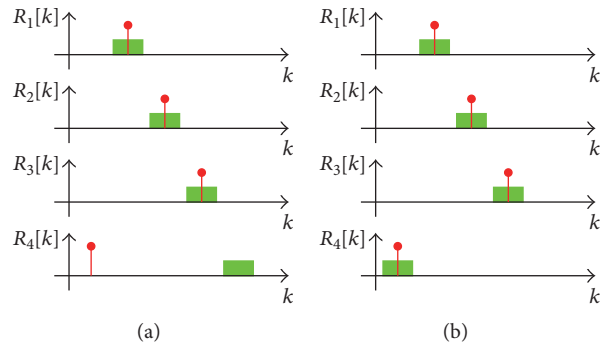

Figure 1: Illustration of the ability of the SW technique to estimate the correct loudspeakers' emission order. Example of windows (in green) that (a) do not have the correct emission order and (b) have the correct emission order. Note that windows are delayed in distinct CCFs in order to compensate for the transmission delay between different loudspeakers.

The key idea underlying the SW technique is to jointly search for the direct-path peaks across all CCFs simultaneously. This joint optimization is conducted by adding the selected peaks within fixed-duration windows defined in order to account for possible errors due to acoustic impairments, such as reverberation, interference, and LOS obstruction. The duration of those windows satisfies physical constraints based on room dimensions (e.g., the maximum lag can be upper bounded by the room diameter—maximum distance between points within the room—divided by the speed of sound). Moreover, the knowledge of the delay between consecutive emissions of different loudspeakers as well as the probe signals' durations is used to constrain the distance between windows from the corresponding CCFs.

Figure 1 illustrates the way SW technique works in a simplified setup. Consider an anechoic room where four loudspeakers emit cyclical impulses as probe signals. Assume there is a fixed delay between emissions from different loudspeakers, which is handy in practical reverberant environments for it helps to deal with signal superpositions, although not mandatory in this toy-example. The first step is to define time-windows (depicted as green boxes in Figure 1) based on the maximum admissible propagation delays within the room. After that, the windows (one for each CCF) are spaced apart based on the fixed delay between consecutive emissions of probe signals. Then, the values of the highest CCF peak within each window are added and stored. A sliding windows process takes place in order to evaluate the initial time-index for which the aforementioned accumulated peak values achieve their maximum. Figure 1(a) shows the windows in a position where the accumulated peak values are not maximal; the sliding process continues in a cyclical fashion, accounting for the cyclical emission of probe signals, as illustrated in Figure 1(b), where the maximum is finally achieved. Thus, the SW technique also has the ability to blindly detect the emission order of the probe signals in the acquired signal, regardless of the order ambiguity induced by the asynchrony among transmitters and receivers.

Mathematically, the SW technique yields a set of time-windows $\mathscr{W}_s$, in which the sth TOF should be located, defined as [23]

$$\mathscr{W}_s \triangleq \{k_{\max} + (u_s - 1)\,\delta - K, \ldots, k_{\max} + (u_s - 1)\,\delta + K\}, \tag{21}$$

where $k_{\max}$ is related to the start of the windows that (one expects to) contain the actual TOFs, δ is the fixed delay between windows (easily defined based on the duration of a complete probe signal cycle), K is dependent on the dimensions of the room and defines the duration of the windows, and $[u_1 \;\; u_2 \;\; \cdots \;\; u_S]$ is a right-cyclical shift of vector $[1 \;\; 2 \;\; \cdots \;\; S]$.

It is noteworthy that the highest peak inside $\mathscr{W}_s$ often corresponds to the desired sth TOF and, when this does not happen, the search for the correct peak should be done inside the window $\mathscr{W}_s$, whose length is usually much shorter than the sth CCF/GCC support, significantly reducing the computational burden of forthcoming stages that try to correct misestimated TOFs. In fact, choosing the correct peak within those windows may still be rather challenging depending on the ASL environment. This fact calls for nonlinear processing (see Section 4.2) that goes beyond the linear processing implemented by CCF-based approaches.

4.2. Matching Pursuit. Using the lag associated with the highest CCF peak as TOF estimate is equivalent to employing matched filters for the TOF-estimation task [38]. As seen before, multipath propagation, along with other acoustic phenomena, modifies the desired correlation properties of the probe signals, hampering the TOF-estimation task. A way to circumvent such issues consists of performing a precise estimation of the channel-impulse response (CIR), which can be done if the probe signals are known. The time stamp of the first nonzero coefficient of the estimated CIR should be the desired TOF in an ideal setup.

A natural candidate for CIR estimation is the maximum likelihood (ML) estimate, whose solution, under some mild hypotheses, is equivalent to an unconstrained and nonregularized least-squares (LS) estimate, in the absence of noise measurement and interferences [39] (in fact, the equivalence remains even under normally distributed noise [40]).

