FEBRUARY 01 2014

A Bayesian direction-of-arrival model for an undetermined number of sources using a two-microphone array

Jose Escolano; Ning Xiang; Jose M. Perez-Lorenzo; Maximo Cobos; Jose J. Lopez

J. Acoust. Soc. Am. 135, 742–753 (2014)
https://doi.org/10.1121/1.4861356

A Bayesian direction-of-arrival model for an undetermined number of sources using a two-microphone array

Jose Escolano[a)]
Sophandrey Research, Brooklyn, New York 11223

Ning Xiang
Graduate Program in Architectural Acoustics, School of Architecture, Rensselaer Polytechnic Institute, Troy, New York 12180

Jose M. Perez-Lorenzo
Multimedia and Multimodal Processing Research Group, University of Jaén, 23700, Linares, Spain

Maximo Cobos
Computer Science Department, University of Valencia, 46100, Burjassot, Spain

Jose J. Lopez
Institute for Telecommunication and Multimedia Applications, Universidad Politécnica de Valencia, 46021, Valencia, Spain

(Received 9 July 2013; revised 27 November 2013; accepted 23 December 2013)

Sound source localization using a two-microphone array is an active area of research, with considerable potential for use with video conferencing, mobile devices, and robotics. Based on the observed time-differences of arrival between sound signals, a probability distribution of the location of the sources is considered to estimate the actual source positions. However, these algorithms assume a given number of sound sources. This paper describes an updated research account on the solution presented in Escolano *et al*. [J. Acoust. Am. Soc. **132**(3), 1257–1260 (2012)], where nested sampling is used to explore a probability distribution of the source position using a Laplacian mixture model, which allows both the number and position of speech sources to be inferred. This paper presents different experimental setups and scenarios to demonstrate the viability of the proposed method, which is compared with some of the most popular sampling methods, demonstrating that nested sampling is an accurate tool for speech localization. © *2014 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4861356]

PACS number(s): 43.60.Jn, 43.72.Ne [ZHM] Pages: 742–753

I. INTRODUCTION

Sound source localization has many applications in acoustic communication, such as video conferencing, robotics, speech enhancement, noise reduction, and sound source separation.[1–4] Direct localization approaches are based on computing a cost function over a set of candidate locations and take the most likely source positions, whereas the so-called *indirect approaches* are based on the estimation of relative signal time delays to estimate the direction-of-arrival (DOA) and in some cases, also the position of the source.[5] Most traditional approaches are based on the generalized cross correlation (GCC) method,[6] which calculates the correlation function using the inverse Fourier transform of the cross-power spectral density function multiplied by a proper weighting function.

Different microphone array configurations may be used, and usually the performance in DOA estimation improves with the number of microphones. In fact, most microphone arrays today have more than two microphones. However, there are still some applications where the use of two-microphones is required such as humanoid robotics,[7] binaural hearing,[8] hearing aids,[9] phone devices, or modeling of psychophysical studies.[10] Because of the ability of the human auditory system to locate sounds, research dealing with two-microphones using binaural models of computational auditory scene analysis is still a growing field.[11–15] Among the different available variants of GCC, GCC-PHAT (*phase transformation*) is the most popular.[6] According to the literature, this modification improves the GCC algorithm in environments with relatively high reverberation time when the noise presence is low.[16] Also, it has been proven to reduce the spreading effect that occurs when uncorrelated noise appears at both microphones.

However, the usual scenario for using GCC-PHAT has been mostly limited to localization of one source and the multiple-source case has been rarely described.[17] Even in those few instances, the number of sources is always a known parameter. For more practical applications, it seems appropriate to define a multi-source framework to implement the histogram-based GCC-PHAT localization algorithm for an unknown number of sources.

This paper presents an extension of Bayesian inference method for speech localization using nested sampling to infer the number of active sources and their DOA parameters.[18] With this aim, a Laplacian mixture model is used to

[a)]Author to whom correspondence should be addressed. Electronic mail: joseescolanocarrasco@gmail.com

 0001-4966/2014/135(2)/742/12/$30.00

model the delay histogram generated by GCC-PHAT algorithm,[20] where the number of those Laplacian functions will provide the number of sources, and their parameters, their corresponding direction-of-arrivals. This contribution demonstrates that the source location with an unknown number of sources requires two levels of inference. Bayesian two-level inference provides a systematic complement to the histogram-based GCC-PHAT localization algorithm to be extended for an unknown number of sources. Moreover, it has to be validated in a wider range of real scenarios, a topic barely studied in the technical literature.[21]

This paper deals with experimental measurements in real scenarios. These scenarios are featured with different room-acoustic environments, characterized by different reverberation times and signal-to-noise ratios (SNR) in the enclosure under investigation. Several speaker configurations are also used to validate the algorithm. Moreover, a comparison with other popular sampling methods within the scope of Bayesian inferential analysis is also provided to justify the advantages of this method in terms of efficiency and accuracy. This paper demonstrates that the use of nested sampling together with Jeffrey's scale for the model selection provides accurate results down to an SNR of approximately 20 dB in an autonomous way.

This paper is organized as follows: first, Sec. II briefly surveys the source localization model based on GCC-PHAT. Section III describes the basics of Bayesian model selection and parameter estimation. Section IV further discusses different approaches for sampling implementation, including nested sampling. Finally, Sec. V thoroughly describes results using the proposed methodology under different scenarios, when faced with different SNRs, reverberation times and degrees of partially correlated noise.

II. SOUND SOURCE LOCALIZATION

A. Signal model

Given a two-microphone array in an anechoic scenario,[15] let us consider signals $x_1(t)$ and $x_2(t)$ defined as

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - \tau_{mn}), \quad m = 1, 2, \tag{1}$$

where N is the number of sources, $s_n(t)$ are the time-domain source signals, a_{mn} are weighting coefficients, representing the signal amplitude variation in propagating from source n to microphone m, and τ_{mn} are their corresponding source-to-sensor time delays. If the sources are assumed to be located in the far field, the DOA of the sources can be directly related to the inter-sensor time delays by

$$\tau_n = \tau_{2n} - \tau_{1n} = (d/c)\cos\hat{\theta}_n, \tag{2}$$

where d is the inter-microphone distance, c is the speed of sound, and $\hat{\theta}_n$ is the DOA angle of the nth source (see Fig. 1 for details).

B. Generalized cross-correlation

Considering the above model, the DOA is implicitly found in the time difference τ of the two microphone signals

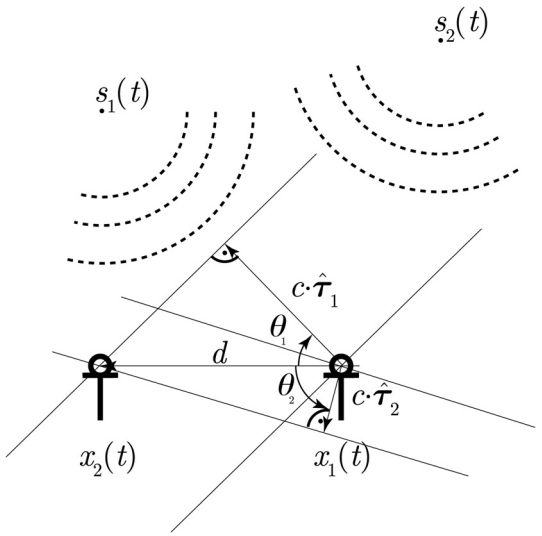

FIG. 1. Angles of incidence of two plane waves at two microphones associated with sources $s_1(t)$ and $s_1(t)$. Each plane wave produces a time delay dependent only on its corresponding source location.

$x_1(t)$ and $x_2(t)$; the estimated time delay $\hat{\tau}_n$, can be obtained with a generalized cross-correlator as

$$\hat{\tau}_n = \arg\max_{\tau_n} E\{[x_1(t) * w_1(t)][x_2(t + \tau_n) * w_2(t)]\}, \tag{3}$$

with $E\{\cdot\}$ being the statistical average over time, assuming ergodic signals, and * is the linear time convolution. Impulse responses $w_1(t)$ and $w_2(t)$ are the weighting functions applied to each microphone signal, respectively. Using this correlator, the estimated delay $\hat{\tau}$ converges to the time difference of arrival between both signals. This is then related to the angle of arrival. Given a fixed time interval for the prediction of an angle of arrival, these angles are used to construct a histogram, representing a probability distribution $H(\theta)$ of where a source is situated. When multiple speech sources are measured, based on the superposition principle, each source is represented by a certain area of the histogram.

