# Localizing multiple sound sources by associating sub-band TDOA features for biodiversity monitoring

Manuel Alejandro Jaramillo Rodríguez*, Randall Ali†, Toon van Waterschoot*
*Dept. of Electrical Engineering (ESAT), STADIUS, KU Leuven, Leuven, Belgium
†Institute of Sound Recording, Music & Media, University of Surrey, Guildford, UK

***Abstract*—Estimating the position of a single sound source within the biodiversity monitoring context can be achieved via an estimation of the time difference of arrival (TDOA) between microphone pairs among several single-channel autonomous recording units spatially distributed across a geographic area to be surveyed. When multiple sound sources are simultaneously active, however, each microphone pair estimates multiple TDOAs leading to an ambiguity of which TDOAs from different pairs correspond to the same physical source, i.e., an association problem. Although several methods have been proposed to address this problem, most rely on the availability of a large number of microphone pairs, whereas in biodiversity monitoring, the number of available microphones is limited. In this paper, we propose a method for localizing multiple sound sources based on sub-band TDOA estimations. The approach facilitates the definition of a sub-band feature vector, which characterizes the spectral content of sources, and is used to solve the association problem. This leads to a decomposition of the multiple source localization problem into independent single source localization tasks. Preliminary simulations demonstrate the potential of the method in scenarios with a limited number of microphones and hence its prospect for biodiversity monitoring.**

***Index Terms*—sound source localization, biodiversity monitoring, time difference of arrival, data association problem.**

## I. Introduction

Estimating the positions of terrestrial wildlife species using acoustic signals captured by outdoor autonomous recording units (ARUs) [1] can provide valuable information for the purpose of biodiversity monitoring in a non-invasive manner. For instance, knowledge of species' locations over time can be used to infer animal population densities [2] or for the study of animal behavior [3]. Despite considerable advances in sound source localization approaches [4], estimating wildlife species' locations using outdoor ARUs remains challenging, particularly when there are multiple active sources overlapping in time and frequency, a limited number of ARUs (i.e, few microphones) in relation to the number of active sources, and unpredictable environmental conditions.

A common practice is to deploy multiple synchronized single-channel ARUs (e.g., [5], [6]) spatially distributed across the area of interest, which are programmed to record during a determined schedule. Audio data is then collected and analyzed offline on a local workstation. In these setups, source localization is usually based on estimating the time difference of arrival (TDOA) of the source signals between microphone pairs [1]. Essentially, each of the estimated TDOAs is used to constrain source positions to a hyperbola in 2D space, (or a hyperboloid in 3D space), and multiple hyperbolas are used to find the source position by solving a minimization problem to find their approximate intersection [4], [7]. This general approach is, however, not always robust, especially when there is simultaneous activity of multiple sources due to the so-called association problem [8]–[11]. As there are different possible ways in which multiple hyperbolas originating from various sources and microphone pairs can intersect, an ambiguity arises resulting in false source positions or what we refer to as 'ghost sources'.

Multiple approaches have been proposed to address this TDOA association problem [9]–[11] albeit for indoor scenarios. In [9], the authors avoid explicit TDOA association across microphone pairs by firstly identifying spatial regions that are geometrically consistent with the observed TDOA measurements, and subsequently estimating each source position independently within the detected regions. Similarly, in [10], the authors proposed a delay density map approach, where multiple delay estimates across microphone pairs are used to construct a spatial likelihood representation of the source positions. While effective in avoiding explicit TDOA association, these approaches rely on spatial discretization, resulting in a trade-off between resolution and computational complexity. Moreover, the performance highly depends on the availability of a large number of microphones, as reliable results require redundant constraints across many microphone pairs. A different methodology is proposed in [11], where the actual TDOA associations are computed by grouping multiple TDOA candidates, extracted from generalized cross-correlations (GCCs), across microphone pairs using a multi-dimensional assignment model. Each candidate association is scored via a single-source least-squares localization cost, and the globally optimal set is found through Lagrangian relaxation. However, the cost relies only on geometric consistency from single-source least-squares fitting and ignores signal-level reliability measures. As a consequence, incorrect TDOA combinations can still yield low-cost solutions unless there is sufficient constraint redundancy.

