IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received July 21, 2021, accepted July 27, 2021, date of publication August 18, 2021, date of current version September 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3105650

Analytical Model for the Relation Between Signal Bandwidth and Spatial Resolution in Steered-Response Power Phase Transform (SRP-PHAT) Maps

GUILLERMO GARCÍA-BARRIOS, JUANA M. GUTIÉRREZ-ARRIOLA, (Member, IEEE), NICOLÁS SÁENZ-LECHÓN, VÍCTOR JOSÉ OSMA-RUIZ, AND RUBÉN FRAILE
CITSEM, Universidad Politécnica de Madrid—Campus Sur, 28031 Madrid, Spain

Corresponding author: Rubén Fraile (r.fraile@upm.es)

This work was supported in part by the Universidad Politécnica de Madrid through its Programa Propio de I + D + I, specifically the Mentorship and Predoctoral Calls.

ABSTRACT An analysis of the relationship between the bandwidth of acoustic signals and the required resolution of steered-response power phase transform (SRP-PHAT) maps used for sound source localization is presented. This relationship does not rely on the far-field assumption, nor does it depend on any specific array topology. The proposed analysis considers the computation of a SRP map as a process of sampling a set of generalized cross-correlation (GCC) functions, each one corresponding to a different microphone pair. From this approach, we derive a rule that relates GCC bandwidth with inter-microphone distance, resolution of the SRP map, and the potential position of the sound source relative to the array position. This rule is a sufficient condition for an aliasing-free calculation of the specified SRP-PHAT map. Simulation results show that limiting the bandwidth of the GCC according to such rule leads to significant reductions in sound source localization errors when sources are not in the immediate vicinity of the microphone array. These error reductions are more relevant for coarser resolutions of the SRP map, and they happen in both anechoic and reverberant environments.

INDEX TERMS Acoustic signal processing, microphone arrays, signal sampling, sound source localization, steered-response power maps.

I. INTRODUCTION

Sound source localization based on steered-response power (SRP) maps computed using the generalized cross-correlation (GCC) function with phase transform (PHAT), i.e. SRP-PHAT, has been reported to perform robustly against noise and, especially, reverberation [1], [2]. The PHAT applied to the GCC function has the effect of narrowing its maxima, hence allowing a more precise identification of the time difference of arrival (TDOA) between microphones [3]. However, this increased precision can only be exploited by correspondingly reducing the spatial resolution[1] of SRP maps, which turns out to be one of the main drawbacks of sound source localization based on SRP-PHAT [1] since it implies higher computational costs.

Therefore, implementing a sound source localization system based on SRP involves finding a balance between computational cost and precision. To present, this challenge has been approached in several ways. One of them has consisted in performing calculations at several resolution levels, from coarsest to finest, and limiting the extent of the map each time the resolution is decreased. This hierarchical search can be implemented, for instance, by defining rectangular and regular grids whose cells are iteratively decomposed into finer grids [4]–[7]. Instead of conducting the hierarchical search using regular grids, some researchers have proposed grouping regions by TDOA [8], or decreasing resolution mainly in regions where the SRP function is expected to vary most abruptly [9]. Other approaches try to avoid iterative processes, thus keeping resolution fixed,

The associate editor coordinating the review of this manuscript and approving it for publication was Huaqing Li.

[1] Herein, resolution is defined as the distance between contiguous points in the map grid. Therefore, the lowest resolutions correspond to the finest map grids, and the highest resolutions are associated with the coarsest grids.

while computational cost is maintained at an affordable level by restricting the search space only to regions where the sound source is expected to be, according to some *a priori* information [10], [11].

When coarse spatial resolutions are used for generating SRP maps based on spiky functions such as the GCC-PHAT, two risks are taken. On the one hand, the narrow peak corresponding to the global maximum of the GCC may not be adequately sampled; on the other hand, spurious local maxima of the GCC may be reflected in the SRP maps. These two effects can distort localization estimates, an effect that is more likely to happen at the first stages of hierarchical searches, thus leading to severe errors in the overall results. In order to avoid such errors several approaches have been proposed so far, such as stochastic region contraction (SRC), which involves performing a stochastic search of the highest peaks in the SRP map [12] before decreasing map resolution and reducing map extent; calculating the integral of the GCC-PHAT along an interval of time delay values defined by the position of each grid point and the spatial resolution of the map [13]; or designing the map grid considering the specific geometry of the microphone array [14]. Alternative approaches based on deep learning have also been proposed to reduce the number of local maxima in the SRP map by either post-processing the GCC [15], or the map itself [16].

Qualitatively, the width of peaks in the GCC are known to be related to the spectral content of the audio signal. Thus, signal spectrum or, more specifically, signal bandwidth is not independent of the spatial map resolution required to obtain good localization estimates. This relation can be used to design the afore-mentioned hierarchical search considering the bandwidth of the specific signal being processed [4], and it is also implicit in proposals such as integrating the GCC-PHAT [13] (integration is equivalent to low-pass filtering) or applying multi-band analysis to reduce the effects of spatial aliasing [17].

The peak narrowing in the GCC becomes particularly relevant for large inter-microphone distances when the sound source is likely to be near the microphone array, or even inside it. For this reason, distributed microphone arrays potentially allow better precision in source localization [18], but at the cost of higher computational load, as reasoned before. For these particular cases, a quantitative rule relating signal bandwidth and inter-microphone distances has been proposed in order to avoid the appearance of spurious secondary lobes in the beam pattern of the array. Specifically, it is commonly assumed that the acoustic wavelength for far-field measurements using microphone arrays should be larger than twice the inter-microphone distance (e.g. [19]). However, such rule does not consider map resolution.

Considering all the previous questions as a whole, it is straightforward to conclude that there is a relationship between inter-microphone distance or array size, signal bandwidth, and the spatial resolution required to avoid under-sampling the GCC. In this paper, we present a rule that quantifies this relation. The analysis leading to this rule does not rely on the far-field assumption and it is not dependent on any specific array topology. The rule can be applied to hierarchical searches at every resolution level to avoid the emergence of spurious maxima at the corresponding SRP maps, hence achieving lower errors in sound source localization estimates. Furthermore, it provides an alternative interpretation, based on basic signal processing theory, of algorithms involving GCC integration [13], design of map grids with reduced resolution in certain areas [14], or adjustment of grid resolution as a function of signal bandwidth [4].

The adopted approach considers the computation of a SRP map as a process of sampling a set of GCC functions, each one corresponding to a different microphone pair. This theoretical analysis is presented in sections II and III, while its implications for SRP map calculations are discussed in section IV. Section V shows how to incorporate the previous theoretical results into the process of calculating SRP maps by limiting the bandwidth of the GCC specifically for each point in the map. Results obtained using this approach presented in section VI indicate that it can provide significant error reductions in the estimation of source positions. The conditions in which such improvements can be achieved are discussed in section VII.

II. PROBLEM STATEMENT

Sound source localization consists in estimating the position $\vec{r}_s$ of an acoustic source with respect to a certain coordinate reference, given the corresponding acoustic signals captured at a set of K microphones whose positions are known. When a SRP algorithm is used, the source position is estimated as [2]:

$$\vec{r}_s \approx \arg\max P\left(\vec{r}\right), \tag{1}$$

where $P\left(\vec{r}\right)$ is the value of the SRP map at position $\vec{r}$. This can be calculated as:

$$\begin{aligned} P\left(\vec{r}\right) &= 2\pi \sum_{k=1}^{K}\sum_{l=1}^{K} R_{kl}\left(\tau_l\left(\vec{r}\right) - \tau_k\left(\vec{r}\right)\right) \\ &= 2\pi \sum_{k=1}^{K}\sum_{l=1}^{K} R_{kl}\left(\tau_{kl}\left(\vec{r}\right)\right), \end{aligned} \tag{2}$$

where $\tau_k\left(\vec{r}\right)$ is the propagation delay between position $\vec{r}$ and the position of the k^{th} microphone, and $R_{kl}\left(\tau_{kl}\left(\vec{r}\right)\right)$ is the GCC function between the sound signals captured at microphones k and l, respectively $s_k\left(t\right)$ and $s_l\left(t\right)$, evaluated at time lag $\tau_{kl}\left(\vec{r}\right)$. When the PHAT weighting is used, the GCC can be calculated as:

$$R_{kl}\left(\tau\right) = \int_{-\infty}^{\infty} \frac{S_k\left(\omega\right) S_l^*\left(\omega\right) \cdot \mathrm{e}^{j\omega\tau}}{2\pi\left|S_k\left(\omega\right) S_l^*\left(\omega\right)\right|} \mathrm{d}\omega, \tag{3}$$

being $S_k\left(\omega\right)$ the Fourier transform of $s_k\left(t\right)$, and j the imaginary unit. This calculation is problematic when the integral spans over frequencies for which the signal-to-noise ratio (SNR) corresponding to $s_k\left(t\right)$ and $s_l\left(t\right)$ is low [3], due to the division in (3). In the case of passband signals, whose SNR is high only within a certain frequency interval

