SEPTEMBER 12 2012

A Bayesian inference model for speech localization (L)

José Escolano; José M. Perez-Lorenzo; Ning Xiang; Máximo Cobos; José J. López

J. Acoust. Soc. Am. 132, 1257–1260 (2012)
https://doi.org/10.1121/1.4740489

A Bayesian inference model for speech localization (L)

José Escolano[a] and José M. Perez-Lorenzo
Multimedia and Multimodal Processing Research Group, University of Jaén, 23700, Linares, Spain

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

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

José J. López
Institute for Telecommunication and Multimedia Applications, Universidad Politécnica de Valencia, 46021, Valencia, Spain

(Received 23 January 2012; revised 28 June 2012; accepted 4 July 2012)

The localization of active speakers with microphone arrays is an active research line with a considerable interest in many acoustic areas. Many algorithms for source localization are based on the computation of the Generalized Cross-Correlation function between microphone pairs employing phase transform weighting. Unfortunately, the performance of these methods is severely reduced when wall reflections and multiple sound sources are present in the acoustic environment. As a result, estimating the number of active sound sources and their actual directions becomes a challenging task. To effectively tackle this problem, a Bayesian inference framework is proposed. Based on a nested sampling algorithm, a mixture model and its parameters are estimated, indicating both the number of sources—model selection—and their angle of arrival—parameter estimation, respectively. A set of measured data demonstrates the accuracy of the proposed model.
 [http://dx.doi.org/10.1121/1.4740489]

PACS number(s): 43.60.Jn, 43.72.Ne [ZHM] Pages: 1257–1260

I. INTRODUCTION

Sound source localization in multi-source environments is a challenging task. To this end, microphone arrays are commonly employed in many signal processing tasks such as source tracking, source separation, speech enhancement, and noise reduction. Algorithms for sound source localization using microphone arrays can be broadly divided into indirect and direct approaches.[1] Indirect approaches usually follow a two-step procedure where they first estimate the time difference of arrival[2] between microphone pairs and, afterwards, they estimate the source position based on the geometry of the array and the estimated delays. On the other hand, direct approaches compute a cost function over a set of candidate locations and take the most likely source positions. Most of these algorithms are based on the Generalized Cross-Correlation (GCC) method,[3] which calculates the correlation function by using the inverse Fourier transform of the cross-power spectral density function multiplied by a proper weighting function. The most widely used weighting function is the phase transform (PHAT), which has been shown to be optimal in reverberant environments.[4]

Localization of multiple sound sources using only two microphones has been receiving increasing attention in the last years. In this context, while a number of sparse methods working in the time-frequency domain have been developed by exploiting the sparsity properties of speech in this domain,[5] algorithms for multi-source localization based on GCC-PHAT have been rarely described. This is due to the fact that most of these algorithms are based on a single-source signal model and are not suitable to localize multiple sources. As a result, a multi-source localization framework based on GCC-PHAT analysis seems to be specially meaningful to tackle the problem from the well-known GCC perspective, even when the number of sources is *a priori* unknown.

This letter presents a multiple source localization system based on GCC-PHAT and Bayesian inference, allowing one to determine both the number of sound sources and their actual directions of arrival (DOAs).

II. SOUND SOURCE LOCALIZATION

A. Signal model

Consider two microphone signals $x_1(t)$ and $x_2(t)$ following the anechoic mixture model,[4]

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

where N is the number of sources, $s_n(t)$ are the time-domain source signals, a_{mn} are scalar coefficients, and τ_{mn} are the source-to-sensor time delays. If the sources are assumed to

[a]Author to whom correspondence should be addressed. Electronic mail: escolano@ujaen.es

be located in the far field, the DOAs of the sources can be directly related to the inter-sensor time delays $\tau_n = \tau_{2n} - \tau_{1n} = (d/c)\cos\hat{\theta}_n$, 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.