Besides its lack of robustness under non-Gaussian perturbations [41], the ML solution is not feasible for real-time localization systems due to its high computational complexity. The authors in [20] proposed a greedy pursuit for CIR estimation, specifically the *matching pursuit* (MP) algorithm [42]. Such strategy was motivated by [43], which developed a CIR estimation algorithm for multiuser environments in code division multiple access (CDMA) systems, aiming at TOF-based radio localization. The authors in [43] compared the ML and MP techniques, eventually concluding that the last outperforms the former. The inferior ML performance is attributed to the underlying overparameterization of the CIR, which makes the detection of the actual TOF unreliable [43].

The MP algorithm relies on sparse representations of the signals of interest [44], working by progressively isolating the signal structures which are coherent with a predefined dictionary of signals. In the context of ASL, the works [12, 20] employ the MP algorithm to describe an excerpt of the recorded signal as a linear combination of delayed versions of one probe signal. Such decomposition permits to infer the direct-path delay even when its CCF/GCC peak is highly attenuated compared to peaks corresponding to reflected-path delays.

(1) **procedure** TofEstimation($s, N_p, \mathscr{P}_s$)
(2) $\quad\overline{y}[n] \longleftarrow y[n]$
(3) $\quad n_{\text{It}} \longleftarrow 0$
(4) $\quad$**while** $n_{\text{It}} < N_p$ **do**
(5) $\quad\quad\forall k \in \mathscr{P}_s : R_s[k] \longleftarrow \sum_{n\in\mathbb{N}} \overline{y}[n]x_s[n-k]$
(6) $\quad\quad k_p[n_{\text{It}}] \longleftarrow \underset{k}{\operatorname{argmax}}|R_s[k]|$
(7) $\quad\quad\alpha[n_{\text{It}}] \longleftarrow R_s[k_p[n_{\text{It}}]]$
(8) $\quad\quad\overline{y}[n] \longleftarrow \overline{y}[n] - \alpha[n_{\text{It}}]x_s[n - k_p[n_{\text{It}}]]$
(9) $\quad\quad n_{\text{It}} \longleftarrow n_{\text{It}} + 1$
(10) $\quad$**end while**
(11) $\quad\widehat{t}_s \longleftarrow \dfrac{(\min\{k_p[m]\})}{F}$
(12) $\quad$**return** $\widehat{t}_s$ (sth estimated TOF)
(13) **end procedure**

Algorithm 1: TOF estimation by MP algorithm.

Suppose one has good reasons to believe that the early arrivals of the signal emitted by the sth loudspeaker occur at the set of indexes $\mathscr{P}_s$ (e.g., $\mathscr{P}_s = \mathscr{W}_s$ given in (21)) of the recorded signal $y[n]$. Assuming that the N_p most significant channel coefficients extracted by the MP contain the actual TOF, the estimation of the sth TOF could be performed as described in Algorithm 1. Note that this algorithm can easily provide a set of N_p candidate TOF estimates corresponding to the sth loudspeaker.

When the N_p parameter employed in Algorithm 1 is too large, or when probe signals are not completely orthogonal, some spurious components may appear before the direct-path delay, as illustrated in Figure 2. One can circumvent this issue by establishing a detection threshold, as proposed in [20].

It should be pointed out that, even when Algorithm 1 fails, it yields a number (N_p) of TOF candidates that is much smaller than the number of CCF/GCC peaks; this is key to further performing a refined search for the correct peak without significantly increasing the computational burden (it is assumed here that $\mathscr{P}_s$ contains the actual TOF). A possible refinement step is described in the next section.

4.3. TOF Selection. As explained before, reverberation and LOS obstruction between loudspeaker and sensor nodes (and other issues) may severely impair the TOF estimate and, therefore, the overall localization procedure [45]. One way of tackling those issues consists of wisely selecting the TOFs from the peaks of the CCFs/GCCs when they are not the highest ones, or even discarding some CCF/GCC information when the actual TOF cannot be found in it

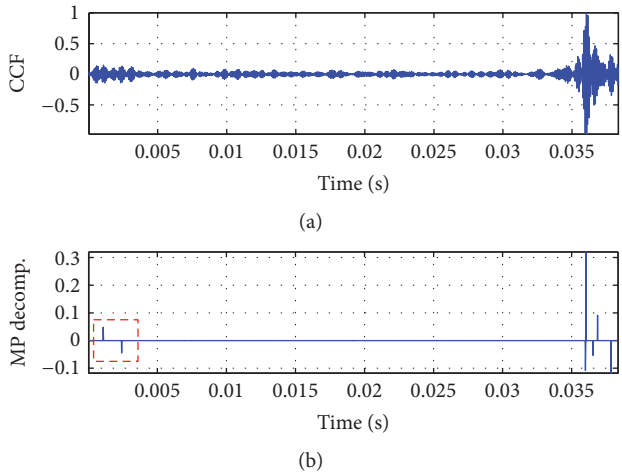

FIGURE 2: Example of MP decomposition. (a) Original CCF; (b) the first 7 components extracted with Algorithm 1. The spurious components that appear before the actual TOF are depicted inside the red box.

