AUGUST 24 2022

Source separation with an acoustic vector sensor for terrestrial bioacoustics

Irina Tolkova ; Holger Klinck

J Acoust Soc Am 152, 1123 (2022)
https://doi.org/10.1121/10.0013505

JASA ARTICLE

Source separation with an acoustic vector sensor for terrestrial bioacoustics

Irina Tolkova[1,a] and Holger Klinck[2]
[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA
[2]K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, Ithaca, New York 14850, USA

ABSTRACT:
Passive acoustic monitoring is emerging as a low-cost, non-invasive methodology for automated species-level population surveys. However, systems for automating the detection and classification of vocalizations in complex soundscapes are significantly hindered by the overlap of calls and environmental noise. We propose addressing this challenge by utilizing an acoustic vector sensor to separate contributions from different sound sources. More specifically, we describe and implement an analytical pipeline consisting of (1) calculating direction-of-arrival, (2) decomposing the azimuth estimates into angular distributions for individual sources, and (3) numerically reconstructing source signals. Using both simulation and experimental recordings, we evaluate the accuracy of direction-of-arrival estimation through the active intensity method (AIM) against the baselines of white noise gain constraint beamforming (WNC) and multiple signal classification (MUSIC). Additionally, we demonstrate and compare source signal reconstruction with simple angular thresholding and a wrapped Gaussian mixture model. Overall, we show that AIM achieves higher performance than WNC and MUSIC, with a mean angular error of about 5°, robustness to environmental noise, flexible representation of multiple sources, and high fidelity in source signal reconstructions.
 https://doi.org/10.1121/10.0013505
(Received 11 May 2022; revised 20 July 2022; accepted 24 July 2022; published online 24 August 2022)
[Editor: Karim G. Sabra] Pages: 1123–12

I. INTRODUCTION

To design and implement effective conservation policies, there is a need for a thorough understanding of large-scale population dynamics. However, monitoring through manual observation is infeasible to perform at the temporal and spatial scales necessary to track changes in global biodiversity. To this end, recent years have brought a rise in automated technologies such as camera traps,[1–3] drone footage,[4–6] or even satellite imagery[7] for species-level identification and population estimation. Yet, while such vision-based methods can be effective under certain conditions, particularly for large mammals,[8] their general use is limited by occlusion, constrained by animal size, and primarily confined to terrestrial environments. An alternative emergent approach is passive acoustic monitoring (PAM). PAM is a relatively low-cost, non-invasive methodology which can be applied for highly diverse taxa (birds, cetaceans, elephants, frogs, and others) across terrestrial and marine ecosystems and which has proven to be successful for both ecological study and wildlife conservation efforts.[9–13] However, there are several challenges limiting the performance and robustness of PAM. A particular difficulty is the occurrence of overlapping calls, or the overshadowing of calls with high-amplitude noise. As PAM classifiers are frequently trained on "focal recordings"—recordings with a high signal-to-noise ratio (SNR) made in close proximity to individual animals—such systems experience a domain shift when deployed in complex natural soundscapes, further exacerbating this challenge.[14]

One approach to address overlap could be to separate an acoustic recording into multiple components corresponding to different sound sources prior to further analysis. Source separation over monaural audio classically relies on matrix decomposition methods, such as independent component analysis (ICA), principal component analysis (PCA), or non-negative matrix factorization (NMF), which calculate a lower-dimensional basis with specified properties for the observed signal.[15–17] More recently, studies have leveraged deep learning to differentiate sources.[18,19]

A fundamental challenge of single-channel source separation is that the problem is under-determined; there are many possible choices of a basis for vocalizations such that mixtures would yield the observed signal.[15] This issue could be addressed with an array of microphones, with which an individual sound event can be localized by cross-correlating different channels to find the time-difference-of-arrival across microphones, triangulating to determine source location, and beamforming to reconstruct the true signal.[13,20] Localization with acoustic arrays has been used to study animal movement, behavior, population densities, and anomalous sound events such as gunshots; a thorough review is given by Rhinehart *et al.*[21] While array processing can be very fruitful, it carries some limitations. First, array systems

[a]Electronic mail: itolkova@g.harvard.edu

are more expensive than monaural microphones and require more resources to deploy and maintain. Additionally, microphone arrays are prone to clock drift—de-synchronization of time across different recorders—resulting in inaccurate time difference estimates and therefore errors in localization.[13] An intermediate approach between monaural analysis and microphone array processing is the use of co-located microphones—such as stereo or quadraphonic systems—which capture spatial acoustic information while requiring a single clock and fewer hardware components. In this work, we consider acoustic vector sensors (AVSs), devices that measure acoustic pressure and particle velocity,[22] commonly implemented as systems of three or four co-located capsules.[23] AVSs enable streamlined direction-of-arrival (DoA) estimation through calculation of active intensity vector statistics, which makes it possible to spatially distinguish different sound sources.

In underwater acoustics, the use of AVS-integrated hydrophones dates back to the development of directional low-frequency analysis and recording sonobuoys (DIFARs) by the United States Navy in 1965.[24] DIFARs and the successive directional autonomous seafloor acoustic recorders (DASARs) have since become an established tool within marine bioacoustics.[25–27] In particular, Thode *et al.* (2019)[28] demonstrated how active-intensity-based DoA estimation with DIFAR recordings can be used to visualize whale calls through "azigrams," representations of spectrograms indicating directionality for each pixel. Azigrams give an additional dimension to time-frequency signal analysis and bridge the gap between image processing techniques and beamforming. Furthermore, active intensity methods with arrays of AVSs have been leveraged for two-dimensional localization and tracking of bowhead and humpback whales and of coral reef ecosystems.[26,29–31] Yet, despite the adoption of AVSs in marine bioacoustics, applications of co-located arrays in the terrestrial domain have been very limited.

In this work, we propose using an AVS to perform DoA estimation and separate overlapping vocalizations, with a focus on bird calls in terrestrial environments. We build on recent work in marine bioacoustics and compare the performance of the active intensity method (AIM) against two established DoA estimation algorithms: white noise gain constraint beamforming (WNC) and multiple signal classification (MUSIC). We start with the theoretical foundation and implementation details of DoA estimation and source signal reconstruction (Sec. II). We then evaluate angular accuracy and reconstruction fidelity through both simulation and controlled outdoor experiments (Sec. IV). We report results for one and two sources and describe the effects of signal characteristics, algorithm parameters, and environmental factors on DoA estimation (Sec. V). Finally, we discuss challenges and possibilities for further analysis, such as the visualization of vocal activity across both time and azimuth angles (Sec. VI).

II. DOA ESTIMATION

In general, DoA estimation with microphone arrays is achieved through beamforming: a family of methods for assigning weights to coherently combine measurements across different sensors. Alternatively, eigenspace-based techniques estimate the dominant directional modes within the data, subject to the constraints of the array architecture. While these methods can be applied to any microphone array, the particular properties of an AVS enable empirical calculation of the active intensity vector, and the corresponding DoA, for each time sample. In this work, we compare the performance of a beamforming method (white noise gain constraint beamforming), an eigenspace method (multiple signal classification), and an active-intensity-based method.

Let R be the number of sensors (for our microphone, $R=4$), N be the number of sampled time points, and K be the number of signal sources. We assume that the recorded audio can be described by a linear model,

$$X = \sum_{k=1}^{K} m(\phi_k, \theta_k) S_k + W, \tag{1}$$

where $X \in \mathbb{C}^{R\times N}$ is the matrix of measured signals, $\phi_k \in [-\pi, \pi]$ is the azimuth angle of source k, $\theta_k \in [-\pi/2, \pi/2]$ is the elevation angle of source k, $m(\phi_k, \theta_k) \in \mathbb{R}^{R\times 1}$ is the array response vector, $S_k \in \mathbb{C}^{1\times N}$ is the signal emitted by source k, and $W \in \mathbb{C}^{R\times N}$ is a noise matrix.

