



Exploiting spatial diversity for increasing the robustness of sound source localization systems against reverberation

Guillermo García-Barrios^{a,*}, Eduardo Latorre Iglesias^b, Juana M. Gutiérrez-Arriola^a, Rubén Fraile^a, Nicolás Sáenz-Lechón^a, Víctor José Osma-Ruiz^a

^aCentro de Investigación en Tecnologías Software y Sistemas Multimedia para la Sostenibilidad (CITSEM), Universidad Politécnica de Madrid, Campus Sur, 28031 Madrid Spain

^bDep. Ingeniería Audiovisual y Comunicaciones, Universidad Politécnica de Madrid, Campus Sur, 28031 Madrid Spain

ARTICLE INFO

Article history:

Received 27 April 2022

Received in revised form 20 October 2022

Accepted 17 November 2022

Available online 8 December 2022

Keywords:

Acoustic signal processing
Microphone arrays
Sound source localization
Steered-response power maps
Acoustic reverberation
Spatial diversity

ABSTRACT

Acoustic reverberation is one of the most relevant factors that hampers the localization of a sound source inside a room. To date, several approaches have been proposed to deal with it, but have not always been evaluated under realistic conditions. This paper proposes exploiting spatial diversity as an alternative approach to achieve robustness against reverberation. The theoretical arguments supporting this approach are first presented and later confirmed by means of simulation results and real measurements. Simulations are run for reverberation times up to 2 s, thus providing results with a wider range of validity than in other previous research works. It is concluded that the use of systems consisting of several, sufficiently separated, small arrays leads to the best results in reverberant environments. Some recommendations are given regarding the choice of the array sizes, the separation among them, and the way to combine SRP-PHAT maps obtained from diverse arrays.

© 2022 Published by Elsevier Ltd.

1. Introduction

While sound source localization (SSL) has been an active research topic for a long time, during the last years the development of both wireless sensor networks [1] and computational analysis of sounds [2] has renewed its interest for some applications, such as surveillance [3]. Developing robust SSL systems in order to make these applications feasible is still an open research issue [4]. Reverberation is one of the factors that most significantly compromises the robustness of these systems, even in the case of short reverberation times [5].

1.1. Problem statement: Effect of reverberation on sound source localization using the GCC

SSL algorithms can be grouped into three broad types [e.g. 6]: one-stage beamforming, two-stage time delay, and high-resolution spectral estimation-based methods. The first one is based on maximizing the sound source power over an evaluated region, the second one is based on calculating the time difference of arrival (TDOA) for each pair of microphones as a first stage, and the third one implies calculating eigenvalues of multiple signal

correlation matrices (e.g. MUSIC). In complex acoustic scenarios where the audio signals are harmed by multi-path reflections due to reverberation, the performance of all these algorithms is degraded.

Being able to estimate the TDOA of the acoustic signal to two different microphones is at the core of sound source localization algorithms, being it either explicitly as in two-stage algorithms, or implicitly as in both one-stage and spectral estimation schemes. One of the most widely used tools for estimating the TDOA is the generalized cross-correlation (GCC) [7,8]. Therefore, analyzing the effect of reverberation on the GCC can lead to conclusions that are valid for the majority of SSL algorithms.

Given a sound signal $s(t)$ generated by an acoustic source placed at position \vec{r}_s , the sound captured by such microphones, i and k , can be expressed as:

$$\begin{aligned} m_i(t) &= h_{s,i}(t) * s(t) \\ m_k(t) &= h_{s,k}(t) * s(t), \end{aligned} \quad (1)$$

where only the signal distortion caused by the acoustic transfer function h has been considered. Under anechoic conditions $h_{s,i}(t) = \delta(t - \tau_{s,i})$, where $\tau_{s,i}$ is the propagation delay between the source and the microphone i :

$$\tau_{s,i} = c \cdot \|\vec{r}_s - \vec{r}_i\|, \quad (2)$$

* Corresponding author.

E-mail address: guillermo.garcia.barrios@upm.es (G. García-Barrios).

being c the sound velocity, \vec{r}_i the position of microphone i , and $\|\cdot\|$ the Euclidean norm. The same definitions apply to microphone k . Thus, under such conditions, the following identities hold true:

$$\begin{aligned} m_i(t) &= s(t - \tau_{s,i}) \\ m_k(t) &= s(t - \tau_{s,k}) = m_i(t - \Delta\tau_{ik}), \end{aligned} \quad (3)$$

where $\Delta\tau_{ik} = \tau_{s,k} - \tau_{s,i}$ is the TDOA, which can be estimated from the cross-correlation, i.e. the GCC, between $m_i(t)$ and $m_k(t)$.

However, the response of the acoustic channel in reverberant environments cannot be assumed to be a mere delay. Instead, the sound signal undergoes some delay spreading, and each channel impulse response can be written as the sum of a direct path plus a reverberant component:

$$\begin{aligned} h_{s,i}(t) &= \delta(t - \tau_{s,i}) + h_{s,i}^r(t - \tau_{s,i}) \\ h_{s,k}(t) &= \delta(t - \tau_{s,k}) + h_{s,k}^r(t - \tau_{s,k}), \end{aligned} \quad (4)$$

where $h_{s,i}^r(t)$ and $h_{s,k}^r(t)$ are delay spread models and are assumed to be null for $t < 0$. In general, $h_{s,i}^r(t)$ and $h_{s,k}^r(t)$ will be different, since both microphones are not placed in the same position, and the identities in (3) are not valid:

$$\begin{aligned} m_i(t) &= s(t - \tau_{s,i}) + s(t - \tau_{s,i}) * h_{s,i}^r(t) \\ m_k(t) &= s(t - \tau_{s,k}) + s(t - \tau_{s,k}) * h_{s,k}^r(t) \neq m_i(t - \Delta\tau_{ik}). \end{aligned} \quad (5)$$

The GCC between signals $m_i(t)$ and $m_k(t)$ is defined as [8]:

$$R_{ik}(\tau) = \int_{-\infty}^{\infty} \frac{M_i(\omega)M_k^*(\omega)}{\psi(\omega)} \cdot e^{j\omega\tau} d\omega, \quad (6)$$

where $M_i(\omega)$ and $M_k(\omega)$, respectively, are the Fourier transforms of the microphone signals $m_i(t)$ and $m_k(t)$, $\psi(\omega)$ is a frequency weighting function, $*$ means complex conjugation, and j is the imaginary unit. The use of the phase transform (PHAT) weighting has been shown to be advantageous in reverberant environments [9]. If this weighting is used, then the GCC evaluated at time lag τ can be calculated as:

$$R_{ik}(\tau) = \int_{-\infty}^{\infty} \frac{M_i(\omega)M_k^*(\omega)}{2\pi|M_i(\omega)M_k(\omega)|} \cdot e^{j\omega\tau} d\omega, \quad (7)$$

Under anechoic conditions, the microphone signals satisfy (3). Therefore:

$$\begin{aligned} R_{ik}(\tau) &= \int_{-\infty}^{\infty} \frac{M_i(\omega)M_k^*(\omega)}{2\pi|M_i(\omega)|^2} \cdot e^{j\omega(\tau+\Delta\tau_{ik})} d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{j\omega(\tau+\Delta\tau_{ik})} d\omega \\ &= \delta(\tau + \Delta\tau_{ik}), \end{aligned} \quad (8)$$

where $\delta(\tau)$ is the Dirac delta function. The shape of $R_{ik}(\tau)$ under anechoic conditions is illustrated in Fig. 1. However, the GCC in reverberant conditions cannot be assumed to be an impulse, according to the model in (5):

$$\begin{aligned} R_{ik}^r(\tau) &= \int_{-\infty}^{\infty} \frac{S(\omega)(1+H_{s,i}^r(\omega))S^*(\omega)(1+H_{s,k}^{r*}(\omega))}{2\pi|S(\omega)|^2|1+H_{s,i}^r(\omega)||1+H_{s,k}^{r*}(\omega)|} \cdot e^{j\omega(\tau+\Delta\tau_{ik})} d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{(1+H_{s,i}^r(\omega))(1+H_{s,k}^{r*}(\omega))}{|1+H_{s,i}^r(\omega)||1+H_{s,k}^{r*}(\omega)|} \cdot e^{j\omega(\tau+\Delta\tau_{ik})} d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{(1+H_{s,i}^r(\omega)+H_{s,i}^{r*}(\omega)+H_{s,i}^r(\omega)H_{s,i}^{r*}(\omega))(1+H_{s,k}^r(\omega)+H_{s,k}^{r*}(\omega)+H_{s,k}^r(\omega)H_{s,k}^{r*}(\omega))}{|1+H_{s,i}^r(\omega)+H_{s,i}^{r*}(\omega)+H_{s,i}^r(\omega)H_{s,i}^{r*}(\omega)||1+H_{s,k}^r(\omega)+H_{s,k}^{r*}(\omega)+H_{s,k}^r(\omega)H_{s,k}^{r*}(\omega)|} \cdot e^{j\omega(\tau+\Delta\tau_{ik})} d\omega, \end{aligned} \quad (9)$$

being $S(\omega)$ and $H_{s,i}^r(\omega)$ Fourier transforms of the acoustic signals $s(t)$ and $h_{s,i}^r(t)$, respectively. Reverberation has a negative impact on the estimation of relative time delays because the delay spread introduced by the acoustic channels causes secondary peaks in the GCC, due to the fact that $m_k(t) \neq m_i(t - \Delta\tau_{ik})$, and these additional peaks can lead to wrong estimations of the TDOA $\Delta\tau_{ik}$ [5,10–12]. This effect is illustrated in Fig. 1, where the GCC function for two pairs of microphones is plotted under both anechoic and reverberant conditions. Note that the presence of reverberation causes the appearance of secondary peaks in the GCC (left plot), and it may even lead to a significant shift of the main peak (right).

