Direction-of-Arrival Data Association for Wildlife Acoustic Localization

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 animals over time provides useful additional information for understanding animal behavior and for ecology studies in general. A common approach for this task is to deploy microphone arrays (nodes) and use the acoustic signals to estimate the direction of arrival (DOA) of the sound source. DOAs from different nodes are then intersected to find the source's position. However, when multiple sources are active, the DOA association problem (AP) arises as it becomes unclear which DOAs correspond to the same source. This problem is further exacerbated in bioacoustical scenarios where large distances increase the error in the DOA estimates, and sounds often overlap in both time and frequency. In this paper, we propose a method to tackle the DOA AP in such challenging environments. In particular, we beamform to each of the estimated DOAs and extract features that characterize each of the detected sources, then, we associate features from different nodes based on their similarity, resulting in groups of DOAs that belong to the same source. Preliminary simulations suggest the potential of the proposed method for scenarios with missed detections and unknown number of sources, even when the number of microphones available at each node is limited.**

***Index Terms*—sound source localization, wildlife monitoring, direction of arrival, data association problem.**

I. Introduction

Wildlife acoustic localization (WAL) is the task of estimating the position of one or more animal individuals using acoustic signals obtained from microphones that have been deployed in outdoor environments. This has become a useful task within the ecological context as the knowledge of animals' positions over time has contributed to understanding animal interactions, quantifying species densities, and observing animal responses to various disturbances, among several other purposes [1], [2]. Although there has been significant research into sound source localization in general [3], WAL presents a number of unique challenges, including but not limited to simultaneous activity of multiple sound sources overlapping in time and frequency, low signal-to-noise ratios of acoustic signals due to propagation attenuation and environmental noise, and sparse microphone arrays with little to no synchronization.

A common scenario for WAL is the deployment of spatially distributed microphone arrays (also referred to as nodes) across the region of interest. In the single-source case, each node estimates a direction of arrival (DOA) of the sound, after which the source's position can be estimated by intersecting these DOAs [4], [5]. When multiple sources are active, however, each node will estimate multiple DOAs, resulting in DOA intersections that yield ambiguous source locations, a problem referred to as the direction-of-arrival association problem (DOA AP) [6]–[8]. In this paper, we assume that DOA estimates are available from some number of microphone nodes, and focus on tackling the DOA AP within the wildlife monitoring context.

To address the DOA AP in general, one possible solution is to extract features from detected sources and then use them to guide the DOA association. For instance, the authors in [7] proposed to construct a histogram, based on how the frequency components of the signals at each node are distributed among detected sources. These histograms were subsequently grouped based on their similarity, resulting in a grouping of DOAs corresponding to the same source. However, this method assumes a known number of sources, and signals that satisfy the window-disjoint orthogonality, making it unsuitable for bioacoustic scenarios [1], [9]. An alternative approach was proposed in [6], where the estimated associations were obtained by using only the DOA estimates. To achieve this, the authors selected the optimal groupings of DOAs by maximizing the ratio of two likelihoods: one representing the probability that the detections come from real sources and another assuming they are all false alarms. The optimization problem was then re-framed as a source-destination assignment problem and solved using Lagrange relaxation. While it does not rely on the previously mentioned assumptions, its performance degrades in low signal-to-noise ratio (SNR) conditions, specially when the number of sources increases.

In this paper, we propose yet an alternative method to approach the DOA AP, but one that is more suitable for a wildlife monitoring context, taking into account the aforementioned challenges. The proposed method consists of two steps. Firstly, we perform a beamforming-based feature extraction step. This involves applying a beamformer to the available DOAs from each node in order to obtain an estimate of the corresponding source signal. These signals are then fed into a pre-trained species classification neural network to extract the vector embeddings. In the second step, we use the obtained embeddings to compute a similarity metric that quantifies the quality of a candidate association, after which the correct associations are found by solving a fractional multi-dimensional

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.

assignment problem (F-MDAP). Monte Carlo simulations are used to gauge the potential of the approach in relation to the DOA estimation error, the number of microphones used for per node, and background noise levels. The results suggest that high association accuracies are possible for several relevant cases and warrants further evaluation on more realistic data.