$\omega_{\min} \leq \omega \leq \omega_{\max}$, this can be solved by limiting the integration to the same interval:

$$R_{kl}\left(\tau\right) = \int_{\omega_{\min} \leq |\omega| \leq \omega_{\max}} \frac{S_k\left(\omega\right) S_l^*\left(\omega\right) \cdot \mathrm{e}^{j\omega\tau}}{2\pi \left|S_k\left(\omega\right) S_l^*\left(\omega\right)\right|} \mathrm{d}\omega. \tag{4}$$

Therefore, $P\left(\vec{r}\right)$ is a non-linear function of a three-dimensional variable $\vec{r}$. Its maximization is commonly performed by evaluating it on a set of predefined points (usually a grid) in the area of interest, and selecting the point yielding the highest value [2]. Considering (2), this approach can be seen as a sampling of $P\left(\vec{r}\right)$ in which each sample is obtained by combining certain samples of the GCC functions $R_{kl}\left(\tau\right)$. When jumping from one of these predefined points $\vec{r}$ to a contiguous one in the grid $\vec{r} + \Delta\vec{r}$, the time lags at which the GCC functions need to be evaluated change from $\tau_{kl}\left(\vec{r}\right)$ to $\tau_{kl}\left(\vec{r} + \Delta\vec{r}\right)$, hence missing all intermediate values of the GCC functions. In cases for which $\left|\tau_{kl}\left(\vec{r}\right) - \tau_{kl}\left(\vec{r} + \Delta\vec{r}\right)\right|$ is large enough, some narrow peaks of the GCC may be missed, leading to localization errors like those illustrated in [20, Figs. 7 and 8].

According to this approach, the calculation of SRP maps can be understood as a compound sampling process of the GCC functions corresponding to all microphone pairs. The research question faced here is whether some basic sampling theory can be applied to model this process and derive an equation that relates GCC bandwith to grid resolution, and whether such a model could be useful for improving the localization performance of SRP-PHAT algorithms by modifying the calculation of GCC functions instead of making use of GCC integration at a later stage as in [13], or designing point grids specific for each scenario, like in [14].

III. GEOMETRICAL ANALYSIS

Let's consider the simple case of two microphones k and l and one point $\vec{r} = (x, y, z)$ for which $P\left(\vec{r}\right)$ needs to be evaluated (Fig. 1). Without loosing generality, let's further suppose that both microphones are symmetrically arranged around the origin of coordinates, so microphone k is placed at position $\vec{r}_{\mathrm{m}} = (x_{\mathrm{m}}, y_{\mathrm{m}}, z_{\mathrm{m}})$, and the position of microphone l is $-\vec{r}_{\mathrm{m}}$. Given the value of the sound velocity c, the TDOA between microphones k and l associated to point $\vec{r}$ is:

$$\tau_{kl}\left(\vec{r}\right) = \frac{1}{c}\left(r_l - r_k\right) = \frac{1}{c}\left(\|\vec{r} + \vec{r}_{\mathrm{m}}\| - \|\vec{r} - \vec{r}_{\mathrm{m}}\|\right), \tag{5}$$

where $\|\cdot\|$ is the Euclidean norm, and $r_k = \|\vec{r}_k\|$. We are interested in studying the sampling process of $R_{kl}\left(\tau\right)$, so we analyze how the sampling time τ_{kl} changes as the potential source position changes from one grid point to a contiguous one. Specifically, according to a first-order Taylor approximation, given $\tau_{kl}\left(\vec{r}\right)$, the TDOA at a contiguous point $\vec{r} + \Delta\vec{r}$ can be approximated as [21, chap.11]:

$$\tau_{kl}\left(\vec{r} + \Delta\vec{r}\right) \approx \tau_{kl}\left(\vec{r}\right) + \nabla\tau_{kl}\left(\vec{r}\right) \cdot \Delta\vec{r}, \tag{6}$$

where $\nabla\tau_{kl}\left(\vec{r}\right)$ is the gradient of the TDOA and $\cdot$ is the dot product. The interval between adjacent samples of $R_{kl}\left(\tau\right)$ can then be estimated as:

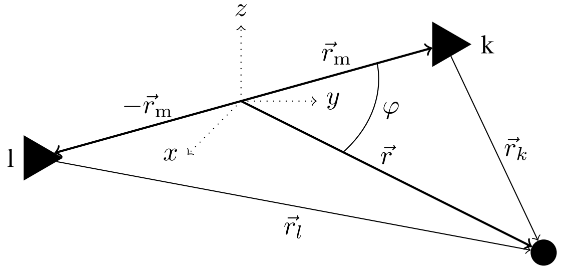


FIGURE 1. Simplified scenario comprising two microphones (triangles) and one position (circle).

$$\Delta\tau_{kl} = \left|\tau_{kl}\left(\vec{r} + \Delta\vec{r}\right) - \tau_{kl}\left(\vec{r}\right)\right| \approx \left|\nabla\tau_{kl}\left(\vec{r}\right) \cdot \Delta\vec{r}\right|. \tag{7}$$

According to the properties of the dot product:

$$\Delta\tau_{kl} \approx \left|\nabla\tau_{kl}\left(\vec{r}\right) \cdot \Delta\vec{r}\right| \leq \left\|\nabla\tau_{kl}\left(\vec{r}\right)\right\| \cdot \Delta r, \tag{8}$$

where $\Delta r = \|\Delta\vec{r}\|$. Therefore, the maximum sampling interval of $R_{kl}\left(\tau\right)$ is bounded by the product between the resolution of the SRP map and the modulus of the gradient of the TDOA. The resolution of the SRP map is defined as the distance between any point in the map grid and its closest surrounding points. It is mathematically represented by Δr, previously defined as the distance between contiguous points in the grid. This resolution is constant for regular grids, and position-dependent for irregular grids. In what follows, no assumption is made with respect to this issue. That is, the subsequent formulation is valid for both regular and irregular grids.

Considering (5), the gradient of the TDOA can be calculated as:

$$\nabla\tau_{kl}\left(\vec{r}\right) = \left(\frac{\partial\tau_{kl}}{\partial x}, \frac{\partial\tau_{kl}}{\partial y}, \frac{\partial\tau_{kl}}{\partial z}\right), \tag{9}$$

with

$$\begin{aligned}
\frac{\partial\tau_{kl}}{\partial x} &= \frac{1}{c}\left(\frac{x + x_{\mathrm{m}}}{\|\vec{r} + \vec{r}_{\mathrm{m}}\|} - \frac{x - x_{\mathrm{m}}}{\|\vec{r} - \vec{r}_{\mathrm{m}}\|}\right), \\
\frac{\partial\tau_{kl}}{\partial y} &= \frac{1}{c}\left(\frac{y + y_{\mathrm{m}}}{\|\vec{r} + \vec{r}_{\mathrm{m}}\|} - \frac{y - y_{\mathrm{m}}}{\|\vec{r} - \vec{r}_{\mathrm{m}}\|}\right), \\
\frac{\partial\tau_{kl}}{\partial z} &= \frac{1}{c}\left(\frac{z + z_{\mathrm{m}}}{\|\vec{r} + \vec{r}_{\mathrm{m}}\|} - \frac{z - z_{\mathrm{m}}}{\|\vec{r} - \vec{r}_{\mathrm{m}}\|}\right).
\end{aligned}$$

And the square of its Euclidean norm is:

$$\begin{aligned}
\left\|\nabla\tau_{kl}\left(\vec{r}\right)\right\|^2 &= \left(\frac{\partial\tau_{kl}}{\partial x}\right)^2 + \left(\frac{\partial\tau_{kl}}{\partial y}\right)^2 + \left(\frac{\partial\tau_{kl}}{\partial z}\right)^2 \\
&= \frac{1}{c^2}\left(\frac{\|\vec{r} + \vec{r}_{\mathrm{m}}\|^2}{\|\vec{r} + \vec{r}_{\mathrm{m}}\|^2} + \frac{\|\vec{r} - \vec{r}_{\mathrm{m}}\|^2}{\|\vec{r} - \vec{r}_{\mathrm{m}}\|^2}\right. \\
&\quad \left. - 2\frac{x^2 - x_{\mathrm{m}}^2 + y^2 - y_{\mathrm{m}}^2 + z^2 - z_{\mathrm{m}}^2}{\|\vec{r} + \vec{r}_{\mathrm{m}}\| \cdot \|\vec{r} - \vec{r}_{\mathrm{m}}\|}\right) \\
&= \frac{2}{c^2}\left(1 - \frac{r^2 - r_{\mathrm{m}}^2}{\|\vec{r} + \vec{r}_{\mathrm{m}}\| \cdot \|\vec{r} - \vec{r}_{\mathrm{m}}\|}\right),
\end{aligned} \tag{10}$$