Therefore, reverberation poses the challenge for SSL systems of producing localization estimates that are robust against the distortion introduced in the GCC or, more generally, in algorithms for calculating TDOA.

1.2. State of the art

Calculating steered-response power (SRP) maps has shown to be one of the sound source localization algorithms providing the highest robustness against reverberation [9,13], especially when the phase transform is used to calculate the GCC function [12,14,15]. Note that this approach does not explicitly rely on TDOA estimates, it is a one-stage algorithm. Instead, SRP maps are built directly from the GCC function. This eliminates the impact of erroneous TDOA estimation, though secondary peaks of the GCC still affect the localization results. It is known that in general circumstances the robustness of SRP-based algorithms can be enhanced by increasing the number of microphones in the array [9], and by averaging frame-based GCCs in the case of speech signals [13].

Given that reverberation is to a greater or lesser extent present in all real acoustic environments, a number of research works have

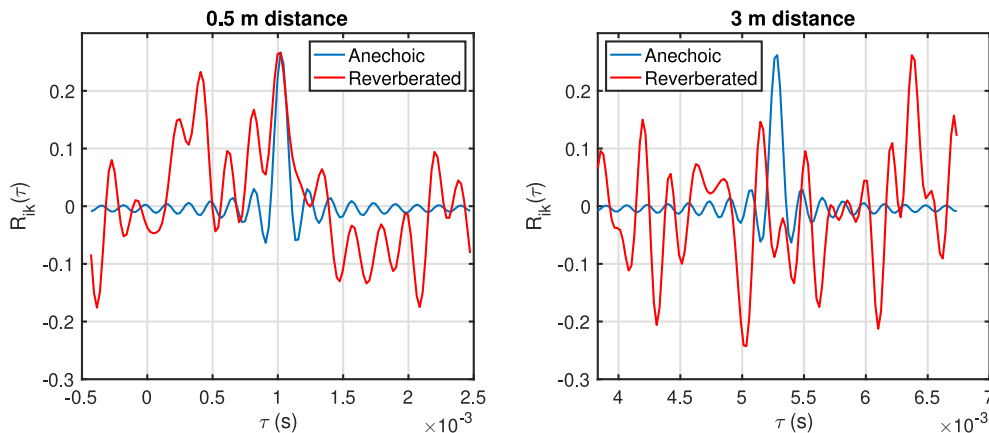


Fig. 1. Comparison of the GCCs for the same pair of microphones and sound source position in anechoic and reverberant conditions (reverberation time, $RT = 0.8$ s) for a microphone separation equal to 0.5 m (left) and 3 m (right).

been targeted at improving SSL robustness against this effect. These may be approximately classified into three great groups: those trying to compensate the effect of the reverberant component of the acoustic channels $h_{s,i}^r(t)$ on the microphone signals $m_i(t)$, those attempting to reduce the relevance of the secondary peaks in the GCC or, alternatively, reducing their effect on the localization estimates, and those combining TDOA estimates from several microphone arrays.

The first one of the previously mentioned groups of approaches aims at estimating the acoustic channel between the sound source and the different microphones to compensate for the effect of reverberation in the original signal. One of the firstly proposed techniques was based on cepstral pre-filtering before calculating the generalized cross-correlation function [16]. The cepstral filter was calculated based on the assumption that the delay spreading filters modelling reverberation have minimum phase. The same algorithm was later applied to binaural estimation of the direction of arrival (DOA) [17]. Operating in the cepstral domain is computationally expensive; for this reason an alternative all-pole modeling of the acoustic channel was proposed by Parisi et al. [18]. Alternative approaches in this group involve adaptive processing of both signals $m_i(t)$ and $m_k(t)$ to estimate a “de-reverberated” GCC when the sound signal is stationary [19]. Later developments propose reducing the reverberant components of the microphone signals by processing them in the time–frequency plane [20], or by applying iterative optimization algorithms [21,22].

The second group of approaches addresses the problem of reverberation similarly to noise, by proposing or modifying GCC estimators. This is the case of [15,23], where a new version of the maximum likelihood (ML) weight for the GCC using a circular array was introduced. Yet, different articles have reported the out-performance of the PHAT weighting function over ML [9,10,24] in several conditions. For this reason, a new GCC estimator that consisted of a combination of both was presented in [25]. Yet another estimator, called PHAT- β , was designed to improve the accuracy of SSL systems for narrowband and broadband signals [26]. Some additional algorithms have been proposed later OR in later contributions for post-processing the GCC in order to smooth it [27], to optimize the information extracted from GCC peaks [28,29], or to select the components of the GCC most reliable for estimating the DOA using a diffuseness mask obtained from a dereverberation technique [30].

The idea that using systems with a large number of microphone pairs (i, k) could be used to generate a large number of TDOA estimates, subsequently discarding the most inconsistent ones (outliers), was proposed some decades ago [31]. This exploitation of spatial diversity for achieving good localization results has also been implicit in later proposals involving distributed arrays [e.g. 32] or even moving arrays [33]. Apart from discarding inconsistent TDOA estimates, some other algorithmic refinements profiting from spatial diversity have also been developed, such as improving the weighting of consistent peaks of the GCCs obtained from diverse arrays [34], diminishing the relevance of the signals captured by microphones more likely to be suffering from reverberation effects [35], or applying a transform to the GCC before using it for estimating localization [36].

1.3. Limitations of previously published experiments

The performance analysis of sound source localization systems carried out so far has suffered from several weaknesses. One of such weaknesses is that many simulations have been run under low reverberation conditions. The magnitude of reverberation is commonly quantified by means of the reverberation time (RT). Typical reverberation times in real acoustic environments range

Table 1

Typical reverberation times in diverse types of facility [38].

Type of facility	RT at mid frequencies
Broadcast studio	0.5 s
Classroom Conference room	1.0 s
Multipurpose auditorium	1.3 s to 1.5 s
Contemporary church Opera house	1.4 s to 1.6 s
Rock concert hall	1.5 s
Symphony hall	1.8 s to 2.0 s
Cathedral	3.0 s or higher

from 0.5 s to 3 s (Table 1). However, except for the thorough evaluation reported by Pérez-Lorenzo et al. [12], in which the RT of the evaluated scenarios reached 2 s, and the works of Zannini et al. [28] and Comanducci et al. [36], who considered reverberation times up to 1.5 s and 1.7 s respectively, the majority of the remaining published results consider scenarios in which the RT is usually below 0.5 s (we do not consider here the results in [21], as they correspond to a small room and position was estimated in a 2D plane). For instance, acoustic conditions simulated by Champagne et al. [11] correspond to an estimated maximum RT equal to 0.5 s; results reported by DiBiase et al. [9] correspond to RT up to 0.2 s; Zhang et al. simulated conditions corresponding to RT equal to 0.1 s and 0.5 s [15]; Lee et al. simulated RT values from 0.2 s to 0.6 s [30]. Some related works have considered longer RT values, but they aimed at estimating DOA instead of source position [17,20,22,37]. Therefore, there still is a need to do further research on the performance of SSL systems in both typical and hard reverberation conditions, i.e. with longer reverberation times.

An additional issue that hampers the practical implementation of sound source localization systems is the requirement of *a priori* information about the acoustic channel associated with some proposed algorithms, such as that proposed by Parisi et al. [18]. One last question that merits further research is the effect of the spatial layout of the microphones within the array. To the best of our knowledge, it seems that only Yu and Silverman [39] have reported a systematic analysis of the performance of DOA estimation as a function of microphone separation. They came to the conclusion that large aperture arrays required over 40 cm separation between microphones to achieve low angle quantization errors, and that excessive separation (over 100 cm) could lead to performance degradation due to the differences between $h_{s,i}^r(t)$ and $h_{s,k}^r(t)$ negatively affecting the resulting GCC. Among the research works cited previously, the effect of modifying the number of microphones is only studied in [36]. To the best of our knowledge, the remaining publications proposing the use of several microphone arrays in reverberant environments do not specifically and systematically analyze the effect of spatial diversity.