[23, 45]. Note that such strategy, henceforth called "TOF selection" or simply TS, does not perform an exhaustive search over a grid of spatial points nor requires evaluations of complex functionals; therefore, it is expected to demand fewer numerical operations than SRP-inspired algorithms.

Each CCF/GCC usually provides many peaks and, consequently, a combinatorial search for the best combination of peaks may be infeasible in some applications. In order to perform a computationally efficient search, the TS technique should rely on some physically plausible heuristics. One possible heuristics example, for instance, makes use of the fact that spurious CCF/GCC peaks generated by multipath propagation often occur after the correct peak, because non-LOS paths are longer [23]. This assumption motivates a search among peaks that occur before the current one if its correctness is under question. Another example of heuristics takes into account that the magnitude of an erroneously detected peak is slightly larger than the magnitude of the correct CCF/GCC peak in a high-SNR regime without LOS obstruction [23]. A more common hypothesis states that the highest CCF/GCC peak is often associated with the desired TOF information.

Such heuristics are not sufficient for reliable TOF estimation in practical environments. In general, they should be connected with an objective function that incorporates geometric constraints in order to assess a specific set of candidate TOFs. Such assessment, which takes into account available *a priori* information, is key to solving inverse problems [31]. One example of such geometric-based function is the maximum discrepancy between the position obtained by feeding the localization procedure with the TOFs (or TDOFs) estimated from the CCFs/GCCs and the TOFs (or TDOFs) that would have been observed if the sensor were indeed at the estimated position [23]. Another example is the coherence of the current set of estimated TOFs (or TDOFs) with one or more spatial regions [45] (see Section 3.3 for more details).

The TS strategy begins with an evaluation of the objective function, whose result should meet a stop criterion. Such criterion can test whether a threshold is not violated and/or whether the maximum number of iterations is not exceeded. While such criterion is not satisfied, a "Refine" procedure should be used. Such refinement represents the core procedure of this method, which updates the current set of TOF (or TDOF) candidates and may even include a procedure for discarding some CCFs/GCCs from which one cannot extract a reliable TOF (or TDOF) estimate [23].

5. Proposed ASL System

As the localization techniques described in Section 3 present complementary benefits, the proper choice of the technique to be used strongly depends on the requirements of the particular ASL application. In addition, the recent advances described in Section 4 have proved to work in practice and their combinations with the techniques in Section 3 open up a myriad of exciting research directions that have not been fully explored in the ASL context. This section contains an example of how one can put together the advantage of all techniques in Section 4 along with the region-based search proposed in Section 3.3, giving rise to a novel ASL system.

As mentioned before, SRP-inspired searches present high computational complexity due to the required evaluation of many complex functionals. Converting the search into a selection of correct candidate TOFs among the CCF/GCC peaks is an interesting alternative as long as the number of such peaks is small. This is not the case, however, if each CCF/GCC presents many peaks. Resorting to MP-based algorithms is a convenient way of circumventing such problem—recall that Algorithm 1 may return a variable number (N_p) of candidate CCF/GCC peaks. Nevertheless, the MP algorithm should be fed with a set of indexes $\mathscr{P}_s$, for the sth loudspeaker, in which the early arrival of the related probe signal is likely to be (see Section 4.2). This set $\mathscr{P}_s$ must have a small cardinality; otherwise the computational complexity becomes prohibitively high. Fortunately, a robust estimate of a relatively small set $\mathscr{P}_s = \mathscr{W}_s$ for each loudspeaker can be performed by the SW technique described in Section 4.1; indeed, the SW technique can deliver a small excerpt of the recorded signal to the MP-based sth TOF-estimation algorithm, which may then return N_p candidate TOFs for each loudspeaker. Using the common assumption that the highest CCF/GCC peak (or, equivalently, the highest coefficient extracted by the MP method) is often related to the desired TOF, a reasonable strategy consists of using the highest coefficients obtained by the MP as a preliminary starting point for the set of estimated TOFs. This set cannot be directly used to infer the sensor location with LS techniques (like T*ab* or D*a*) because such techniques are very sensitive to TOFs/TDOFs errors [22], and one cannot assure beforehand that the initial set of TOFs is accurate (one may only expect that *most of* TOF/TDOF information is correct). At this point, the new robust region-based search mechanism—see Section 3.3— that solves the inverse problem of inferring the sensor location from the data by progressively partitioning some candidate regions (i.e., the winner ones) takes place.