In this paper, we draw inspiration from sub-band TDOA estimation [12] to develop a general method for localizing multiple sound sources consisting of three main stages. Firstly,

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of KU Leuven internal funds C14/21/075, and FWO Research Project G0A0424N. This project has received funding from the European Union's Horizon 2023 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101116715.

for each microphone pair, multiple TDOAs are estimated from sub-band regions over a fixed number of time frames to yield a TDOA distribution, where peaks correspond to the candidate source TDOAs. Each candidate source TDOA is then linked to a sub-band feature that captures the spectral content of the source that generated it. These features are consequently used to group TDOAs across microphone pairs based on a similarity metric between their corresponding features. Since the same sound source is expected to result in a similar frequency distribution across microphone pairs, this grouping provides a data-driven estimate of the underlying TDOA associations. Finally, the spatial locations of the sources are estimated by solving a single source localization problem within each associated group. Although this approach can be used for sound source localization tasks in general, it is particularly relevant to the biodiversity monitoring context as limited spatial diversity resulting from few single-channel ARUs can be effectively augmented with spectro-temporal information of sources to improve the estimation of species' locations.

The remainder of the paper is structured as follows. Section II outlines the TDOA association problem. Section III provides a detailed description of the proposed method. The performance of the proposed approach is evaluated by means of simulations in Section IV and Section V concludes the paper.

## II. Problem Statement

We consider $M$ synchronized microphones ($M$ single-channel ARUs) at known positions distributed over a geographic area of interest for biodiversity monitoring and $S$ sound sources which can be simultaneously active at unknown positions. Synchronization among the ARUs can be achieved through GPS modules, which are available on some ARUs [5]. For simplicity, in this work, we also consider a 2D scenario, but it should be noted that the approach is not restricted to this dimensionality and is equally applicable for 3D scenarios. Fig. 3 illustrates an example scenario (which will also be the simulated scenario in Section IV) with $M = 3$ and $S = 4$. As can be observed, there are multiple hyperbolas corresponding to the multiple TDOA estimations from each microphone pair (the colors of which will be explained later on). The ambiguity that arises from the TDOA association problem is that we do not know which hyberbolas (TDOAs) from each microphone pair corresponds to the same physical source (green bird silhouettes). As shown in Fig. 3, the result is that intersections of the hyperbolas can erroneously estimate the locations of ghost sources (grey bird silhouettes) as opposed to/in addition to the true sources. Our objective in this work is to resolve this association problem by correctly assigning estimated TDOAs to the corresponding physical source, and subsequently estimating the unknown true source positions.

## III. Proposed method

The discrete-time signals received at each microphone are firstly transformed into the time-frequency domain by means of a short-time Fourier transform (STFT). For each time frame, we define a sub-band region $\Omega$, that spans a fixed number of frequency bins, analogously to the approaches presented in [12] and [13]. The proposed method, consisting of three stages as described in the following sub-sections, is built on the assumption that for the majority of sub-band regions in each microphone pair, there exists a dominant source whose contribution determines the TDOA observation at that region. This is a relaxation of the W-disjoint orthogonality assumption typically used in sound source localization tasks [14], as it does not require sources to be isolated at the level of individual time-frequency bins.

Two aspects of the proposed method are worth emphasizing. First, source dominance is only required locally, i.e., within a given sub-band region of a particular time frame for a specific microphone pair. Thus, even with spectral and temporal source overlap, reliable TDOA estimates are still possible as long as one source dominates the cross-correlation within that region. Second, source dominance is evaluated per microphone pair, meaning that propagation effects such as distance-dependent attenuation due to the different source-microphone distances can make different sources dominant for different pairs, even when their emitted signals have similar spectral content and power.