1.4. Research objective

Considering the previously reported literature review, the objective of the research presented in this paper is two-fold. On the one hand, exploitation of spatial diversity in order to improve SRP-PHAT performance in reverberant environments is explored. Specifically, it is shown that combining information from diverse arrays can provide more robustness against reverberation than some other techniques mentioned before. Moreover, the performance of algorithms that do not require *a priori* information about the acoustic channel is compared with that of SSL systems using the standard SRP-PHAT but with microphone arrays separated at several distances. Secondly, the effect of reverberation in SSL performance is analyzed for RT values up to 2 s. This allows assessing the feasibility of sound source localization applications in realistic scenarios.

The SRP-PHAT algorithm is chosen as a reference because it has consistently shown to provide good performance in reverberation when systematically compared to other approaches. This is true even for some of the most recent experiments involving deep learning approaches [36]. However, this analysis begins by evaluating the impact of microphone distance on the GCC (Section 2), which is at the core of many SSL algorithms. After that, the subsequent impact on SRP-PHAT maps is studied (Section 3). The validity of these analyses is confirmed by both simulations (Section 4) and measurements (Section 5). The discussion of the obtained results is presented in Section 6.

2. Impact of microphone distance on the GCC

2.1. Impact related to signal sampling

The GCC corresponding to two microphone signals captured by a microphone array operating in ideal conditions has a peak at a time delay corresponding to the TDOA (see Fig. 1). When the sound source is sufficiently far from the array, each value of TDOA corresponds to two different DOAs in two-dimensional scenarios. These directions correspond to a certain angle $\pm\theta$ with respect to the straight line connecting both microphones. Thus, identifying the time delay associated with the peak of the GCC is equivalent to estimating the angle of arrival θ . According to the geometrical reasoning presented by Yu and Silverman [39], the root mean square error in estimating θ due to the sampling of audio signals can be approximated as:

$$\sigma_\theta = \left| \arcsin \left(\sin(\theta) + \frac{c}{f_s \cdot d_{ik} \sqrt{12}} \right) - \theta \right|, \quad (10)$$

where f_s is the sampling frequency, and d_{ik} is the distance between both microphones. Fig. 2 shows the values of σ_θ as a function of this distance for several DOAs and for $f_s = 44.1$ kHz. It can be noticed that σ_θ is a decreasing function of distance, so the microphones should be as separated as possible in order to minimize the error in the DOA estimation caused by signal sampling.

2.2. Impact related to reverberation

Since both microphones are placed in the same environment, the mean square value of the reverberant components of their corresponding acoustic channels is expected to be similar [40]:

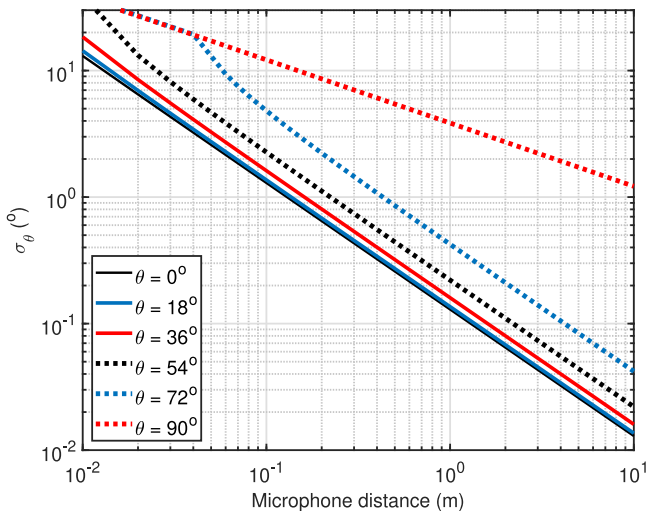


Fig. 2. Root mean square error in the estimation of the DOA as a function of microphone distance for a sampling frequency equal to 44.1 kHz.

$$E \left\{ \left(h_{s,i}^r(t) \right)^2 \right\} \approx E \left\{ \left(h_{s,k}^r(t) \right)^2 \right\} \approx \frac{1-\alpha}{\pi S \alpha}, \quad (11)$$

where $E\{\cdot\}$ is the expectation operator, S is the surface of the room in which the acoustic source and the microphones are placed, and α is the average wall absorption coefficient. An approximate relation between α and the reverberation time of the room T_{60} is given by Sabine's formula [chap.8] [41]:

$$T_{60} \approx 0.163 \cdot \frac{V}{S \alpha}, \quad (12)$$

being V the volume of the room.

Assuming that the mean square value of both $h_{s,k}^r(t)$ and $h_{s,i}^r(t)$ is the same, the reverberant response $h_{s,k}^r(t)$ can be written as a combination of two components, one proportional to $h_{s,i}^r(t)$ and another one independent from it:

$$h_{s,k}^r(t) = \rho_{ik} h_{s,i}^r(t) + (1 - \rho_{ik}) \widetilde{h_{s,k}^r}(t), \quad (13)$$

where $E\{h_{s,i}^r(t) \widetilde{h_{s,k}^r}(t)\} = 0$. ρ_{ik} is the correlation coefficient for both reverberant responses, $h_{s,i}^r(t)$ and $h_{s,k}^r(t)$. It can be approximated by [42]:

$$\rho_{ik} \approx \frac{\sin(kd_{ik})}{kd_{ik}}, \quad (14)$$

where k is the wave number corresponding to the center of the signal bandwidth. Considering (13), the numerator in (9) can be written as:

$$\begin{aligned} & 1 + H_{s,i}^r(\omega) + H_{s,k}^{r*}(\omega) + H_{s,i}^r(\omega) H_{s,k}^{r*}(\omega) \\ &= 1 + H_{s,i}^r(\omega) + \rho_{ik} H_{s,i}^{r*}(\omega) + (1 - \rho_{ik}) \widetilde{H_{s,k}^r}^*(\omega) \\ & \quad + \rho_{ik} H_{s,i}^r(\omega) H_{s,i}^{r*}(\omega) + (1 - \rho_{ik}) H_{s,i}^r(\omega) \widetilde{H_{s,k}^r}^*(\omega) \\ &= \left(1 + H_{s,i}^r(\omega) + H_{s,i}^{r*}(\omega) + H_{s,i}^r(\omega) H_{s,i}^{r*}(\omega) \right) \\ & \quad + (1 - \rho_{ik}) \cdot \left(-H_{s,i}^{r*}(\omega) + \widetilde{H_{s,k}^r}^*(\omega) - H_{s,i}^r(\omega) H_{s,i}^{r*}(\omega) + H_{s,i}^r(\omega) \widetilde{H_{s,k}^r}^*(\omega) \right) \\ &= \left(1 + 2 \cdot \text{Re}\{H_{s,i}^r(\omega)\} + |H_{s,i}^r(\omega)|^2 \right) \\ & \quad + (1 - \rho_{ik}) \cdot \left(-H_{s,i}^{r*}(\omega) + \widetilde{H_{s,k}^r}^*(\omega) - |H_{s,i}^r(\omega)|^2 + H_{s,i}^r(\omega) \widetilde{H_{s,k}^r}^*(\omega) \right). \end{aligned} \quad (15)$$

Note that for $\rho_{ik} = 1$ the second term is null, and the integral in (9) becomes:

$$\begin{aligned} R_{ik}^r(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\left(1 + 2 \cdot \text{Re}\{H_{s,i}^r(\omega)\} + |H_{s,i}^r(\omega)|^2 \right)}{\left| 1 + 2 \cdot \text{Re}\{H_{s,i}^r(\omega)\} + |H_{s,i}^r(\omega)|^2 \right|} \cdot e^{j\omega(\tau + \Delta\tau_{ik})} d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\left(1 + \text{Re}\{H_{s,i}^r(\omega)\} \right)^2 + \left(\text{Im}\{H_{s,i}^r(\omega)\} \right)^2}{\left| \left(1 + \text{Re}\{H_{s,i}^r(\omega)\} \right)^2 + \left(\text{Im}\{H_{s,i}^r(\omega)\} \right)^2 \right|} \cdot e^{j\omega(\tau + \Delta\tau_{ik})} d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{j\omega(\tau + \Delta\tau_{ik})} d\omega, \end{aligned} \quad (16)$$

where $\text{Re}\{\cdot\}$ and $\text{Im}\{\cdot\}$ refer to the real and the imaginary parts, respectively. Since the numerator is always positive, because it is the sum of two squares, the integrand equals 1 and the GCC in the time domain is a delayed impulse $R_{ik}^r(\tau) = \delta(\tau - \Delta\tau_{ik})$, as in the case of anechoic conditions. Consequently, in the ideal case where the reverberant responses of the acoustic channels corresponding to both microphones were proportional to each other, reverberation would not have a negative impact on the GCC, nor on TDOA estimation. However, this would imply both microphones

occupying the same position ($d_{ik} = 0$), as indicated by (14), which is not possible.