The remaining of the paper is organized as follows. Section II formulates the DOA AP for a general scenario. Section III describes the proposed method in detail. The evaluation of the proposed method via simulation is presented in Section IV. Finally, Section V provides the conclusions.

II. Problem Statement

We consider a set of N microphone nodes, each consisting of M microphones, deployed in the region of interest. The geometry of the arrays, and the positioning of each node can be any configuration that suits the specific requirements of the wildlife monitoring task. For each node, we assume that we have access to a set of estimated DOAs from multiple locally detected sources. In this work we consider 2D localization, meaning that only one angle is needed to represent each DOA. Figure 1 illustrates a typical scenario with four nodes (light blue circles) each consisting of a different number of microphones (smaller dark blue circles). Nodes 1, 3, and 4 have two estimated DOAs, while node 4 has only one. The true number of sources in this scenario is three, depicted by the orange bird silhouettes. There is, however, an ambiguity as to which DOAs from each node correspond to the same true source (the so-called DOA AP). The consequence of this ambiguity is shown in Fig. 1, where the pairwise intersections of the estimated DOAs from each node yields a number of ghost sources (grey bird silhouettes) in addition to the true sources. Our goal in this work is therefore to determine the correct associations among DOAs from different nodes in order to ultimately determine a source's location within an outdoor environment.

III. Proposed Method

Our proposed method to address the DOA AP consists of a feature extraction step, followed by a multi-dimensional assignment problem (MDAP) to be solved. As it is challenging to separate multiple sources due to overlap in both time and frequency (for instance, when sources correspond to the same species), each node firstly applies spatial processing (beamforming) to the direction(s) specified by that node's estimated DOA(s). These beamformed signals are then fed into a pre-trained species classification neural network to extract the vector embeddings. In this work, we use BirdNet [10] for this purpose, however any other relevant species classifier can be substituted. In the second step of the proposed method, these embeddings are used as features to compute an association similarity score. By maximizing the total similarity score, the estimated associations can then be determined. For this optimization problem, we use a variation of the classical MDAP [11] that allows for variable assignment cardinality.

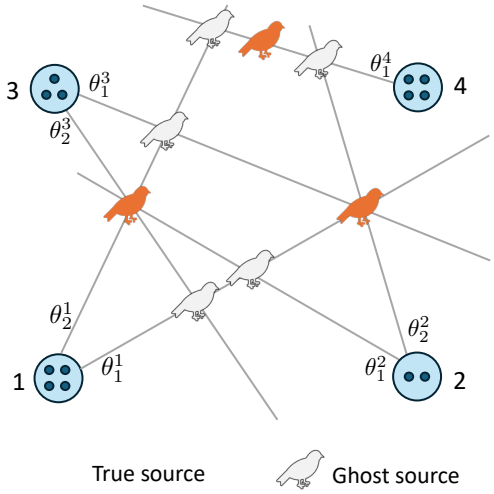


Fig. 1: Example of a common multi-source scenario in which the DOA AP arises. Nodes are illustrated as the light blue circles, each of which consists of a different number of microphones depicted by the smaller dark blue circles.

A. Feature Extraction

We start from a set of noisy DOA estimates $\left\{\hat{\theta}_{i_n}^n(t)\right\}$ at a time instant t, from each node n $(n = 1, ..., N)$. We assume that each of the DOA estimates is modeled by

$$\hat{\theta}_{i_n}^n(t) = \begin{cases} \theta_{i_n}^n(t) + \eta_{i_n}^n - 2\pi, & \theta_{i_n}^n(t) + \eta_{i_n}^n > 2\pi \\ \theta_{i_n}^n(t) + \eta_{i_n}^n + 2\pi, & \theta_{i_n}^n(t) + \eta_{i_n}^n < 0 \\ \theta_{i_n}^n(t) + \eta_{i_n}^n, & \text{otherwise} \end{cases}, \quad (1)$$