### A. Multi-source TDOA estimation

The first stage of the proposed method involves estimating the candidate TDOAs for each source. To this end, *local* TDOAs are initially estimated in each time frame and each sub-band region from a subset of microphone pairs. This strategy of microphone pair selection is done to limit the computational complexity associated with evaluating all possible pairs. We use the generalized cross-correlation with phase normalization to estimate these local TDOAs, which for a microphone pair $(i, j)$ is defined as [12]

$$\Psi_{ij}(l, \Omega, n) = \frac{1}{N} \sum_{k \in \Omega} \frac{X_i(l,k) X_j^*(l,k)}{|X_i(l,k)||X_j(l,k)|} e^{j \frac{2\pi k n}{N}}, \quad (1)$$

where $l$ represents the time-frame index, $k$ is the frequency bin index, $\Omega$ is the sub-band region of some pre-determined bandwidth composed by a set of subsequent frequency bins, $X_i(l,k)$ and $X_j(l,k)$ are the STFT-domain signals corresponding to microphones $i$ and $j$ respectively, $N$ is the time-frame length, and $n$ indicates the lag in the time domain. The local TDOA (in samples) for the microphone pair $(i, j)$ in time frame $l$, and sub-band region $\Omega$, is then estimated by

$$\hat{\tau}_{ij}(l, \Omega) = \arg\max_n \Psi_{ij}(l, \Omega, n). \quad (2)$$

We then compute a histogram from these locally estimated TDOAs across $T$ consecutive time frames and all sub-band regions. An example of such a histogram for one microphone pair in a scenario with four active sources is shown in Fig. 1. Since these histograms are computed from a large number of locally estimated TDOAs, we expect them to exhibit pronounced peaks in the vicinity of the true TDOAs of the active sources, while erroneous estimates, due to noise, reverberation,

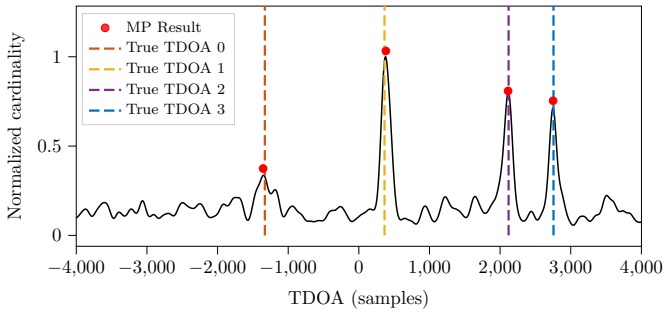


Fig. 1: Example of the local TDOAs distribution for a single microphone pair in a scenario with four active sources. The image also shows the peaks that were selected by the matching pursuit algorithm (MP result).

and fake detections are generally expected to be random, and assumed to add to the noise floor in the histogram.

To capture the pronounced peaks, we employ a matching pursuit algorithm similar to the one presented in [13]. This procedure yields a set of $S$ TDOA estimates per microphone pair, which we denote as $\hat{\tau}_{ij}^{s}$ $(s = 0, 1, ..., S-1)$ and are referred to as the candidate source TDOAs. In general, $S$ is an unknown parameter that can be estimated either in a preliminary stage or by appropriately tuning the stopping criterion of the matching pursuit algorithm [13], but for simplicity in this paper we assume that $S$ is known a priori.

Note that, in practical scenarios, certain microphone pairs may fail to detect one or more sources due to a low signal-to-noise ratio and/or if multiple sources generate very close TDOAs for which their peaks in the TDOA histogram cannot be distinguished. In such cases, spurious peaks may be selected by the matching pursuit algorithm. Nevertheless, these fake detections are subsequently eliminated during the association stage, as will be detailed in the following section.

### *B. Feature-based TDOA association*

We now need to associate candidate source TDOAs across microphone pairs, such that grouped TDOAs correspond to the same physical source. For this task, we propose a feature-based association step, and to realize it, each of the candidate source TDOAs $\hat{\tau}_{ij}^{s}$, must be linked to a feature vector. We make use of previously obtained sub-band information to define this vector as $\mathbf{h}_{ij}^{s} = [c_{ij}^{s}(\Omega_0), c_{ij}^{s}(\Omega_2), ..., c_{ij}^{s}(\Omega_{N-1})]^{\top}$, where $c_{ij}^{s}(\Omega)$ is a count of how many times a local TDOA estimated from the sub-band region $\Omega$, (from (2)) has contributed to the neighborhood around the peak located at the candidate source TDOA $\hat{\tau}_{ij}^{s}$, defined by an interval of size $W$ centered at $\hat{\tau}_{ij}^{s}$.