In the realistic case of $\rho_{ik} \neq 1$, the integral becomes:

$$R_{ik}^r(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\frac{\left(1 + \operatorname{Re}\{H_{s,i}^r(\omega)\}\right)^2 + \left(\operatorname{Im}\{H_{s,i}^r(\omega)\}\right)^2}{\left|1 + H_{s,i}^r(\omega) + H_{s,k}^{r*}(\omega) + H_{s,i}^r(\omega)H_{s,k}^{r*}(\omega)\right|} \right. \\ \left. + (1 - \rho_{ik}) \cdot \frac{-H_{s,i}^{r*}(\omega) + \widetilde{H_{s,k}^r}^*(\omega) - |H_{s,i}^r(\omega)|^2 + H_{s,i}^r(\omega)\widetilde{H_{s,k}^r}^*(\omega)}{\left|1 + H_{s,i}^r(\omega) + H_{s,k}^{r*}(\omega) + H_{s,i}^r(\omega)H_{s,k}^{r*}(\omega)\right|} \right) \\ \cdot e^{j\omega(\tau + \Delta\tau_{ik})} d\omega \\ = \frac{1}{2\pi} \int_{-\infty}^{\infty} A_{ik}(\omega) \cdot e^{j\omega(\tau + \Delta\tau_{ik})} d\omega \\ + \frac{1 - \rho_{ik}}{2\pi} \int_{-\infty}^{\infty} B_{ik}(\omega) \cdot e^{j\omega(\tau + \Delta\tau_{ik})} d\omega. \quad (17)$$

$A_{ik}(\omega)$ is a real positive function of ω , whose value is not 1 in this case because the numerator is not equal to the denominator. $A_{ik}(\omega) \cdot e^{j\omega\Delta\tau_{ik}}$ is a Fourier transform with linear phase. Therefore, the component of $R_{ik}^r(\tau)$ corresponding to its inverse transform will be a symmetric signal around $\tau = \Delta\tau_{ik}$ [chap.5][43]. In other words, the ideal delayed impulse $\delta(\tau - \Delta\tau_{ik})$ is widened as an effect of reverberation. On the opposite, $B_{ik}(\omega)$ is a complex-valued function of ω . Therefore, the inverse Fourier transform of $B_{ik}(\omega) \cdot e^{j\omega\Delta\tau_{ik}}$ may be asymmetric and may include several peaks in the time domain. Thus, a second effect of reverberation is the loss of symmetry in the GCC around $\tau = \Delta\tau_{ik}$, and the emergence of secondary peaks.

Note that the relevance of the term including $B_{ik}(\omega)$ diminishes as ρ_{ik} approaches 1, and that $A_{ik}(\omega)$ also becomes closer to 1 in this event. Therefore, the impact of reverberation on the GCC is expected to become worse as the distance between microphones increases. This behavior is opposite to that of the DOA estimation error due to signal sampling (recall Fig. 2). Thus a compromise value for microphone distance has to be carefully chosen to keep both effects bounded.

3. SRP maps with spatial diversity

When the GCC-PHAT functions (7) corresponding to all possible microphone pairs within a given array are available, the corresponding SRP map $P(\vec{r})$ can be built as [9]:

$$P(\vec{r}) = 2\pi \sum_{i=1}^K \sum_{k=1}^K R_{ik}(\tau_k(\vec{r}) - \tau_i(\vec{r})) \quad (18) \\ = \sum_{i=1}^K \sum_{k=1}^K \int_{-\infty}^{\infty} \frac{M_i(\omega)M_k^*(\omega)}{|M_i(\omega)M_k(\omega)|} \cdot e^{j\omega(\tau_k(\vec{r}) - \tau_i(\vec{r}))} d\omega,$$

where K is the number of microphones, \vec{r} is the geometrical position, and $\tau_i(\vec{r})$, or $\tau_k(\vec{r})$, is the propagation delay between position \vec{r} and the i^{th} , or k^{th} , microphone. It is well known that this sample-and-sum process can lead to localization errors due to the frequency aliasing problem that was already discussed in [44]. In low-noise and reverberant conditions $P(\vec{r})$ can be interpreted as a log-likelihood function of the position of the acoustic source [15]. Consequently, the best estimate for such position is:

$$\vec{r}_s \approx \arg \max P(\vec{r}). \quad (19)$$

Note that this log-likelihood function results from the addition of terms that can be interpreted as the log-likelihoods of the source positions obtained from the information available in each pair of microphones (i, k) . According to the reasoning in the previous sec-

tion, these additive terms $R_{ik}(\tau_k(\vec{r}) - \tau_i(\vec{r}))$ have the following characteristics:

- The reliability of each term as a likelihood function strongly depends on the distance between microphones d_{ik} : the shorter the distance, the higher the correlation between reverberant responses ρ_{ik} and consequently, the smaller the widening of the main peak of the GCC and the lower the chance of secondary peaks emerging. This effect is illustrated in Fig. 1 for two different microphone distances. It can be seen that in the case of the shorter distance the main peak of the reverberated GCC matches the main peak of the anechoic case corresponding to the true TDOA. Secondary peaks have emerged due to the presence of reverberation, but they do not exceed the height of the main peak. In contrast, for longer microphone distances the height of secondary peaks may exceed that of the main peak, which may even disappear. This results in an evident degradation of the GCC as an estimator of TDOA.
- When several terms corresponding to microphone pairs (i, k) with ρ_{ik} values near 1 are added, the log-likelihood of the true source position should be increased due to the addition of the peaks corresponding to the first term in (17), the one associated to $A_{ik}(\omega)$.
- However, if the same microphone pairs are in nearby positions, the values corresponding to the second term in (17), the one associated to $B_{ik}(\omega)$, should not be expected to be independent among them, since the function $H_{s,i}^r(\omega)$ will be similar for all pairs. This implies that secondary peaks and other distortions appearing in the GCC due to reverberation are not likely to be canceled by adding terms corresponding to different microphone pairs; instead, they might be reinforced.

Therefore, the strategy for selecting the additive terms in (18) should be two-fold. On the one hand, microphone pairs with the lowest possible distance between microphones d_{ik} are preferred, as they yield the lowest distortions in the GCC due to reverberation. On the other hand, if the summation includes terms corresponding to diverse microphone pairs placed at distant positions, the distortions in the GCC due to reverberation are more likely to be compensated when adding these terms. In other words, being $P(\vec{r})$ the SRP map corresponding to one microphone array and one sound signal generated at a certain source position, and being $Q(\vec{r})$ the SRP map corresponding to another array and the same sound source, our hypothesis is that:

- The log-likelihood functions of the source position $P(\vec{r})$ and $Q(\vec{r})$ are distorted by reverberation, and such distortions can be minimized by reducing the distance between the microphones in the corresponding arrays. An example of this effect is represented in Fig. 3, where the maximum peak of the SRP-PHAT map using an array with a short microphone distance is near the actual position of the sound source. However, when the microphone distance is increased, the lack of correlation between the reverberation components of both acoustic channels results in a distorted SRP-PHAT map that whose maximum is far from the position of the sound source.
- The distortions experienced by $P(\vec{r})$ and $Q(\vec{r})$ are more independent among them as the distance between both arrays becomes longer, so $P(\vec{r}) + Q(\vec{r})$ is a less distorted log-likelihood function than either $P(\vec{r})$ or $Q(\vec{r})$. Fig. 4 shows the case of SRP-PHAT maps corresponding to two separated arrays. While the maximum value of each map does not provide a good estimate of source position, the addition of both SRP-PHAT maps reinforces the relevance of the GCC peaks corresponding to the actual TDOAs, and diminishes the relevance of spurious peaks.

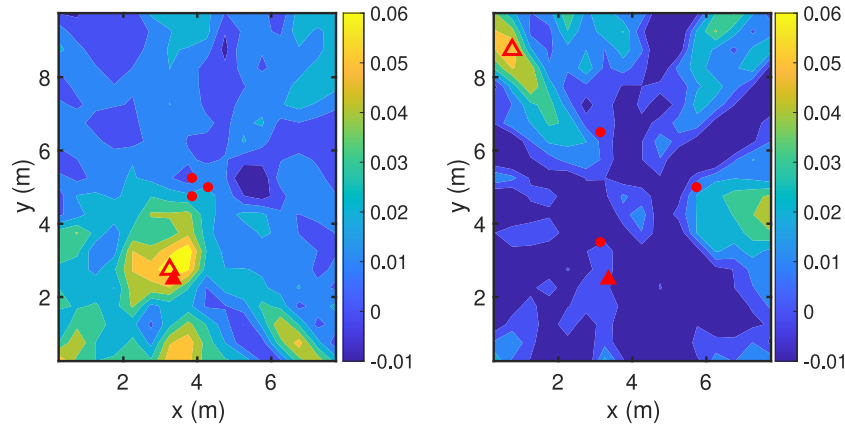


Fig. 3. SRP-PHAT maps generated for a small (left) and a large (right) microphone array. The red points indicate the simulated microphone positions, the filled triangles mark the simulated source position, and the empty triangles show the estimated sound source position. This is a 2D representation at the height of the estimated position. The simulated room has a reverberation time equal to 1.8 s.