where $\theta_{i_n}^n$ are the true (but unknown) DOAs from the node n to the i_n-th $(i_n = 1, ..., P_n)$ detected source, P_n is the number of sources detected by the n-th node, and $\eta_{i_n}^n$ is some DOA estimation error, modeled by a Gaussian distribution $\mathcal{N}(0, \sigma_\eta^2)$, with standard deviation σ_η. Note that $\theta_{i_n}^n$ is in the range $[0, 2\pi]$, and the conditions $\theta_{i_n}^n(t) + \eta_{i_n}^n > 2\pi$ and $\theta_{i_n}^n(t) + \eta_{i_n}^n < 2\pi$ guarantee that the noisy estimates $\hat{\theta}_{i_n}^n$ remain in that same interval. Since BirdNet operates on three-second-long input signals, we extract the corresponding three-second segment of microphone signals from each node centered around the detection instant, t. These signals are transformed to the short-time Fourier transform (STFT) domain, resulting in the complex-valued signal, $x_m^n(l, k)$, where $l = 1, \ldots, L$ is the time-frame index with L frames, $k = 1, \ldots, K$ is the frequency bin index with K frequency bins, and $m = 1, \ldots, M$ is the microphone index with M microphones. In the following steps, we omit the node index n since the process is the same for every node.

For each node, and time-frequency bin index, we apply the narrowband minimum power distortionless response (MPDR) beamformer [12]

$$y_{i_n}(l, k) = \mathbf{w}_{i_n}^H(k)\mathbf{x}(l, k), \quad (2)$$

where $\{.\}^H$ is the Hermitian transpose, $\mathbf{x}(l, k) = [x_1(l, k), \ldots, x_M(l, k)]^T$ is the vector of microphone signals from each node ($\{.\}^T$ is the transpose), and $\mathbf{w}_{i_n}(k) =$

$[w_{i_n,1}(k), \ldots, w_{i_n,M}(k)]^T$ is the corresponding MPDR beamformer given by

$$\mathbf{w}_{i_n}(k) = \frac{\mathbf{R}^{-1}(k)\mathbf{a}(\hat{\theta}_{i_n})}{\mathbf{a}^H(\hat{\theta}_{i_n})\mathbf{R}^{-1}(k)\mathbf{a}(\hat{\theta}_{i_n})}, \tag{3}$$

where $\mathbf{R}(k)$ is an estimated covariance matrix using all time frames, L, at frequency bin, k, with components given by $R_{i,j}(k) = \frac{1}{L}\sum_{l=1}^{L} x_i(l,k)x_j^*(l,k)$. $\mathbf{a}(\hat{\theta}_{i_n}) = \left[a_1(\hat{\theta}_{i_n}), \ldots, a_n(\hat{\theta}_{i_n})\right]^T$ is the steering vector [12] of the array towards the direction $\hat{\theta}_{i_n}$ with

$$a_m(\hat{\theta}_{i_n}) = e^{-j2\pi\frac{kf_s}{L}\tau_m(\hat{\theta}_{i_n})}; \quad m = 1, ..., M, \tag{4}$$

where τ_m is the time delay at the m-th microphone relative to a fixed reference time and f_s is the sampling frequency. Under far-field assumption, τ_m is determined by the array geometry.

Performing this process across all frequency bins k and all nodes n, results in the STFT representation of the broadband beamformed signals for each of estimated DOA of every node, denoted as $Y_{i_n}^n(l,k)$. These signals are then fed into BirdNet, where the vector embedding of the final hidden layer is used as a feature vector. We denote this as $\mathbf{E}_{i_n}^n$, the feature vector extracted from the i_n-th detected source of the n-th node.