In this paper, a PHAT weighting function is used, defined as

$$W_1(\omega)W_2(\omega) = |\Phi_{x_1x_2}(\omega)|^{-1}, \tag{4}$$

where $\Phi_{x_1x_2}(\omega)$ is the cross-power spectral density function and $W_1(\omega)$ and $W_2(\omega)$ are the Fourier transforms of $w_1(t)$ and $w_2(t)$, respectively.

Although this method is described for anechoic environments, the same formulation can be used for reverberant environments, with $x_1(t)$ and $x_2(t)$ being the result of convolution of the source signals with their corresponding room impulse responses. It has been proven to work very well under low noise environments, even when the reverberation of the room is high.[16]

C. Histogram model

The probability function of source localization, or histogram, $H(\theta)$, can be modeled by a mixture of Laplacian

distributions $M(\theta)$ to represent the angles resulting from a scatter plot of a two-channel mixture,[20] computed from delay estimates (see Fig. 2 for an example of histograms for illustrative purposes)

$$M(\theta) = \sum_{n=1}^{N} A_n e^{-|\theta_k - \mu_n|/\sigma_n}, \tag{5}$$

where $\sum_{n=1}^{N} A_n/2\sigma_n = 1$ and N represents the number of sound sources.

Note that available information is only given for discrete angles θ_k in the angular space $\boldsymbol{\theta}$. Therefore, $H(\theta)$ and M are actually a probability mass function.

One of the key points of this work is to determine the number of sound sources N. A different N represents a different model $M(\theta, \Theta)$, where Θ represents the parameters of the model; in Bayesian analysis, this corresponds to the high-level inference, termed *model selection*. Given a maximum number of potential sources, model selection will estimate the number of sources consistent with the data and prior information. At the same time, the parameters $\Theta = \{\boldsymbol{\mu}; \boldsymbol{\sigma}; \mathbf{A}\}$, where $\boldsymbol{\mu}$ is the vector of means of angles to be estimated, $\boldsymbol{\sigma}$ is the vector of angular variances and $\mathbf{A}$ of Laplacian amplitudes will be also estimated, which can be used to determine where the speech sources are situated. In Bayesian analysis, this is known as *parameter estimation* or low-level inference.

In these real scenarios, $H(\theta)$ should account for noise presence. Depending on the geometrical/reflective particularities of the scenario, the noise signals in each microphone can be either correlated or uncorrelated, having different effects on $H(\theta)$.[21] In the ideal case the noise should be uncorrelated. The degree of correlation in real rooms depends mainly on the scattering effects of the surrounding walls and microphone separation. Unfortunately, uncorrelated noise is a highly ideal case, where a perfectly diffuse sound field is required, and the sound field in any real room differs in fundamental aspects from a diffuse field.[19] On the contrary, correlated noise signals might be evidenced as new speech sources coming from some particular directions, thus being hard to distinguish from real speech sources in some cases.[21] This will mainly depend on the level of noise correlation. It also depends, at the same time, on the geometrical/ reflective characteristics of the room under investigation. Section V B 2 will elaborate on this fact.

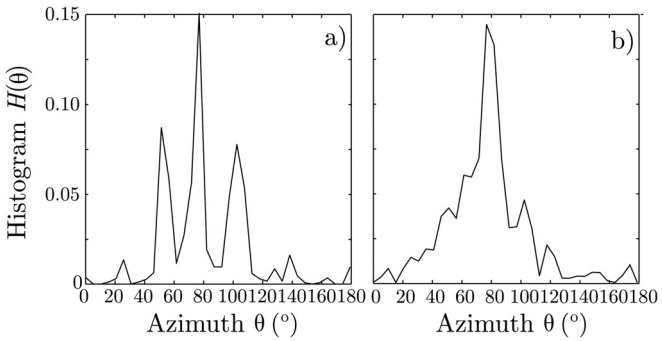

FIG. 2. Histograms obtained from real measurements with three active sources and (a) SNR = 40 dB and (b) 10 dB.

D. Additional remarks

In building an appropriate model, some additional consideration should be taken into account to reduce the search space and to speed up the algorithms explained in the following sections. In particular, since just two microphones are used, the system is unable to distinguish the ambiguity for those sources located at $\theta > 180°$. Therefore, the histogram will be built on the basis of $M(\theta) = M(\theta + 180°)$. As a consequence, $\mu_n \in [0,180]\ \forall n$.

In some applications, especially for those involving human speakers, i.e., video conference, it is straightforward to assume there will be a minimum angular separation between them, meaning there will not be any strong overlap between Laplacian functions. Therefore, $||\mu_i - \mu_j|| > \gamma$, $\forall i \neq j$, with γ being the minimum separation, which may be considered for certain applications.

III. BAYESIAN INFERENCE

A. Parameter estimation

Bayesian inference is extensively based on the Bayes' theorem. For a given model M and a given dataset $\mathbf{D}$, being a vector with K components as a function of an arrival angle vector $\boldsymbol{\theta}$, the *posterior probability distribution* of the model parameters $\Theta = \{\boldsymbol{\mu}; \boldsymbol{\sigma}; \mathbf{A}\}$ is calculated as follows:

$$p(\boldsymbol{\Theta}|\mathbf{D}, \mathrm{M}) = \frac{p(\mathbf{D}|\boldsymbol{\Theta}, M)p(\boldsymbol{\Theta}|M)}{p(\mathbf{D}|M)}. \tag{6}$$

The term $p(\mathbf{D}|\boldsymbol{\Theta}, M)$ represents the *likelihood* function, indicating the resemblance of the data $\mathbf{D}$ for a given parameter set $\boldsymbol{\Theta}$ to the model M. Prior to analysis, the error components are only known to be of a finite amount of energy. With this being the only information available, applying the principle of maximum entropy[22] leads to an assignment of $p(\mathbf{D}|\boldsymbol{\Theta}, M)$ as a Gaussian distribution of data $\mathbf{D}$. The maximum-entropy assignment of a Gaussian probability density function follows the fact that no further information about errors is available, other than a finite variance of the error.[22] After marginalizing over an unknown error variance, the assigned likelihood function becomes a Student t-distribution[23]

$$p(\mathbf{D}|\Theta, M) \equiv \mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{2}\Gamma\left(\frac{K}{2}\right)\left(\frac{\zeta(\boldsymbol{\Theta})}{2\pi}\right)^{-K/2}, \tag{7}$$

where $\zeta(\boldsymbol{\Theta}) = \sum_{k=1}^{K} ||D(\theta_k) - H(\theta_k)||^2, \Gamma(\cdot)$ is the gamma function, and θ_k is the kth element of angle vector $\boldsymbol{\theta}$.

The term $p(\boldsymbol{\Theta}|M)$ corresponds to the prior distribution of the parameters. This distribution is usually assigned uniform to avoid any subjective preference. In this particular problem, if there exist some bounds on the parameters values, as remarked in Sec. II D, they have to be introduced to the problem through the prior distribution, i.e., $0° \leq \mu \leq 180°$.

The term $p(\mathbf{D}|M)$ is known as a marginal likelihood, or *Bayesian evidence* or just evidence. In most parameter estimation problems, the evidence is a normalization constant,

but it plays a fundamental role in model selection, as will be elaborated in the following subsection. In order to act as a normalization constant, evidence Z is calculated as

$$p(\mathbf{D}|M) \equiv Z = \int_{\mathbf{\Theta}} p(\mathbf{D}|\mathbf{\Theta}, M) p(\mathbf{\Theta}|M) d\mathbf{\Theta}. \quad (8)$$

Once the posterior probability function is calculated, the mean parameters $\langle \mathbf{\Theta} \rangle$ are calculated via

$$\langle \mathbf{\Theta} \rangle = \int_{\mathbf{\Theta}} \mathbf{\Theta} p(\mathbf{\Theta}|\mathbf{D}, M) d\mathbf{\Theta}. \quad (9)$$

B. Model selection

This section discusses some basic ideas regarding the model selection in Bayesian inference. Two approaches to solve the model selection problem are reviewed. The following section will expose some of their characteristics and advantages. Additional details about both approaches for the model selection can be found in Ref. 24.