Each of these vectors represents the spectral content of the candidate source that generated its corresponding TDOA, and hence can be used for associating the candidate source TDOAs across microphone pairs to the same physical source. Fig. 2 shows the feature vectors $\mathbf{h}_{ij}^{s}$, for the candidate source TDOAs of three different microphone pairs in a scenario with four active sources. In this example, we observe that the spectral content from different microphone pairs shows a certain degree of similarity, suggesting that they were generated from the

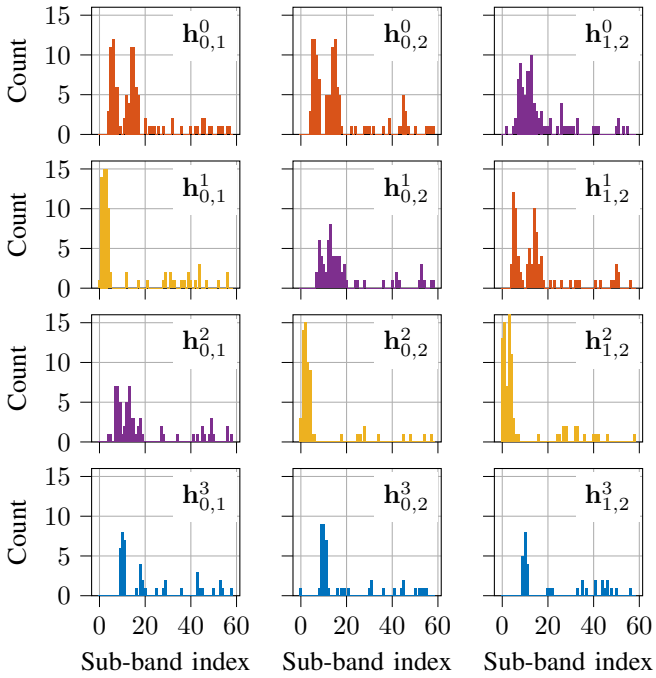


Fig. 2: Feature vectors $\mathbf{h}_{ij}^{s}$, for all the candidate source TDOAs in a scenario with four active sources and three microphones (three microphone pairs in total). Colors are indicative of which feature vectors have been associated.

same physical source. For this reason, an intuitive step to follow is to associate candidate source TDOAs based on their feature vector's similarities.

We therefore use a cosine similarity score between two feature vectors defined as

$$\mathrm{sim}\left(\mathbf{h}_{ij}^{s}, \mathbf{h}_{i'j'}^{s'}\right) = \frac{\mathbf{h}_{ij}^{s}(\Omega)^{\top}\, \mathbf{h}_{i'j'}^{s'}(\Omega)}{\|\mathbf{h}_{ij}^{s}(\Omega)\|\ \|\mathbf{h}_{i'j'}^{s'}(\Omega)\|}, \tag{3}$$

where $\|.\|$ is the $\ell_2$ norm. To respect the physical constraint that a single source can generate at most one candidate source TDOA per microphone pair, similarities are only computed between feature vectors from different microphone pairs. Two feature vectors are considered as 'prospective' if $\mathrm{sim}\left(\mathbf{h}_{ij}^{s}, \mathbf{h}_{i'j'}^{s'}\right) \geq \beta$, where $\beta$ is a predefined threshold. This thresholding eliminates isolated feature vectors whose spectral content is inconsistent with that of any physical source. Prospective pairs of feature vectors are then iteratively clustered into different groups by merging different pairs if they have a feature vector in common. This is done while still enforcing the constraint that each group contains at most one TDOA per microphone pair. In cases where this constraint is not satisfied, only the pair with the highest similarity is accepted. This procedure is repeated until no further merges are possible, resulting in $\hat{S}$ groups of feature vectors, where $\hat{S}$ is the estimated number of active sources. In general, as $\hat{S}$ can differ from the parameter $S$, when a certain source is not detected by at least two pairs, it will be considered a spurious detection and will end up not being associated with any group (resulting in $\hat{S} \leq S$). Recall that since each feature vector