Since an initial estimate of the source signal is obtained by beamforming to a specific estimated direction, the extracted feature can be interpreted as a representation of the source signal received from that direction. This means that if we compare feature vectors from different nodes, those corresponding to the same source are expected to be similar. Consequently, if we associate features based on their similarity, we can obtain an estimate of the DOA associations.

B. Multi-dimensional Assigment Problem

In this Section, we frame the DOA AP as an MDAP. Our goal is to associate DOAs among different nodes that correspond to the same real source using the extracted features $\mathbf{E}_{i_n}^n$ ($i_n = 1, ..., P_n$). To handle missed detections, we add a dummy feature $\mathbf{E}_0^n$ to each node, so that if a source is missed by a node, the dummy can be assigned instead. Hence we refer to $\mathbf{E}_{i_n}^n$ as real features and $\mathbf{E}_0^n$ as dummy features. The problem translates into finding the set of feature vector associations such that all of the feature vectors corresponding to the same source are associated together. Defining an association as the N-tuple $A_{i_1,...,i_N} = \left\{\mathbf{E}_{i_n}^n\right\}_{n=1}^N$, which means that the features $\mathbf{E}_{i_1}^1$ from node 1, $\mathbf{E}_{i_2}^2$ from node 2,..., and $\mathbf{E}_{i_N}^N$ from node N were assigned together to the same source, the goal is to find one association per active source. We call the complete set of associations an assignment. As an example, in the scenario illustrated in Fig. 1, the correct assignment involves the associations $A_{2,1,2,0}$, $A_{1,2,1,0}$, and $A_{0,0,0,1}$.

Let us start by defining the similarity between features in a possible association $A_{i_1,...,i_N}$. Let $\{1, ..., n', ..., N'\}$ be the subset of nodes contributing with real features to the association, then, the similarity value $T_{i_1,...,i_N}$ for the association $A_{i_1,...,i_N}$, with N' real features and $N_d = N - N'$ dummy features is given by

$$T_{i_1,...,i_N} = \frac{2W(N_d)}{N'(N'-1)}\sum_{n'=1}^{N'}\sum_{m'=n'+1}^{N'} S(\mathbf{E}_{i_{n'}}^{n'}, \mathbf{E}_{i_{m'}}^{m'}), \tag{5}$$

where $S(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A}.\mathbf{B}}{\|\mathbf{A}\|.\|\mathbf{B}\|}$ is the pairwise cosine similarity, and $W(N_d)$ is a weighting function used to penalize the use of dummies in the association. A deeper analysis on this function is provided in Section IV. If there is just one real feature used in the association, its similarity is set to the maximum similarity obtained in the rest of associations.

In a classical MDAP framework, the goal would be to find the assignment that maximizes the sum of similarities between associated features. However, in our approach, the number of associations is variable, as it depends on the number of active sources P, which is unknown. This makes the sum of similarities a suboptimal objective function. Instead, we propose to use the average similarity between assigned associations. By doing this, the algorithm equally considers every possible number of active sources. The problem translates into the fractional programming optimization problem:

$$\begin{aligned}
\max_{x_{i_1,\ldots,i_N}} \quad & \frac{\sum_{i_1=0}^{P_1}\cdots\sum_{i_N=0}^{P_N} T_{i_1,...,i_N}x_{i_1,...,i_N}}{\sum_{i_1=0}^{P_1}\cdots\sum_{i_N=0}^{P_N} x_{i_1,...,i_N}} \\
\text{s.t.} \quad & \sum_{i_2=0}^{P_2}\cdots\sum_{i_N=0}^{P_N} x_{i_1,...,i_N} = 1, i_1 = 1, ..., P_1 \\
& \sum_{i_1=0}^{P_1}\cdots\sum_{i_N=0}^{P_N} x_{i_1,...,i_N} = 1, i_2 = 1, ..., P_2 \\
& \qquad\qquad \vdots \\
& \sum_{i_1=0}^{P_1}\cdots\sum_{i_{N-1}=0}^{P_{N-1}} x_{i_1,...,i_N} = 1, i_N = 1, ..., P_N,
\end{aligned} \tag{6}$$

where $x_{i_1,...,i_N}$ are the decision variables defined as

$$x_{i_1,...,i_N} = \begin{cases} 1 & \text{if } A_{i_1,...i_N} \text{ is assigned} \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

The constraints serve to ensure that only feasible assignments are considered. A feasible assignment meets the following conditions:

1) Every real feature must be assigned once,
2) Each association can contain only one feature per node,
3) Dummy features can be assigned multiple times.