B. GCC

Considering the above-presented model, a time delay estimate of the predominant sound source, $\tau_{\hat{n}}$, can be obtained by means of the GCC as

$$\hat{\tau}_{\hat{n}} = \arg\max_{\tau} E\{[x_1(t) \bigstar w_1(t)][x_2(t+\tau) \bigstar w_2(t)]\}, \quad (2)$$

where $E\{\cdot\}$ is the statistical average over time and $\bigstar$ denotes time convolution. The impulse responses $w_1(t)$ and $w_2(t)$ are the weighting functions applied to each microphone signal, respectively. In practice, $\tau_{\hat{n}}$ is computed via the Fourier transforms of the microphone signals and the weighting functions. To make the estimator robust to reverberation, the well-known PHAT weighting is used,

$$W_1(\omega)W_2^*(\omega) = \|X_1(\omega)X_2(\omega)\|^{-1}. \quad (3)$$

This time delay, and therefore, its associated angle, is calculated every time frame and each result is used to construct an histogram that represents a probability function $H(\theta)$ of where a source is situated. When multiple speech sources are measured, based on the superposition principle, each source is represented by a particular area of the histogram.

III. BAYESIAN INFERENCE

A. Parameter estimation

The starting point of Bayesian inference is the Bayes' theorem. For a given model H and a given dataset $\mathbf{D}$, formed by a vector with K components as a function of an angle vector $\boldsymbol{\theta}$, the *posterior probability distribution* of the model parameters Θ is calculated as follows:

$$p(\Theta|H,\mathbf{D},I) = \frac{p(\mathbf{D}|\Theta,H,I)p(\Theta|H,I)}{p(\mathbf{D}|H,I)}, \quad (4)$$

where I is the relevant background information encapsulating that model H represents the data $\mathbf{D}$ well. The term $p(\mathbf{D}|\Theta,H,I)$ represents the *likelihood* function, indicating the resemblance of the data $\mathbf{D}$ and the model H for a given parameter set Θ, i.e., it grows when the difference error decreases. This distribution is assigned according to the background information I. Appealing to the maximum entropy and after marginalizing about an unknown variance error, the likelihood distribution may be assigned to a Student t-distribution as follows:[6]

$$p(D|\Theta,H,I) \equiv \mathcal{L}(\Theta) = \frac{1}{2}\Gamma\left(\frac{K}{2}\right)\left(\frac{E(\Theta)}{2\pi}\right)^{-K/2}, \quad (5)$$

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

The term $p(\Theta|H,I)$ corresponds to the prior distribution of the parameters. This distribution is usually assigned uniformly to avoid any subjective preference. The term $p(\mathbf{D}|H,I)$ is known as a marginal likelihood or Bayesian *evidence*. In most parameter estimation problems, the evidence is a normalization constant, but it plays a fundamental role in the model selection, as will be shown in Sec. III B. In order to act as a normalization constant, evidence Z is calculated as

$$p(\mathbf{D}|H,I) \equiv Z = \int_{\Theta} p(\mathbf{D}|\Theta,H,I)p(\Theta|H,I)d\Theta. \quad (6)$$

B. Model selection

According to Bayes' theorem, the posterior probability of a model H_i, given data $\mathbf{D}$ and relevant background information I is given by

$$p(H_i|\mathbf{D},I) = \frac{p(\mathbf{D}|H_i,I)p(H_i|I)}{p(\mathbf{D}|I)}. \quad (7)$$

The idea behind model selection is to compare the posterior probability of a set of competitive models and to select the one with the highest posterior probability to the data.