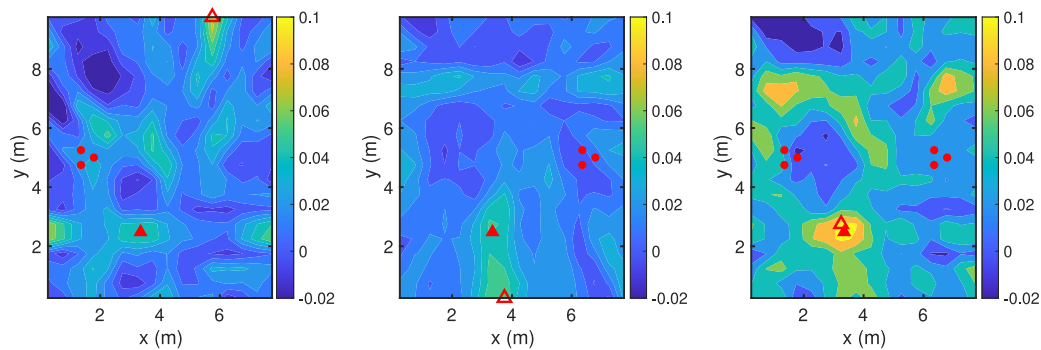


Fig. 4. SRP-PHAT maps generated for two different small microphone arrays (left and middle), and the SRP-PHAT map resulting from combining the previous ones (right). Simulation conditions are the same as in Fig. 3.

4. Simulations and results

4.1. Acoustic environment

The hypothesis stated above was evaluated by running a set of experiments similar to those reported in [44]. The acoustic environment consisted of a $8\text{ m} \times 10\text{ m} \times 4\text{ m}$ room in which wave propagation was simulated using the image method proposed by Allen and Berkley in [45], as implemented in Matlab® by Habets [46]. The absorption coefficients of the walls were adjusted using Sabine's formula (12) to yield reverberation times from 0 to 2 s in 0.2 s steps. The sound speed was assumed equal to 343 m/s.

4.2. Audio events

1000 uniformly distributed source positions were randomly selected inside the room. Four sound events were simulated at each source position, thus generating a total of 4000 simulated sound events. The sound source signals corresponded to the door slam, keys dropping, phone ringing and speech events from the database of the DCASE 2016 *Sound event detection in synthetic audio task* [47]. These events were selected because they have different shapes in their spectra [48]: noisy non-harmonic low-pass (door slam), harmonic low-pass with resonances (speech), noisy flat (keys dropping), and harmonic with flat envelope (phone). For each event, signals were randomly selected among all available for the same type of event. The signal bandwidth was assumed to be between 100 Hz and 6000 Hz, since the signal-to-noise ratio

beyond 6000 Hz is poor for most of these signals [48]. In all cases, the sound signals were digitized with 16 bits per sample at a rate of 44100 samples per second. The duration of the recordings ranged from 0.13 s to 3.34 s. Since the focus of this research is reverberation, no additional background noise was added to the utilized audio recordings.

4.3. Microphone arrays

Simulations were carried out for two different microphone arrays. Both were formed by 4 microphones placed in the corners of a regular tetrahedron whose central point was located at the center of the room (see Fig. 5, up). This number of microphones was selected because it is the minimum needed to allow the localization of the sound source in three dimensions using SRP-PHAT. The length of the tetrahedron edges was 0.5 m in one case (small array) and 3 m in the other (large array). For some experiments, two arrays were simulated simultaneously. In those cases, both arrays were placed symmetrically with respect to the center along the length of the room (see Fig. 5, down).

4.4. Signal processing

The audio signal corresponding to each event in the database was processed as follows. First of all, sound activity detection was performed, as suggested in [27]. Specifically, the audio signal was split into 50 ms frames, and the average power was calculated for each frame. The frame that produced the highest

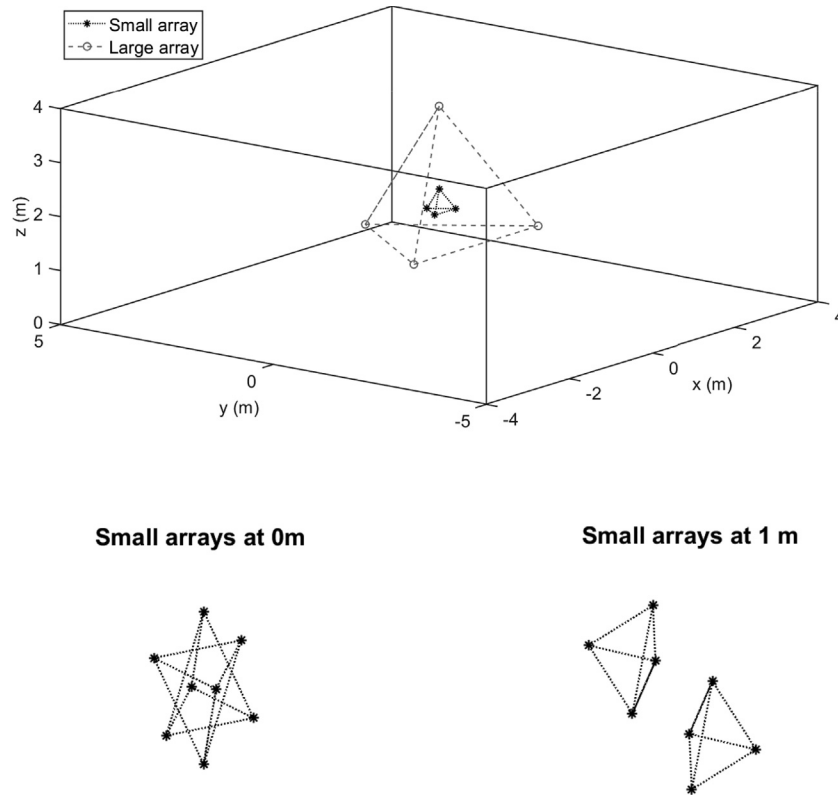


Fig. 5. Array topology and position within the simulated room (up), and relative array orientations when two arrays are simulated simultaneously (down).

average power was selected as the reference one, and all frames with an average power below 10 % of that reference were classified as silent frames. Only non-silent frames underwent subsequent processing.

Consecutive audio frames with average power above the threshold were concatenated after activity detection to generate audio segments. Sound source localization based on SRP-PHAT maps was carried out for each of these segments, with the map function $P(\vec{r})$ (18) being evaluated in the nodes of a regular grid with a sampling distance equal to 0.5 m. In the reference or standard set-up, the simulated microphone signals $m_i(t)$ corresponding to each audio event and each microphone position were used for calculating $P(\vec{r})$. The band limitation scheme described in [44] was applied to avoid spatial aliasing.

Among all the approaches proposed so far to improve localization performance in reverberant environments, and mentioned in Section 1.2, the next two were chosen and simulated according to the criteria of not requiring any *a priori* information about the acoustic environment, not involving iterative processes, and not being specifically suited to any signal type:

Cepstral prefiltering proposed in [16] for equalizing the effect of the acoustic channels $h_{s,i}(t)$ on microphone signals $m_i(t)$. Cepstral prefiltering was configured according to the values recommended in [16] for static sources: splitting audio segments into frames with duration equal to 0.6 s, using non-overlapped rectangular windows, and setting the memory parameter to 0.06. If the simulated audio segment was shorter than 0.6 s, then we used a frame length that corresponded to the half signal duration.

Averaging along several frames the GCC $R_{ik}(\tau)$ estimated for each microphone pair [13]. For averaging, each audio segment was split in 25 ms frames with a 50% overlap between consecutive frames.

4.5. Results

The Euclidean distance between the estimated and the actual source position, i.e. the localization error, for each of the 4000 simulated events was chosen as the performance indicator for each sound source localization approach. The evolution of the median localization errors with reverberation time is depicted in Fig. 6 for both the small and the large arrays in Fig. 5 (up), and for each one of the signal processing approaches mentioned before: standard, with cepstral prefiltering, and with GCC averaging. Note that the 99% confidence intervals for these median values are very small compared to the scale of the plots: less than ± 0.12 m for the small array, and less than ± 0.20 m for the large array.

The effect of introducing spatial diversity was analyzed by carrying out simulations with two small microphone arrays instead of a single one. Both arrays had the same topology, although they were oriented symmetrically (see Fig. 5, down). Localization performance against reverberation was evaluated for inter-array distances ranging from 0 m to 5 m. As before, the median localization error was used as a performance indicator for each configuration. The results are plotted in Fig. 7. The performance of a single array including all 8 microphones in the same positions as in the case of two arrays sharing the same center has also been included in the plot for reference purposes. In this case, only the standard algorithm was simulated. The 99% confidence intervals for these median values are less than ± 0.05 m in all cases.