The proposed F-MDAP can be solved by applying the Dinkelbach's Algorithm [13], which is NP-hard as the number of possible associations grows exponentially with the number of dimensions (nodes in our case). In the wildlife monitoring scenario, however, it is expected that the number of nodes used to detect the same sources within some area is limited. Consequently, it is possible to find the exact solution to (6).

When the total similarity is maximized, we end up with a set of $\hat{P}$ associations $\mathcal{A} = \left\{A_{i_1,...,i_N}^j\right\}_{j=1}^{\hat{P}}$, where $\hat{P}$ is the estimated number of active sources. Each association in this

set tells us which are the features, and therefore, the DOAs from each node that belong to the same real source.

IV. Results

To evaluate the proposed method, we conducted a series of Monte Carlo simulations in a $30\,\mathrm{m}^2$ free-field environment with 4 nodes arranged in a $10\,\mathrm{m}\times10\,\mathrm{m}$ grid (similar to Fig. 1) centered in the middle of the environment. Each node is equipped with a circular microphone array with M microphones spaced $0.05\,\mathrm{m}$ apart (the impact of M is observed in Sec. IV-C). In each simulation, the sources were randomly selected from a pool of eight sound recordings taken from xeno-canto[1]. The pool consists of four individuals of the same species Anthus trivialis, one Tetrao urogallusm, one Vulpes vulpes, one Poecile montanus, and, to include external sources, one aircraft. It is important to highlight that we included three types of sounds that are not present in the training dataset of BirdNet (Vulpes vulpes, Poecile montanus, and the aircraft). For each experiment, we performed 100 Monte Carlo simulations, randomly sampling the source positions from a uniform distribution across the environment.

We evaluated the performance under different noise conditions. To achieve this, we introduce uncorrelated Gaussian noise at each microphone, with noise power level L_{noise} ranging from $50\,\mathrm{dB}$ sound pressure level (SPL) to $60\,\mathrm{dB}$ SPL. Source signals were scaled to $80\,\mathrm{dB}$ SPL before propagation, leading to significantly varying received source signal power levels at each microphone, as they depend on the propagation distance. Consequently, the resulting SNRs vary for each source-microphone pair. Missed detections are added to the simulations when two sources have an angular separation lower than the node's angular resolution (which is related to the standard deviation of the DOA estimation error, σ_η), or when a source is masked by background noise or other by sources, based on the received power at the microphones. The code used to implement the proposed method and conduct the experiments is available online[2].

The similarity metric defined in (5) requires the choice of the weighting function $W(N_d)$. We parametrize $W(N_d)$ with the parameter α as:

$$W(N_d) = 1 - \alpha N_d, \quad N_d = 1, 2, 3. \tag{8}$$

High values of α strongly penalize the use of dummies, which is beneficial when the percentage of missed detections is low, i.e., when it is expected that the majority of nodes detected the same source. Conversely, if missed detections are expected to be frequent, a lower value of α reduces the penalty, making the function more tolerant to missed detections. During the experiments, we fix α to a value of 0.03, as this provides a robust balance across different scenarios without requiring prior knowledge of the missed detection rate.

