HAL
archives-ouvertes.fr

Evaluation of an Open-Source Implementation of the SRP-PHAT Algorithm within the 2018 Locata Challenge

Romain Lebarbenchon, Ewen Camberlein, Diego Di Carlo, Clément Gaultier, Antoine Deleforge, Nancy Bertin

HAL Id: hal-02187964
https://hal.archives-ouvertes.fr/hal-02187964

Submitted on 18 Jul 2019

EVALUATION OF AN OPEN-SOURCE IMPLEMENTATION OF THE SRP-PHAT ALGORITHM WITHIN THE 2018 LOCATA CHALLENGE

Romain Lebarbenchon[1], Ewen Camberlein[1], Diego di Carlo[1],
Clément Gaultier[1], Antoine Deleforge[2], Nancy Bertin[1]

[1]Univ Rennes 1, Inria, CNRS, IRISA, F-35000 Rennes, France
[2]Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

ABSTRACT

This short paper presents an efficient, flexible implementation of the SRP-PHAT multichannel sound source localization method. The method is evaluated on the single-source tasks of the LOCATA 2018 development dataset, and an associated Matlab toolbox is made available online.

1. INTRODUCTION

Many source localization methods are based on the estimation of time-difference-of-arrival (TDOA) between microphones. In the two-microphone case, the thorough evaluation carried out in [1] highlighted the good performance of the Generalized Cross-Correlation with PHAse Transform (GCC-PHAT) algorithm [2] among several other methods. A general principle for extending this method to the multichannel setting, named SRP-PHAT, was proposed in [3]. However, no reference implementation of this family of algorithms (including all its possible variants) was available to date, to the best of the authors' knowledge. In particular, choices of parameters (pooling in time and frequency, grid search resolution) and management of the multiple coordinate systems between pairs of microphones and the whole antenna have not been systematically documented.

We present here our participation to the 2018 LOCATA challenge [4]. This aims at evaluating our implementation of SRP-PHAT, whose good performance was already probed in realistic environments in [5], but which was not confronted to other methods on a common, realistic dataset so far. It is here applied to task 1 (single, static source and static array), task 3 (single moving source and static array) and task 5 (single moving source and moving array) of the challenge. Evaluation of the method on the development dataset shows encouraging results, confirming the legitimate place of SRP-PHAT among state-of-the-art methods for multichannel source localization in reverberant environments.

2. PRINCIPLES

2.1. TDOA and angular spectrum

The general principle of SRP-PHAT is to compute a function $\Phi(\theta, \varphi)$ termed "angular spectrum", where θ and ϕ are azimuth and elevation variables, which is expected to exhibit local maxima in the directions of active sources. More precisely, the following steps are followed:

- build in each time-frequency (t, f) bin a local angular spectrum *function* $\phi(t, f, \theta, \varphi)$ that is large for directions (θ, φ) which are compatible with the observed signal at (t, f) and small otherwise,
- *integrate* (or *pool*) this function over the time-frequency plane, leading to a global angular spectrum,
- find the *peaks* of this spectrum above a certain *threshold* and distant from a certain *minimum angle*.

2.2. GCC-PHAT

The GCC algorithm [2] generalizes cross-correlation by multiplying the cross-spectral density (CSD) of two signals by complex weights in the frequency domain. The PHAT weighting consists of normalizing the CSD to have unit amplitude at all frequencies, thus considering phase differences only and minimizing the influence of the source's power spectral density. In the case of a single broadband source, two microphones and under far-field (source-array distance much larger that the array aperture h), free-field (only the direct source-to-microphone sound paths are considered) and noiseless assumptions, the asymptotic GCC-PHAT of the microphone signals in the discrete frequency domain is $e^{-2i\pi f\tau \cos(\alpha)}$ where f is the frequency index, α is the angle of arrival (AOA) of the source signal at the array and τ is the maximum observable delay in samples. We have $\tau = hF_s/C$ where C denotes the speed of sound and F_s is the frequency of sampling. This leads to the following natural local angular spectrum for a microphone pair:

$$\phi^{\mathrm{GCC-PHAT}}(t, f, \alpha) = \Re\left(\frac{x_1(t,f)x_2^*(t,f)}{|x_1(t,f)x_2^*(t,f)|}e^{-2i\pi f\tau \cos(\alpha)}\right) \quad (1)$$

where x_1, x_2 denote the microphone signals in the short-time Fourier transform (STFT) domain and α is a *local* azimuth in the coordinate system defined by the microphone pair. Note that this spectrum does not depend on elevation.

2.3. Extension to multichannel

SRP-PHAT's principle is to first compute local angular spectra for each microphone pair, then to aggregate them across pairs (bringing them back in the global coordinate system) before being pooled and maximized. The computation consists of the following steps:

1. Define the search space, *i.e.* a grid of possible DOAs (θ_j, φ_k) for which we want to evaluate Φ in the global coordinate system;

2. For each microphone pair n:

(a) Compute the corresponding AOAs $\{\alpha_{jk}^{(n)}\}_{jk}$ with respect to the microphone pair;

(b) Resample $\{\alpha_{jk}^{(n)}\}_{jk}$ into a smaller set $\{\alpha_i^{(n)}\}_i$ in order to reduce computational time;

(c) Compute the GCC-PHAT local angular spectrum ϕ_n at angles $\{\alpha_i^{(n)}\}_i$ for microphone pair n according to formula (1);

(d) Linearly interpolate ϕ_n back to the original angle resolution and global coordinate system;

3. Compute the global spectrum $\Phi(\theta_j, \varphi_k)$ by pooling the local angular spectra ϕ_n over all time-frequency bins (t, f) and across all microphone pairs n. Pooling methods (such as maximum or sum) and their order over each of the f, t, n indexes must be chosen for this purpose.