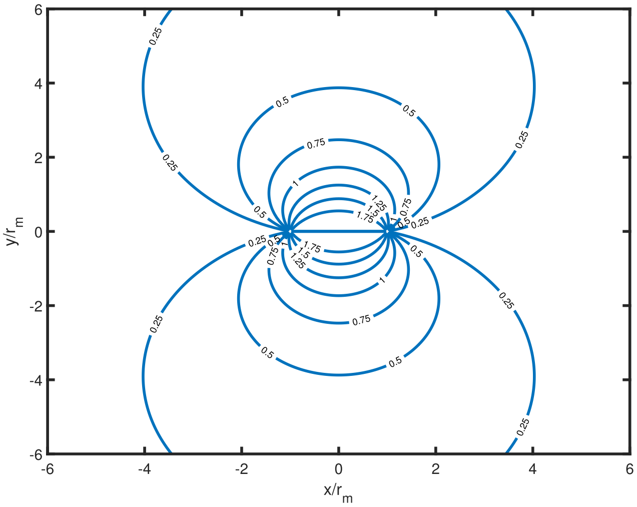


FIGURE 2. Contour plot of $c\,\|\nabla\tau_{kl}\,(\vec r)\|$ in the horizontal plane when both microphones are in that plane.

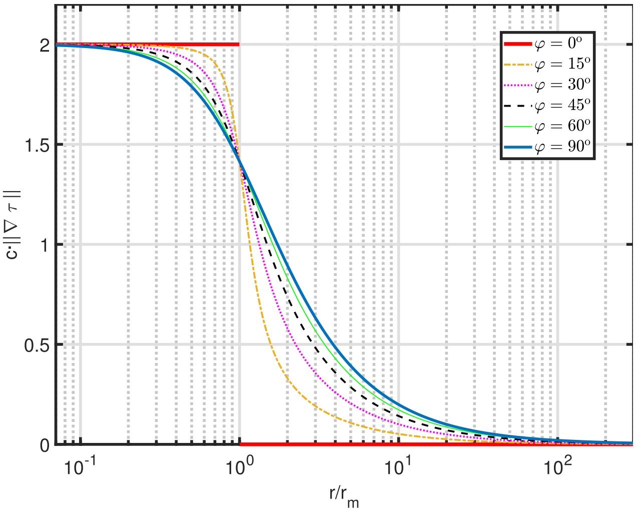


FIGURE 3. Plot of $c\,\|\nabla\tau_{kl}\,(\vec r)\|$ as a function of distance for several angles.

where, similarly as before, $r = \|\vec r\|$ and $r_{\mathrm{m}} = \|\vec r_{\mathrm{m}}\|$. According to the law of cosines:

$$r_k = \|\vec r - \vec r_{\mathrm{m}}\| = \sqrt{r^2 + r_{\mathrm{m}}^2 - 2\,r r_{\mathrm{m}}\cos\varphi}, \qquad (11)$$

where φ is the angle indicated in Fig. 1. Analogously:

$$r_l = \|\vec r + \vec r_{\mathrm{m}}\| = \sqrt{r^2 + r_{\mathrm{m}}^2 + 2\,r r_{\mathrm{m}}\cos\varphi}. \qquad (12)$$

Now, substituting (11) and (12) in (10):

$$\begin{aligned}\|\nabla\tau_{kl}\,(\vec r)\|^2 &= \frac{2}{c^2}\left(1 - \frac{r^2 - r_{\mathrm{m}}^2}{\sqrt{r^2 + r_{\mathrm{m}}^2 + 2\,r r_{\mathrm{m}}\cos\varphi}} \cdot \frac{1}{\sqrt{r^2 + r_{\mathrm{m}}^2 - 2\,r r_{\mathrm{m}}\cos\varphi}}\right)\\ &= \frac{2}{c^2}\left(1 - \frac{r^2 - r_{\mathrm{m}}^2}{\sqrt{\left(r^2 + r_{\mathrm{m}}^2\right)^2 - 4\,r^2\,r_{\mathrm{m}}^2\cos^2\varphi}}\right).\end{aligned} \qquad (13)$$

Thus, the Euclidean norm of the gradient is:

$$\begin{aligned}\|\nabla\tau_{kl}\,(\vec r)\| &= \frac{1}{c}\sqrt{2 - \frac{2\left(r^2 - r_{\mathrm{m}}^2\right)}{\sqrt{\left(r^2 + r_{\mathrm{m}}^2\right)^2 - 4\,r^2\,r_{\mathrm{m}}^2\cos^2\varphi}}}\\ &= \frac{1}{c}\sqrt{2 - \frac{2\left(\left(\frac{r}{r_{\mathrm{m}}}\right)^2 - 1\right)}{\sqrt{\left(\left(\frac{r}{r_{\mathrm{m}}}\right)^2 + 1\right)^2 - 4\left(\frac{r}{r_{\mathrm{m}}}\right)^2\cos^2\varphi}}}.\end{aligned} \qquad (14)$$

This expression shows that the norm of the gradient depends on the distance to the centre of the microphone array, relative to half the inter-microphone distance $\left(\frac{r}{r_{\mathrm{m}}}\right)$, and on the angle φ. The contour plot in Fig. 2 shows that the largest gradients occur near the centre of the array and for angles near 90°, with a maximum at the segment linking both microphones. This result is consistent with the simulation results on SRP sensitivity illustrated in [14]. Fig. 3 depicts the relation between the norm of the gradient and $\left(\frac{r}{r_{\mathrm{m}}}\right)$ for several angles. This graph shows that the maximum value of $c\,\|\nabla\tau_{kl}\,(\vec r)\|$ is 2, which happens between both microphones, and that the largest differences for diverse values of φ happen approximately for $0.2 < \left(\frac{r}{r_{\mathrm{m}}}\right) < 20$, i.e. for cases in which the difference between the size of the array and the distance between the array itself and the source positions is one order of magnitude at most. On the opposite, when $\left(\frac{r}{r_{\mathrm{m}}}\right)$ is large (far field) the influence of φ vanishes.

IV. IMPLICATIONS FOR THE CALCULATION OF THE SRP MAP

According to the approach introduced in the previous section, the calculation of a SRP-PHAT map (2) basically consists in a sample-and-sum process that includes sampling of several GCC-PHAT functions (4) with variable sampling intervals (8) whose values depend on the resolution of the SRP map, on the specific position being evaluated, and on the microphone positions. This sample-and-sum process leads to erroneous results when the selected samples of the GCC cannot represent some narrow peaks of the function. As stated by the sampling theorem [22, chap.8], if such loss of information (due to aliasing) is to be avoided, then the inverse of the sampling time should be greater than twice the bandwidth of the signal:

$$2\cdot\frac{\omega_{\max}}{2\pi} < \frac{1}{\Delta\tau_{kl}} \implies \Delta\tau_{kl} < \frac{\pi}{\omega_{\max}}, \qquad (15)$$

where it has been implicitly assumed that bandpass sampling is not feasible or, in other words, that $\omega_{\max} > 2\cdot\omega_{\min}$. If condition (15) is to be met in all cases, then taking (8) into account one can derive a sufficient condition that allows obtaining a SRP map that does not suffer from aliasing in the sampling of GCCs, given a specific microphone array and the corresponding audio signals:

$$\|\nabla\tau_{kl}(\vec{r})\| \cdot \Delta r < \frac{\pi}{\omega_{\max}}. \tag{16}$$

This relationship between distance r, array size r_{m} (both implicit in $\|\nabla\tau_{kl}(\vec{r})\|$), map resolution Δr, and signal bandwidth $\omega_{\max}$ can be exploited in several ways, depending on which of these magnitudes are defined by the scenario where the localization system is to be deployed and which ones are adjustable:

- For distributed microphone arrays in which the sound source is likely to be placed somewhere between the microphones, this implies $r \lesssim r_{\mathrm{m}}$ and in this case $\sqrt{2} \lesssim c\,\|\nabla\tau_{kl}(\vec{r})\| \leq 2$ (see Fig. 3). Therefore, the required map resolution is:

$$\Delta r < \frac{c\pi}{2 \cdot \omega_{\max}}. \tag{17}$$

- When the distance from the source to the array is known to be larger than the array size, then the TDOA gradient is bounded by the case $\varphi = 90^{\mathrm{o}}$ (see Fig. 3), thus:

$$\Delta r \sqrt{1 - \frac{\left(\frac{r}{r_{\mathrm{m}}}\right)^2 - 1}{\left(\frac{r}{r_{\mathrm{m}}}\right)^2 + 1}} < \frac{c\pi}{\sqrt{2}\cdot\omega_{\max}}, \tag{18}$$

  and the map resolution can be estimated from the minimum value expected for r or, alternatively, a map resolution dependent on r can be set.
- In any of both cases, if the map resolution is not adjustable, the corresponding conditions can be used for setting the upper limit in the integral used for calculating the GCC (4).
- Alternatively, condition (16) can also be used to calculate a SRP map with predefined resolution and variable GCC bandwidth. This implies limiting the GCC bandwidth in those points of the SRP map for which the TDOA gradient is high. Note that the time-domain effect of reducing the GCC bandwidth is similar to that of an integration of the GCC function, which is the operation proposed in [13]. Further details about this approach are given in the next section.

V. CALCULATION OF SRP MAPS WITH VARIABLE GCC BANDWIDTH

As pointed out before, when SRP maps with predefined resolution are to be generated, condition (16) can be used to generate them while avoiding aliasing in the sampling of GCC functions. This can be done following the next procedure:

1) Obtain the coordinates of the points in the grid used for generating the SRP map. Such grid will typically be characterized by its boundaries and a certain resolution Δr.
2) For each point in the grid, the SRP will be obtained as the summation in (2). After initializing this summation, the next actions should be performed for each pair of microphones in the array (k, l):
   a) Obtain the vectors linking the grid point and the microphone positions $\vec{r}_k$ and $\vec{r}_l$ (see Fig. 1).
   b) Calculate $\vec{r} = 0.5 \cdot (\vec{r}_l + \vec{r}_k)$ and $\vec{r}_{\mathrm{m}} = 0.5 \cdot (\vec{r}_l - \vec{r}_k)$.
   c) Also compute $\cos\varphi = (\vec{r} \cdot \vec{r}_{\mathrm{m}}) \,/\, (\|\vec{r}\| \cdot \|\vec{r}_{\mathrm{m}}\|)$.
   d) Use the previous results to calculate the norm of the gradient of the TDOA, as in (14).
   e) Knowing Δr and $\|\nabla\tau_{kl}(\vec{r})\|$, estimate the maximum frequency $\hat{\omega}_{\max}$ that guarantees avoiding aliasing according to (16).
   f) Calculate the GCC-PHAT as in (4), setting the upper limit to the integral equal to $\hat{\omega}_{\max}$.

$$\hat{R}_{kl}(\tau) = \int_{\omega_{\min}\leq|\omega|\leq\hat{\omega}_{\max}} \frac{S_k(\omega)\,S_l^*(\omega)\cdot e^{j\omega\tau}}{2\pi\,\left|S_k(\omega)\,S_l^*(\omega)\right|}\,d\omega. \tag{19}$$

   g) Evaluate the resulting GCC for $\tau_{kl} = c\,(\|\vec{r}_l\| - \|\vec{r}_k\|)$, and add the result to the SRP value.

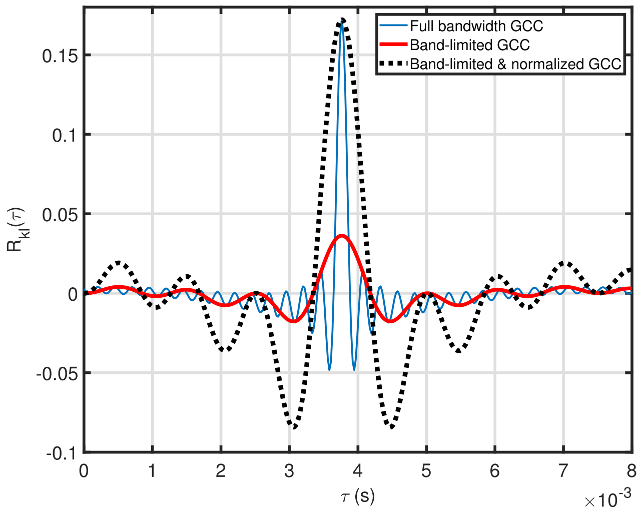


FIGURE 4. Cross correlation (GCC-PHAT) between two exactly equal speech signals taken from the dataset described in section VI-A, having a 3.76 ms delay between them. The thin line depicts the GCC-PHAT calculated by integration along the 200 Hz-4000 Hz band, the continuous thick line shows the result of limiting this interval to 200 Hz-1000 Hz, and the dotted line shows the effect of applying the proposed normalization to the band-limited GCC-PHAT.

The SRP-PHAT function in (2) can be interpreted as a likelihood function that should be maximized to find the best possible estimate for the sound source position [23]. Fig. 4 illustrates the effect that the bandwidth limitation specified in step 2.f has on the resulting GCC (band-limited GCC). Apart from the expected effect of reducing the frequency of the oscillations in the GCC, and increasing the width of its main peak, limiting the bandwidth has the consequence of reducing the height of that peak. While increasing the width of the peak has the positive effect of reducing the aliasing when the GCC is sampled to generate a SRP map, reducing its height may reduce the likelihood associated to the true source position when evaluating (2), thus altering the position of the maximum value of the SRP map. This reduction on the amplitude of the GCC peak can be compensated by normalizing the band-limited GCC proportionally to the bandwidth reduction, as follows:

$$\tilde{R}_{kl}(\tau) = \frac{\omega_{\max} - \omega_{\min}}{\hat{\omega}_{\max} - \omega_{\min}} \cdot \int_{\omega_{\min} \leq |\omega| \leq \hat{\omega}_{\max}} \frac{S_k(\omega) S_l^*(\omega) \cdot e^{j\omega\tau}}{2\pi \left|S_k(\omega) S_l^*(\omega)\right|} d\omega, \qquad (20)$$

where $\omega_{\min}$ and $\omega_{\max}$ are the limits of the signal bandwidth, and $\hat{\omega}_{\max}$ is the maximum frequency estimated in step 2.e. This normalization has the effect of keeping the value of the GCC peak unaltered, as shown in Fig. 4, at the cost of amplifying the oscillations of the function when τ moves away from the peak position.

The qualitative effect of applying the band limitation procedure proposed before is illustrated in Figs. 5 and 6. Both correspond to simulation in fully anechoic conditions of the acoustic propagation of a speech signal taken from the database described in section VI-A. In both cases a triangular microphone array has been supposed, with the sound source placed in the same plane, in a nearby position in Fig. 5, and in a further position in Fig. 6. The left plot in both figures shows the standard SRP-PHAT map, while the middle and right plots show the SRP-PHAT maps calculated with the procedure proposed here, both without (middle) and with (right) the normalization in (20). One noticeable effect of limiting the band of the GCC is a reduction on the number and relative relevance of the local maxima in the resulting SRP map, which makes it more robust against changes in spatial resolution. For source positions far from the centre of the array, the norm of the TDOA gradient (Fig. 3) is low, which results in little band limitation and, consequently, similar results are expected in the estimation of source positions; this is the case illustrated in Fig. 6. However, for source positions near the microphone array, greater differences in the estimated source positions are expected, as the case in Fig. 5.

Calculating SRP maps is computationally expensive, and the fact that calculating band-limited GCCs, as specified in (19) and (20), increases such computational cost cannot be overlooked. To present, several strategies have been proposed to speed up SRP calculation, such as decomposing the SRP map in spatial basis functions [24], or look-up tables for TDOA values [25]. These strategies can be extrapolated to the case of using band-limited GCCs by running a spatial analysis before calculating GCCs in order to identify the required bandwidths, computing and storing GCCs, and using them as look-up tables when building SRP maps. However, it should be stressed that implementation issues are beyond the scope of this research.

VI. EXPERIMENTS AND RESULTS

The procedure proposed in section V to calculate SRP maps has been incorporated into some simulation experiments in order to evaluate its potential impact on the source localization performance of systems based on SRP maps.