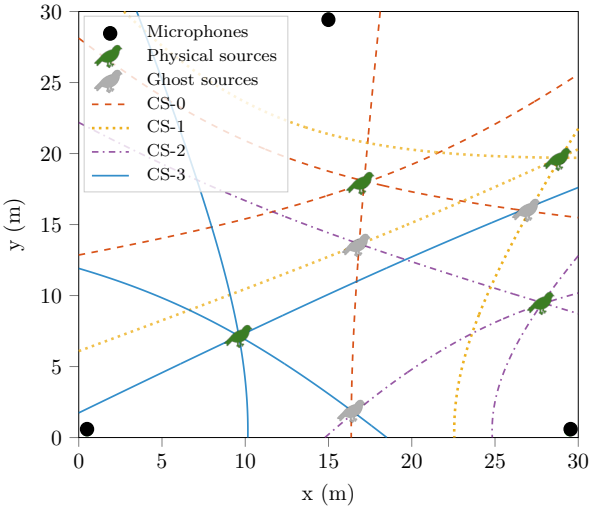


Fig. 3: Localization result for a scenario with four active sources and three microphones.

was linked to a single candidate source TDOA, the grouped feature vectors now results in a grouping of the candidate source TDOAs that correspond to the same physical source.

### C. Source localization

The $\hat{S}$ groups of candidate source TDOAs from the previous stage can now be employed to perform $\hat{S}$ independent single-source localizations. In this paper, we solve a nonlinear least-squares optimization problem, following the approach proposed in [7], however, any single-source position estimator can be used. This yields the final source position estimates $\hat{\mathbf{p}}_s = [\hat{x}_s, \hat{y}_s]$ $(s = 0, 1, \ldots, \hat{S}-1)$. Fig. 3 shows an example of the localization result for the scenario with three microphones and four active sources. Each green bird silhouette marks a true source position, and each gray silhouette a ghost source. The lines (CS-$i$, with $i = 0, 1, 2, 3$) are hyperbolas corresponding to the candidate source TDOAs, where hyperbolas sharing color and line style belong to the same group. The estimated positions $\hat{\mathbf{p}}_s$ are the intersections of these hyperbolas.

## IV. Results

To assess the performance of the proposed method, we conducted a series of Monte Carlo simulations within a $30\,\mathrm{m}^2$ environment, considering varying numbers of microphones (from 3 to 6) and simultaneously active sources (from 2 to 8). The microphone signals were synthesized using a forest impulse response simulator [15]. The microphones were positioned over the region on a uniform grid. In each simulation, the acoustic sources were randomly drawn from a pool of 16 sound recordings obtained from xeno-canto[1]. This pool comprises recordings from 10 individuals belonging to 10 distinct species, as well as 6 individuals of the same species (Emberiza citrinella). For each simulation configuration, we carried out 50 Monte Carlo trials, in which the source positions were independently sampled from a uniform spatial distribution over the environment.

We evaluated the performance of the proposed method under varying acoustic noise conditions. Specifically, spatially uncorrelated Gaussian noise was added independently to each microphone with two different noise power levels $L_N$; moderate ($L_N = 40\,\mathrm{dB}$ sound pressure level (SPL)) and high ($L_N = 60\,\mathrm{dB}$ SPL) at 1m. Source signals were normalized to $80\,\mathrm{dB}$ SPL at 1m, resulting in substantially different received source power levels at the individual microphones due to propagation-distance-dependent attenuation. As a consequence, the resulting signal-to-noise ratios (SNRs) differ across source–microphone pairs.