The form of m is dictated by array architecture and preprocessing. Notably, if we take X to be the B-format signal of a tetrahedral AVS, m simplifies to

$$m(\phi, \theta) = \begin{bmatrix} \sqrt{3} \\ \cos(\phi)\cos(\theta) \\ \sin(\phi)\cos(\theta) \\ \sin(\theta) \end{bmatrix}. \tag{2}$$

Please see Figure 1 for a visualization, and Appendix A for a derivation. The DoA estimation problem then consists of determining K, the ϕ's, and the θ's, and the goal of source separation is to find S.

In this work, we discuss DoA estimation in both the azimuth (horizontal) angle ϕ and elevation (vertical) angle θ but focus on calculating and evaluating the azimuth estimates. Additionally, we consider DoA estimation in both the time-domain and time-frequency-domain. For the former, we take X to be the $4 \times N$ matrix of B-format measurements, where N is the number of time samples. For the latter, we take X to be the $4 \times (N_t * N_f)$ matrix of vectorized spectrograms, where N_t is the number of spectrogram bins in time and N_f is the number of frequency bins.

A. WNC

WNC can be considered a regularized adaptation of minimum variance distortionless response (MVDR) beamforming.[32,33] MVDR aims to optimally minimize output variance under the constraint of unity gain in the look direction. Specifically, the MVDR weight vector for data X for a particular (azimuth, elevation) pair (ϕ, θ) is

https://doi.org/10.1121/10.0013505

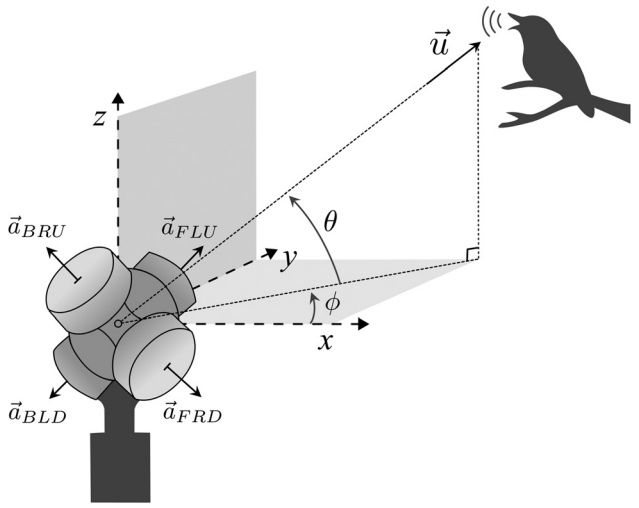

FIG. 1. A first-order ambisonic microphone, comprised of four co-located capsules, is one design for an AVS. The four cardioid capsules are oriented in a tetrahedron, with axes $\vec{a}^T_{FLU}$ (front-left-up), $\vec{a}^T_{FRD}$ (front-right-down), $\vec{a}^T_{BLD}$ (back-left-down), and $\vec{a}^T_{BRU}$ (back-right-up). The azimuth angle ϕ and elevation angle θ define the direction-of-arrival unit vector $\vec{u}$ relative to the microphone.

$$w(\phi,\theta) = \frac{C_{ni}^{-1} m(\phi,\theta)}{m(\phi,\theta)^H C_{ni}^{-1} m(\phi,\theta)}, \tag{3}$$

where C_{ni} is the covariance of noise and interference within the environment, and ^{H} indicates the conjugate transpose. C_{ni} can be approximated with the time-averaged sample covariance of a segment of the recording, X_{ni}, selected to lie outside of the target signal,

$$C_{ni} = \frac{1}{N} X_{ni} X_{ni}^H. \tag{4}$$

However, C_{ni} may be poorly conditioned, causing instability in the calculation of C_{ni}^{-1} and therefore of $w(\phi,\theta)$. To mitigate this issue, WNC replaces C_{ni} with the regularized expression $C_{ni} + \epsilon I$, where $\epsilon > 0$ is a manually chosen parameter balancing between data adaptivity (smaller ϵ) and stability (larger ϵ). Overall, the WNC weight vector takes the form

$$w(\phi,\theta) = \frac{(C_{ni} + \epsilon I)^{-1} m(\phi,\theta)}{m(\phi,\theta)^H (C_{ni} + \epsilon I)^{-1} m(\phi,\theta)}, \tag{5}$$

and the output is

$$Y_{wnc}(\phi,\theta) = ||w(\phi,\theta)^H X||_2. \tag{6}$$

We use $\epsilon = 0.001$ and manually select regions of the recordings outside of the target signals to provide X_{ni}.

B. MUSIC

The MUSIC algorithm, introduced in 1986, is a common eigenspace-based approach for DoA estimation.[34,35] MUSIC considers sound sources to occupy a K-dimensional subspace within the R-dimensional space of sensor measurements, with the complementary $(R-K)$-dimensional subspace occupied only by noise. More specifically, given a measurement matrix X, we construct the time-averaged sample covariance matrix $C = (1/N)XX^H$ and take an eigenvalue decomposition, $C = U\Lambda U^H$. Then we can consider the dominant K vectors of U to represent a basis for the signal DoA subspace (denoted U_{signal}) and the last $R-K$ to represent a basis for the noise DoA subspace (denoted U_{noise}). For each pair of azimuth and elevation angles, the squared projection of $\vec{m}(\phi,\theta)$ onto the noise DoA subspace would be $||U^H_{noise}\vec{m}(\phi,\theta)||^2$. We are looking for $\vec{m}(\phi,\theta)$, which is closest to the signal DoA subspace and therefore has a minimal projection onto the noise DoA subspace. Equivalently, we can maximize the metric,

$$Y_{music}(\phi,\theta) = \frac{1}{\vec{m}(\phi,\theta)^H U_{noise} U^H_{noise} \vec{m}(\phi,\theta)}. \tag{7}$$

The source directions can then be estimated by selecting the azimuth and elevation values at which Y_{music} attains local maxima. Note that the number of sources K must be chosen prior to computation and is limited to $K < R$ by the nature of the algorithm.

C. AIM

An alternative approach for DoA estimation is built on empirical calculation of the active intensity vector.[28,36] Let $p(f,\ t)$ and $\vec{v}(f,t)$ denote short-term Fourier transforms of the acoustic pressure and particle velocity vector at frequency f and time t. For a sufficiently distant source, a signal emitted by a point source will be received at the microphone as a plane wave, satisfying

$$\vec{v}(f,t) = -\frac{p(f,t)}{\rho c}\vec{u}, \tag{8}$$

where ρ is the ambient density, c is the speed of sound in the medium (air), and $\vec{u}$ is a direction-of-arrival unit vector. Note that the plane wave assumption holds when the source distance is large relative to the signal wavelength λ; since the lowest frequency we consider is 200 Hz ($\lambda = 1.7$ m) and bird calls are usually above 1 kHz ($\lambda = 0.34$ m), this is a reasonable expectation for avian monitoring. As described in Appendix A, the tetrahedral arrangement of cardioid microphones has a pickup pattern proportional to $p\vec{u}$ when converted to B-format and yields measurements of both p and $\vec{v}$. Now we can obtain the acoustic active intensity vector,

$$\vec{I} = \mathrm{Re}(\overline{p(f,t)}\vec{v}(f,t)). \tag{9}$$

Since $\vec{I}$ has direction $\vec{u}$, we can compute the corresponding azimuth ϕ and elevation θ, which define the direction-of-arrival,

$$\phi(f,t) = \tan^{-1}(I_y/I_x), \tag{10}$$

$$\theta(f,t) = \tan^{-1}\left(I_z \Big/ \sqrt{I_x^2 + I_y^2}\right). \tag{11}$$

The AIM entails calculating the azimuth and elevation for each time sample (in the time-domain) or spectrogram pixel (in the time-frequency-domain) and then constructing a histogram of both angle estimates. For robustness against low-intensity noise, all samples are weighted by the magnitude of intensity within the histogram calculation. The source directions can be obtained from the maxima of these distributions or from decomposing the histogram into components as described in Sec. III.