A. AUDIO DATA

The signals used for the simulation experiments corresponded to several acoustic events included in the *Sound event detection in synthetic audio* task of the DCASE 2016 challenge [26]. Its associated dataset includes audio files corresponding to 11 types of sound events. According to the spectral analysis reported in [27], these types of sound events can be grouped into several categories taking into account their spectra. Specifically, the shape of the spectra can be classified in the following four categories:

- *Noisy (non-harmonic) low-pass spectra*, which includes the cases of door slams, opening or closing drawers, typing, door knocking, and page turning.
- *Low-pass spectra with resonances due to the human vocal tract*, as in the case of clearing one's throat, coughing, laughing, and speaking.
- *Noisy flat spectra*, which is the case for key dropping events.
- *Harmonic spectra with flat envelope*, as in phone ringing.

In order to cover all the four classes of spectral shape, one event type from each class was selected for running the simulations, namely door slams, speaking, key dropping, and phone ringing. All the 20 recordings corresponding to each type included in the development dataset of the challenge were used, which resulted in a total of 80 recordings. In all cases, the sampling rate was equal to 44.1 kHz and the sound was sampled with a resolution of 16 bits. The duration of the recordings ranged from 0.13 s to 3.34 s.

B. EXPERIMENTS

The aforementioned sound events were simulated to happen in a 8 m × 10 m × 4 m room. Specifically, 1000 source positions inside the room were randomly selected with uniform probability distribution. For each source position one audio recording corresponding to each event type was randomly selected, thus resulting in a total of 4000 simulated sound events. The sound propagation between the source positions and the microphones was simulated by delaying each signal according to the corresponding propagation distance. The sound speed was assumed equal to 343 m/s.

Simulations were carried out for two different microphone arrays. Both of them were formed by 4 microphones placed in the corners of a regular tetrahedron whose central point was located at the centre of the room. The length of the tetrahedron edges were 0.5 m in one case (small array) and 3 m in the other (large array).

The position of the sound source in each case was estimated from the simulated microphone signals according to the algorithm in V. The resolution chosen for generating the SRP

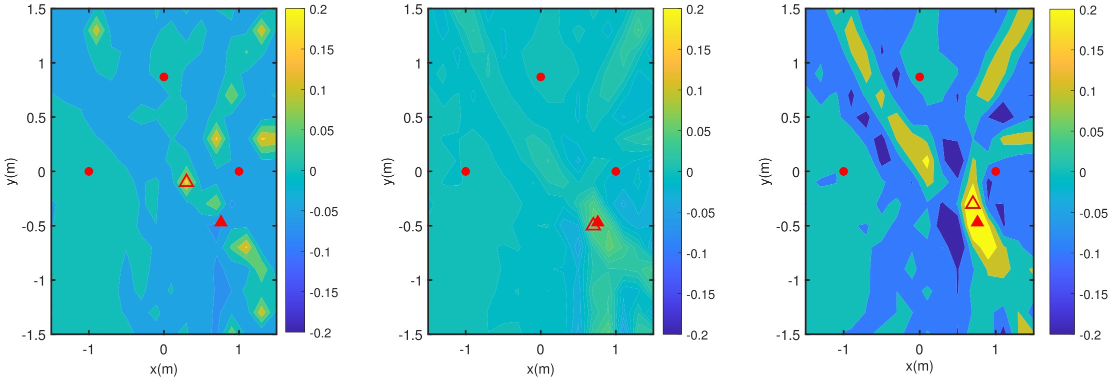


FIGURE 5. SRP-PHAT maps generated according to the standard procedure (left), applying (19) for limiting the bandwidth of the GCC (middle), and adding the normalization in (20) (right). Red points indicate the simulated microphone positions, the filled triangles mark the simulated source position, and the empty triangles show the maximum peaks of the SRP maps, i.e. the estimated source positions. Anechoic conditions have been assumed, and the audio signal used for simulation is the same as in fig. 4.

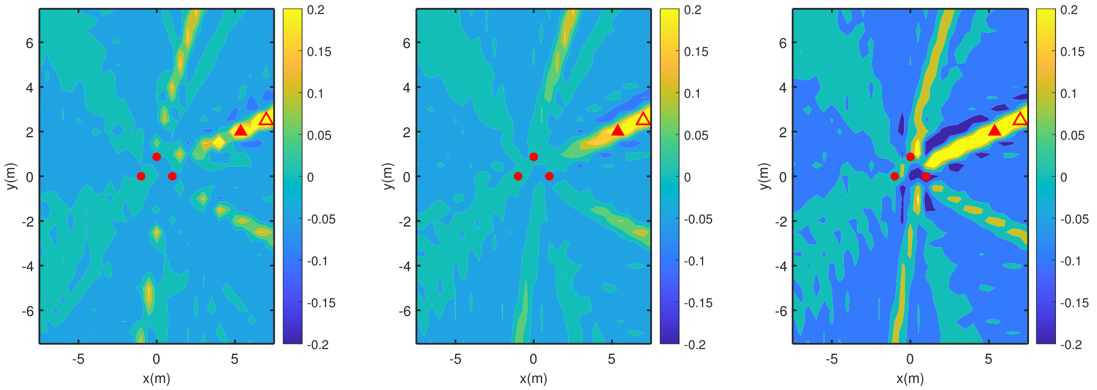


FIGURE 6. Same as fig. 5, but with the simulated source further from the microphone array.

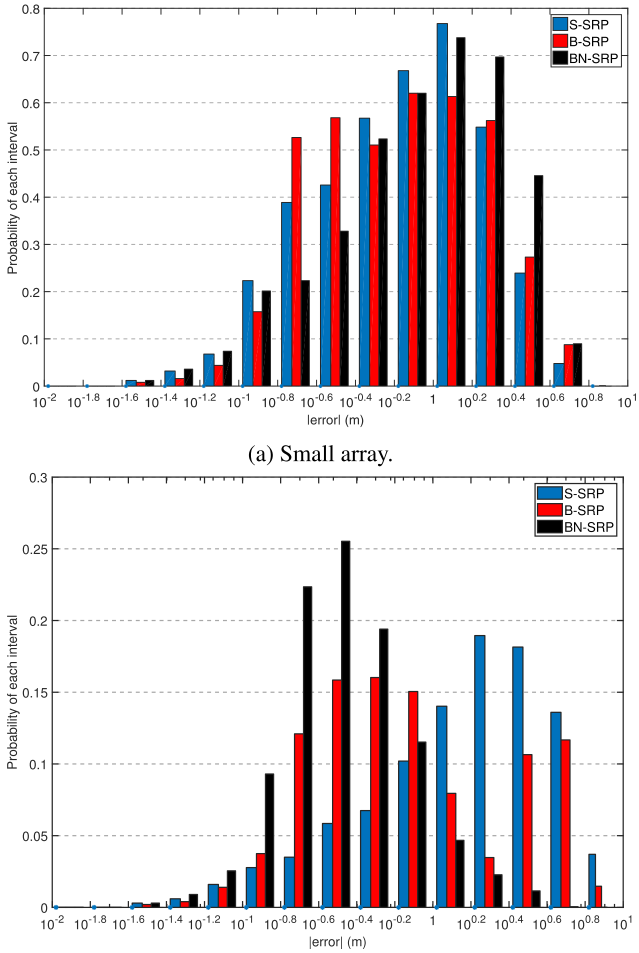
maps was 0.5 m for all experiments except for one performed with 1 m resolution for the sake of assessing the effect of increasing resolution. Each experiment involved simulating all 4000 sound events mentioned before. The signal bandwidth was assumed to be between 100 Hz and 6000 Hz. According to the analysis in [27], the signal-to-noise ratio beyond 6000 Hz was poor for low-pass signals. The localization error was calculated as the absolute value of the difference between the estimated source positions and the actual simulated positions.

Two different acoustic conditions were simulated: anechoic conditions and reverberant conditions with reverberation time equal to 0.6 s, corresponding to a realistic low-reverberant environment [28]. Simulations of the reverberant room were performed using the image method proposed by Allen and Berkley [29], as implemented in Matlab® by Habets [30].

C. RESULTS

Fig. 7 shows the histograms representing the distributions of localization errors for the anechoic scenario mentioned before, and for SRP maps calculated using the standard GCC (4) (S-SRP), the band limited GCC (19) (B-SRP), and the normalized band-limited GCC (20) (BN-SRP). The plot in Fig. 7a shows the histograms for the small array, while the plot in Fig. 7b corresponds to the large array. At first sight, the use of the band-limited GCC does not produce results significantly different to those of the standard GCC. Furthermore, the normalization proposed in (20) produces a moderate worsening of the localization performance. But in the case of the large array (Fig. 7b) the band limitation in the GCC produces a relevant reduction in localization error and the magnitude of this reduction is higher when the normalization in (20) is applied.