(1) **procedure** TofSelection($\mathbf{k}_1, \ldots, \mathbf{k}_S, \bar{\mathbf{t}}_c, \mathbf{c}_w, N_p$)
(2) **for** $s \coloneqq 1$ **to** S **do**
(3) **for** $m \coloneqq 1$ **to** N_p **do**
(4) $\tilde{t}_s[m] \longleftarrow \dfrac{k_s[m]}{F}$
(5) **end for**
(6) **end for**
(7) **for** $s \coloneqq 1$ **to** S **do**
(8) **fitness** $\longleftarrow [0, \ldots, 0]_{1\times N_p}$
(9) **for** $m \coloneqq 1$ **to** N_p **do**
(10) $\bar{\mathbf{t}} \longleftarrow \bar{\mathbf{t}}_c$
(11) $\bar{t}[m] \longleftarrow \tilde{t}_s[m]$
(12) **for** $i \coloneqq 1$ **to** $\#\mathbf{c}_w$ **do**
(13) fitness$[m] \longleftarrow$ fitness$[m] + \mathscr{L}_{c_w[i]}(\bar{\mathbf{t}})$
(14) **end for**
(15) **end for**
(16) $\widehat{m} \longleftarrow \arg\max$ **fitness**$[\cdot]$
(17) $\widehat{t}[s] \longleftarrow \bar{t}_s[\widehat{m}]$
(18) **end for**
(19) **end procedure**

Algorithm 2: TOF selection by geometric constraints.

A simple choice is the use of cuboids, thus generating $2^3 = 8$ children cuboids from the parent one by dividing each dimension into two equal parts. For each candidate cuboid, one can evaluate the objective function (19), whose computational complexity is very low. Cuboids that maximize such functional are selected for further decompositions. Due to the "principle of least commitment" [35] it is expected that the winner cuboids contain the sensor location, which imposes additional geometric constraints to the TOF/TDOF data. Such constraints can be employed to perform a smart selection of the TOFs—remember that (18) indicates whether some TDOF data is coherent with a given cuboid. This selection is rather favored by the previous application of the MP-based TOF detection step, since it delivers only N_p candidate TOFs for each loudspeaker, with N_p being typically smaller than 10 [12, 20].

Algorithm 2 describes the proposed TOF-selection scheme. Vector $\mathbf{c}_w$ contains the indexes $c_w[i]$ of current winner cuboids, vectors $\mathbf{k}_s$, with $s \in \mathcal{S}$, contain the N_p sample indexes $k_s[m]$ of the sth loudspeaker candidate peaks, and $\bar{\mathbf{t}}_c$ contains the current set of TOF estimates. Note that the core calculation of this algorithm is performed in line 13, which employs (19), whose computational cost is negligible since it does not require multiplication nor division operations (only comparisons and sums). Algorithm 2 returns vector $\widehat{\mathbf{t}}$, which contains a set of updated TOF data. After the geometric-based selection of the next set of candidate TOFs, one may again proceed to a new decomposition step, until some criterion is met (typically, until the prescribed maximum number of iterations is reached). Although such recursive procedure is not pointwise, one may easily obtain a point estimation for the sensor location by evaluating the mass center of the last winner cuboid (it should be stressed that at the final step of the described procedure only one cuboid remains, in general). Figure 3 presents a block diagram of the proposed method.

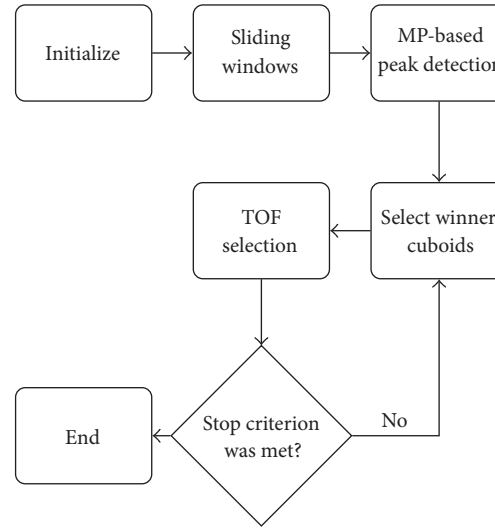

Figure 3: Block diagram of the proposed ASL system.

Table 1: Coordinates (in meters) of each loudspeaker.

Index	x	y	z
1	0.3520	5.5430	2.2570
2	0.3880	5.5620	1.7065
3	0.6160	0.8275	2.0105
4	0.3120	2.8905	1.6430
5	0.3420	2.8725	0.9850
6	4.5720	3.0250	1.6020
7	4.5630	3.0185	1.0075
8	2.3315	0.5235	1.1710
9	4.5330	5.5830	2.2605
10	4.4740	5.5505	1.6875
11	4.1700	0.9000	2.0160

6. Experimental Results

The main goal of this section is to present an evaluation of the proposed method (described in Section 5) in a real-world scenario, as a proof of concept.