III. SOURCE SEPARATION

After obtaining a DoA distribution with AIM, we seek to reconstruct the signals corresponding to individual sources. We approach this problem by decomposing the DoA distribution of the acoustic mixture into DoA distributions of the constituent sources, through simple angular thresholding or through a Gaussian mixture model. We can then assign spectrogram pixels to different components to obtain both spectrogram- and time-domain source reconstructions. We compare the reconstruction fidelity obtained with both pixel assignment methods.

A. Angular thresholding

The locations of the sound sources can be inferred from the peaks of the DoA distributions. Consequently, a simple approach to reconstructing source signals could entail calculating a collection of peaks and then masking the original spectrogram to include only pixels with azimuth angles within a pre-specified range of these maxima. In practice, to represent a recording as a sum of K components, we calculate $K-1$ maxima and leave one component to incorporate all complementary azimuth angles, thereby capturing background noise.

B. Wrapped Gaussian mixture model

As an alternative to angular thresholding, we consider fitting the DoA estimates with a mixture model: a sum of simple probability distributions representing individual components. In particular, the Gaussian mixture model (GMM), a weighted sum of Gaussian probability densities, has been extensively studied and is a common choice for statistical modeling.[37] A GMM can be fit to data through the iterative expectation-maximization (EM) algorithm; a detailed description of EM-GMM is provided in Appendix B. To adapt this algorithm to angular data, which are characterized by circular wraparound of the azimuth angle, we modify the calculations of the distances and averages of data samples. First, we replace the linear distance $\phi_1 - \phi_2$ with the angular distance,

$$\mathrm{dist}(\phi_1, \phi_2) = (\phi_1 - \phi_2 + 180^\circ) \bmod (360^\circ) - 180^\circ. \tag{12}$$

Additionally, one step of the EM algorithm requires calculating means of data samples weighted by "member weights" $\{w_{ik}\}$, which represent the probability that the ith azimuth sample ϕ_i belongs to component k. Instead of the standard mean $\sum_{i=1}^{N} w_{ik}\phi_i$, we calculate this weighted azimuthal mean by converting a set of angles to unit vector form, summing them, and converting back to angular form,

$$\mu_k \leftarrow \tan^{-1}\left(\frac{\sum_{i=1}^{N} w_{ik}\sin(\phi_i)}{\sum_{i=1}^{N} w_{ik}\cos(\phi_i)}\right). \tag{13}$$

Last, to reduce runtime, we fit the GMM to a random subset of size $\tilde{N} = 10\,000$ rather than to all N azimuth samples.

With the GMM, source separation can be achieved by multiplying the pressure spectrogram pixel-wise by the member weights for the kth component to reconstruct the kth source signal. Since the member weights represent probabilities, we automatically enforce that the sum of all reconstructions is equal to the pressure spectrogram. Furthermore, if a time-domain representation of the signal is required, we can obtain a rough audio reconstruction by performing pixel assignment over the complex spectrogram prior to applying an inverse short-time Fourier transform.

IV. METHODS

Through analysis of simulated and experimental audio recordings, we aim to evaluate DoA estimation across variable signal characteristics (such as source frequency and recording duration), environmental factors (such as noise level), and algorithm parameters (such as angular resolution). Examination of simulated narrowband signals allows us to isolate estimation error associated with the AVS model, while outdoor experiments enable assessment of expected accuracy under realistic conditions for passive acoustic monitoring.

For our simulation, we consider a tetrahedral cardioid array with an axis-capsule distance of 1.5 cm, accounting for associated time delays. We use single-frequency (monochromatic) sound sources of duration 0.5 s and add white noise to each A-format capsule measurement for an overall SNR of 20 dB, as described in Sec. V D.

For our experiments, we used the AMBEO VR Mic (Sennheiser Electronic GMBH & Co. KG, Wedemark, Germany) and the MixPre-10 recorder (Sound Devices LLC, Reedsburg, WI). The microphone has a flat frequency response from 20 Hz to 20 kHz, with a sensitivity rating of 31 mV/Pa (–30 dBV); the recorder has a 24-bit depth with a sampling rate of 48 kHz and preamplifier gain set to 50 dB. Signals were played through Aomais Real Sound Portable Bluetooth speakers, with a stereo pair used for the two-source measurements. To simplify measurements of relative distances and angles between the microphone and speakers, we conducted the experiments outdoors on an artificial turf sports field, with the microphone positioned at the center mark and mounted on a tripod at a height of about 1.5 m. For distance measurements, we took advantage of accurate

https://doi.org/10.1121/10.0013505

markings for a 10-yard (9.1-m) center circle and lines. To determine 60° angles, we geometrically inscribed a hexagon within the center circle. For smaller angle increments, we calculated corresponding chord lengths and measured them manually along the circle. Since we estimate the error in length measurement and tripod positioning to be within 10 cm, the associated angular error would be within 0.1/9.1 radians, or about 1°. As our source signals, we used cropped, filtered, and repeated versions of four recordings from Xeno-Canto: blue jay (*Cyanocitta cristata*, XC571792), Carolina wren (*Thryothorus ludovicianus*, XC556630), American robin (*Turdus migratorius*, XC464766), and song sparrow (*Melospiza melodia*, XC480068) songs.

We pre-process recordings with a 5th-order Butterworth high pass filter at 200 Hz to remove low-frequency noise and calculate spectrograms with a Hann window of size 1024 and a shift of 50%, trimmed to a maximum frequency of 10 kHz. Note that, for a 1-s recording, the number of samples N is equal to 48 000 (the sampling rate) for the time-domain methods and 20 330 ($N_x = 95$, $N_f = 214$) for the time-frequency-domain methods.

V. RESULTS

A. Examples of DoA estimation

First, we demonstrate and compare the output of each algorithm across a set of sample recordings with varied characteristics. Figure 2(A) shows four recordings: a simulation with one 3 kHz source (top left), a simulation with two non-overlapping sources at 2 and 4 kHz (top right), an experimental recording of a speaker playing a blue jay call (bottom left), and an experimental recording of two speakers playing overlapping American robin and song sparrow calls (bottom right). Both single-source recordings are emitted from an azimuth of −60°, while both two-source recordings have a separation angle of 45°. For each recording, Fig. 2(B) shows the associated azimuth values calculated by AIM for each time-frequency bin (pixel). Despite the lack of intensity information, the calls are still recognizable and can be visibly separated from each other and from the background. Finally, Fig. 2(C) shows the outcome of DoA estimation by plotting the output of all algorithms in both the time-domain and time-frequency-domain: $Y_{wnc}(\phi, 0)$, as defined in Eq. (6), for WNC; $Y_{music}(\phi, 0)$, as defined in Eq. (7), for MUSIC; and the weighted histogram of azimuth values for AIM. We indicate the time-domain variant of each algorithm with "(time)" and the time-frequency-domain variant with "(t-f)". Additionally, for ease of visual comparison, we scale all outputs to a maximum value of 1. For a single simulated source, all algorithms identify the source direction at −60° with high accuracy. However, for the case of a real recording with a single source, WNC exhibits an inaccurate response, suggesting a lack of robustness to directional noise in the environment. The presence of two sources reveals greater differences between algorithms. Both in simulation and in practice, only AIM produces a bimodal distribution; MUSIC and WNC result in metrics that peak either at one of the sources or in between. Overall, in these examples, AIM outperforms WNC and MUSIC through higher robustness to noise in real-world data and more flexible representation of multiple sources.