1. Bayesian evidence

According to Bayes' theorem, the posterior probability of a model M_i, given data $\mathbf{D}$ is given by

$$p(M_i|\mathbf{D}) = \frac{p(\mathbf{D}|M_i) p(M_i)}{p(\mathbf{D})}. \quad (10)$$

The idea behind the model selection is to compare the posterior probability of a set of competing models and to select the one with the highest posterior probability given the data. Given two models M_i and M_j, the posterior ratio or Bayes' factor B_{ij}, is defined as

$$B_{ij} = \frac{p(M_i|\mathbf{D})}{p(M_j|\mathbf{D})} = \frac{p(\mathbf{D}|M_i) p(M_i)}{p(\mathbf{D}|M_j) p(M_j)} = \frac{p(\mathbf{D}|M_i)}{p(\mathbf{D}|M_j)}, \quad (11)$$

when assigning the competing models equal prior probability, i.e., no model hypothesis is favored against the other, $p(M_i) = p(M_j)$; the model selection is determined in terms of the likelihood function $p(\mathbf{D}|M_i)$. The likelihood function in the model selection is exactly the marginal likelihood function or Bayesian evidence term in the parameter estimation task [see Eq. (8)]. Therefore, model selection can be carried out just comparing evidences obtained within the effort of parameter estimation.

However, in any estimation of the evidence value it is necessary to define the Bayesian ratio to favor one model over another. For this purpose, Jeffreys' scale was introduced as a way of quantifying whether or not one model is significantly better than another.[25] In practical applications, it is often considered that a model M_i is favored over a model M_j when the former model overcomes the latter by at least 10 decibans, i.e., 10 $\log_{10} B_{ij}$, which in terms of Jeffrey's scale (see Table I) indicates the evidence of model M_i is sufficiently stronger than the one obtained by model M_j. In any case, Jeffreys' scale has to be interpreted not as a calibration of the Bayes' factor rather as a qualitative, descriptive statement about standards of evidence in scientific investigations. Jeffreys' categorization has to be considered and interpreted on the context of its applicability. As it will be demonstrated later, the model selection based on the strongest evidence is the appropriate distinguishing indicator.

TABLE I. Jeffreys' scale for the Bayesian ratio expressed in *decibans*, which are defined as 10 $\log_{10} B_{ij}$.

Evidence (decibans)	*Strength of evidence*
<0	Negative (supports M_j)
0–5	Barely worth mentioning
5–10	Substantial
10–15	Strong
15–20	Very strong
>20	Decisive

In order to simplify model selection via the evidence, the following procedure is proposed:

(1) Select the model $M|_{\max Z(\text{decibans})}$ with the highest evidence expressed in decibans.
(2) For those simpler models (less parameters) which differ from $M|_{\max Z(\text{decibans})}$ by less or equal to 10 decibans, choose the simplest model.

The same procedure can be followed using different log-scales of Bayes' factor, i.e., in nepers and in this case, a model is stronger compared to another one if $\log(B_{ij}) > 2.3$.

Bayesian model selection favors simpler models over overly complex models which fit data better, which intrinsically represents a quantitative implementation of the principle of Ockham's razor.[26]

2. Bayesian information criteria

The main difficulty of dealing with evidence lies in the intractability of Eq. (8), since it cannot be solved analytically even when the model has a small number of parameters. An alternative approach can be used to rank models. The Bayesian information criterion (BIC) or Schwarz criterion is another criterion for model selection. Schwarz derived BIC to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. In large-sample settings, the fitted model favored by BIC ideally corresponds to the candidate model which is *a posteriori* the most favorable; i.e., the model which is rendered most plausible by the available data. Given a finite set of models, it is possible to increase the likelihood just by adding parameters, but this may result in overfitting. The BIC resolves this problem by introducing a penalty term proportional to the number of parameters in the model. This criterion is defined as

$$\text{BIC}_i = 2 \log p(\mathbf{D}|\mathbf{\Theta}, M_i)_{\max} - \eta_i \log K, \quad (12)$$

where η_i is the number of parameters used by the model M_i, and $p(\mathbf{D}|\mathbf{\Theta}, M_i)_{\max}$ is the maximized likelihood. The higher the BIC, the higher the probability the data $\mathbf{D}$ were generated by this model. Note this definition differs from the usual definition by a negative sign,[27] in order to facilitate the

comparison with the evidence (see Sec. III B 1). The BIC may be regarded as an empirical approximation to the log-Bayes' factor and its computation does not require the specification of priors. Thus, BIC has appeal in many Bayesian modeling problems.

The BIC approach assumes, however, that the likelihood distribution of interest can be approximated by a multi-variant Gaussian in the vicinity of the global extreme. For many applications this is not the case, particularly multimodal distributions. If the shape of the likelihood distribution deviates drastically from a multi-variant Gaussian distribution, the BIC will hardly be able to correctly rank the models; therefore, in order to use the BIC, one needs to be sure there are not multiple modes near the global extreme.

Another limitation of this method is that it is necessary to calculate $p(\mathbf{D}|\boldsymbol{\Theta}, M_i)_{\max}$. Since this value is obtained via sampling the likelihood function, it may be difficult to obtain precisely when the number of samples is not large enough; the asymptotic approximation assumed by the BIC is critically sensitive to the maximum posterior estimation.

As it will be presented in the following section, some of the most popular sampling algorithm implementations for Bayesian analysis (Metropolis-Hastings and importance sampling) perform the model selection based on ranking BIC values. In general, the model selection is based on selecting the model with the highest BIC value; in addition, a similar selection criterion as described in the previous section based on Jeffreys' scale can be applied since Kass and Wasserman[28] have shown that a Bayes' factor can be approximated in terms of BIC as follows:

$$B_{ij}(\text{decibans}) \approx 10\log_{10}(e^{\text{BIC}_i}) - 10\log_{10}(e^{\text{BIC}_j}). \quad (13)$$

IV. SAMPLING METHODS

Bayesian calculation may be computationally challenging due to the fact that the evidence integral in Eq. (8) is defined over a high-dimensional parameter space. A number of Monte Carlo sampling methods can be used to cope with the computational challenges.

A. Metropolis-Hastings algorithm

The Metropolis-Hastings[29] algorithm is a Markov-chain Monte Carlo method for obtaining a sequence of random samples from a complex, non-standard probability distribution for which direct sampling is difficult, usually due to the fact that those probability distributions are in most cases, multi-dimensional.

This algorithm generates dependent samples Θ from the posterior probability distribution $p(\Theta|\mathbf{D}, H)$ using only knowledge of the likelihood and the prior distribution. The main motivation of using this approach lies on the fact that posterior distributions are not usually simple to sample; Metropolis-Hastings algorithm provides a simple method to implement and converges to a stationary distribution proportional to the posterior.

Starting from a set of random samples $\boldsymbol{\Theta}_l$, with $l = 1,\ldots,L$, the Metropolis-Hastings algorithm generates a sequence of new samples, e.g., using an uniform distribution. It uses each existing sample to generate one new sample, and this is repeated many times producing a Markov chain. A fundamental property of Markov chains is that a new sample depends only on the previous sample. At each step of the algorithm, a new sample $\boldsymbol{\Theta}^*{}_l$ in the search space is chosen with probability distribution given by $q(\boldsymbol{\Theta}^*{}_l, \boldsymbol{\Theta}_l)$ which is a candidate-generating density that is irreducible and aperiodic, in addition to a user-defined proposal distribution. In other words, $q(\boldsymbol{\Theta}^*{}_l, \boldsymbol{\Theta}_l)$ represents the probability of proposing a sample $\boldsymbol{\Theta}^*{}_l$ given a previous sample $\boldsymbol{\Theta}_l$. The new sample $\boldsymbol{\Theta}^*{}_l$ is accepted, i.e., $\boldsymbol{\Theta}_{l+1} = \boldsymbol{\Theta}^*{}_l$ with a probability

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{\Theta}_l^*|\mathbf{D}, M)q(\boldsymbol{\Theta}_l, \boldsymbol{\Theta}_l^*)}{p(\boldsymbol{\Theta}_l|\mathbf{D}, M)q(\boldsymbol{\Theta}_l^*, \boldsymbol{\Theta}_l)}\right\}, \quad (14)$$

but if it is not accepted, the sample remains the same, i.e., $\boldsymbol{\Theta}_{l+1} = \boldsymbol{\Theta}_l$. Finally, the parameter estimation is carried out by averaging the accepted samples. To perform the model selection in terms of the Metropolis-Hastings algorithm, it makes use of the BIC (see Sec. III B 2).