5. Measurements and results

5.1. Measurements

In order to validate the previous simulated experiments, real recordings were performed using a set-up similar to that of the

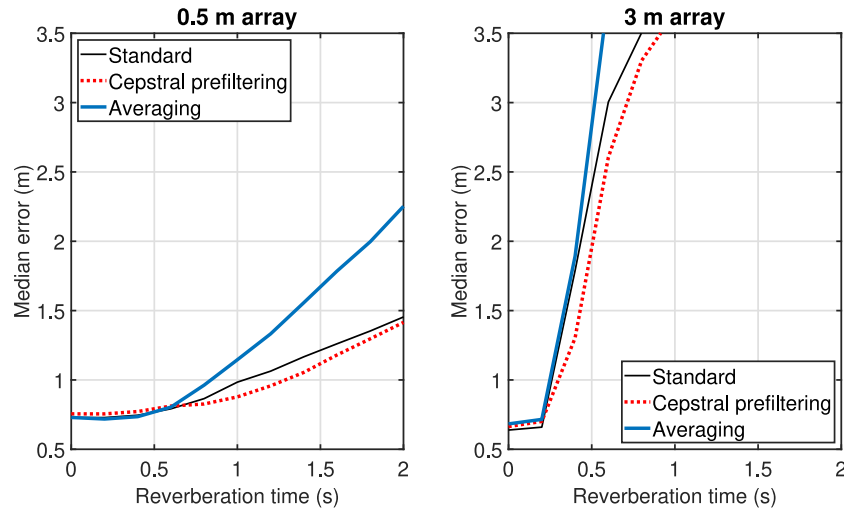


Fig. 6. Median localization error as a function of reverberation time for the small (left) and the large (right) arrays. 99% confidence intervals for the median were shorter than ± 0.12 m for the small array, and shorter than ± 0.20 m for the large array.

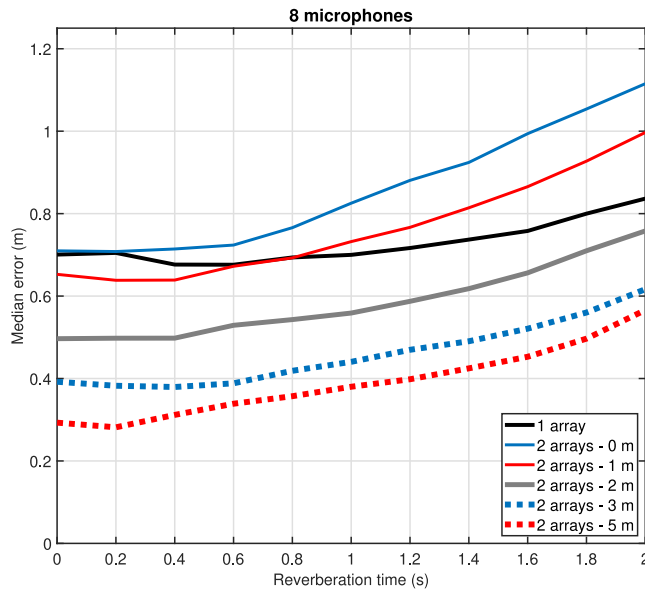


Fig. 7. Median localization error as a function of reverberation time for a small array with 8 microphones and for two small arrays of 4 microphones at several distances. 99% confidence intervals for the median were shorter than ± 0.05 m in all cases.

simulation experiments. In this case, an empty quiet office of dimensions $7.05 \text{ m} \times 5.64 \text{ m} \times 2.84 \text{ m}$ and a reverberation time of 0.7 s was selected as the recording environment. The acoustic signals were captured by 8 microphones arranged in two different microphone arrays of tetrahedral shape with side length equal to 0.5 m (small array). The acoustic signals were captured using Superlux ECM99 omni-directional condenser microphones and a Behringer UMC1820 audio interface. The selected audio events were the same as in the simulation, played using a Yamaha Msp5 speaker. In the same way as the simulations, no background noise was artificially generated.

The placement of the microphones and the speaker was performed using an OptiTrack system made up of four Flex 3 cameras. This allowed us to cover a region of $4 \text{ m} \times 4 \text{ m} \times 2 \text{ m}$ with a calibration error of 0.681 mm . In this case, the signal processing was the same as in Subsection 4.4, except for the regular grid size, which was set to 0.1 m as the evaluated space was smaller. For

each microphone array configuration, 40 source positions were distributed uniformly in the horizontal plane considering two possible heights, resulting in 80 sound source positions. Taking into account that 4 audio events were generated per each position, that made a total of 320 different recordings for each microphone array arrangement.

5.2. Results

Results plotted in the following figures show the whole distribution of localization errors for each case. These distributions are represented using box plots. The segment at the center of each box marks their median value, while the width of the notch around each median value indicates its 95% confidence interval. Lower and upper box limits correspond to the 25th and 75th percentiles, respectively. The length of the whiskers (dashed lines) is 1.5 times the inter-quartile difference, and values beyond the whiskers may be considered outliers. Fig. 8 shows the distribution of localization errors for one array (4 microphones) and the same algorithms as in Fig. 6.

Similarly as in the case of the simulated experiments, the effect of introducing spatial diversity was analyzed by using all 8 microphones arranged in a single array, and in two arrays with a growing distance between them. In this case, the scenario that considered a 5 m distance between array centers was not feasible due to the dimensions of the room. The localization performance is represented in Fig. 9.

6. Discussion

Regarding the proposed techniques for facing reverberation effects, it is shown in Fig. 6 that the cepstral prefiltering technique enhanced localization accuracy when the reverberation time was longer than 0.6 s and 0.4 s for the small and the large array, respectively. Despite that, the improvement in the median localization error was not greater than 0.1 m . On the contrary, the GCC averaging degraded significantly the performance of the algorithm for the small array scenario when the reverberation time was longer than 0.6 s . Therefore, there seems to be no advantage in splitting audio segments into short frames to perform GCC averaging afterwards. However, cepstral prefiltering provides some improvement in performance, though such improvement may not be relevant enough to justify the additional computational effort required.

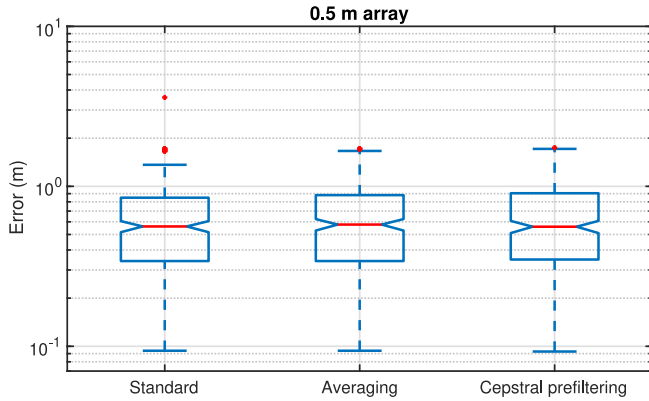


Fig. 8. Distribution of localization error for the small array using the standard SRP-PHAT, the averaging and the cepstral prefiltering techniques in a real room with a reverberation time of 0.7 s.

Localization based on real measurements confirmed this trend (Fig. 9). Note that the RT of the room (0.7 s) corresponds to the point in Fig. 6 where performances begin to differ, but they are still similar. Fig. 8 shows that the median localization errors for all three methods do not differ significantly, although cepstral prefiltering provides a slightly lower value. In addition, the magnitude of localization errors for both measurements and simulations is similar, which suggests the validity of results obtained after simulation.

Incidentally, both plots in Fig. 6 show that the degradation of localization performance in reverberant environments mainly happens for reverberation times over 0.4–0.6 s. This suggests the limited value of studies in which measured or simulated reverberation does not go beyond this limit, as pointed out in Section 1.3.

Regarding the size of the array, ergo the inter-microphone distance, it is shown that the large array performed slightly better than the small one for short reverberation times. In fact, for anechoic conditions, the median error of the standard algorithm was 0.64 m for the large array and 0.73 m for the small one. This is consistent with the plot in Fig. 2 indicating that larger microphone distances imply improved angular resolutions, thus lower localization errors. However, for longer reverberation times, the lower correlation between acoustic channels ρ_{ik} (14) in the large array has a negative impact on localization performance, as indicated in (17), which completely masks the improved angular resolution. Consequently, small arrays seem to provide performances more robust to reverberation, even if they have poorer angular resolution.

From another point of view, if we consider the length of the diagonal of a cubic grid ($0.5 \cdot \sqrt{3} \approx 0.87$ m), when errors are below this value, it means that the algorithm is estimating the source position with an error that is less than the largest distance between adjacent points in the SRP map grid. For the small array, this happens in the majority of cases for reverberation times up to 0.8 s approximately, while the large array yields larger errors for reverberation times larger than 0.2 s.