B. Effect of angular resolution

The angular resolution of the DoA estimate is set by azimuth discretization for WNC and MUSIC and by the number of histogram bins for AIM. For the first two methods, we expect accuracy to increase with resolution, with

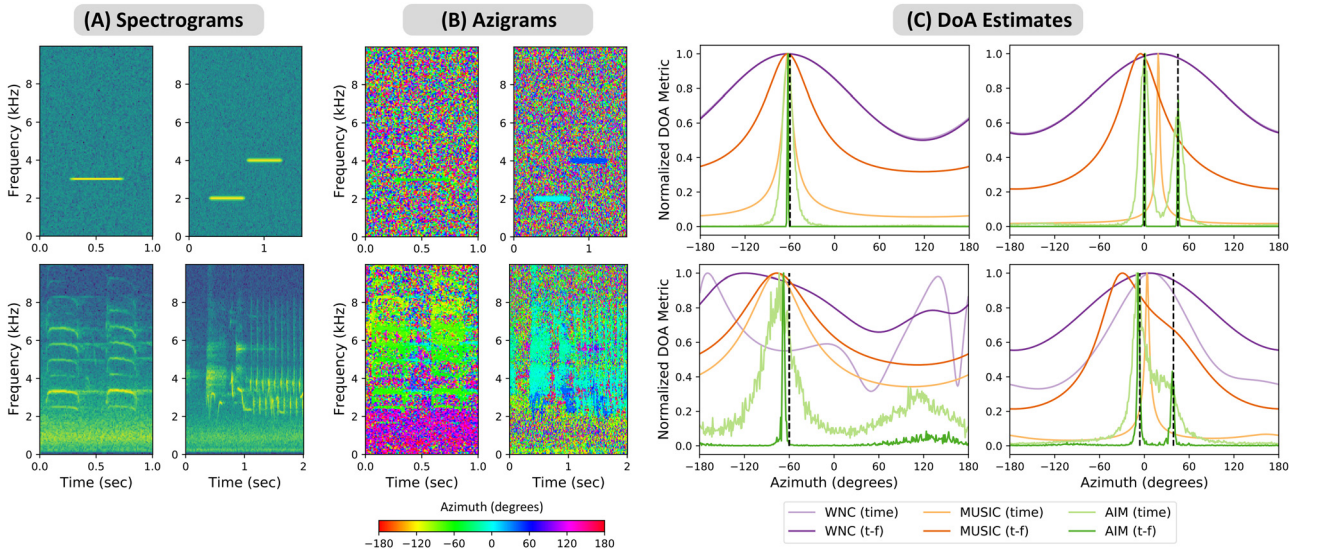

FIG. 2. (Color online) (A) Spectrograms of four sample recordings: a simulation of one monochromatic source at −60° (top left), a simulation of two monochromatic sources separated by 45° (top right), experimental data for one speaker playing a blue jay call at −60° (bottom left), and experimental data for two speakers playing overlapping American robin and song sparrow calls separated by 45° (bottom right). (B) Azigrams corresponding to the four sample signals; pixel color indicates azimuth in degrees as measured by AIM in the time-frequency domain. (C) DoA estimates returned by all algorithms for the four sample signals. Dashed lines indicate speaker angles for the single-source examples and 45° separation for the two-source examples.

the only disadvantage being a proportional increase in computation. For AIM, we might expect a trade-off between angular resolution and robustness. With a numerical evaluation, we find that AIM (t-f) achieves best performance in the 100–1000 bin range per a 1-s recording (200–20 pixels/bin). As anticipated, a small bin count (below 30 bins) yields a consistent estimate across trials but higher error due to the coarse-grained resolution. As bin count increases, the mean angular error remains stable, suggesting robustness to sample sparsity. Finally, past about 1000 bins, error variance increases. Given the low sensitivity of the algorithm to this parameter, we simply choose to use 360 bins for an angular resolution of 1° and apply the same discretization for WNC and MUSIC.

C. Effect of recording duration

Since natural soundscapes are dynamic, and the directionality of vocalizations may change at short time scales, we consider how recording duration affects DoA estimation. To evaluate this parameter, we analyze a recording of a Carolina wren call played back at 0° by a speaker positioned 9.1 m (10 yards) from the microphone, with the dominant source of background noise emitted by highway traffic located at about 180°. The angular error resulting from DoA estimation over the call cropped to varying durations is shown in Fig. 3(A). At the shortest limit, all DoA estimates become dominated by noise. AIM (t-f) is the most robust to short durations, yielding an accurate estimate even with a 0.05-s analysis window, while the other methods show poor performance below a duration of 0.5 s. In practice, spectrograms of complete calls are probably at least 0.5 s long, but short recording windows are relevant for temporal analysis of DoA, as shown in Fig. 8 and discussed in Sec. VI.

D. Effect of noise

Vocalizations in natural soundscapes will occur alongside ambient environmental and anthropogenic noise. In this section, we analyze the robustness of WNC, MUSIC, and AIM against SNR by keeping source level fixed and varying the distance of the speaker from the microphone.

We compute SNR as $20\log_{10}(A_s/A_n)$, where A_s and A_n are the root mean square time-domain amplitudes of segments of the B-format pressure channel corresponding to the signal and noise, respectively. The results are thus comparable to SNR for an omnidirectional microphone. For the experimental recordings, we do not have true signal samples that are isolated from noise and interference, so we approximate the signal sound level by selecting a narrow time window around a bird call and filtering below 2 kHz. To represent background noise, we manually select a nearby portion of the recording that excludes the calls and interfering sound sources. Note that we do not band-limit the noise sample to the range of a particular bird call, since all frequencies affect DoA calculations.

Figure 3 shows the angular error of DoA estimation for a blue jay call played back from a speaker positioned at 0° at distances ranging from 4.6 to 37 m (5–40 yards) at increments of 4.6 m (5 yards). Once again, the dominant noise source was highway traffic located at about 180° relative to the microphone. We find that all algorithms yield consistent accuracy for a SNR above about 2 dB. Below this threshold, the directionality of background noise overshadows the signal for WNC and MUSIC. AIM (t-f) appears to be most robust to noise, even for an SNR close to 0 dB. Since pixels are weighted by acoustic intensity, a small proportion of high-amplitude signal pixels with consistent directionality can outweigh a large proportion of noise pixels with scattered directionality while maintaining similar root mean square amplitudes.

Note that frequency-domain filtering will impact DoA distributions. For example, Fig. 4 shows an example of the spectrograms and DoA distributions for the same recording, pre-processed with a 200 Hz and 2 kHz high pass filter. With 200 Hz filtering, the noise dominates all DOA estimates except AIM (t-f), which exhibits a broad distribution for the noise and a narrow distribution for the target source. With 2 kHz filtering, all DoA estimates indicate the target source. Consequently, limiting analysis to the target frequencies can substantially improve DoA estimates if the noise occupies disjoint frequency bands.