B. Importance sampling

The point of importance sampling is that sampling from a uniform distribution can be very inefficient and it can be much better to concentrate non-uniform sampling on high probability (important) regions of the parameter space. To this end, the parameter estimates using the importance sampling are defined by

$$\langle\boldsymbol{\Theta}\rangle \approx \frac{1}{L}\sum_{l=0}^{L-1}\boldsymbol{\Theta}_l w(\boldsymbol{\Theta}_l), \quad (15)$$

with $w(\boldsymbol{\Theta}_l) = p(\boldsymbol{\Theta}_l|\mathbf{D}, M)/g(\boldsymbol{\Theta}_l)$.

The key point of Eq. (15) is to sample a simpler function $\boldsymbol{\Theta}_l \sim g(\boldsymbol{\Theta})$ instead of $p(\boldsymbol{\Theta}|\mathbf{D}, M)$. The only condition is that $g(\boldsymbol{\Theta})$ should have the same support as $p(\boldsymbol{\Theta}|\mathbf{D}, M)$. However, the main difficulty lies in finding the appropriate function $g(\boldsymbol{\Theta})$, especially in high dimensions.[30]

This difficulty lies, in practice, in the lack of prior knowledge on the actual posterior probability density function shape.[23] Therefore, selecting $g(\boldsymbol{\Theta})$ without any prior information will often lead to a failure in solving Eq. (15). Since the region of high probability becomes exponentially small in higher dimensions, it takes a very large number of samples to get several that actually contribute to the sum.

C. Nested sampling

An alternative approach is to calculate the Bayesian evidence using a sampling algorithm, termed *nested sampling*.[31] The nested sampling approximates these marginalization integrals, while at the same time sampling the posterior

distribution $p(\boldsymbol{\Theta}|\mathbf{D}, H)$. Comprehensive tutorials and practical details on the nested sampling may be found in.[31–33]

The basic idea behind the nested sampling is to rearrange Eq. (8) as a one-dimensional integral, by considering a constrained *prior mass*, $\xi(\lambda) \in [0,1]$, that represents the amount of prior in the region where the likelihood is greater than a certain value λ.

One of the main contributions of the nested sampling lies in taking into account that prior mass ξ can be accumulated from its differential elements $d\xi$, so let us define

$$\xi(\lambda) = \int_{\mathcal{L}(\boldsymbol{\Theta})>\lambda} p(\boldsymbol{\Theta}|H)d\boldsymbol{\Theta} \tag{16}$$

as the cumulant prior mass covering all likelihood values greater than λ. As λ increases, the enclosed mass ξ decreases from 1 to 0. The evidence may be rewritten,[32]

$$Z = \int_0^1 \mathcal{L}(\xi)d\xi. \tag{17}$$

This one-dimensional integral can be solved numerically,

$$Z \simeq \sum_{i=1}^{\infty} \mathcal{L}_i \Delta\xi_i \quad \text{with} \quad \Delta\xi_i = \xi_{i-1} - \xi_i, \tag{18}$$

where $\xi_0 = 1$, $\xi_\infty = 0$, and $\mathcal{L}_\infty = \mathcal{L}_{\max} < \infty$. This procedure of solving an integral is analogous to Lebesgue integration in the mathematical literature.[33]

Starting with a set of L initial random samples, $\boldsymbol{\Theta}_l$ sampled from the prior distribution, and their associated likelihoods $\mathcal{L}_l$, where $l \in [1, L]$, the parameters with the lowest likelihood value, labeled $[\boldsymbol{\Theta}_1, \mathcal{L}_1]$, is stored and replaced by a new random parameter $\boldsymbol{\Theta}_{\text{new}}$ under the only constraint $\mathcal{L}_{\text{new}} > \mathcal{L}_1$, leaving L samples.[32] The process is repeated iteratively selecting a new sample with the lowest likelihood and generating a new one with higher likelihood. Repeating this process, the evidence is accumulated according to Eq. (18). At the same time the evidence is accumulated, in each iteration a new sample $\boldsymbol{\Theta}_l$ is generated, i.e., the parameter samples can be easily sampled from standard distributions, e.g., from a uniform distribution, with the only restriction of $\mathcal{L}(\boldsymbol{\Theta}_l) > \mathcal{L}(\boldsymbol{\Theta}_{l-1})$. This should be considered as an important advantage over other methods since the nested sampling efficiently performs the model comparison. Usually the BIC-based evidence evaluation is at least an order of magnitude more costly than the parameter estimation,[34] especially for those very sharp likelihood function, where the accuracy on estimating $p(\mathbf{D}|\boldsymbol{\Theta}, M)_{\max}$ is critical. The main advantage of this method lies in enabling the two levels of inference at the same time.

For practical implementations, elementary prior mass $\Delta\xi_i$ can be statistically approximated by $\Delta\xi_i \approx e^{-1/L}$. Equation (18) will keep accumulating up to $\log(Z_i) - \log(Z_{i-1}) < \delta$, with $Z_i = \mathcal{L}_i\Delta\xi_i$, with δ being a predefined threshold value.

The bottleneck of this algorithm lies, however, in generating new samples under the hard constraint $\mathcal{L}(\boldsymbol{\Theta}_l) > \mathcal{L}(\boldsymbol{\Theta}_{l-1})$. An efficient way to generate new samples under that constraint is using a *ellipsoidal nested sampling*,[34] consisting of a clustering nested sampler which is capable of detecting and isolating multiple separated regions of high likelihood, fitting separate ellipsoidal bounds around each region; or *multimodal nested sampling*,[35,36] based also on ellipsoidal clustering, but using a X-means clustering algorithm to estimate the number of modes in the likelihood distribution, instead of using K-means where the number of modes needs to be previously known.

Finally, the repeatability of the nested sampling's efficacy will be demonstrated through several runs. Furthermore, the uncertainty of the log-evidence can be also estimated in a single run.[31] The log-evidence uncertainty is inversely proportional to $\sqrt{L}$.

V. EXPERIMENTAL RESULTS

In this section, several experiments are performed to evaluate the nested sampling algorithm under different scenarios, by varying speaker configurations, SNRs and reverberation times. Moreover, aspects related to the use of correlated noise are investigated, together with a qualitative comparison to other sampling methods. It should be clarified that practical applications, e.g., teleconferencing, only deal with a limited number of potential sources at the same time. For this reason and for computational cost reduction, this paper deals with a maximum of five simultaneous sources.

A. Experimental setup

In order to validate the methodology described above two rooms were used during the investigation: the first room [see Fig. 3(a)] with a volume of 248.64 m^3 and its measured reverberation time is 1 s; whereas the second room has a

FIG. 3. Rooms used to perform the experiments: (a) Multimedia and Multimodal Processing research laboratory and (b) office room in the University of Jaén (Spain).

volume of 115.92 m^3 and 0.52 s of reverberation time [see Fig. 3(b)]. A two-microphone array with a separation of 13.5 cm between microphones was placed in the middle of each room. The array consists of two omni-directional AKG-C417PP microphones, and their output signals are sampled at 44.1 kHz. In both scenarios, the same source–receiver configuration is used. Four speakers were distributed over an angular range between 0 and 180 deg, following the scheme presented at Table II. In order to provide different configurations, different numbers of speakers were used.