Given the limited improvement achieved with strategies such as cepstral prefiltering, and considering the reasoning exposed in Section 3, the potential impact of spatial diversity was assessed by analyzing the performance of combining SRP-PHAT maps from two different arrays, so referred as $P(\vec{r})$ and $Q(\vec{r})$ in Section 3. The results plotted in Fig. 7 show that using two arrays instead of one provides a relevant improvement in performance with respect to the single array case.

At first sight, one may reasonably argue that the main improvement comes from the fact of using 8 microphones instead of 4. In fact, the graph labeled as “8 mics” in Fig. 7 shows the performance of an 8 microphone array that has the topology shown on the left of Fig. 5 (down). This performance is significantly better than that of a single 4 microphone array (Fig. 6). When the 8 microphones are organized into two arrays, separated 0 m, two different SRP maps are computed, one per array, and later summed to produce the resulting map. In this last case, there is less information about the true contribution of the sound source for the SRP map estimation as the number of microphones is 4, and the GCCs for some microphone pairs are not considered. Consequently, the performance worsens when the microphones are separated into two arrays.

However, as the distance between microphone arrays increases, the localization error decreases for all simulated reverberation times. When the distance between arrays was 5 m, the reduction of the median error was between 0.4 m and 0.5 m approximately compared with the case with no separation between arrays. In this case, advantage is taken from a short distance between microphones in each array, and a large distance between microphone arrays. Then, the calculated GCCs of each array avoid the arising of secondary peaks, and the distortion between both SRP maps is more independent proving the analysis performed in Section 3. Note that for the lowest relevant frequency of the simulated events (100 Hz, see Subsection 4.2), the value of kd for 5 m is approximately 9.16. For values above that one, ρ_{ik} in (14) does not reach values over 0.13, which implies that only some limited reduction in its value can be expected by increasing the distance between arrays.

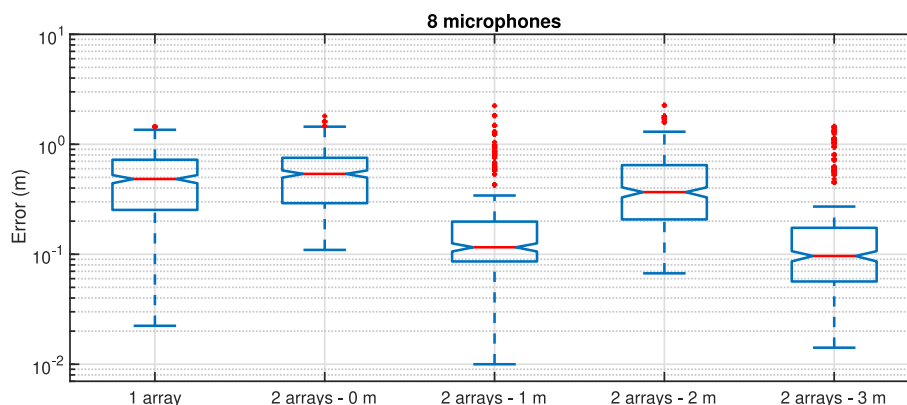


Fig. 9. Distribution of localization error for a small array with 8 microphones and for two small arrays of 4 microphones at several distances in a real room with a reverberation time of 0.7 s.

Similar results can be observed for the real experiments (view Fig. 9). On the one hand, there is a little worsening of results when the 8 microphones are arranged into two arrays placed around the same point, instead of a single array. The magnitude of this worsening is approximately 0.1 m in the median error, and the difference between both cases is in the limit of statistical significance. However, when the distance between microphone arrays increases the median error is significantly reduced. Specifically, there is a reduction in the median error of 0.44 m between the arrays separated 3 m and those centered around the same point. The magnitude of this improvement is in the same range as that plotted in Fig. 7. When the arrays are separated 1 m, some significant improvement is obtained, but lower than when separation is 3 m. The only atypical behavior is the case of the arrays separated 2 m, which produces worse results than when separation is 1 m, although a significant improvement is obtained with respect to the case of no spatial diversity (0 m separation). We attribute this atypical behavior to the specific acoustic characteristics of the room.

7. Conclusions

Several approaches have been proposed so far for reducing the negative impact of reverberation on the performance of sound source localization systems inside a room. However, many of them have been tested in reverberant environments with short reverberation times, typically below 0.6 s, which are not representative of real acoustic environments.

An alternative approach to increase the robustness against reverberation is proposed using two microphone arrays and exploiting the spatial characteristics of the acoustic channels. A theoretical analysis has shown that reverberation affects more large microphone arrays than smaller ones, and combining the information obtained from diverse arrays may be advantageous. The performed simulations using 4000 audio events with different lengths and spectral shapes and two arrays with four microphones have confirmed the achieved theoretical conclusions showing that smaller arrays significantly outperform large ones for reverberation times above 0.4 s. Although the median error of source localization shows more robustness when the number of microphones of a single array is increased from four to eight in the simulations, the most relevant results show that separating the two arrays is more advantageous than simply adding more microphones to a single array. Localization results obtained after real measurements confirm the same conclusions.

This study shows that combining information from several arrays, thus taking advantage of spatial diversity, provides more robust sound source localization estimates in reverberant conditions; this approach being easier to apply than increasing the complexity of the signal processing algorithms aimed at reducing the impact of reverberation on the audio signals. Such a combination of information can be implemented through the addition of the SRP maps corresponding to all microphone arrays. The size of each array should be chosen so that the correlation coefficient among the acoustic channels is as close to one as possible, while the distance between the arrays should be decided so that the same coefficient is as low as possible.

CRediT authorship contribution statement

Guillermo García-Barrios: Methodology, Investigation, Software, Data curation. **Eduardo Latorre Iglesias:** Software, Validation, Writing - review & editing. **Juana M. Gutiérrez-Arriola:** Supervision, Investigation. **Rubén Fraile:** Conceptualization, Formal analysis, Writing - original draft. **Nicolás Sáenz-Lechón:** Writ-

ing - review & editing. **Víctor José Osmá-Ruiz:** Writing - review & editing.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Universidad Politécnica de Madrid through its Programa Propio de I+D+I, specifically the Pre-doctoral Call. The authors gratefully acknowledge the Universidad Politécnica de Madrid for providing computing resources on Magerit Supercomputer.

References

- [1] Cobos M, Antonacci F, Alexandridis A, Mouchtaris A, Lee B. A survey of sound source localization methods in wireless acoustic sensor networks. *Wireless Commun Mobile Comput* 2017;2017. <https://doi.org/10.1155/2017/3956282>.
- [2] Virtanen T, Plumbley MD, Ellis D. *Computational Analysis of Sound Scenes and Events*. Springer 2018;2018. <https://doi.org/10.1007/978-3-319-63450-0>.
- [3] Crocco M, Cristani M, Trucco A, Murino V. Audio surveillance: A systematic review. *ACM Comput Surv* 2016;48(4):52:1–52:46. <https://doi.org/10.1145/2871183>.
- [4] Chandrakala S, Jayalakshmi SL. Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies. *ACM Comput Surv* 2019;52(3):1–34. <https://doi.org/10.1145/3322240>.
- [5] T. Gustafsson, B.D. Rao, M. Trivedi, Source localization in reverberant environments: Modeling and statistical analysis, *IEEE Trans. Speech Audio Process.* 11 (6) (2003) 791–803 (2003). doi:10.1109/TSA.2003.818027.
- [6] Velasco J, Martín-Arguedas CJ, Macías-Guarasa J, Pizarro D, Mazo M. Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios. *Signal Process* 2016;119(2016):209–28.
- [7] P.R. Roth, Effective measurements using digital signal analysis, *IEEE Spectr.* 8 (4) (1971) 62–70 (1971). doi:10.1109/MSPEC.1971.5218046.
- [8] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, *IEEE Trans. Acoust., Speech, Signal Process.* 24 (4) (1976) 320–327 (1976). doi:10.1109/TASSP.1976.1162830.
- [9] J.H. DiBiase, H.F. Silverman, M.S. Brandstein, Robust localization in reverberant rooms, in: M. Brandstein, D. Ward (Eds.), *Microphone Arrays*, Springer, 2001, pp. 157–180 (2001). doi:10.1007/978-3-662-04619-7_8.
- [10] S. Bédard, B. Champagne, A. Stéphenne, Effects of room reverberation on time-delay estimation performance, in: *IEEE Internat. Conf. Acoust. Speech, & Signal Process.*, Vol. 2, 1994, pp. 261–264 (1994). doi:10.1109/ICASSP.1994.389670.
- [11] B. Champagne, S. Bédard, A. Stéphenne, Performance of time-delay estimation in the presence of room reverberation, *IEEE Trans. Speech Audio Process.* 4 (2) (1996) 148–152 (1996). doi:10.1109/89.486067.
- [12] J.M. Pérez-Lorenzo, R. Viciano-Abad, P. Reche-Lopez, F. Rivas, J. Escolano, Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments, *Applied Acoust.* 73 (8) (2012) 698–712 (2012). doi:10.1016/j.apacoust.2012.02.002.
- [13] J.H. DiBiase, A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays, Ph.D. thesis, Brown University (2000).
- [14] Dmochowski J, Benesty J, Affes S. Direction of arrival estimation using the parameterized spatial correlation matrix. *IEEE Trans Audio, Speech, Language Process* 2007;15(4):1327–39.
- [15] C. Zhang, D. Florêncio, Z. Zhang, Why does PHAT work well in low noise, reverberative environments?, in: *IEEE Internat. Conf. Acoust. Speech, & Signal Process.*, 2008, pp. 2565–2568 (2008). doi:10.1109/ICASSP.2008.4518172.
- [16] A. Stéphenne, B. Champagne, A new cepstral prefiltering technique for estimating time delay under reverberant conditions, *Signal Process.* 59 (3) (1997) 253–266 (1997). doi:10.1016/S0165-1684(97)00051-0.
- [17] R. Parisi, F. Camoes, M. Scarpiniti, A. Uncini, Cepstrum prefiltering for binaural source localization in reverberant environments, *IEEE Signal Process. Lett.* 19 (2) (2012) 99–102 (2012). doi:10.1109/LSP.2011.2180376.
- [18] R. Parisi, R. Gazzetta, E.D. Di Claudio, Prefiltering approaches for time delay estimation in reverberant environments, in: *IEEE Internat. Conf. Acoust. Speech, & Signal Process.*, Vol. 3, 2002, pp. III/2997–III/3000 (2002). doi:10.1109/ICASSP.2002.5745279.