Table 1 provides a quantitative description of the distributions of localization errors. Specifically, the average error for each case, and the mean deviation of errors are given. It is apparent from Fig. 7 that distributions cannot be assumed to be Gaussian (e.g. the distributions of errors for B-SRP is bimodal). For this reason, nonparametric tests were chosen to evaluate the statistical significance of differences among the means and the dispersions (mean deviations) of distributions. Namely, the Wilcoxon test was used to evaluate differences in the mean values, and a permutation test of deviances to evaluate differences in the dispersions [31]. For a given distribution of N estimation errors e_n, the mean



(a) Small array.

(b) Large array.

FIGURE 7. Histograms of localization errors for the small and large arrays. Results are given as the probability of each interval in the *x* axis for SRP maps calculated using the standard GCC (4) (S-SRP), the band-limited GCC (19) (B-SRP), and the normalized band-limited GCC (20) (BN-SRP).

TABLE 1. Average value and mean deviation of the distributions of localization errors for the anechoic scenario.

	Mean error (m)			Mean deviation (m)		
	S-SRP	B-SRP	BN-SRP	S-SRP	B-SRP	BN-SRP
Small array	1.024*	1.062*	1.268	0.6845	0.7653	0.8308
Large array	2.2405	1.478	0.4693	1.485	1.383	0.2929

The difference between quantities marked with * is not statistically significant ($p > 0.01$).

TABLE 2. Average value and mean deviation of the distributions of localization errors for the anechoic scenario discriminated for three different distance intervals.

	Mean error (m)			Mean deviation (m)		
	S-SRP	B-SRP	BN-SRP	S-SRP	B-SRP	BN-SRP
Small array $\left(\frac{r}{r_{\mathrm{m}}} \leq 5\right)$	2.149	3.210	0.8203	1.151*	1.196*	0.6731
Large array $\left(\frac{r}{r_{\mathrm{m}}} \leq 5\right)$	2.241	1.478	0.4693	1.485	1.383	0.2929
Small array $\left(5 < \frac{r}{r_{\mathrm{m}}} \leq 10\right)$	1.313	1.637	1.049	0.8179	0.9324	0.5759
Small array $\left(10 < \frac{r}{r_{\mathrm{m}}}\right)$	0.8428	0.7066	1.3811	0.5433	0.4536	0.9285

The difference between quantities marked with * is not statistically significant ($p > 0.01$).

or average error is defined as:

$$\mu_e = \frac{1}{N}\sum_{n=1}^{N} e_n \tag{21}$$

and the mean deviation is:

$$\delta_e = \frac{1}{N}\sum_{n=1}^{N} |e_n - \mu_e| \,. \tag{22}$$

Results in Tab. 1 confirm the observations that S-SRP and B-SRP perform similarly for the small array (non-significant differences in the mean errors and similar values for dispersions) but not for the large array, B-SRP performs better in that case, and BN-SRP provides a significant performance improvement for the large array, while it performs poorly for the small array.

A deeper insight into the previous results can be obtained if they are segmented by the distance of the simulated sources to the centre of the array. This can be done using the relative measure r/r_{m}. Table 2 shows the average errors and their mean deviances discriminated for three intervals: $r/r_{\mathrm{m}} \leq 5$; $5 < r/r_{\mathrm{m}} \leq 10$; and $10 < r/r_{\mathrm{m}}$. These results indicate that the performance of all three algorithms is not as dependent from the array size as from the relative distance between the source and the array centre. Note that all 1000 simulated source positions comply with the condition $r/r_{\mathrm{m}} \leq 5$ in the case of the large array, while the distribution for the small array is:

- 40 points in the $r/r_{\mathrm{m}} \leq 5$ interval,
- 274 points in the $5 < r/r_{\mathrm{m}} \leq 10$ interval,
- and 686 points in the $10 < r/r_{\mathrm{m}}$ interval.

It can be observed that the localization error using the standard GCC diminishes as the distance between the centre of the array and the source position is increased. The performance of the estimator based on the band-limited GCC is also increased for longer distances, but this estimator differs from the previous one mainly in two aspects: the average reduction in the localization error is greater as distance is increased, and the dispersion of the localization errors is also the lowest for the longest evaluated distances. The SRP map based on the band-limited and normalized GCC provides estimations that behave oppositely with respect to the other two cases. This method provides better estimations for short distances between array and sound source, and its performance is negatively affected by increases in these distances.

TABLE 3. Average value and mean deviation of the distributions of localization errors for the reverberant scenario discriminated for three different distance intervals.

	Mean error (m)			Mean deviation (m)		
	S-SRP	B-SRP	BN-SRP	S-SRP	B-SRP	BN-SRP
Small array $\left(\frac{r}{r_\mathrm{m}} \leq 5\right)$	2.222	3.193	0.9436	1.196*	1.204*	0.6189
Large array $\left(\frac{r}{r_\mathrm{m}} \leq 5\right)$	3.009	3.306$^\diamond$	3.111$^\diamond$	1.770	2.112	1.575
Small array $\left(5 < \frac{r}{r_\mathrm{m}} \leq 10\right)$	1.263	1.638	2.028	0.7881	0.9091	0.4071
Small array $\left(10 < \frac{r}{r_\mathrm{m}}\right)$	0.9822	0.8276	3.976	0.6682	0.5754	0.6686

The differences between quantities marked with *, and $^\diamond$ are not statistically significant ($p > 0.01$).

TABLE 4. Average value and mean deviation of the distributions of localization errors for the reverberant scenario and the small array with coarser map resolution (1 m).

	Mean error (m)			Mean deviation (m)		
	S-SRP	B-SRP	BN-SRP	S-SRP	B-SRP	BN-SRP
Small array $\left(\frac{r}{r_\mathrm{m}} \leq 5\right)$	2.565	2.957	0.9977	1.006*	1.080*	0.5866
Small array $\left(5 < \frac{r}{r_\mathrm{m}} \leq 10\right)$	2.056$^\diamond$	1.840†	1.7016$^{\diamond,\dagger}$	1.110‡	1.020‡	0.4139
Small array $\left(10 < \frac{r}{r_\mathrm{m}}\right)$	1.665	0.9344	3.651	1.150	0.5468	0.6794

The differences between quantities marked with *, $^\diamond$, †, and ‡, are not statistically significant ($p > 0.01$).

The results of running the same experiments but with reverberation time equal to 0.6 s (Tab. 3) indicate that, in general terms, the presence of reverberation tends to negatively affect localization results. In fact, all mean errors in Tab. 3 are higher than the corresponding values in Tab. 2, except for those relative to S-SRP and B-SRP being applied to the few points with $r/r_\mathrm{m} \leq 5$ in the small array case. Such increase in average error happens more prominently for BN-SRP. Another relevant aspect of these results is that the growth in average error is less relevant for the most distant sources ($10 < r/r_\mathrm{m}$), and that B-SRP still provides the best performance for this case in the reverberant scenario.