Assigning the competing models equal prior probability,i.e., no model is favored against the other, the model selection is determined in terms of the marginal likelihood function. However, it should be observed that the above-presented likelihood function equals the evidence term in the parameter estimation task [see Eq. (6)]. Therefore, the model selection can be carried out just by comparing evidence obtained within the effort of the parameter estimation. Bayesian model selection favors a simpler model instead of the model that better fits data, which equivalently represents a quantitative implementation of Ockham's razor.[7]

IV. MODEL IMPLEMENTATION

A. Nested sampling

The main difficulty in Bayesian model selection lies in the analytical intractability of Eq. (6), since it is a generally multi-dimensional integral and the computation burden becomes prohibitive when the model becomes more complex. Therefore, an approximation is necessary to overcome this handicap. An alternative to calculate this integral is to use the *nested sampling* algorithm.[8] The nested sampling algorithm was developed specifically to approximate these marginalization integrals, and it has the added benefit of generating samples from the posterior distribution $p(\Theta|H,\mathbf{D},I)$. Comprehensive tutorials and practical details on the nested sampling may be found in Refs. 8 and 9.

The basic idea behind the nested sampling is to rearrange Eq. (6) as a one-dimensional integral, just 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. Then, the evidence may be rewritten as[9]

$$Z = \int_0^1 \mathcal{L}(\xi)d\xi. \quad (8)$$

This one-dimensional integral can be solved numerically,

$$Z \simeq \sum_{i=1}^{\infty} \mathcal{L}_i \Delta\xi_i \quad \text{with} \quad \Delta\xi_i = \xi_{i-1} - \xi_i, \tag{9}$$

where $\xi_0 = 1$, $\xi_\infty = 0$, and $\mathcal{L}_\infty = \mathcal{L}_{\max} < \infty$.

Starting with a set of M initial random samples Θ_m from the prior and their associated likelihoods $\mathcal{L}_m$, where $m \in [1, M]$, the parameters with the lowest likelihood value, labeled as $[\Theta_1, \mathcal{L}_1]$, are stored and replaced by a new random parameter Θ_{new} under the constraint $\mathcal{L}_{\text{new}} > \mathcal{L}_1$, remaining again M samples.[9] In the ith iteration, the same process is repeated with a new selected sample with the lowest identified likelihood, $[\Theta_i, \mathcal{L}_i]$. Repeating this process, the evidence is accumulated according to Eq. (9).

For practical implementations, elementary prior mass $\Delta\xi_i$ can be statistically approximated by $\Delta\xi_i \approx e^{-1/i}$. Equation (9) will keep accumulating up to $\log(Z_i) - \log(Z_{i-1}) < \delta$, with $Z_i = \mathcal{L}_i \Delta\xi_i$.

B. Histogram model

The histogram can be modeled by a mixture of Laplacian distributions to represent the angles resulting from a scatter plot of a two-channel mixture.[10] Therefore, the model used is $H(\theta) = \sum_{n=1}^{N} A_n e^{-\|\theta-\mu_n\|/\sigma_n}$, where μ_n is the mean, σ_n is the variance, and A_n is the amplitude of each one of the Laplacian functions.

One of the key points of this work is to determine the number of sound sources N, which determines the model. Therefore, given a number of potential sources, the model selection will estimate the number of sources consistent with data and prior information, and the parameter estimation $\Theta = \{\mu; \sigma; \mathbf{A}\}$ with information where speech sources are located.

V. EXPERIMENTAL SETUP AND RESULTS

In order to validate the above-described methodology, a two-microphone array with a separation of 13.5 cm between microphones was placed in the middle of a room. The array consists of two omni-directional AKG-C417PP microphones. The room has a volume of 248.64 m^3 and the measured reverberation time is approximately 1 s. Four speakers were distributed between 0° and 180°, following the scheme presented in Table I.

A text is read by all the speakers at the same time and then recorded simultaneously. Two scenarios were used: *Two speakers* located at angles $\hat{\theta}_2$ and $\hat{\theta}_4$ (labeled as E_1) and *four speakers* located at all the angles described in Table I (labeled as E_2). In both cases, the signal-to-noise ratio was found to be approximately 30 dB. Once the recording was finished, the two microphone signals were processed using the PHAT algorithm in order to obtain the histogram[4] using a Hann window of 25 ms and 50% overlap for each observation. The number of histogram bins, or equivalently, 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 particular case, the number of bins equals $K = 30$.