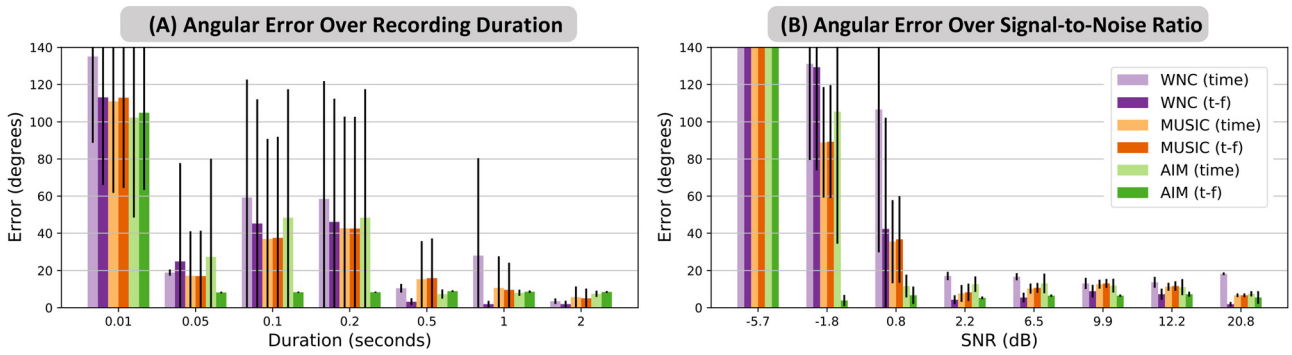

FIG. 3. (Color online) (A) Mean ± standard deviation of the angular error of DoA estimates as a function of the recording duration. For this experiment, a speaker played back a Carolina wren call from a distance of 9.1 m (10 yards), repeated for eight trials. (B) Mean ± standard deviation of the angular error of DoA estimates as a function of the SNR. For this experiment, a speaker played back a blue jay call from distances of 4.6–37 m (5–40 yards) in increments of 4.6 m (5 yards), repeated for six trials.

https://doi.org/10.1121/10.0013505

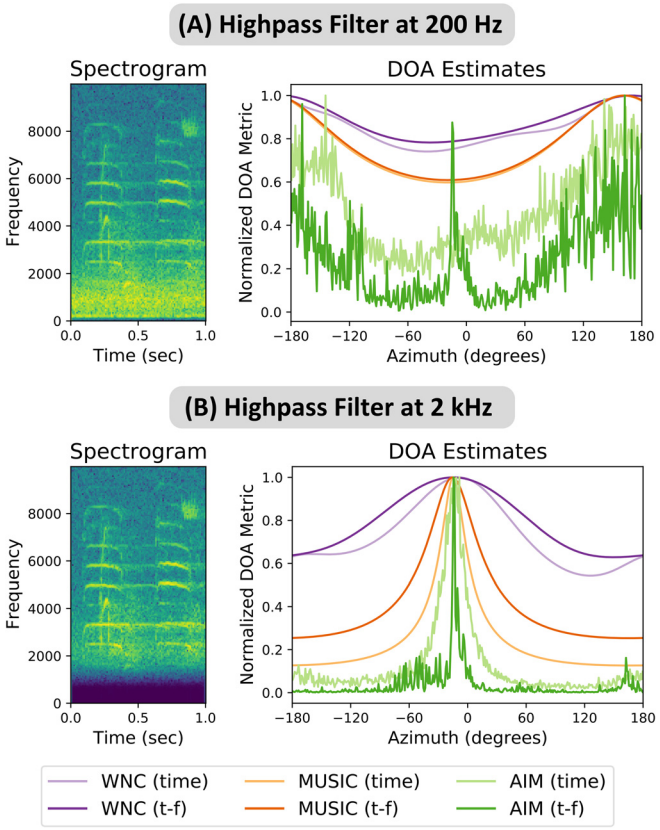

FIG. 4. (Color online) (A) A recording of a blue jay call is played back from a speaker located at about 0° at a distance of 37 m (40 yards) from the microphone and is high pass-filtered at 200 Hz prior to DoA estimation. The DoA estimates are largely dominated by low-frequency noise from a highway located at about 180°. (B) The same recording is high pass-filtered at 2 kHz prior to DoA estimation. The resulting DoA estimates are now all centered on the source direction.

E. Effect of signal frequency

In addition to DoA estimation error associated with environmental noise and inaccuracy in measurement of speaker positions, the underlying model of the AVS is itself not exact. Specifically, the array response vector given in Eq. (2) is an approximation of the true array response and assumes that the four capsules of the tetrahedral array are located at a point. This assumption is reasonable for low frequencies but worsens as the signal wavelength becomes comparable to the finite spacing between capsules. In this work, the microphone has a capsule-axis distance of about 1.5 cm. To evaluate the magnitude of error contributed by violation of this co-location assumption, we simulate the recording associated with monochromatic signals and perform DoA estimation with AIM (t-f) across different azimuthal directions. We set the SNR to 60 dB to remove error associated with noise.

The result of this experiment is presented in Fig. 5. As expected, we find that DoA estimation error increases with frequency. This error is deterministically defined by source DoA; moreover, at particular source azimuth angles (specifically, at multiples of 45° relative to the microphone axis), the model error drops to 0. Overall, for this microphone architecture,

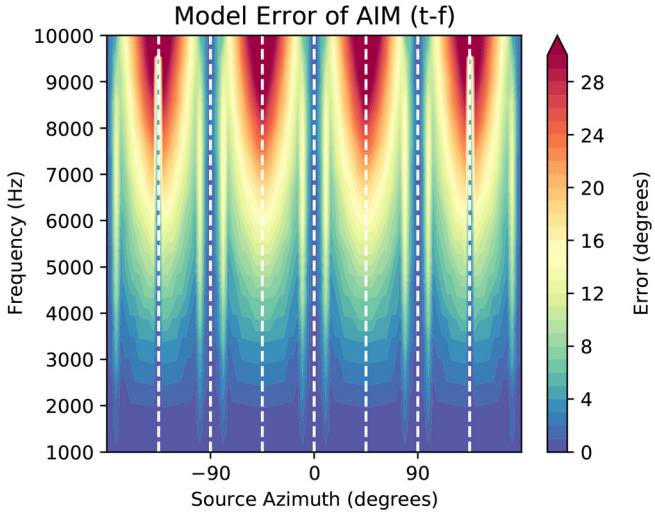

FIG. 5. (Color online) To evaluate DoA estimation error associated with the co-location assumption, we simulate monochromatic signals across source frequency and azimuth and calculate the angular error of AIM (t-f). White dashed lines indicate multiples of 45°. We observe substantial error for wavelengths comparable to the 1.5 cm capsule spacing.

we suggest focusing on sound sources below about 5 kHz to limit maximum angular error to below 10°. In practice, we find that the challenge posed by model error to DoA estimation and source separation is tempered by the rest of the algorithm: high-frequency narrowband calls should still yield a coherent (albeit inaccurate) source direction and should therefore be amenable to directionality-based source separation, and in broadband calls, higher frequencies will experience greater attenuation and therefore receive less weight than lower-frequency components.

F. Overall accuracy and resolution

Finally, after considering the effects of different parameters on DoA estimation error, we summarize comparative algorithm performance in Table I. We focus on two main metrics: angular accuracy for a single source and the minimum separation angle necessary to distinguish two sources.

When considering one source in simulation at frequencies randomly sampled between 0 and 5 kHz, we find that all algorithms have similar accuracy at around 2.5° mean error. Note that we are controlling for experimental inaccuracies, and this error stems from the co-location assumption. Next, we consider the accuracy for one source with speaker playback experiments, which reveals greater differences between algorithms. In particular, both the time- and time-frequency-domain variations of WNC show greater error than MUSIC and AIM. This reduction in performance is likely caused by heterogeneity and directionality of background noise: while the simulation contained only omni-directional white noise, the real-world data contain variable, directional broadband noise together with interfering sound sources that we could not control for (such as real birds vocalizing in the background).

Next, we consider an environment with multiple sound sources. Specifically, we evaluate the minimum angular

TABLE I. We evaluated two key metrics: the DoA estimation accuracy with a single source and the minimum angular separation necessary to distinguish two sources. We evaluated error in simulation across 20 trials of sources at frequencies randomly sampled between 0 and 5 kHz, with non-directional background noise for an SNR of 20 dB. To evaluate error in real-world experiments, we played back a 1-s blue jay call from a speaker located at a distance of 9.1 m (10 yards) from the microphone, at repeated increments of 60°, repeated for eight trials. Next, we evaluated the minimum angular separation across algorithms in simulation by calculating DoA estimates for two 3 kHz sources at randomly sampled azimuth angles. For corresponding real-world experiments, we position two speakers at a radial distance of 9.1 m (10 yards) from the microphone and separation angles of 0°, 6°, 11°, 22°, and 45° and play back American robin and song sparrow recordings overlapping in time and frequency.