4. Find the indexes j and k of the largest peak (single source case) or peaks (multiple source case) of $\Phi(\theta_j, \varphi_k)$, yielding the estimated source azimuth(s) θ_j and elevation(s) φ_k.

A Matlab toolbox allowing easy and flexible implementations of SRP-PHAT as well as 7 other angular spectrum-based was made freely available online under the name Multichannel BSS Locate[1].

3. EXPERIMENTS AND RESULTS

We now evaluate our implementation of SRP-PHAT on tasks 1, 3 and 5 of the LOCATA challenge [4] with the *robot head* and the *Eigenmike* arrays for azimuth and elevation estimation. The other two antennas of the challenge are discarded since they are near-linear (preventing elevation estimation). Moreover, DICIT is incompatible with our far-field assumption while the dummy head is incompatible with our free-field assumption. All microphones (12) and all microphone pairs (66) are used with the robot head. All microphones are also used with the Eigenmike, but microphone pairs with a curvilinear distance lesser than $90°$ are discarded in order to reduce the overall algorithm complexity, resulting in 240 microphone pairs usage instead of the 496 available pairs. A sphere sampled with $1°$ resolution both in azimuth and elevation is used as the search space (see Sec. 2.3 - 1). For each microphone pair, the evaluated AOAs (see Sec. 2.3 - 2.(b)) are computed with $5°$ resolution. SRP-PHAT is applied every 256 ms to 512 ms signals (downsampled to 16 kHz) using an overlapping sliding analysis window. The STFT is applied to each channel of the signals using 64 ms Fourier frames (1024 samples) with 50% overlap and sine windows. This results in 15 Fourier frames per analysis window. The pooling methods used are first summations over microphone pairs and frequencies and then maximum over time frames. Only the largest value

over the global angular search space is returned as a DOA estimate, as only single source localization is considered. DOA estimates are linearly interpolated over time in order to match the one-estimate-per-time-stamp requirement of LOCATA.

Table 1. shows localization errors obtained with the proposed method. Encouragingly, mean azimuth errors (first row) are 2 to 6 times smaller than those reported for the baseline MUSIC method [4] for the same arrays and tasks. To reduce the effect of gross errors on reported mean values, average *successful localization* errors for different success threshold are also reported. The proposed method localizes the target source with less than $20°$ error in at least 92% of the tests for all tasks and arrays. Nearly 100% correct localization is achieved in the static scenario (task 1). While the robot head array enables better performance in all tasks, average errors never exceed $10°$ regardless of the task and array.

4. CONCLUSION

In this paper, we reported competitive single sound source localization results using the SRP-PHAT implementation of the Multichannel BSS Locate toolbox on the 2018 LOCATA challenge dataset. The toolbox's flexibility allows one to easily adapt the method to arbitrary array geometries, to various types of emitted signals and environments, and to easily generalize any two-channel localization method to more channels. It can also be tuned to find a trade-off between computational time and accuracy, although we focused on accuracy for this challenge. While the toolbox can output multiple source direction estimates for a given input signal, it does not yet incorporate source counting or source tracking solutions, preventing its use in multiple sources scenario. Such extensions will be considered for future releases of the toolbox.

5. REFERENCES

[1] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.

[3] J. Dibiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001, pp. 157–180.

[4] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The locata challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM),(Sheffield, UK)*, 2018.

[5] N. Bertin, E. Camberlein, R. Lebarbenchon, E. Vincent, S. Sivasankaran, I. Illina, and F. Bimbot, "Voicehome-2, an extended corpus for multichannel speech processing in real homes," *to appear in Speech Communication, Elsevier*, 2018.

[1]http://bass-db.gforge.inria.fr/bss_locate/

Table 1: Sound localization results obtained on the LOCATA development dataset using the proposed SRP-PHAT method. For each task and array, different *success thresholds* (in degrees) are considered. A localization is considered successful when both azimuth and elevation estimation errors are below these thresholds. The percentage of successful localizations (suc.) and their average azimuth (az.) and elevation (el.) errors in degrees are showed for each threshold.

Success threshold	Task 1						Task 3						Task 5					
	Robot head			Eigenmike			Robot head			Eigenmike			Robot head			Eigenmike		
	az.	el.	suc.	az.	el.	suc.	az.	el.	suc.	az.	el.	suc.	az.	el.	suc.	az.	el.	suc.
No thresh.	1.51	1.71	—	7.04	4.68	—	4.43	2.66	—	8.79	4.41	—	6.19	3.16	—	9.31	4.37	—
20°	1.43	1.66	99.9	6.95	4.64	99.9	2.48	1.75	95.8	7.82	3.12	92.5	1.76	1.83	94.5	5.84	2.94	94.8
10°	1.43	1.66	99.9	6.95	4.64	99.9	2.35	1.56	93.7	6.27	2.40	71.0	1.64	1.77	93.3	5.43	2.80	89.5