In each scenario, different texts were read by several speakers at the same time and then recorded simultaneously. In each room, the same three scenarios were used: *two speakers* located at angles $\hat{\theta}_1$ and $\hat{\theta}_2$; *three speakers* located at angles $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$; and *four speakers* located at the angles described in Table II. After each recording, the speech signals were low-pass filtered to 5 kHz and in all cases, the SNR was modified by adding an uncorrelated white noise signal in each microphone and modifying its power to obtain different SNRs. For each room and each scenario SNRs were 40, 30, 25, 20, and 10 dB. Each experiment is denoted by $R^n_{r_s}$, where r is the room, s denotes the number of active sources, and n is the SNR in dB. The maximum value of the normalized cross-correlation function is measured with a value of 0.08, meaning both signals are practically uncorrelated.

The two microphone signals are processed using the GCC-PHAT algorithm to obtain the histogram[15] using a Hann window of 25 ms and 50% overlap for each observation. The **D** vector length, has been set as the square root of the number of observations, in order to obtain smooth histograms without losing relevant information, in this case it has been set to $K = 30$. Alternatively, smooth histograms can be constructed using different approaches[28] without varying the methodology introduced in this paper. It should also be highlighted that this does not mean the results have a resolution of 6°; the localization is made on the basis of the means $\boldsymbol{\mu}$, but not on the maxima of the histogram.

B. Evaluation of nested sampling

1. Model selection

Regarding the nested sampling algorithm, the initial number of random samples has been set to $M = 5000$ samples. The purpose of using such number of initial samples is to minimize the effects of uncertainty.[31] The stopping condition has been set for the threshold to be $\delta < 10^{-4}$. Moreover, the sources are assumed to be not closer than $\gamma = 10°$. The models under consideration, i.e., number of sources, are set up to $N = 5$. For each scenario, the algorithm is run 30 times, and the mean and the deviation of log-evidence values are computed. In highlighting the numerical differences between models, Table III lists the log-evidence results. For each experiment, the selected model according to Sec. III B 1 is indicated in bold type when correctly ranked, whereas if the selected model is not properly ranked, it is indicated in italic type.

The results show that the models are correctly ranked for each one of the different number of sources when SNR ∈ [25,40] dB for the first room, whereas for the second room, the range of correct inferred models is SNR ∈ [20,40] dB. From lower SNRs, except for very few cases, the model is not correctly ranked.

TABLE II. Speakers angular distribution ($\hat{\theta}_i$) around the two-microphone array, the distance to the array is indicated between parenthesis.

$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
53° (1.65 m)	76° (2 m)	104° (2 m)	127° (1.65 m)

The reason the model is well ranked in R_2 for SNR = 20 dB whereas it is not for R_1 lies in the differences in room conditions. As it was mentioned previously, the GCC-PHAT method works well with low SNRs and a moderate reverberation; the reverberation affects the histogram displaying a number of peaks increasing with the reverberation time. Although GCC-PHAT is considered a robust method in rooms with relatively high reverberation, a reverberation time of 1 s in the first room is considered long in the GCC-PHAT literature.

Jeffreys' scale provides a decision rule that correctly ranks the models. If the model selection was made on the basis of the highest log-evidence value in Table III, the results would be wrong in most cases. Through the results, it is evidenced how the model selection on the basis of the stronger evidence (10 decibans), it accomplishes this task adequately. This procedure has allowed the quantitative implementation of the principle of Ockham's razor.

2. Effects of partially correlated background noise

The direction-of-arrival literature focuses mainly on evaluating algorithms in the ideal case of uncorrelated noise presence. In this section, it is analyzed through several examples how partially correlated noise signals affect the estimation on the number of sources. In order to perform the experiment, a dodecahedral loudspeaker is used as a noise source and its power is varied to obtain the desired SNR. The loudspeaker is situated at angle 38° and 2.5 m distance. The maximum value of the normalized cross-correlation function between both microphone signals is 0.2527 and 0.2774 for R_1 and R_2, respectively. Both values are considerably low, meaning noise signals are fairly uncorrelated. (See Fig. 4.)

Figure 5 shows the log-evidence for different models with varying the SNR from 30 to 10 dB when different numbers of sources are active. For both rooms, the nested sampling correctly ranks in all cases up to SNR = 25 dB in all cases for both rooms. In some cases it also correctly ranks models when SNR = 20 dB, but it seems not to be generalized. Since both rooms have strong scattering surfaces, making the noise signals nearly uncorrelated, a similar behavior to the uncorrelated noise case is observed. Therefore, all the results exposed in previous sections may be extrapolated to those cases where the noise signals are nearly uncorrelated. However, these results cannot be generalized to cases where correlated noise signals may affect the detection performance (see Ref. 21 for some examples in the uni-modal case).

TABLE III. Average log-evidence values and their corresponding standard deviation for the five competitive models in each experiment. Model correctly ranked are indicated in bold type, whereas wrongly ranked models are indicated in italic type letter. Each experiment is denoted by $R^n_{r_s}$, where r is the room, s denotes the number of active sources, and n is the SNR in dB.

	Room 1				
$R^{\text{SNR}}_{r_s}$ Model	1	2	3	4	5
$R^{40}_{1_2}$	59.26 (0.26)	**152.04 (1.96)**	152.89 (2.12)	153.32 (2.7)	150.43 (1.81)
$R^{40}_{1_3}$	−7.40 (0.22)	37.24 (1.25)	**137.35 (5.15)**	140.22 (2.9)	142.62 (4.02)
$R^{40}_{1_4}$	−16.65 (0.23)	25.57 (0.88)	53.62 (1.27)	**123.73 (5.69)**	128.28 (2.91)
$R^{30}_{1_2}$	92.90 (0.23)	**159.45 (1.30)**	162.14 (3.27)	162.63 (2.48)	158.62 (1.60)
$R^{30}_{1_3}$	9.67 (0.29)	96.30 (1.23)	**135.161 (3.91)**	141.42 (5.40)	142.72 (4.13)
$R^{30}_{1_4}$	−20.95 (0.27)	18.13 (0.76)	49.15 (2.89)	**127.82 (5.38)**	130.06 (6.01)
$R^{25}_{1_2}$	155.73 (0.26)	**256.15 (1.92)**	256.69 (4.40)	259.15 (3.99)	261.18 (5.11)
$R^{25}_{1_3}$	11.95 (0.08)	91.14 (1.53)	**132.44 (5.11)**	140.52 (4.18)	139.15 (0.94)
$R^{25}_{1_4}$	−21.32 (0.26)	25.34 (0.87)	61.08 (0.88)	**116.79 (4.97)**	124.83 (2.39)
$R^{20}_{1_2}$	150.03 (0.33)	201.95 (2.00)	*294.58 (6.07)*	292.88 (6.30)	301.33 (9.18)
$R^{20}_{1_3}$	11.95 (0.31)	85.33 (1.38)	114.05 (2.62)	*166.84 (4.73)*	174.12 (4.21)
$R^{20}_{1_4}$	−46.70 (0.20)	4.48 (0.70)	45.39 (3.21)	84.13 (2.15)	*109.80 (6.83)*
$R^{10}_{1_2}$	47.75 (0.33)	69.69 (1.91)	*76.37 (1.50)*	79.18 (2.56)	80.97 (3.49)
$R^{10}_{1_3}$	−22.05 (0.23)	78.08 (1.20)	99.15 (1.93)	130.08 (3.73)	*141.73 (4.69)*
$R^{10}_{1_4}$	−30.19 (0.32)	30.31 (1.33)	71.89 (3.03)	**104.62 (6.88)**	113.74 (5.09)
	Room 2				
E_N (SNR)/Model	1	2	3	4	5
$R^{40}_{2_2}$	74.14 (0.22)	**245.37 (2.90)**	242.48 (2.56)	238.75 (4.42)	235.85 (3.58)
$R^{40}_{2_3}$	36.95 (0.25)	66.18 (0.80)	**174.52 (3.01)**	177.39 (1.51)	175.16 (4.30)
$R^{40}_{2_4}$	5.06 (0.32)	34.94 (1.02)	86.84 (6.06)	**187.03 (9.82)**	195.78 (5.85)
$R^{30}_{2_2}$	120.78 (0.61)	**236.64 (1.35)**	238.27 (4.65)	238.02 (2.83)	235.83 (3.97)
$R^{30}_{2_3}$	64.17 (0.40)	105.61 (1.49)	**189.79 (6.35)**	191.20 (4.39)	194.50 (3.86)
$R^{30}_{2_4}$	40.92 (0.25)	72.04 (1.17)	120.82 (3.26)	**228.46 (10.33)**	232.78 (7.25)
$R^{25}_{2_2}$	108.01 (0.10)	**142.06 (0.81)**	140.64 (0.60)	137.56 (0.93)	134.09 (0.72)
$R^{25}_{2_3}$	125.99 (0.29)	145.70 (0.84)	**163.68 (2.30)**	160.20 (2.04)	164.70 (4.24)
$R^{25}_{2_4}$	56.36 (0.13)	79.88 (1.21)	101.28 (3.89)	**128.73 (3.16)**	134.98 (5.49)
$R^{20}_{2_2}$	116.66 (0.41)	**142.49 (1.31)**	144.03 (2.40)	141.58 (1.65)	139.73 (3.01)
$R^{20}_{2_3}$	88.65 (0.29)	132.51 (1.85)	**144.95 (2.49)**	146.63 (3.51)	148.10 (3.74)
$R^{20}_{2_4}$	49.37 (0.28)	75.32 (1.12)	107.93 (2.89)	**142.10 (4.51)**	144.99 (6.11)
$R^{10}_{2_2}$	63.67 (0.31)	79.59 (1.05)	*88.44 (3.71)*	92.97 (5.19)	93.96 (3.29)
$R^{20}_{2_3}$	92.20 (0.42)	119.29 (1.65)	131.94 (4.02)	*142.48 (5.15)*	148.07 (3.61)
$R^{20}_{2_4}$	15.25 (0.34)	52.79 (0.64)	62.55 (2.15)	73.00 (1.36)	*84.35 (4.60)*