Algorithm	Error (*simulation*)	Error (*experiment*)	Angular separation (*simulation*)	Angular separation (*experiment*)
WNC (time)	2.4° ± 2.6°	28.3° ± 27.0°	150°	>45°
WNC (t-f)	2.4° ± 2.6°	9.9° ± 7.8°	150°	>45°
MUSIC (time)	2.3° ± 2.6°	5.1° ± 3.4°	90°	>45°
MUSIC (t-f)	2.4° ± 2.6°	5.1° ± 3.3°	90°	>45°
AIM (time)	2.4° ± 2.7°	5.6° ± 3.9°	7°	6°–11°
AIM (t-f)	2.5° ± 2.7°	4.7° ± 2.4°	7°	6°–11°

separation necessary between two sources for the DoA estimation algorithms to distinguish them with two peaks. In simulation, we consider two consecutive non-overlapping 0.5-s 3 kHz sources. We find that both variants of WNC begin to distinguish the sources at a minimum separation angle of 150°, and both MUSIC variants require a separation of at least 90°. In contrast, AIM would yield two peaks with approximately 3° width, resulting in a minimum separation angle of about 7°. For the real-world experiments, both WNC and MUSIC maintain unimodal distributions across the separation angles tested (up to 45°). On the other hand, AIM begins to successfully separate two sources when the separation angle reaches 11°, similarly to the simulated results. Overall, AIM significantly outperforms WNC and MUSIC in the ability to represent environments with multiple sound sources.

On the whole, we find that while the algorithms give similar performance for a single source in high-SNR conditions, AIM (t-f) shows best performance and highest robustness for low SNR, short analysis windows, or multiple sources. Additionally, AIM has the advantage of being highly automated: it can flexibly represent multiple sources with variable DoA distributions and gives stable output across a wide range of the angular resolution hyperparameter. On the other hand, the number of sources that can be represented by MUSIC and WNC is much more limited, and WNC requires manual selection of a region of the recording not containing the signals of interest, which may be undesirable for automated monitoring systems.

G. Source separation

Finally, we complete the source separation pipeline and evaluate the fidelity of spectrogram reconstruction. First, we calculate DoA estimates with AIM (t-f), selected due to its high performance throughout the evaluation in Sec. V F. Next, we apply angular thresholding and fit a wrapped Gaussian mixture model to obtain the azimuth distributions associated with individual sources. Based on the observed width of AIM distributions, we use an angular range of 10° to both sides of an azimuth maxima for angular thresholding. Additionally, to avoid multiple maxima corresponding to the same source, we apply a buffer of 20° between maxima. Finally, for each source, we calculate spectrogram and time-domain reconstructions and compare cross-correlations between the true source audio signals and the reconstructions obtained with both methods.

First, we demonstrate the source separation pipeline for recordings of American robin and song sparrow birdsong played back from stereo speakers. We set the number of components K to 3, to represent the two actual sources as well as remaining background noise. Figure 6 shows the original source signal, measured signal, azigram, DoA distribution, and reconstructions. In this example, the high SNR and non-directional background noise result in a DoA distribution composed of two clearly defined peaks. Consequently, the angles associated with each peak can be successfully recovered by either the GMM, which fits two narrow distributions to the peaks, or with a simple thresholding method. The resulting spectrogram reconstructions closely match the structure of the original calls. To quantitatively evaluate reconstruction fidelity, we cross-correlate the time-domain reconstructions with the true source signals and find that angular thresholding yields the same output as a GMM for the robin call and a 26% improvement for the sparrow call in this example.

Next, we repeat the source separation analysis for a recording of a 1-s blue jay call played from a speaker positioned at 0° at a distance of 37 m (40 yards) from the microphone in the presence of significant directional noise. We set the number of components K to 4, to represent multiple sources along with remaining background noise. The original source signal, measured signal, azigram, DoA distribution, and reconstructions are displayed in Fig. 7. Here, we see a greater difference in the components resulting from angular thresholding and from the GMM. Not only is the DoA distribution now dominated by noise and external sound sources (such as a crow call at about −100°), but the noise and signal have variable widths, and the DoA maxima no longer correspond to unique sound sources. As a result, angular thresholding only partially captures the sound sources present in the acoustic environment. On the other hand, the GMM adaptively fits components to the widths present in the data and is able to separate the blue jay call (component 2) from the low-frequency highway noise (component 0) and even from a crow call (component 1). In this example, the GMM results in a 49% higher cross correlation score for the source signal than angular thresholding.

Overall, while simple angular thresholding can be sufficient for signals with high SNR in a quiet environment, the wrapped GMM enables flexible representation of mixtures of angular distributions with variable characteristics. A possible limitation of this approach is that DoA distributions may be

https://doi.org/10.1121/10.0013505

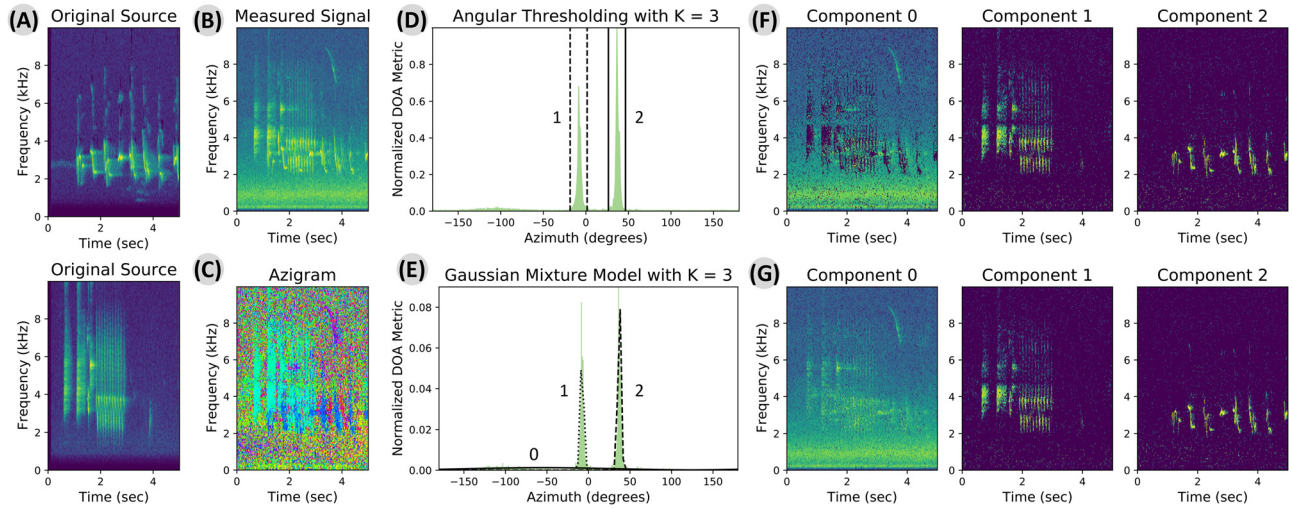

FIG. 6. (Color online) To demonstrate source separation, we consider 5-s recordings of American robin and song sparrow birdsong played back from two speakers separated by 45° at a distance of 9.1 m (10 yards) from the microphone (A). The measured recording (B) is a mixture of these calls. Calculating azimuth estimates for each pixel with AIM (t-f) yields an azigram (C), in which both sources are visibly distinguishable against a backdrop of omnidirectional noise. Next, we compare a baseline angular thresholding approach (D and F) with a wrapped Gaussian mixture model (E and G). Note that the distributions and *y* axes in (D) and (E) differ slightly, as we use a probability density over a subsampled set of azimuth points for the GMM. In this example, we see very similar performance between the two source separation methods.

non-Gaussian and may result in a poor fit for some sources. The GMM is also the most computationally intensive part of the source separation pipeline; runtime depends on the number of subsampled azimuth points, the chosen number of iterations (or convergence threshold), and the number of mixture components. Further development of this approach could incorporate automatic selection of the number of sources. This is commonly achieved by iterating through all relevant values of K, calculating the Akaike or Bayesian information criterion, and choosing the K at which the improvement in the criterion value is below a set threshold.[38,39]

VI. DISCUSSION

All in all, AVSs are a simple, robust, and compact tool for soundscape analysis that can support efficient direction-of-arrival estimation and source separation. While DoA estimation can be performed with conventional beamforming methods like WNC or eigenspace methods like MUSIC, we found best performance with empirical calculation of active intensity. In playback experiments with speakers, AIM (t-f) resulted in angular error of about 5° and could distinguish sources at a separation angle of 11°. Moreover, this performance persisted for recording lengths as short as 0.05 s and for near-zero SNR.