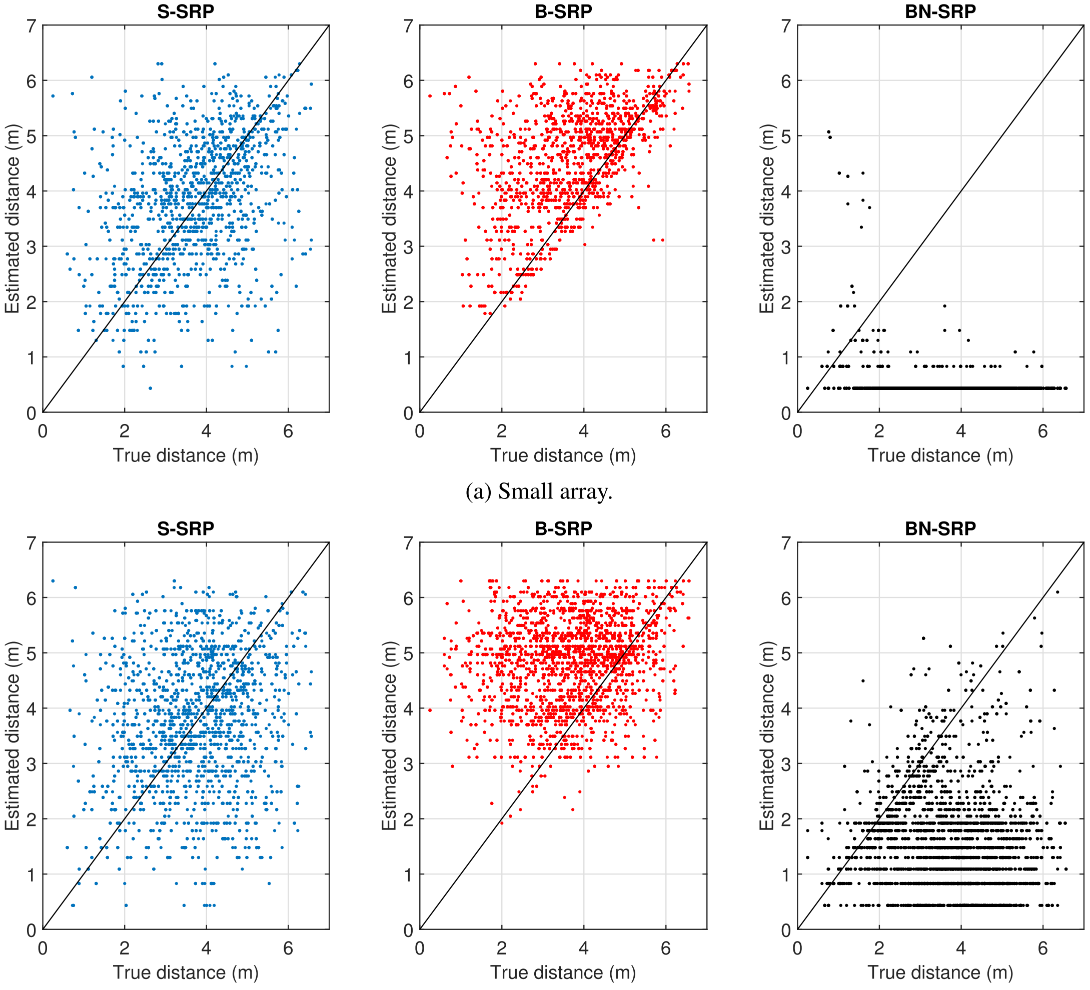
The effect of changing map resolution on the relative performance of all three options for calculating the GCC was assessed by running one additional experiment (i.e. 4000 simulated sound events) with coarser resolution (1 m) in the reverberant scenario. Considering the poor localization performance for short distances in this scenario (Tab. 3), only the small array was simulated in this case. Results summarized in Tab. 4 show that the advantages provided by band-limiting the GCC, either with or without normalization, are more noteworhty in this case. In other words, performance of the SRP based on the standard GCC seems to be more sensitive to increases in map resolution than that of SRP based on the band-limited GCC.

VII. DISCUSSION AND CONCLUSION

The analysis presented in sections III and IV was aimed at calculating SRP maps avoiding the potential aliasing effects that may happen when sampling the GCC function regardless the relation between SRP map resolution and GCC bandwidth. This analysis led to the sufficient condition (16) that allows avoiding such aliasing. However, the inequality in (8) implies that fulfilling this condition is not necessary to avoid aliasing or, in other words, that by applying this condition one can limit the bandwidth of the GCC more than what is strictly necessary. As a consequence, the localization errors produced by B-SRP may be sometimes larger than those of S-SRP, as can be noticed in the histogram in Fig. 7a. According to the numerical results summarized in Tabs. 2 and 3, this worsening of localization performance happens especially for the shortest distances between microphone array and sound source position ($r/r_\mathrm{m} \leq 5$).

One possible explanation for the afore-mentioned worsening of localization performance was hypothesized to be the reduction in the height of the main peak of the GCC that is intrinsically linked to bandwidth limitation (see Fig. 4). As illustrated in Figs. 2 and 3, points near the microphones are associated to the highest TDOA gradients and, consequently, the calculation of their corresponding SRP values is affected the most by the bandwidth limitation of the GCC. This implies a reduction of the height of the main GCC peak and a corresponding reduction of the SRP value. The GCC normalization in (20) was proposed to compensate this effect at the cost of increasing the amplitude of some secondary GCC peaks (Fig. 4). The inclusion of this normalization factor in the calculation of SRP maps has shown to have a very positive impact on localization performance for sound source positions near or even inside the volume occupied by the microphone array (see Fig. 7a and Tab. 2). However, the effect of such normalization on the secondary peaks in the GCC (Fig. 4) is likely to be a key factor in worsening the performance of this approach for reverberant environments (Tab. 3).

For longer source-to-array distances ($10 < r/r_\mathrm{m}$), applying bandwidth limitation to the GCC according to (16) has shown to consistently provide performance improvements over the standard approach for calculating SRP without limiting the bandwidth of the GCC (Tabs. 2 and 3). These improvements involve reductions in both average error and error dispersion. The reason that justifies the improved performance of B-SRP can be explained by looking back to Fig. 5. The left map therein shows a typical localization error of S-SRP maps. This error is mainly caused by two factors: the calculation of the SRP map misses a relevant GCC peak near the actual source position due to this peak being narrower than the corresponding map resolution, and some secondary GCC peaks are added up in a position nearer the



(a) Small array.

(b) Large array.

FIGURE 8. **Scatter plots showing the relation of the actual distance between the sound source and the centre of the microphone array to the distance between the estimated source position and the centre of the array. Plots correspond, left to right, to S-SRP (SRP based on the standard GCC), B-SRP (SRP based on the band-limited GCC), and BN-SRP (SRP based on the band-limited normalized GCC).**

centre of the array, hence producing a peak in the SRP map higher than it should be. As illustrated by Fig. 4, limiting the bandwidth of the GCC has the double effect of widening the main GCC peak and eliminating some secondary peaks. This has the consequence of avoiding errors in which the distance between the sound source and the microphone array is underestimated.

The left and central scatter plots in Fig. 8 represent the relation between the real source-to-array distance and the estimated source-to-array distance for both S-SRP and B-SRP in the case of the reverberant scenario. These plots show that the previously mentioned consequence of limiting the bandwidth of the GCC does not only happen in specific points; instead, it is a general rule for the results of our experiments that limiting the bandwidth of the GCC according to (16) reduces the probability of underestimating source-to-array distances. This explains why the B-SRP performs better for large distances (Tabs. 2 and 3). When source-to-array distances are significantly larger than array size, underestimating this distance is an issue, and B-SRP performs the best for the largest simulated distances.

A reasoning analogous to the previous one leads to the conclusion that the GCC normalization in (20) has the effect of increasing the values of the SRP map in positions near the centre of the array. Thus, it reduces the chance of overestimating the source-to-array distance. This effect is confirmed by the right plot in Fig. 8. However, the presence of reverberation has a very negative impact on localization performance when source-to-array distances are in the range of, or even shorter than the array size ($r/r_\mathrm{m} \leq 5$), which corresponds to the case where the microphones are more distributed in the room. Thus, the negative impact of reverberation masks the potential benefits of using BN-SRP in reverberant scenarios. Yet, note that even in this case BN-SRP yields significantly lower error dispersion for that range of distances than both S-SRP and B-SRP (Tab. 3) for similar average errors.

The high computational cost of calculating SRP with fine spatial resolutions has led several researchers to propose iterative approaches to sound source localization, consisting in a step-by-step decrease in map resolution accompanied by a corresponding reduction in map extent, as mentioned in the introduction. The analysis presented in this paper has made no assumption about specific intervals for map resolution, so it is applicable at any scale in those iterative or hierarchical approaches. To illustrate this, an additional experiment was run with map resolution equal to 1 m instead of 0.5 m. The corresponding results, summarized in Tab. 4, confirmed all the previously stated conclusions. Furthermore, the increased map resolution implies the requirement of a narrower spectrum according to (16) or, from another point of view, coarser resolutions in SRP maps lead to more relevant aliasing effects if the bandwidth of the GCC is not limited. Such increased aliasing leads to a noticeable worsening of localization results for S-SRP (compare results in Tab. 3 to those in Tab. 4). However, the impact of the coarser resolution in the performance of B-SRP and BN-SRP is much lower, to the extent that B-SRP provides significantly better results than S-SRP even for intermediate distances ($5 < r/r_\mathrm{m} \leq 10$), which was not the case when the resolution was 0.5 m.

In conclusion, equations (8) and (15) show that there is a relation between the bandwidth of acoustic signals and the resolution of SRP-PHAT maps calculated for localizing their corresponding source. This relation implies the sufficient condition for an aliasing-free calculation of the SRP map specified by (16). Such calculation can be done according to the algorithm described in section V and limiting the bandwidth of the GCC as indicated in (19). While the fact that integrating (i.e. low-pass filtering) the GCC leads to increased robustness in localization performance of SRP-PHAT maps was already known [13], the analysis presented before provides a theoretical justification for such improvement and an explicit rule that relates GCC bandwidth to the spatial resolution of SRP-PHAT maps.