6.1. Experimental Setup. The test environment is a lecture room with a measured reverberation time ($T60$) of approximately 500 ms and dimensions 5.2 m × 7.5 m × 2.6 m. Eleven loudspeakers with a diameter of 7.6 cm are located according to the positions indicated in Table 1. A mobile device is located at position {1.746, 4.425, 0.748} m with the automatic gain control (AGC) activated. The captured signals, for each estimation, have durations of 0.55 s. The signal is recorded at F = 48 kHz with 24-bit precision.

In order to provide an accurate localization procedure, the probe signals should meet several prerequisites, namely, low audibility, high orthogonality, robustness against interferences, and short duration (which guarantees a high refresh

Table 2: Median localization error (in cm) for different choices of emission power P (in dB, relative to a standard power level). The number of MP decompositions is N_p and the number of hierarchical cuboid decompositions is N_c.

N_p	TS^-	TS^+	TS^-	TS^+	TS^-	TS^+
	$N_c = 4$		$N_c = 6$		$N_c = 8$	
			$P = 0$ dB			
1	50.42	50.42	44.39	44.39	44.37	44.37
3	50.42	18.28	44.39	18.28	44.37	18.16
5	50.42	8.00	44.39	13.79	44.37	11.80
7	50.42	8.00	44.39	13.79	44.37	12.78
			$P = -6$ dB			
1	34.06	34.06	31.04	31.04	33.02	33.02
3	34.06	18.28	31.04	17.50	33.02	17.08
5	34.06	8.00	31.04	16.08	33.02	15.01
7	34.06	8.00	31.04	13.79	33.02	11.45
			$P = -12$ dB			
1	50.42	50.42	43.54	43.54	45.40	45.4
3	50.42	22.76	43.54	26.88	45.40	28.55
5	50.42	18.71	43.54	22.55	45.40	21.5
7	50.42	18.71	43.54	19.18	45.40	17.44
			$P = -18$ dB			
1	50.42	50.42	56.71	56.71	56.72	56.72
3	50.42	34.06	56.71	31.63	56.72	32.65
5	50.42	18.71	56.71	17.91	56.72	18.06
7	50.42	18.71	49.67	26.16	50.1	24.88
			$P = -24$ dB			
1	50.42	50.42	57.73	57.73	58.48	58.48
3	50.42	18.71	57.73	14.04	58.48	12.39
5	50.42	13.14	57.73	15.17	58.48	16.21
7	50.42	18.71	57.73	17.76	58.48	17.98

rate of the sensor location). In the following experiments, a new set of probe signals, called *polyphonic chirps*, were designed to meet such requirements. The starting point of the polyphonic chirps consists of linear chirps, whose bandwidth ranges from 14.0 to 20.0 kHz, with amplitudes following the inverse A-weighting curve (which is the inverse of the human relative loudness [46]), so that there is an increasing gain from 14 to 20 kHz. One can define 4 subbands from the primary chirp signal, namely, from 14.0 to 15.5 kHz, from 15.5 to 17.0 kHz, from 17.0 kHz to 18.5 kHz, and from 17.5 to 20.0 kHz. Within each subband, one can play a chirp with increasing frequency or with decreasing frequency. Bit 0 is associated with the latter and bit 1 with the former. Hence, one can assign to each loudspeaker a 4-bit codeword whose bits, from the most to the least significant, are associated with the subchirps from the lowest to the highest frequency band. Therefore, 4 subchirps are expected to be simultaneously played-back by each loudspeaker, giving birth to a specific polyphonic chirp. Additionally, modifications were heuristically (in the sense that they were based on the audibility level of the resulting probe signals) taken to shorten the probe signals audibility: the two lower-frequency subchirps are attenuated by factors of 20 (subband 14.0–15.5 kHz) and 10 (subband 15.5–17.0 kHz). Further, each polyphonic chirp follows a 5-ms fade-in/fade-out envelope to hide undesirable discontinuities at the start and at the end of the underlying signals. A cyclical emission of 30-ms polyphonic chirps is performed, following an order previously known by the sensor node. There is a 20-ms silence interval between consecutive emissions, aiming at reducing the effects of interference and signal superposition caused by reverberation. It should be emphasized that through informal listening tests under typical environmental conditions, we found that these chirps are inaudible for most people.

The probe signals were emitted with power P dB relative to a standard power level, with $P \in \{0, -6, -12, -18, -24\}$. The purpose of the experiment is to assess the importance of geometric-based TOF-selection procedure (see Algorithm 2) with different levels of emitted power. For each configuration, 100 different excerpts were employed to provide different localization estimates, whose error statistics are presented in the following.