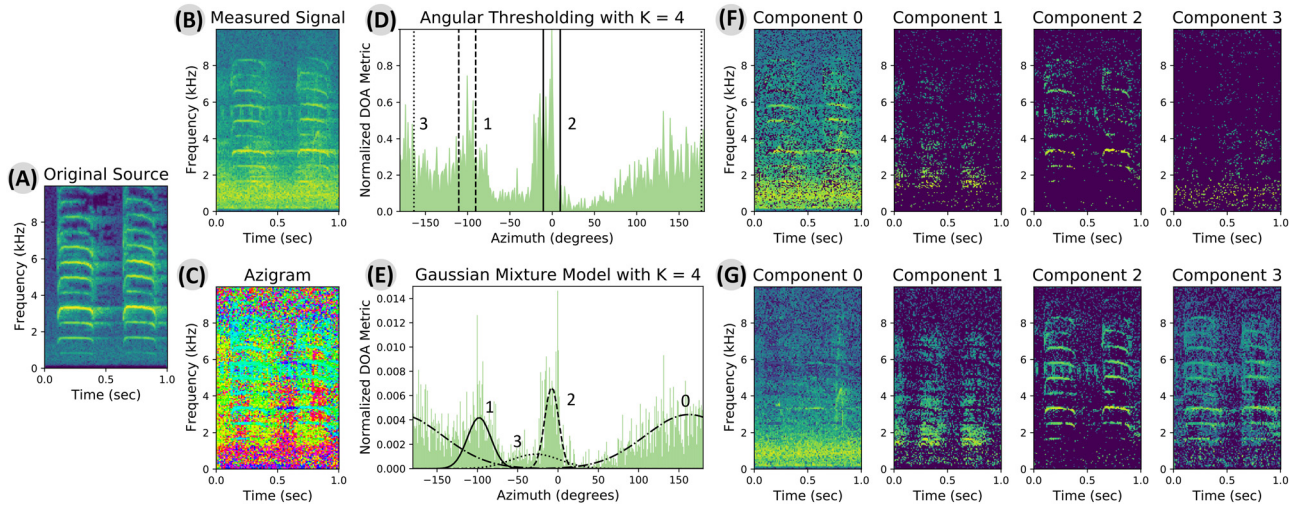

FIG. 7. (Color online) We repeat the source separation pipeline, now with a 1-s blue jay call (A) being played by a speaker located at 37 m. (40 yards) at 0°. The measured recording (B) has low SNR, with significant noise from a highway located at about 180°, a crow call at about −100°, and other interfering sound sources. The azigram (C) displays directional characteristics of both the target signal and noise. Again, we compare a baseline angular thresholding approach [(D) and (F)] with a wrapped Gaussian mixture model in (E) and (G). We see much clearer output from the GMM than from angular thresholding, as the GMM can adaptively fit distributions to represent variable spread in directionality across different sources.

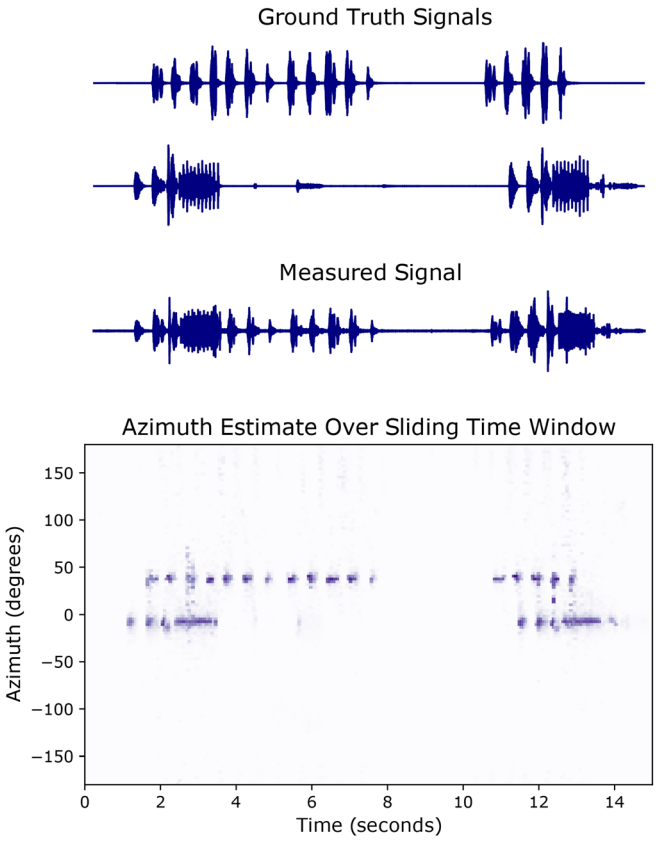

FIG. 8. (Color online) Direction-of-arrival analysis can be enhanced by using a sliding window to visualize both the times at which sources are active and the azimuth angles from which they arrive. In this example, two 15-s recordings of robin and sparrow song were played back from azimuth angles of 45° and 0°, respectively. The time-series representations of the original signals, along with the measured mixture, are shown in the top section of the figure. By applying DoA estimation with AIM (t-f) over 0.1-s windows of the measured signal, stacking the estimates as columns of a matrix, and visualizing with a log-scale pseudocolor plot, we obtain a two-dimensional (2D) representation of a soundscape. We can see that both the angles and times at which the two sources are active closely correspond to the ground-truth.

Unlike a microphone array, an AVS requires minimal calibration, mitigating the need for clock synchronization. Additionally, the only algorithmic parameters are the angular resolution, which has a minor effect on accuracy over a wide range of values, and the number of GMM components or sources, which could be chosen automatically through an information criterion. Furthermore, the proposed pipeline can provide DoA estimates in both the azimuth and elevation angles. Analysis in both dimensions would improve the ability to pinpoint sources and likely increase the number of sounds that can be distinguished within an environment. However, preliminary experiments showed poor results for elevation estimation, likely due to acoustic reflection and scattering from the ground. Further development on these methods could incorporate modeling of acoustic propagation to account for these distortions.[40]

The calculation of azimuth angles advances our understanding of a soundscape by providing an additional dimension for acoustic analysis. For instance, we can apply DoA estimation over a sliding temporal window to visually display both the angles and times at which sources are actively vocalizing. Figure 8 shows this technique for the two-source experiment with 45° angular separation between speakers, using a 0.1-s time window and 50% overlap between consecutive windows. Each column of the figure represents the AIM (t-f) angular histogram for a particular window; darker purple corresponds to higher histogram values. We can clearly see that there are two angles at which sources are active—at about 0° and 45°—along with fine granularity in the times at which they are vocalizing. Comparing against the ground-truth time series, we see that the activity of the 45° signal matches the first source signal, and the activity of the 0° signal matches the second source signal.