- [19] J.M. Yang, C.H. Lee, S. Kim, H.G. Kang, A robust time difference of arrival estimator in reverberant environments, in: *Europ. Signal Process. Conf.*, 2009, pp. 864–868 (2009).
- [20] Guo Y, Wang X, Wu C, Fu Q, Ma N, Brown G. A robust dual-microphone speech source localization algorithm for reverberant environments. *Interspeech* 2016;2016:3354–8. <https://doi.org/10.21437/Interspeech.2016-1063>.
- [21] N. Antonello, T. Van Waterschoot, M. Moonen, P.A. Naylor, Source localization and signal reconstruction in a reverberant field using the FDTD method, in: *Europ. Signal Process. Conf.*, 2014, pp. 301–305 (2014).
- [22] J.R. Jensen, J. Nielsen, R. Heusdens, M. Christensen, DOA estimation of audio sources in reverberant environments, in: *IEEE Internat. Conf. Acoust. Speech, & Signal Process.*, 2016, pp. 176–180 (2016).
- [23] C. Zhang, Z. Zhang, D. Florêncio, Maximum likelihood sound source localization for multiple directional microphones, in: *IEEE Internat. Conf. Acoust. Speech, & Signal Process.*, Vol. 1, 2007, pp. 1/125–1/128 (2007). doi:10.1109/ICASSP.2007.366632.
- [24] B. Lee, T. Kalker, Maximum a posteriori estimation of time delay, in: *2007 2nd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2007, pp. 285–288 (2007). doi:10.1109/CAMSP.2007.4498021.
- [25] Y. Rui, D. Florencio, Time delay estimation in the presence of correlated noise and reverberation, in: *IEEE Internat. Conf. Acoust. Speech, & Signal Process.*, Vol. 2, 2004, pp. II/133–II/136 (2004). doi:10.1109/ICASSP.2004.1326212.
- [26] A. Ramamurthy, H. Unnikrishnan, K.D. Donohue, Experimental performance analysis of sound source detection with SRP PHAT-, in: *IEEE Southeastcon* 2009, 2009, pp. 422–427 (2009). doi:10.1109/SECON.2009.5174117.
- [27] A. Cirillo, R. Parisi, A. Uncini, Sound mapping in reverberant rooms by a robust direct method, in: *IEEE Internat. Conf. Acoust. Speech, & Signal Process.*, 2008, pp. 285–288 (2008).
- [28] C.M. Zannini, A. Cirillo, R. Parisi, A. Uncini, Improved TDOA disambiguation techniques for sound source localization in reverberant environments, in: *IEEE Internat. Symp. Circuits & Syst.*, 2010, pp. 2666–2669 (2010).
- [29] H. Zhu, Z. Li, Q. Cheng, Sound source localization through optimal peak association in reverberant environments, in: *2017 20th International Conference on Information Fusion (Fusion)*, 2017, pp. 1–6 (2017). doi:10.23919/ICIF.2017.8009685.
- [30] Lee R, Kang MS, Kim BH, Park KH, Lee SQ, Park HM. Sound source localization based on GCC-PHAT with diffuseness mask in noisy and reverberant environments. *IEEE Access* 2020;8(2020):7373–82. <https://doi.org/10.1109/ACCESS.2019.2963768>.
- [31] E.E. Jan, J. Flanagan, Sound source localization in reverberant environments using an outlier elimination algorithm, in: *Internat. Conf. Spoken Lang. Process.*, Vol. 3, 1996, pp. 1321–1324 (1996). doi:10.1109/ICSLP.1996.607856.
- [32] M. Arabaci, N. Strickland, Direction of arrival estimation in reverberant rooms using a resource-constrained wireless sensor network, in: *IEEE Internat. Conf. Pervasive Services*, 2007, pp. 29–38 (2007).
- [33] Castellini P, Sassaroli A. Acoustic source localization in a reverberant environment by average beamforming. *Mech Syst & Signal Process* 2010;24(3):796–808.
- [34] R. Parisi, A. Cirillo, M. Panella, A. Uncini, Source localization in reverberant environments by consistent peak selection, in: *IEEE Internat. Conf. Acoust. Speech, & Signal Process.*, Vol. 1, 2007, pp. III/37–III/40 (2007).
- [35] X. Wan, Z. Wu, Improved speech source localization in reverberant environments based on correlation dimension, in: *Internat. Conf. Wireless Commun. & Signal Process.*, 2009, pp. 1–4 (2009). doi:10.1109/WCSP.2009.5371584.
- [36] Comanducci L, Borra F, Bestagini P, Antonacci F, Tubaro S, Sarti A. Source localization using distributed microphones in reverberant environments based on deep learning and ray space transform, *IEEE/ACM Trans. Audio, Speech, Language Process* 2020;28(2020):2238–51.
- [37] M. Sewtz, T. Bodenmüller, R. Triebel, Robust MUSIC-based sound source localization in reverberant and echoic environments, in: *IEEE/RSJ Internat. Conf. Intell. Robots & Syst.*, 2020, pp. 2474–2480 (2020).
- [38] J. Cowan, Building acoustics, in: T. Rossing (Ed.), *Handbook of Acoustics*, Springer, 2007, Ch. 11, pp. 387–425 (2007). doi:10.1007/978-0-387-30425-0_11.
- [39] Y. Yu, H.F. Silverman, An improved TDOA-based location estimation algorithm for large aperture microphone arrays, in: *IEEE Internat. Conf. Acoust. Speech, & Signal Process.*, Vol. 4, 2004, pp. IV/77–IV/80 (2004). doi:10.1109/ICASSP.2004.1326767.
- [40] D.B. Ward, On the performance of acoustic crosstalk cancellation in a reverberant environment, *J. Acoust. Soc. Amer.* 110 (2) (2001) 1195–1198 (2001). doi:10.1121/1.1386635.
- [41] Kuttruff H. *Room Acoustics*. Spon Press 2000;2000. <https://doi.org/10.1201/9781315372150>.
- [42] R.K. Cook, R.V. Waterhouse, R.D. Berendt, S. Edelman, M.C. Thompson Jr, Measurement of correlation coefficients in reverberant sound fields, *J. Acoust. Soc. Amer.* 27 (6) (1955) 1072–1077 (1955). doi:10.1109/89.486067.
- [43] Oppenheim AV, Schaffer RW, Buck JR. *Discrete-Time Signal Processing*. Prentice Hall; 1999. <https://doi.org/10.5555/1795494>.
- [44] García-Barrios G, Gutiérrez-Arriola JM, Sáenz-Lechón N, Osma-Ruiz VJ, Fraile R. Analytical model for the relation between signal bandwidth and spatial resolution in steered-response power phase transform (SRP-PHAT) maps. *IEEE Access* 2021;9(2021):121549–60. <https://doi.org/10.1109/ACCESS.2021.3105650>.
- [45] J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics, *J. Acoust. Soc. Amer.* 65 (4) (1979) 943–950 (1979). doi:10.1121/1.382599.
- [46] Habets EAP. *Room impulse response generator Tech. rep.*. Technische Universiteit Eindhoven; 2006.
- [47] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M.D. Plumbley, Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26 (2) (2018) 379–393 (Feb 2018). doi:10.1109/TASLP.2017.2778423.
- [48] J.M. Gutiérrez-Arriola, R. Fraile, A. Camacho, T. Durand, J.L. Jarrín, S.R. Mendoza, Synthetic sound event detection based on MFCC, in: *Proc. of DCASE 2016 Workshop*, 2016, pp. 30–34 (2016).