TABLE I. Speakers' angular distribution ($\hat{\theta}_i$) around the two-microphone array.[a]

$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
53.1° (2 m)	69.4° (1.70 m)	113.6° (1.78 m)	149° (2.33 m)

[a] The distance to the array is indicated in parentheses.

TABLE II. Average log-evidence values (up) and BIC (down) difference and their corresponding standard deviation for the five competitive models in each experiment, where the selected model has been highlighted.

	1	2	3	4	5
E_1	60 ± 0.3	**76.9** ± 4.9	76 ± 4.2	76.2 ± 5.5	74.9 ± 5.8
E_2	54.7 ± 0.2	70.3 ± 1.8	72.9 ± 5.2	**76.6** ± 7	76.3 ± 7.6
E_1	119.6 ± 0.0	**161.8** ± 11.1	154.2 ± 10.7	149.2 ± 14.5	139.9 ± 17
E_2	109.1 ± 0.0	142.3 ± 2.4	145 ± 25	**148** ± 18.6	143.3 ± 18.8

Regarding the nested sampling algorithm, the initial population has been set to $M = 1000$ samples, and the stop condition has been forced to be $\delta < 10^{-4}$. The models under evaluation, i.e. number of sources, are set from $N = 1$ to $N = 5$. Each model is run 100 times, and the estimated parameters and log-evidence value are stored. From the log-evidence data, some statistics are calculated: First, the average and standard deviation are calculated for each experiment (see Table II). The highest mean corresponds to two speakers in the experiment E_1 and four speakers in the experiment E_2, which evidences the accuracy of the method. It should be mentioned regarding E_1 that from $N = 2$ to $N = 5$ the log-evidence value is quite similar. Despite the corresponding log-evidence value of $N = 2$ being the highest, all these models seem to be competitive; Ockham's razor suggests that one prefers the simplest model. Model $N = 1$ has to be clearly dismissed due to considerable difference regarding the rest. Regarding the second experiment, E_2, clearly the most competitive models are $N = 4$ and $N = 5$; based on the aforementioned criteria, the simplest of the most competitive models is selected, corresponding to the correct number of speakers for that experiment, $N = 4$. For comparison purposes, the *Bayesian Information Criterion* (BIC) is included in Table II. Both techniques show a

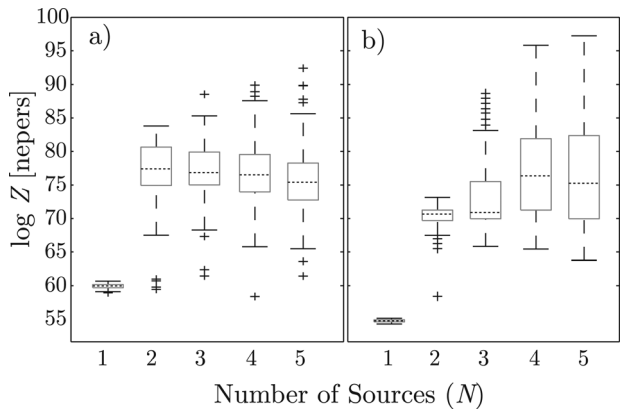

FIG. 1. Log-evidence model probability distribution for each of the models under evaluation corresponding to (a) experiment E_1 and (b) experiment E_2.