Microphone signals were processed with an STFT using a frame length of $N = 3\tau_{\max}$ and a $50\%$ frame overlap, where $\tau_{\max}$ denotes the maximum possible TDOA between microphone pairs, expressed in samples. The factor of three was introduced to ensure that $N$ is sufficiently large to provide adequate signal overlap at large time lags when computing the GCCs. The bandwidth of the sub-band regions $\Omega$, was set to $800$ Hz. To obtain robust feature estimates, multiple consecutive short-time frames were aggregated over an analysis interval of 3 s (corresponding to $T$ frames), during which all sources were active. For each microphone, the subset of microphone pairs was defined by selecting its three nearest neighbors. The size of the interval used for determining the local TDOAs that contributed to the candidate source TDOAs was set to $W = 0.02\tau_{\max}$. Finally, the cosine similarity threshold for grouping feature vectors was fixed at $\beta = 0.7$. The code implementing the proposed method and all simulations is publicly available online.[2]

### A. Metrics

We use two different metrics to evaluate the performance of the proposed method. As a first metric, we compute the percentage of missed detections as $\mathrm{MD} = 100(S - \hat{S})/S$. The second metric quantifies the error of the estimated position using the root mean square error (RMSE) between the source's ground truth positions, and the final position estimates, which for a single Monte Carlo run is defined as

$$\mathrm{RMSE} = \sqrt{\sum_{s=1}^{\hat{S}} \frac{\|\mathbf{p}_s - \hat{\mathbf{p}}_s\|^2}{\hat{S}}}. \tag{4}$$

Both the percentage of missed detections and the RMSE are reported as the average over all Monte Carlo runs within each simulation configuration.

### B. Analysis

Figs. 4 (a) and 4 (b) show the MD and the RMSE respectively as functions of the number of microphones ($M$), for scenarios with $S = 2, 4$, and $8$ simultaneously active sources, and for the two proposed noise power levels. In Fig. 4 (a), we can observe that the MD is reduced with increasing

[1] www.xeno-canto.org

[2] https://github.com/AlejandroMJR/Localizing-multiple-sound-sources-by-associating-sub-band-TDOA-features-for-biodiversity-monitoring

$M$ for both noise scenarios and for any number of active sources. This behavior is expected, as higher $M$ results in better area coverage due to higher microphone density, and lower source-microphone distances increase the probability of detection due to lower distance-dependent attenuation. We can also observe that environments with more active sources benefit from increasing $M$ (e.g., $S = 8$), the MD improves considerably for both noise scenarios (between $20\%$ and $30\%$). We observe that the MD does not exceed $45\%$ even in the worst-case scenario of 3 microphones and 8 simultaneously active sources, demonstrating a considerable reduction of the number of microphones needed for localizing multiple sound sources as compared to other methods [9]–[11].

Conversely, the RMSE does not exhibit a monotonic behavior, as illustrated in Fig. 4 (b). This behavior arises from the definition of RMSE over a Monte Carlo realization. As detailed in Section IV-A, the RMSE can only be evaluated over the set of detected sources $\hat{S}$. Consequently, even when examining the plots for a fixed number of active sources $S$, the effective problem complexity captured by the RMSE corresponds to a configuration with $\hat{S} \leq S$. Thus, two concurrent effects emerge as $M$ increases. Firstly, a larger value of $M$ is expected to yield a lower RMSE, since the number of available constraints grows. Secondly, the MD decreases, implying that more sources are successfully detected. This, however, increases the complexity of the association and localization stages, which tends to increase the RMSE. The interplay between these two opposing effects leads to the non-monotonic RMSE behavior observed in the results. Nevertheless, it can be observed that the RMSE remains below 1 m even for the most complex scenarios, which is relatively small when compared with the overall dimensions of the environment.