6.2. Median Localization Error. This section aims at assessing the impact of different choices of the number of MP decompositions N_p and the number of hierarchical cuboid decompositions N_c on the localization accuracy. Table 2 presents the results with TOF-selection procedure disabled (TS^-) and TOF-selection procedure enabled (TS^+). From this table, one can conclude that the TOF selection is very effective

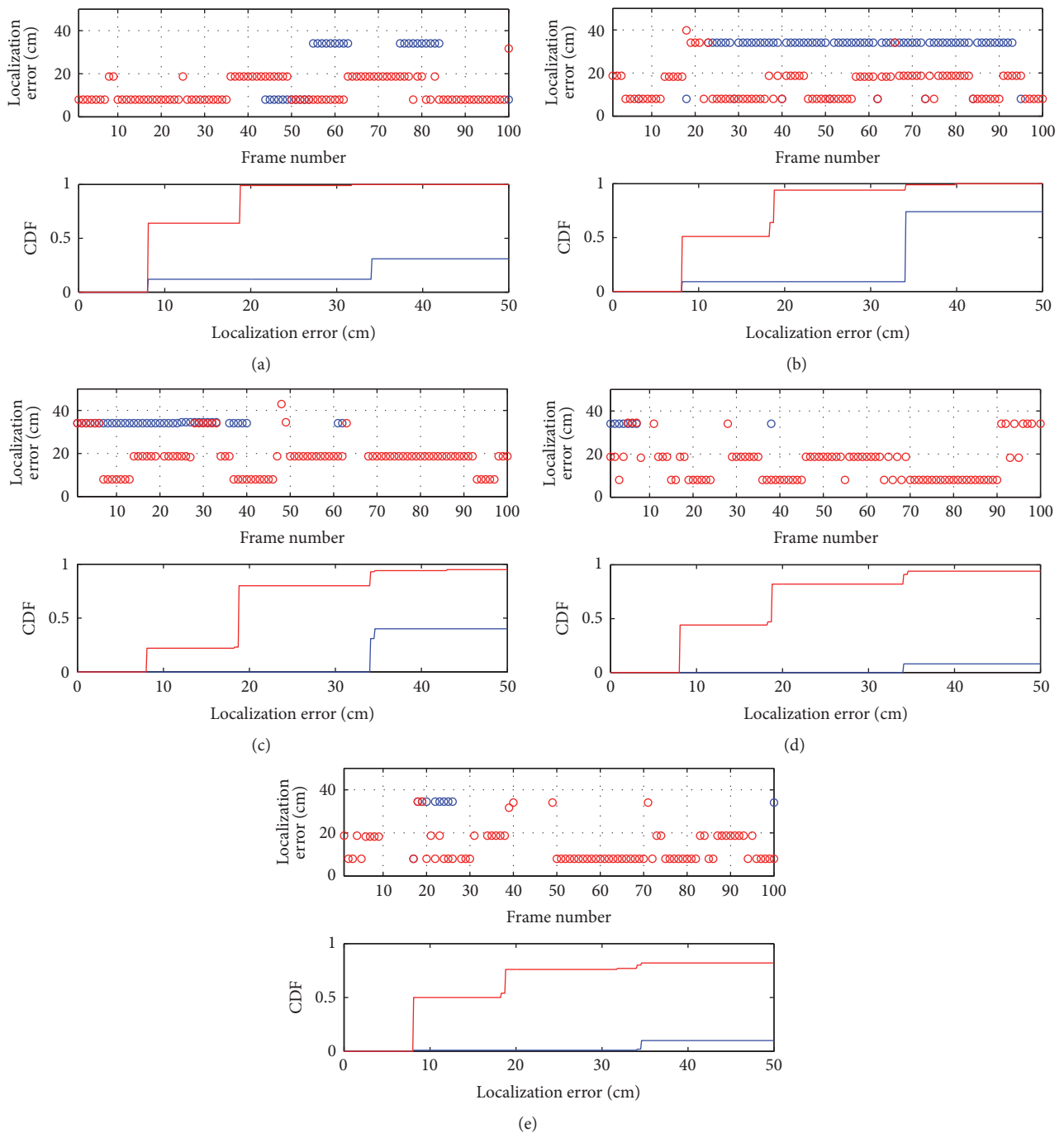

FIGURE 4: Localization error (within a 50 cm range) for each recorded signal excerpt and respective CDF with P = (a) 0, (b) −6, (c) −12, (d) −18, and (e) −24 dB. Blue: localization procedure with TOF selection disabled. Red: localization procedure with TOF selection enabled.

in reducing the median error provided there exists more than one TOF candidate (i.e., $N_p > 1$). The use of such procedure has never increased the median localization error and sometimes reduces it in more than 70%. Hence, from now on it is considered that the TOF selection is always enabled. Regarding the number of MP decompositions, choices of N_p between 5 and 7 (inclusive) appear to achieve a good compromise between accuracy and computational cost. The use of a large number of hierarchical decompositions N_c is usually not a wise strategy, since it increases the probability of selecting an erroneous cuboid, which harms the posterior hierarchical search. It should be noticed that the method works even with extremely low power emission (in general, when $P = -24$ dB, the loudspeaker sound is inaudible).