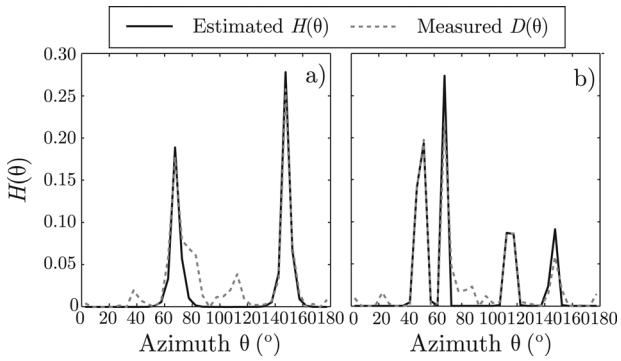

FIG. 2. Measured (dashed grey line) and estimated (solid black line) histogram corresponding to (a) experiment E_1 and (b) experiment E_2.

consistent result within the model selection, supporting the performance of the model selection based on an evidence comparison. However, it should be pointed out that the BIC approach assumes that the likelihood distribution of interest can be approximated by a multi-variant Gaussian in the vicinity of the global extreme, but for many applications this is not the case, particularly multi-modal distributions. If the shape of the likelihood distribution deviates drastically from a multi-variant Gaussian, the estimates can be extremely poor to be able to correctly rank the models.

The algorithm has been validated based on its log-evidence average values. In order to give more support to the validity of the aforementioned conclusions, a *boxplot* representation of the values helps corroborate them and obtain a compact way to represent the log-evidence distribution for each model (see Fig. 1). In both cases, the highest modes confirm two speakers for E_1 and four speakers for E_2.

Regarding parameter estimation, Fig. 2 shows both the measured and the estimated histograms. The estimated histogram with the highest evidence has been chosen from the selected model in each case. Regarding the specific problem of the localization, DOAs are described on the basis of the mean of each Laplacian function. Table III list the resulting means, showing a maximum deviance lower than 3° with respect to the real position and evidencing the considerable resemblance between measured and estimated histograms. For the highest models, reflections are identified as speakers, but the posterior probability is barely changed, making the simplest model the most suitable. This will occur once the histogram area corresponding to reflections, i.e., its energy, is considerably smaller than the speech contributions. For PHAT algorithm, this is valid when the signal-to-noise ratio is higher than 20 dB.

VI. CONCLUSIONS

In this letter, a Bayesian inference model is presented for multi-speaker detection and localization using a two-microphone array. Based on the use of a GCC-PHAT algorithm, a Laplacian mixture is employed to model the histogram resulting from time-delay estimates, using Bayesian nested sampling to determine the number of active speakers and their actual angular distribution. The algorithm has been evaluated in a real scenario, demonstrating the accuracy of the method. The simplicity and elegance of the GCC-PHAT, together with Bayesian inference allows an unsupervised solution, even if the number of sources is unknown *a priori*. The advantage of the nested sampling relies on the fact that the model selection and parameter estimation are performed simultaneously, with no additional effort.

More research should be timely done testing this algorithm in scenarios with different signal-to-noise ratios and to be used using alternative localization methods in applications such as sound source separation and on-line detection.[11] Moreover, a thoughtful comparison with some other evidence estimation methods such as annealed importance sampling[12] should be done in order to definitively evidence nested sampling as the right choice in this application.

TABLE III. Estimated DOAs, $\hat{\theta}_i$, for each experiment, obtained as the estimated mean of each Laplacian function.

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
E_1	–	68.3°	–	148.2°
E_2	50.2°	67.0°	114.0°	146.1°

[1]N. Madhu and R. Martin, *Advances in Digital Speech Transmission*, (Wiley, Chichester, UK, 2008), pp. 135–166.
[2]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).
[3]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).
[4]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).
[5]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).
[6]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).
[7]D. J. C. McKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2003).
[8]J. Skilling, "Nested sampling for general Bayesian computation," Bayesian Anal. **1**, 833–860 (2006).
[9]D. Silvia and J. Skilling, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, New York, 2006).
[10]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).
[11]M. Cobos and J. J. López, "Two-microphone separation of speech mixtures based on interclass variance maximization," J. Acoust. Soc. Am. **127**, 1661–1673 (2010).
[12]R. Neal, "Annealed importance sampling," Stat. Comput. **11**, 125–139 (2001).