[1]www.xeno-canto.org

[2]https://github.com/AlejandroMJR/DOA_Association_WAL

A. Metrics

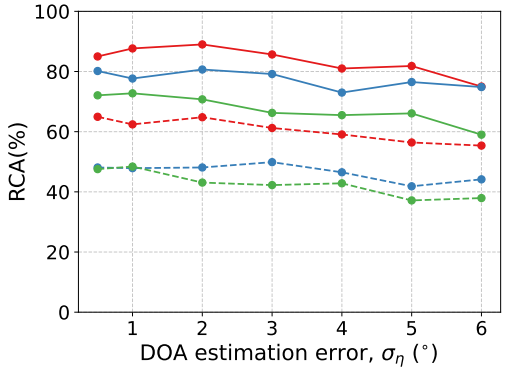
To validate the results, we compute two different metrics. The first metric is the ratio of correct associations (RCA), which evaluates the accuracy of an assignment. It is defined as the ratio between the number of estimated associations in $\mathcal{A} = \left\{A^j_{i_1,\ldots,i_N}\right\}_{j=1}^{P}$, that are present in the set of ground truth associations $\mathcal{G} = \left\{G^j_{i_1,\ldots,i_N}\right\}_{j=1}^{P}$, to the total number of true associations P. However, this metric does not always reflect the true performance of the method. For instance, an outcome with partially correct associations will always be evaluated with a score of 0, even if the majority of DOAs were correctly assigned. To address this limitation, inspired by [7], we introduce the ratio of correct pairwise associations (RCPA), which measures the number of correctly assigned pairs inside every association relative to the total number of possible pairs. This metric provides a more fair evaluation by accounting for partially correct associations.

B. Effect of DOA estimation error

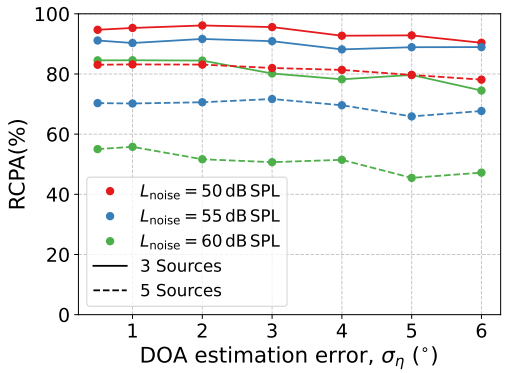
We start by analyzing the impact of the DOA measurement error on the accuracy of the method. Fig. 2 shows the RCA (Fig. 2a) and RCPA (Fig. 2b), as a function of the standard deviation of the DOA measurement error σ_η, using nodes with $M = 4$. Results are presented for 3 and 5 sources, with varying background noise levels. Given that the accuracy of the DOA estimation decreases in large environments, we consider values of σ_η up to 6°. Since σ_η is related to the node's angular resolution, increasing it also rises the percentage of missed detections. As expected, larger values of σ_η reduce the accuracy; however, the effect is not critical, as revealed by the RCPA, which remains relatively constant. These results exhibit robustness not only to large measurement errors, but also to missed detections. In contrast, if we look at the effect of the different noise levels L_{noise}, it is clear that very high noise levels affect the performance. This is attributed to lower SNRs values for each of the received source signals, and to the higher percentage of missed detections due to noise masking.

C. Effect of number of microphones per node

For these simulations we set $\sigma_\eta = 3^\circ$, and tested cases with 3 and 5 sources under varying values of L_{noise}, and varying numbers of microphones per node, M. Fig. 3 depicts the RCA (Fig. 3a) and RCPA (Fig. 3a) as a function of M. We observe that the performance is considerably reduced for $M = 2$, which is presumably due to the low angular resolution of such a small array, and spatial aliasing issues, adding errors to the feature extraction step. However, increasing the number of microphones to $M \geq 3$ results in a substantial improvement, particularly for the case of 3 sources, where the RCPA reaches more than 70% for all noise levels, indicating that the proposed method maintains its performance even when the number of microphones per node is limited. Higher M provides better angular resolution, which is beneficial in cases with a high number of sources, as their angular separation decreases. This improvement is reflected in the results for 5 sources.