3. Parameter estimation

In this section, the accuracy of the parameter estimation is analyzed. When estimating the angle of arrival, it is only necessary to estimate the Laplacian means $\boldsymbol{\mu}$. In some other applications such as sound source separation, $\mathbf{A}$ and $\boldsymbol{\sigma}$ must also be known. In this paper, this estimation will be mainly focused on the mean estimation, since it is the only data experimentally available (see Table II). Table IV lists the DOA results in all configurations and scenarios. The results show a low error up to SNR = 25 dB with a relatively small variance, that increases when SNR decreases. Figure 5 shows an averaged error for all the angles in each room with a particular SNR. It shows that with SNR = 20 dB, despite the averaged error being lower than 5°, the variance becomes significantly high at both rooms. Finally, with only 10 dB SNR, the noise level is too high to consider localization accuracy. Therefore, it seems reasonable to affirm the DOA estimation should be limited to work correctly up to no more than 20 dB. This limit should be slightly higher when increasing the reverberation, for example to ≈25 dB. These facts, together with those exposed in Sec. V B 1 confirms the valid application of Bayesian inference based on the GCC-PHAT model along with nested sampling in DOA estimation in a range of SNR no lower than 20 dB.[18]

C. Comparison to other methods

In this section, some qualitative, but not exhaustive, comparisons are made between the nested sampling, Metropolis-Hastings, and importance sampling, for this particular application.

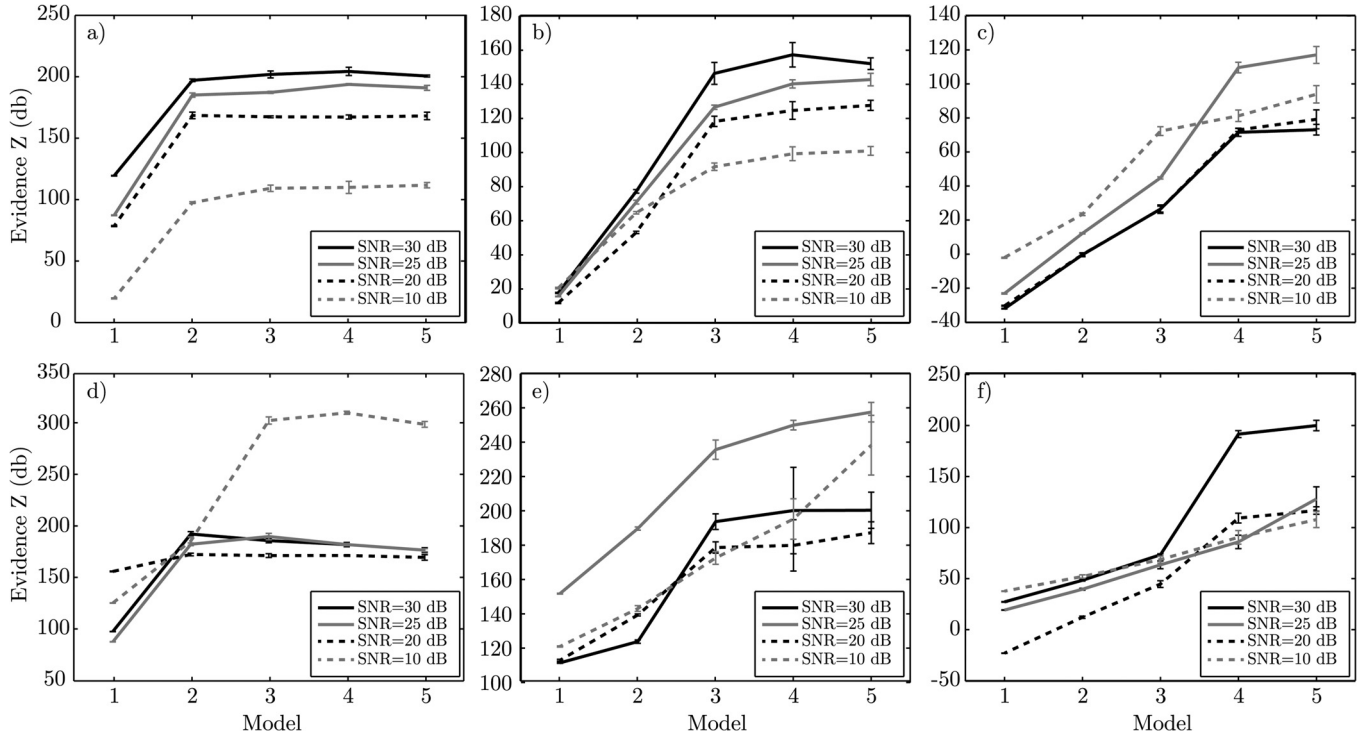

FIG. 4. Evidence at different SNRs when noise signals are partially correlated at different scenarios: (a) R_{1_2}, (b) R_{1_3}, (c) R_{1_4}, (d) R_{2_2}, (e) R_{2_3}, and (f) R_{2_4} where db indicates the unit decibans for the log-evidence.

When using the Metropolis-Hasting algorithm (see Sec. IV A) the candidate-generating density $q(\boldsymbol{\Theta}^*_m, \boldsymbol{\Theta}_m)$ is one decisive parameters to be selected. As previously mentioned, if the distribution is chosen to be symmetric, the methodology is simplified. This results in a *random walk,* which in most cases is considered as a good default option. In this paper, the use of the Metropolis-Hastings algorithm is limited to this configuration.

The main challenge with this algorithm is that, despite the fact that the algorithm eventually converges, it is difficult to estimate when and how fast it does, and the scientific literature has shown that formal convergence criteria seem not to work as expected.[37] In this paper, a selected scenario, $R^{40}_{2_2}$, is used to compare the nested sampling with Metropolis-Hastings. Figure 6 shows different BIC rating according to the number of samples used. In this case, $p(\mathbf{D}|\boldsymbol{\Theta}, H_i)_{\max}$ is calculated by finding the already generated sample with the highest likelihood value. In all cases the acceptance ratio has been ∼30% which is considered as acceptable. The BIC ranks the model correctly when the number of samples increases, and at some point the variance becomes considerably reduced. When compared to the results listed in Table III, similar variances are obtained when 10^7 samples are used in Metropolis-Hastings algorithm; however, the number of total samples used in average on the same example with the nested sampling, is approximately 5×10^4.