6.3. CDF of the Error. In order to evaluate the impact of the emitted power on the localization accuracy, Figure 4 shows the localization error (when it is smaller than 50 cm) and the

cumulative distribution function (CDF) of the localization error with $N_p = 5$ MP extracted components and $N_c = 4$ hierarchical decompositions of cuboids. These results reveal a progressive accuracy degradation with the reduction of the power emission. Moreover, the advantage of employing the proposed TS technique is also clear from those results.

7. Concluding Remarks

This paper serves three purposes. First, it presents a brief review of the vast literature of acoustic sensor localization (ASL) for those beginning in this field. Indeed, a wide range of topics are covered, ranging from the fundamentals of ASL to some state-of-the-art techniques. Second, this paper provides new research directions within the ASL field by explaining how one can borrow some concepts from its dual problem: the sound source localization (SSL). In this way, many research opportunities are opened. Third, this paper proposes a new ASL technique that combines region-based search (which is inspired by some recently proposed SSL techniques that employ hierarchical searches) and matching pursuit estimation of times-of-flight (TOFs).

Another difference from our previous work [23] is that the ASL technique proposed here works with probe signals which are inaudible for most people, as they have low power and contain only high-frequency components. However, the use of such probe signals makes the TOF estimation a much more challenging task. This explains why robust TOF-estimation techniques are thoroughly discussed throughout the paper.

A real-world experiment was conducted in order to demonstrate that the proposed ASL technique is capable of estimating the position of mobile devices with a median localization error below 20 cm. Usually, the ultimate goal of many practical ASL systems is to find the position of someone carrying a mobile device and, therefore, an estimation error inferior to 20 cm is rather reasonable.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Haddad, Lima, Martins, and Biscainho thank Capes, CNPq, and Faperj agencies for funding their research. The authors thank Mr. Mauricio V. M. Costa, Mr. Igor M. Quintanilha, and Mr. Felipe B. da Silva for their valuable support on recording the database used in this work.

References

[1] F. Höflinger, J. Wendeberg, R. Zhang et al., "Acoustic self-calibrating system for indoor smartphone tracking (ASSIST)," in *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN '12)*, pp. 1–9, Sydney, Australia, November 2012.

[2] O. Vinyals, E. Martin, and G. Friedland, "Multimodal indoor localization: an audio-wireless-based approach," in *Proceedings of the 4th IEEE International Conference on Semantic Computing (ICSC '10)*, pp. 120–125, September 2010.

[3] N. Gutierrez, C. Belmonte, J. Hanvey, R. Espejo, and Z. Dong, "Indoor localization for mobile devices," in *Proceedings of the 11th IEEE International Conference on Networking, Sensing and Control (ICNSC '14)*, pp. 173–178, April 2014.

[4] T. Ajdler, I. Kozintsev, R. Lienhart, and M. Vetterli, "Acoustic source localization in distributed sensor networks," in *Proceedings of the Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers (Asilomar '04)*, pp. 1328–1332, Pacific Grove, CA, USA, 2004.

[5] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 439–443, 2013.

[6] K. Hasegawa, N. Ono, S. Miyabe, and S. Sagayama, "Blind estimation of locations and time offsets for distributed recording devices," in *Latent Variable Analysis and Signal Separation*, V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, Eds., vol. 6365, pp. 57–64, Springer, Berlin, Germany, 2010.

[7] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, "BeepBeep: a high accuracy acoustic ranging system using COTS mobile devices," in *Proceedings of the 5th ACM International Conference on Embedded Networked Sensor Systems (SenSys '07)*, pp. 1–14, November 2007.

[8] M. H. Hennecke and G. A. Fink, "Towards acoustic self-localization of ad hoc smartphone arrays," in *Proceedings of the 3rd Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA '11)*, pp. 127–132, Edinburgh, UK, May-June 2011.

[9] A. M. Cavalcante, R. C. D. Paiva, R. Iida, A. Fialho, A. Costa, and R. D. Vieira, "Audio beacon providing location-aware content for low-end mobile devices," in *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN '12)*, November 2012.

[10] P. Pertila, M. Mieskolainen, and M. Hamalainen, "Passive self-localization of microphones using ambient sounds," *in Proc. of the 20th European Signal Processing Conference (EUSIPCO '12)*, pp. 1314–1318, 2012.

[11] K. Liu, X. Liu, L. Xie, and X. Li, "Towards accurate acoustic localization on a smartphone," in *Proceedings of the IEEE INFOCOM 2013 - IEEE Conference on Computer Communications*, pp. 495–499, Turin, Italy, April 2013.

[12] F. J. Álvarez, T. Aguilera, and R. López-Valcarce, "CDMA-based acoustic local positioning system for portable devices with multipath cancellation," *Digital Signal Processing*, vol. 62, pp. 38–51, 2017.

[13] V. Raykar and R. Duraiswami, "Automatic position calibration of multiple microphones," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 69–72, Montreal, Que., Canada.