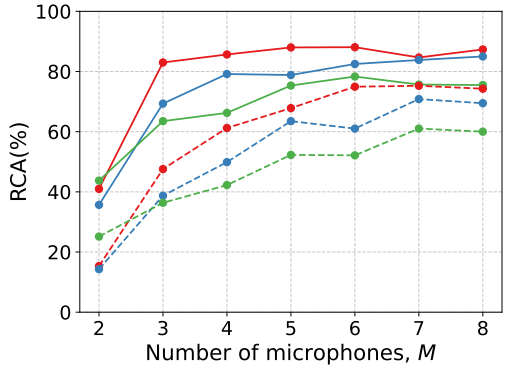


(a)

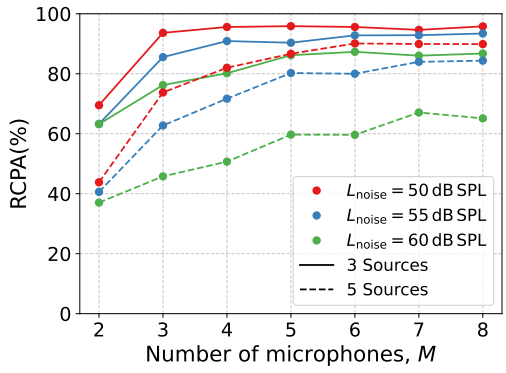


(b)

Fig. 2: RCA (a) and RCPA (b) accuracy as a function of σ_η, for 3 and 5 sources with varying background noise levels.



(a)



(b)

Fig. 3: RCA (a) and RCPA (b) accuracy as a function of M, for 3 and 5 sources with varying background noise levels.

V. Conclusions

In this work, we have proposed a two-step method to address the direction-of-arrival (DOA) data association problem for wildlife acoustic localization in scenarios where spatially distributed microphone arrays (nodes) are deployed across a region of interest for wildlife monitoring. Given noisy DOA estimates from each node, the first step applies a beamformer to obtain an estimate of the source signal, which is subsequently fed into a species classifier, yielding a corresponding vector embedding. Vector embeddings are then used in a second step to define a similarity metric for a fractional multi-dimensional assignment problem, which is solved to retrieve the correct DOA associations, and hence the location of the respective sound source. Simulations in scenarios involving DOA errors, missed detections, and varying numbers of microphones per microphone node have suggested a promising performance of the method. To fully gauge its potential, however, a more extensive evaluation using real data from wildlife passive acoustic monitoring scenarios is required.

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] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Commun. and Mobile Computing*, vol. 2017, no. 1, pp. 1–24, 2017.

[4] L. M. Kaplan, Q. Le, and N. Molnar, "Maximum likelihood methods for bearings-only target localization," in *Proc. 2001 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '01)*, vol. 5, 2001, pp. 3001–3004.

[5] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Processing*, vol. 107, pp. 54–67, 2015.

[6] X. Dang, Q. Cheng, and H. Zhu, "Indoor multiple sound source localization via multi-dimensional assignment data association," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, 2019.

[7] A. Alexandridis and A. Mouchtaris, "Multiple sound source location estimation in wireless acoustic sensor networks using doa estimates: The data-association problem," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 2, pp. 342–356, 2017.

[8] M. Swartling, B. Sällberg, and N. Grbić, "Source localization for multiple speech sources using low complexity non-parametric source separation and clustering," *Signal Processing*, vol. 91, no. 8, 2011.

[9] N. Lin, H. Sun, and X.-P. Zhang, "Overlapping animal sound classification using sparse representation," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '18)*, 2018, pp. 2156–2160.

[10] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.

[11] W. P. Pierskalla, "The multidimensional assignment problem," *Operations Research*, vol. 16, no. 2, pp. 422–431, 1968.

[12] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[13] S. Schaible, "Fractional programming. ii, on dinkelbach's algorithm," *Management science*, vol. 22, no. 8, pp. 868–873, 1976.