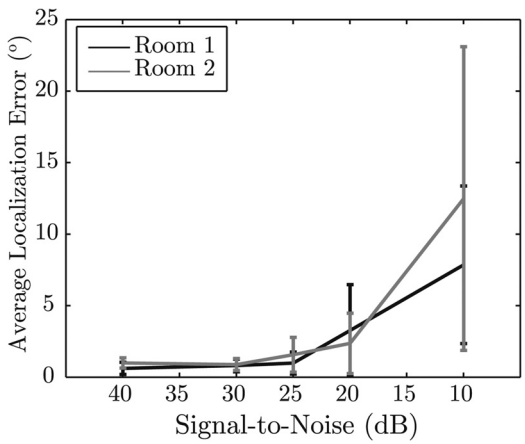

FIG. 5. Averaged errors in position for each different SNRs in each room.

Regarding the importance sampling (see Sec. IV B), the key point is to select the sampling distribution $g(\boldsymbol{\Theta})$ in such a way that it provides samples around the global maximum, i.e., the same support, of the posterior probability distribution function. The auxiliary distribution $g(\boldsymbol{\Theta})$ needs to be a standard distribution that is easy to sample from, and a bad choice of $g(\boldsymbol{\Theta})$ may result in highly inefficient sampling. Unless some background information is available, any choice may result in an inefficient solution. Figure 7 shows several examples of choosing different $g(\boldsymbol{\Theta})$ over a marginal posterior probability density function. Using the same scenario, $R^{40}_{2_2}$, in which two speakers are assumed to be active, Fig. 7 illustrates the marginalized posterior probability density function over the amplitude and variance. The ellipsoids marked in the figure conceptually indicate proposal distributions $g(\mu_1, \mu_2)$ for the importance sampling algorithm. The proposal distribution marked by A, with a support similar to the actual posterior probability density function, situated around the global maximum. This is a good choice for accurate and unbiased estimations using the importance sampling integration. The one marked by B provides a solution with a support different from the actual one, therefore the solution

TABLE IV. Estimated parameters by using nested sampling algorithm in each of the scenarios. Each experiment is denoted by $R^n_{r_s}$, where r is the room, s denotes the number of active sources, and n is the SNR in dB.

	$\tilde{\theta}_1$	$\tilde{\theta}_2$	$\tilde{\theta}_3$	$\tilde{\theta}_4$
$R^{40}_{1_2}$	52.24° (0.08)	76.38° (0.05)	-	
$R^{40}_{1_3}$	52.51° (0.07)	76.55° (0.04)	102.51° (0.01)	
$R^{40}_{1_4}$	53.00° (0.15)	76.90° (0.05)	103.29° (0.06)	127.23° (0.17)
$R^{30}_{1_2}$	51.72° (0.16)	76.57° (0.04)	-	
$R^{30}_{1_3}$	52.49° (0.24)	76.76° (0.08)	102.51° (0.02)	
$R^{30}_{1_4}$	53.04° (0.15)	77.04° (0.06)	102.94° (0.06)	127.60° (0.17)
$R^{25}_{1_2}$	53.00° (0.16)	77.60° (0.01)	-	
$R^{25}_{1_3}$	53.03° (0.31)	77.05° (0.18)	102.51° (0.01)	
$R^{25}_{1_4}$	55.17° (0.13)	76.37° (0.20)	102.57° (0.02)	127.60° (0.18)
$R^{20}_{1_2}$	53.57° (0.20)	77.54° (0.02)	-	
$R^{20}_{1_3}$	59.96° (5.98)	84.46° (6.13)	110.90° (8.77)	
$R^{20}_{1_4}$	52.65° (0.20)	77.15° (0.52)	101.80° (0.22)	128.24° (0.27)
$R^{10}_{1_2}$	55.29° (1.09)	80.60° (0.12)	-	
$R^{10}_{1_3}$	66.42° (7.72)	90.62° (6.30)	120.40° (9.73)	
$R^{10}_{1_4}$	60.02° (0.49)	82.15° (0.15)	105.96° (0.97)	131.22° (1.43)
$R^{40}_{2_2}$	53.75° (0.02)	77.40° (0.00)	-	
$R^{40}_{2_3}$	54.24° (0.05)	77.24° (0.03)	103.44° (0.36)	
$R^{40}_{2_4}$	54.10° (0.24)	76.80° (0.20)	105.35° (0.03)	127.47° (0.00)
$R^{30}_{2_2}$	53.42° (0.11)	77.34° (0.01)	-	
$R^{30}_{2_3}$	53.73° (0.15)	77.21° (0.03)	104.22° (0.21)	
$R^{30}_{2_4}$	53.92° (0.31)	77.47° (0.00)	104.95° (0.01)	127.64° (0.34)
$R^{25}_{2_2}$	52.55° (0.38)	79.21° (0.03)	-	
$R^{25}_{2_3}$	54.38° (0.21)	79.84° (0.03)	104.74° (0.05)	
$R^{25}_{2_4}$	54.46° (0.60)	77.71° (0.18)	103.72° (0.60)	128.04° (0.52)
$R^{20}_{2_2}$	55.06° (0.28)	80.75° (0.03)	-	
$R^{20}_{2_2}$	56.48° (0.79)	82.39° (0.42)	105.40° (0.94)	
$R^{20}_{2_4}$	52.50° (1.08)	77.64° (0.08)	104.01° (0.60)	128.00° (0.40)
$R^{10}_{2_2}$	67.59° (2.28)	90.50° (1.07)	-	
$R^{10}_{2_3}$	45.25° (17.30)	78.04° (8.56)	104.48° (7.67)	
$R^{10}_{2_4}$	16.40° (13.80)	60.52° (9.01)	90.40° (7.89)	119.63° (7.60)

obtained from this sampling distribution is far from being accurate. The last ellipsoid marked by C has partial support, located around a local maximum; the integration will lead to an erroneous solution. Therefore, using the importance sampling for this particular application seems to be highly inefficient.

Clearly, the nested sampling has advantages over the importance sampling since few parameters need to be tuned. However, a combination of both methods may be interesting: once the nested sampling has located an approximate maximum, this could facilitate finding a $g(\boldsymbol{\Theta})$ function with the right support and then to sample around the global maximum, leading to more accurate results in both the parameter estimation and the model selection.

VI. CONCLUSIONS

In this paper, a thorough analysis of a multi-speaker localization method using a two-microphone array based on a combination of the generalized cross-correlation method with phase-transformation (GCC-PHAT) and the nested sampling is presented. This is made through an extensive campaign of measurements in real spaces under different configurations. The main goal of this investigation is to establish a methodology to estimate the DOA when the number of speech sources is an unknown parameter, together with their position. The results obtained have substantiated that the nested sampling method works correctly under a SNR ≤ 20 dB in a relatively high reverberant environment. However, when the reverberation time increases, the histogram progressively degenerates and the range of validity in terms of SNR decreases. The main limitation of this method lies in the limitations of the GCC-PHAT-model itself, rather than the sampling method used in the Bayesian framework. In other words, for low SNR and moderately high reverberation time, the histograms obtained by the GCC-PHAT algorithm cannot be longer modeled as a Laplacian mixture model.

Additionally, it has been empirically demonstrated that in order to avoid detecting strong reflections as additional sources, complex models that differ by less or equal to 10 decibans in evidence should not be considered; in other words, unless the evidence of a new source was "strong"

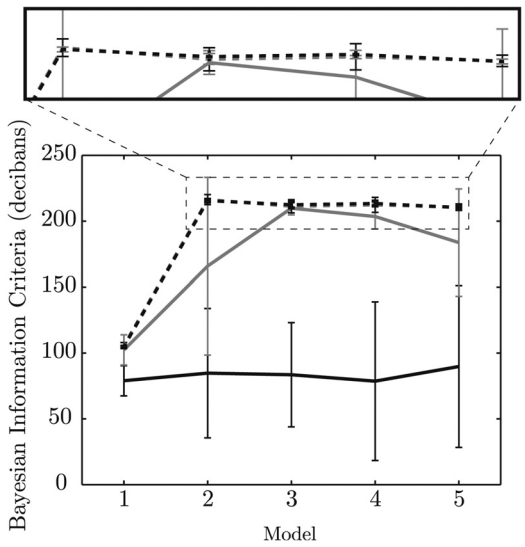

FIG. 6. Bayesian information criteria (BIC) ranks depending on the numbers of samples initialized. Solid black line represents that 10^4 samples have been used, solid gray line represents when using 10^5 samples, dashed black line when using 10^6 samples, and dashed gray line when using 10^7 samples. Note that in average cases with 10^6 and 10^7 are nearly the same; however the deviation is considerably reduced for the case with 10^7 samples, as evidenced in the zoom part of the figure.

enough to be considered, the simplest model should be selected.