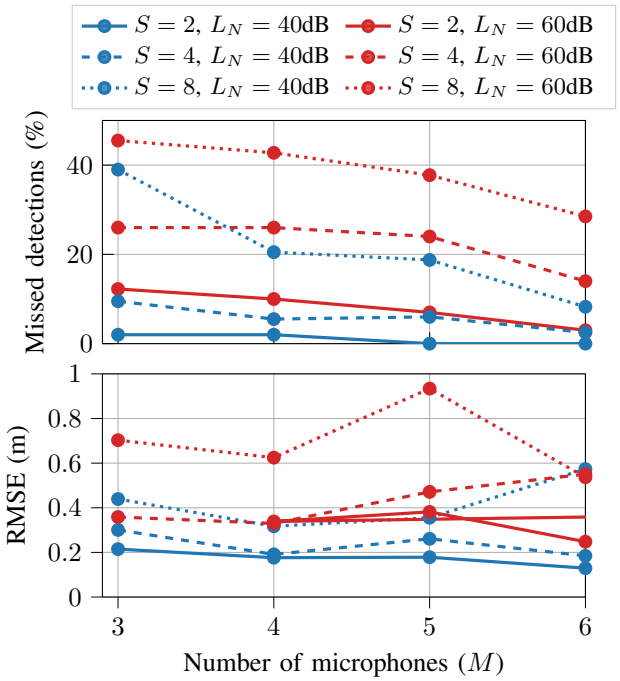


Fig. 4: (a) Percentage of missed detections and (b) RMSE, as a function of number of microphones. Results are shown for 2, 4, and 8 active sources under two different noise levels.

## V. Conclusions

We have presented a method for localizing multiple sound sources that makes use of a sub-band feature vector to associate estimated TDOAs across microphone pairs. As this association allows for the multiple source localization problem to be decomposed into independent single source localization tasks, it reduces the number of microphones required for reliable operation. This makes the method promising for the biodiversity monitoring context, where limited number of single-channel autonomous recording units are usually deployed over an area to be surveyed. However, a more thorough evaluation on realistic data is needed to fully gauge its potential.

## References


[1] T. A. Rhinehart, L. M. Chronister, T. Devlin, and J. Kitzes, "Acoustic localization of terrestrial wildlife: Current practices and future opportunities," *Ecology and Evolution*, vol. 10, no. 13, pp. 6794–6818, 2020.

[2] C. Pérez-Granados and J. Traba, "Estimating bird density using passive acoustic monitoring: a review of methods and suggestions for further research," *Ibis*, vol. 163, no. 3, pp. 765–783, 2021.

[3] L. P. Fitzsimmons, J. R. Foote, L. M. Ratcliffe, and D. J. Mennill, "Frequency matching, overlapping and movement behaviour in diurnal countersinging interactions of black-capped chickadees," *Animal Behaviour*, vol. 75, no. 6, pp. 1913–1920, 2008.

[4] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, 2017.

[5] A. P. Hill, P. Prince, J. L. Snaddon, C. P. Doncaster, and A. Rogers, "Audiomoth: A low-cost acoustic device for monitoring biodiversity and the environment," *HardwareX*, vol. 6, p. e00073, 2019.

[6] C. M. Wood, F. Günther, A. Rex, D. F. Hofstadter, H. Reers, S. Kahl, M. Z. Peery, and H. Klinck, "Real-time acoustic monitoring facilitates the proactive management of biological invasions," *Biological Invasions*, vol. 26, no. 12, pp. 3989–3996, 2024.

[7] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer speech & language*, vol. 11, no. 2, pp. 91–126, 1997.

[8] M. A. J. Rodríguez, R. Ali, and T. van Waterschoot, "Direction-of-arrival data association for wildlife acoustic localization," in *2025 33rd European Signal Process. Conf. (EUSIPCO)*, 2025, pp. 211–215.

[9] H. Sundar, T. V. Sreenivas, and C. S. Seelamantula, "Tdoa-based multiple acoustic source localization without association ambiguity," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 1976–1990, 2018.

[10] R. Boora and S. K. Dhull, "A TDOA-based multiple source localization using delay density maps," *Sādhanā*, vol. 45, no. 1, p. 204, 2020.

[11] X. Dang, H. Zhu, and Q. Cheng, "Multiple sound source localization based on a multi-dimensional assignment model," in *2018 21st International Conf. on Inf. Fusion (FUSION)*. IEEE, 2018, pp. 1732–1737.

[12] M. Cobos, F. Antonacci, L. Comanducci, and A. Sarti, "Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1270–1281, 2020.

[13] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[14] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I–529.

[15] S. Kaneko and H. Gamper, "A fast forest reverberator using single scattering cylinders," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021, pp. 1–5.