[14] J. M. Sachar, H. F. Silverman, and W. R. Patterson, "Microphone position and gain calibration for a large-aperture microphone array," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 42–52, 2005.

[15] S. T. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1025–1034, 2005.

[16] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *IEEE Transactions on*

Audio, Speech and Language Processing, vol. 16, no. 3, pp. 666–670, 2008.

[17] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 70–83, 2005.

[18] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "Self-localization of ad-hoc arrays using time difference of arrivals," *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 1018–1033, 2016.

[19] R. Pfeil, M. Pichler, S. Schuster, and F. Hammer, "Robust acoustic positioning for safety applications in underground mining," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 11, pp. 2876–2888, 2015.

[20] F. J. Álvarez and R. López-Valcarce, "Multipath cancellation in broadband acoustic local positioning systems," in *Proceedings of the 9th IEEE International Symposium on Intelligent Signal Processing (WISP '15)*, pp. 1–6, IEEE, Siena, Italy, May 2015.

[21] M. Cobos, J. J. Perez-Solano, Ó. Belmonte, G. Ramos, and A. M. Torres, "Simultaneous ranging and self-positioning in unsynchronized wireless acoustic sensor networks," *IEEE Transactions on Signal Processing*, vol. 64, no. 22, pp. 5993–6004, 2016.

[22] D. B. Haddad, L. O. Nunes, W. A. Martins, L. W. P. Biscainho, and B. Lee, "Closed-form solutions for robust acoustic sensor localization," in *Proceedings of the 14th IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '13)*, pp. 1–4, October 2013.

[23] D. B. Haddad, W. A. Martins, M. D. V. M. Da Costa, L. W. P. Biscainho, L. O. Nunes, and B. Lee, "Robust acoustic self-localization of mobile devices," *IEEE Transactions on Mobile Computing*, vol. 15, no. 4, pp. 982–995, 2016.

[24] L. O. Nunes, W. A. Martins, M. V. Lima et al., "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5171–5183, 2014.

[25] 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, 1976.

[26] J. Chen, J. Benesty, and Y. Huang, "An acoustic MIMO framework for analyzing microphone-array beamforming," in *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pp. 25–28, Honolulu, HI, USA, April 2007.

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

[28] A. Beck, P. Stoica, and J. Li, "Exact and approximate solutions of source localization problems," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1770–1778, 2008.

[29] P. Stoica and J. Li, "Source Localization from Range-Difference Measurements," *IEEE Signal Processing Magazine*, vol. 23, no. 6, pp. 63–66, 2006.

[30] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 1, pp. 121–124, April 2007.

[31] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 499–508, 2004.

[32] M. V. S. Lima, W. A. Martins, L. O. Nunes et al., "A volumetric SRP with refinement step for sound source localization," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1098–1102, 2015.

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

[34] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*, Springer, New York, NY, USA, 2007.

[35] S. Birchfield and D. Gillmor, "Acoustic source direction by hemisphere sampling," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, pp. 3053–3056, Salt Lake City, UT, USA, 2001.

[36] 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, 2007.

[37] B. Gendron and T. G. Crainic, "Parallel branch-and-bound algorithms: survey and synthesis," *Operations Research*, vol. 42, no. 6, pp. 1042–1066, 1994.

[38] D. Middleton, "On new classes of matched filters and generalizations of the matched filter concept," *IRE Transactions on Information Theory*, vol. 6, no. 3, pp. 349–360, 1960.

[39] T. J. Abatzoglou, J. M. Mendel, and G. A. Harada, "The constrained total least squares technique and its applications to harmonic superresolution," *IEEE Transactions on Signal Processing*, vol. 39, no. 5, pp. 1070–1087, 1991.

[40] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, "Parameter estimation in the presence of bounded data uncertainties," *SIAM Journal on Matrix Analysis and Applications*, vol. 19, no. 1, pp. 235–252, 1998.

[41] R. R. Wilcox, *Introduction to robust estimation and hypothesis testing*, Academic Press, Amsterdam, The Netherlands, 1st edition, 2011.

[42] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[43] S. Kim and R. A. Iltis, "A matching-pursuit/GSIC-based algorithm for DS-CDMA sparse-channel estimation," *IEEE Signal Processing Letters*, vol. 11, no. 1, pp. 12–15, 2004.

[44] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[45] D. B. Haddad, W. A. Martins, L. W. P. Biscainho, M. D. V. M. Da Costa, and K.-H. Kim, "Choosing coherent times of flight for improved acoustic sensor localization," in *Proceedings of the International Telecommunications Symposium (ITS '14)*, August 2014.

[46] J. Parmanen, "A-weighted sound pressure level as a loudness/annoyance indicator for environmental sounds - Could it be improved?" *Applied Acoustics*, vol. 68, no. 1, pp. 58–70, 2007.