In addition to the evaluation of the nested sampling applied to the sound source detection, it has also been compared to other popular sampling methods such a Metropolis-Hastings and importance sampling. The nested sampling outperforms these methods, not only in terms of computational cost, but also simplicity of setting tuning parameters and prior information. The nested sampling produces a normalized posterior, suitable for both levels of inference.

Although real-time applications are beyond the scope of this paper, nested sampling has some advantages and may make this method suitable. Moreover, fewer parameters need to be tuned and they can be selected to reduce the computational time. Furthermore, modern parallel processors will be suitable to perform several model evaluations at the same time. This issue should be addressed in future research efforts.

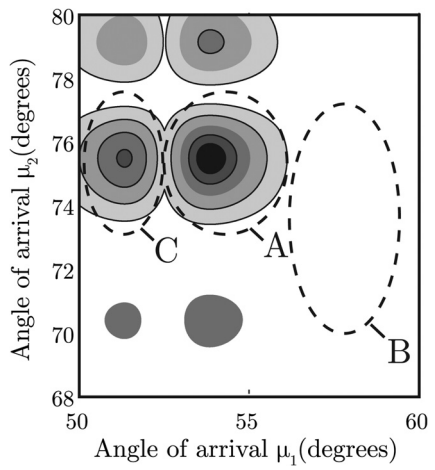

FIG. 7. Marginalized posterior probability density function over amplitude and variance, using E_2(40 dB) at R_2. Dashed ellipsoids indicating proposal distributions for importance sampling integration. Proposal distributions marked by *A*, with a support similar to the posterior probability density function results in accurate results. Proposal distribution marked by *B* is not longer working since it has a different support, whereas *C* has a support around a local maximum: both approaches will lead to inaccuracies.

ACKNOWLEDGMENTS

The authors are highly grateful to Dr. Jonathan Botts, Dr. Tomislav Jasa, Torben Pastore, and Cameron Fackler for their insightful comments. The work of J.E., M.C., and J.J.L. has been supported by the Spanish Ministry of Economy and Competitiveness supported this work under the projects TEC2012-37945-C02- 01/01 and TEC2012-37945-C02-01/02. The work of J.M.P.L. has been supported by the Spanish Ministry of Economy and Competitiveness under the project TIN2012-38079-C03-03.

[1]N. Madhu and R. Martin, "Acoustic source localization with microphone arrays," *Advances in Digital Speech Transmission* (Wiley, UK, 2008), pp. 135–166.
[2]S. E. Dosso and M. J. Wilmut, "Bayesian multiple-source localization in an uncertain ocean environment," J. Acoust. Soc. Am. **129**, 3577–3589 (2011).
[3]S. E. Dosso and M. J. Wilmut, "Bayesian tracking of multiple acoustic sources in an uncertain ocean environment," J. Acoust. Soc. Am. **133**, EL274–EL280 (2013).
[4]Z.-H. Michalopoulou, "Multiple source localization using a maximum a posteriori Gibbs sampling approach," J. Acoust. Soc. Am. **120**, 2627–2634 (2006).
[5]J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," EURASIP J. Appl. Signal Process. **2006**, 1–19 (2006).
[6]C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust., Speech Signal Process. **24**, 320–327 (1976).
[7]C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," Signal Process. **92**, 1950–1960 (2012).
[8]C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," J. Acoust. Soc. Am. **108**, 1888–1905 (2000).
[9]T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," IEEE Trans. Audio, Speech Lang. Process. **19**, 1–13 (2011).
[10]C. Faller and J. Merimaa, "Source localization in complex listening situations: selection of binaural cues based on interaural coherence," J. Acoust. Soc. Am. **116**, 3075–3089 (2004).
[11]W. Zhang and B. Rao, "A two microphone-based approach for source localization of multiple speech sources," IEEE Trans. Audio Speech Lang. Process. **18**, 1913–1928 (2010).
[12]M. Cobos and J. J. López, "Two-microphone separation of speech mixtures based on interclass variance maximization," J. Acoust. Soc. Am. **127**, 1661–1672 (2010).
[13]M. Cobos and J. J. López, "Two-microphone multi-speaker localization based on a Laplacian mixture model," Digit. Signal Process. **21**, 66–76 (2011).
[14]S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," J. Acoust. Soc. Am. **123**, 2136–2147 (2008).
[15]T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," IEEE Trans. Speech Audio Process. **11**, 791–803 (2003).
[16]C. Zhang, D. Floréncio, and Z. Zhang, "Why does PHAT work well in low noise, re verberative environments?" in *International Conference in Audio, Speech and Signal Processing (ICASSP 2008)*, pp. 2565–2568.
[17]B. Kwon, Y. Park, and Y. Park, "Analysis of the GCC-PHAT technique for multiple sources," in *Proc. of Control Automation and Systems (ICCAS 2010)*, pp. 2070–2073.

[18]J. Escolano, J. M. Pérez, N. Xiang, M. Cobos, and J. J. López, "A Bayesian inference model for speech localization (L)," J. Acoust. Soc. Am. **132**, 1257–1260 (2012).
[19]F. Jacobsen and T. Roisina, "The coherence of reverberant sound fields," J. Acoust. Soc. Am. **108**, 204–210 (2000).
[20]N. Mitianoudis and T. Stathaki, "Batch and online underdetermined source separation using Laplacian mixture models," IEEE Trans. Audio Speech Lang. Process. **15**, 1818–1832 (2007).
[21]J. M. Perez-Lorenzo, R. Viciana, P. J. Reche, F. Rivas, and J. Escolano, "Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments," Appl. Acoust. **73**, 698–712 (2012).
[22]P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge University Press, UK, 2005), pp. 192–200.
[23]T. Jasa and N. Xiang, "Efficient estimation of decay parameters in acoustically coupled spaces using slice sampling," J. Acoust. Soc. Am. **126**, 1269–1279 (2009).
[24]L. Wasserman, "Bayesian model selection and model averaging," J. Math. Psychol. **44**, 92–107 (2000).
[25]H. Jeffreys, *Theory of Probability*, 3rd ed. (Oxford University Press, UK, 1965), pp. 193–244.
[26]D. J. C. McKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, UK, 2003), pp. 343–356.
[27]P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," IEEE Signal Process. Mag. **21**, 36–47 (2004).
[28]R. E. Kass and L. Wasserman, "A reference Bayesian test for nested hypotheses and its relationship to the Schwartz criterion," J. Am. Stat. Assoc. **90**, 928–934 (1995).
[29]W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," Biometrika **57**, 97–109 (1970).
[30]C. C. Robert and G. Casella, *Monte Carlo Statistical Methods* (Springer Verlag, New York, 1999), pp. 79–122.
[31]J. Skilling, "Nested sampling for general Bayesian computation," Bayesian Anal. **1**, 833–860 (2006).
[32]D. Silvia and J. Skilling, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, UK, 2006), pp. 181–208.
[33]T. Jasa and N. Xiang, "Nested sampling applied in Bayesian room-acoustics decay analysis," J. Acoust. Soc. Am. **132**, 3251–3262 (2012).
[34]J. R. Shaw, M. Bridges, and M. P. Hobson, "Efficient Bayesian inference for multimodal problems in cosmology," Mon. Not. R. Astron. Soc. **378**, 1365–1370 (2007).
[35]F. Feroz and M. P. Hobson, "Multimodal nested sampling: an efficient and robust alternative to Markov chain Monte Carlo methods for astronomical data analyse," Mon. Not. R. Astron. Soc. **384**, 449–463 (2008).
[36]K. H. Knuth, "Optimal data-based binning for histograms," arXiv:physics/0605197v2 (2013).
[37]G. O. Roberts and J. S. Rosenthal, "Optimal scaling for various Metropolis-Hastings algorithms," Stat. Sci. **16**, 351–367 (2001).