The reported experiments show that this approach leads to improved source localization results for source positions far from the microphone array, since the probability of underestimating the source-to-array distance is reduced. It has also been tested that the proposed approach is robust against reverberation, since it provides similar advantages in both anechoic and reverberant scenarios. Last, it should be stressed that the use of condition (16) to avoid aliasing effects in the calculation of SRP maps is fully compatible with hierarchical localization algorithms in which map resolution is iteratively changed. Moreover, it should significantly contribute to obtain more robust results at the coarsest resolution levels.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Universidad Politécnica de Madrid for providing computing resources on Magerit Supercomputer.

REFERENCES

[1] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 1–52, 2016.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. New York, NY, USA: Springer, 2001, pp. 157–180.

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

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

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

[6] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, and M. V. M. Costa, "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5171–5183, Oct. 2014.

[7] M. V. S. Lima, "A volumetric SRP with refinement step for sound source localization," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1098–1102, Aug. 2015.

[8] D. Yook, T. Lee, and Y. Cho, "Fast sound source localization using two-level search space clustering," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 20–26, Jan. 2016.

[9] D. Salvati, C. Drioli, and G. L. Foresti, "Sensitivity-based region selection in the steered response power algorithm," *Signal Process.*, vol. 153, pp. 1–10, Dec. 2018.

[10] S. Astapov, J. Berdnikova, and J.-S. Preden, "Optimized acoustic localization with SRP-PHAT for monitoring in distributed sensor networks," *Int. J. Electron. Telecommun.*, vol. 59, no. 4, pp. 383–390, Jan. 2013.

[11] M. A. Awad-Alla, A. Hamdy, F. A. Tolbah, M. A. Shahin, and M. A. Abdelaziz, "A two-stage approach for passive sound source localization based on the SRP-PHAT algorithm," *APSIPA Trans. Signal Inf. Process.*, vol. 9, pp. 1–12, Oct. 2020.

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

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

[14] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement," *J. Acoust. Soc. Amer.*, vol. 141, no. 1, pp. 586–601, 2017.

[15] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Acoustic source localization with deep generalized cross correlations," *Signal Process.*, vol. 187, Oct. 2021, Art. no. 108169.

[16] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 300–311, 2021.

[17] A. D. Firoozabadi, P. Irarrazaval, P. Adasme, H. Durney, and M. S. Olave, "A novel quasi-spherical nested microphone array and multiresolution modified SRP by GammaTone filterbank for multiple speakers localization," in *Proc. Signal Process., Algorithms, Archit., Arrangement, Appl. (SPA)*, Sep. 2019, pp. 208–213.

[18] H. F. Silverman and W. R. Patterson, "Visualizing the performance of large-aperture microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1999, pp. 969–972.

[19] J. McDonough and K. Kumatani, "Microphone arrays," in *Techniques for Noise Robustness in Automatic Speech Recognition*. Hoboken, NJ, USA: Wiley, 2012, pp. 109–157.

[20] J. Velasco, C. J. Martín-Arguedas, J. Macias-Guarasa, D. Pizarro, and M. Mazo, "Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios," *Signal Process.*, vol. 119, pp. 209–228, Feb. 2016.

[21] R. C. Wrede and M. R. Spiegel, *Theory and Problems of Advanced Calculus*. New York, NY, USA: McGraw-Hill, 2010.

[22] A. V. Oppenheim, A. S. Willsky, and I. Young, *Signals System*. Upper Saddle River, NJ, USA: Prentice-Hall, 1983.

[23] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 538–548, Apr. 2008.

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

[25] Y. Cho, D. Yook, S. Chang, and H. Kim, "Sound source localization for robot auditory systems," *IEEE Trans. Consum. Electron.*, vol. 55, no. 3, pp. 1663–1668, Aug. 2009.

[26] (2016). *Task 2—Sound Event Detection in Synthetic Audio*. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/challenge

[27] J. M. Gutiérrez-Arriola, R. Fraile, A. Camacho, T. Durand, I. N. Jarr, and S. R. Mendoza, "Synthetic sound event detection based on MFCC," in *Proc. DCASE*, 2016, pp. 30–34.

[28] J. Cowan, "Building acoustics," in *Handbook Acoustics*, T. Rossing, Ed. Cham, Switzerland: Springer, 2007, pp. 387–425.

[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[30] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Eindhoven Univ. Technol., Eindhoven, The Netherlands, Jun. 2007. [Online]. Available: https://pure.tue.nl/ws/portalfiles/portal/1972985/200710970.pdf

[31] J. J. Higgins, *Introduction to Modern Nonparametric Statistics*. Pacific Grove, CA, USA: Thomson Brooks-Cole, 2004.

GUILLERMO GARCÍA-BARRIOS was born in Madrid, Spain, in 1994. He received the B.S. degree in sound and image engineering and the M.S. degree in systems and services engineering for the information society from the Universidad Politécnica de Madrid (UPM), in 2017 and 2018, respectively, where he is currently pursuing the Ph.D. degree in acoustic signal processing.

His research interests include sound source localization algorithms and systems, acoustic simulation, machine learning for automatic sound signal recognition, and wireless acoustic sensor networks.

JUANA M. GUTIÉRREZ-ARRIOLA (Member, IEEE) was born in Santander, Spain, in 1971. She received the B.S. and M.S. degrees in telecommunications engineering from the Universidad de Cantabria, in 1993 and 1994, respectively, and the Ph.D. degree from the Universidad Politécnica de Madrid (UPM), in 2008.

Currently, she is a Senior Lecturer with the Telematics and Electronics Engineering Department, UPM, where she belongs to the Research Group on Acoustics and Multimedia Applications (GAMMA) and carries out research on biomedical image processing, speech analysis, and sound source localization. She has authored over 25 scientific papers in those fields.

Dr. Gutiérrez-Arriola has been the Secretary of the Escuela Universitaria de Ingeniería Técnica de Telecomunicació, UPM, from 2004 to 2008, and the Head of the Department of Circuits and Systems Engineering, UPM, from 2011 to 2014. She is currently the Secretary of the Research Center on Software Technologies and Multimedia Systems for Sustainability and also with UPM.

NICOLÁS SÁENZ-LECHÓN was born in Barcelona, Spain, in 1972. He received the B.S., M.S., and Ph.D. degrees in telecommunications engineering from the Universidad Politécnica de Madrid (UPM), Spain, in 1996, 2001, and 2010, respectively.

Since 2019, he has been a Lecturer with the Department of Audiovisual Engineering and Communications, UPM. His research interests include biomedical image processing and speech analysis for clinical applications. He is the author of over 20 journal articles and more than 50 conference papers in those fields.

VÍCTOR JOSÉ OSMA-RUIZ was born in Cuenca, Spain, in 1973. He received the B.S., M.S., and Ph.D. degrees in telecommunications engineering from the Universidad Politécnica de Madrid (UPM), Spain, in 1995, 2001, and 2010, respectively.

He gives lectures at the Escuela Técnica Superior de Ingeniería y Sistemas de Telecomunicación, UPM, since 1999, where he is currently a Senior Lecturer at the Telematics and Electronics Engineering Department. His research interests include biomedical image processing, speech analysis for clinical applications, gamification, and virtual reality. He has published 20 journal articles and another 60 scientific papers in conferences and congresses in those fields.

RUBÉN FRAILE was born in Barakaldo, Spain, in 1972. He received the M.S. level degree in telecommunications engineering from the Universidad Politécnica de Valencia (UPV), in 1995, and the Ph.D. degree in mobile communications from the UPV, in 2000.

Before completing his doctorate, he worked as a Radio-Planning Engineer with Retevisión Móvil (Spanish mobile network operator), and he later taught several subjects in a secondary school, from 2000 to 2007. Simultaneously, he stayed as a Postdoctoral Researcher with the UPV, from 2001 to 2007. During this period, he made research in the simulation of wireless communication networks. He got a position as a Senior Lecturer with the Universidad Politécnica de Madrid (UPM), in 2007, after which he re-oriented his research interests towards the field of discrete-time signal processing applications. He belongs to the Research Group on Acoustics and Multimedia Applications (GAMMA), UPM.

Dr. Fraile was the Coordinator of the scientific programme in the international workshop on Advanced Voice Function Assessment (AVFA), in 2009, and during his career has held two management positions in educational institutions: he coordinated the quality management system (ISO-9000) in a secondary school, from 2004 to 2006, and he was an adjunct to the Vice-Principal of Academic Staff at the CEU Cardenal Herrera University, Valencia, Spain, from 2010 to 2013.