The ability to perform DoA estimation and source separation within acoustic mixtures opens a number of opportunities within passive acoustic monitoring. First, this approach can be integrated with a machine-learning-based classifier to disentangle individual vocalizations and thereby improve accuracy of species-level identification and count estimates. Furthermore, localization and tracking algorithms with AVS arrays established in marine bioacoustics could be adapted and further developed for terrestrial environments.[30,31] Finally, behavioral studies of vocal behavior could also benefit from identifying source directions and removing undesired sound sources.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF)-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (Award No. 1764269) and the Harvard Quantitative Biology Initiative. We thank Dr. Elena Tolkova for thoughtful suggestions and support in conducting experiments as well as Dr. Aaron Thode for his time and thorough feedback on our manuscript. We also thank Dr. L. Mahadevan and Dr. Steven Gortler for helpful discussion. Finally, we are very grateful to two anonymous reviewers for their valuable comments.

APPENDIX A: AVS PICKUP PATTERN

In three dimensions, the pickup pattern of a cardioid microphone is given by $g(\theta) = \frac{1}{2} + \frac{1}{2}\cos(\theta)$, where θ is the angle between the microphone's axis $\vec{a}$ and the sound direction of propagation $\vec{u}$.[41] Equivalently,

$$g(\vec{u}) = \frac{1}{2} + \frac{1}{2}\vec{a}^T\vec{u}.$$

An ambisonic microphone is composed of four cardioid microphones in a tetrahedral arrangement, as shown in Fig. 1. The unit vectors indicating the axes of each capsule are given as

$$\begin{bmatrix} \vec{a}_{FLU}^T \\ \vec{a}_{FRD}^T \\ \vec{a}_{BLD}^T \\ \vec{a}_{BRU}^T \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}. \quad \text{(A1)}$$

https://doi.org/10.1121/10.0013505

Therefore, the A-format response would be

$$g_A(\vec{u}) = \frac{1}{2}\begin{bmatrix}1\\1\\1\\1\end{bmatrix} + \frac{1}{2}\cdot\frac{1}{\sqrt{3}}\begin{bmatrix}1 & 1 & 1\\1 & -1 & -1\\-1 & 1 & -1\\-1 & -1 & 1\end{bmatrix}\vec{u}.$$

The different components of $g_A(\vec{u})$ contain mixed combinations of the elements of $\vec{u}$. However, this pickup pattern can be greatly simplified by converting to B-format through a linear transformation,

$$\begin{aligned} g_B(\vec{u}) &= \begin{bmatrix}1 & 1 & 1 & 1\\1 & 1 & -1 & -1\\1 & -1 & 1 & -1\\1 & -1 & -1 & 1\end{bmatrix} g_A(\vec{u})\\ &= \frac{1}{2}\begin{bmatrix}4\\0\\0\\0\end{bmatrix} + \frac{1}{2\sqrt{3}}\begin{bmatrix}0 & 0 & 0\\4 & 0 & 0\\0 & 4 & 0\\0 & 0 & 4\end{bmatrix}\vec{u}\\ &= \frac{2}{\sqrt{3}}\begin{bmatrix}\sqrt{3}\\ \vec{u}\end{bmatrix}. \end{aligned}$$

Note that the relative gain between the first and other rows depends on architecture and data format convention.

Under the assumption that the distance to a source is much greater than its wavelength, its signal will be received at the microphone as a plane wave, for which

$$\vec{v}(t) = -\frac{p(t)}{\rho c}\vec{u},$$

where $\vec{v}(t)$ is the acoustic particle velocity, $p(t)$ is the pressure, ρ is the density of the medium of propagation, and c is the speed of sound. Therefore, if the AVS yields B-format readings $g_B(\vec{u})s(t)$, where $s(t)$ is the incident source pressure, then the first B-format entry will be proportional to acoustic pressure, and the next three components will be proportional to the acoustic particle velocity. Note that this derivation depends on the approximation that the microphone capsules are coincident. For high frequencies, for which the wavelength is comparable to the capsule spacing, this may no longer hold.[42] Fortunately, performance at high frequencies can be improved through non-coincidence correction filters.[43]

APPENDIX B: GAUSSIAN MIXTURE MODEL

We fit a GMM to decompose an azimuth probability distribution into constituent parts representing different sound sources. Given N azimuth measurements $\Phi = \{\phi_1, \phi_2, \ldots, \phi_N\}$, a GMM with K components has the form

$$P(\phi|\alpha,\mu,\sigma) = \Sigma_{k=1}^{K}\alpha_k P_k(\phi|\mu_k,\sigma_k), \tag{B1}$$

where P_k is a Gaussian distribution with mean μ_k and standard deviation σ_k, and α_k is a mixture weight representing the contribution of the kth component to the overall distribution. Note that by definition, α_k satisfies $\Sigma_{k=1}^{K}\alpha_k = 1$. Each mixture component has the form

$$P_k(\phi|\mu_k,\sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma^2}\mathrm{dist}(\phi,\mu_k)^2\right). \tag{B2}$$

For a standard GMM, the distance function $\mathrm{dist}(\phi_1,\phi_2)$ is simply equal to $\phi_1 - \phi_2$. To adapt this method for azimuth samples, we instead use circular distance as given by Eq. (12).

A GMM can be fit to data through the iterative EM algorithm. The EM algorithm determines the optimal μ_k, σ_k, and α_k by alternating between the *expectation step*, which assigns probabilities to the data points, and the *maximization step*, which updates the means and variances of the mixture components. In the expectation step, we update the member weights w_{ik}, which represent the probability that a sample ϕ_i belongs to component k,

$$w_{ik} \leftarrow \frac{\alpha_k P_k(\phi_i|\mu_k,\sigma_k)}{\Sigma_{j=1}^{K}\alpha_j P_j(\phi_i|\mu_j,\sigma_j)}. \tag{B3}$$

Note that $\Sigma_{k=1}^{K} w_{ik} = 1$, as every point must belong to one of the components. We use " $\leftarrow$" to indicate variable assignment rather than algebraic equality.

In the maximization step, we use the data and the member weights to update distribution parameters. We define N_k to be the effective number of data points assigned to component k,

$$N_k \leftarrow \Sigma_{i=1}^{N} w_{ik}.$$

Then α_k will be the effective proportion of data points assigned to component k,

$$\alpha_k \leftarrow \frac{N_k}{N}.$$

Next, we update the distribution parameters by calculating means and variances, but with all points weighted by their associated member weight. For a vanilla GMM, this would be

$$\mu_k \leftarrow \frac{1}{N_k}\Sigma_{i=1}^{N} w_{ik}\phi_i,$$

$$\sigma_k^2 \leftarrow \frac{1}{N_k}\Sigma_{i=1}^{N} w_{ik}\mathrm{dist}(\phi_i,\mu_k)^2.$$

Once again, for our application, we modify the calculation of the mean, as given by Eq. (13). The algorithm iterates between the expectation and maximization steps until convergence. Specifically, we terminate the algorithm when the difference in consecutive mixture means is less than 0.005, or at a maximum of 100 iterations.

[1]J. M. Rowcliffe and C. Carbone, "Surveys using camera traps: Are we looking to a brighter future?" Anim. Conserv. **11**(3), 185–186 (2008).
[2]R. Steenweg, M. Hebblewhite, R. Kays, J. Ahumada, J. T. Fisher, C. Burton, S. E. Townsend, C. Carbone, J. M. Rowcliffe, J. Whittington, J. Brodie, J. A. Royle, A. Switalski, A. P. Clevenger, N. Heim, and L. N. Rich, "Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors," Front. Ecol. Environ. **15**(1), 26–34 (2017).
[3]A. F. O'Connell, J. D. Nichols, and K. U. Karanth, *Camera Traps in Animal Ecology: Methods and Analyses* (Springer, New York, 2010).
[4]J. Jiménez López and M. Mulero-Pázmány, "Drones for conservation in protected areas: Present and future," Drones **3**(1), 10 (2019).
[5]J. C. van Gemert, C. R. Verschoor, P. Mettes, K. Epema, L. P. Koh, and S. Wich, "Nature conservation drones for automatic localization and counting of animals," in *Computer Vision—ECCV 2014 Workshop*, edited by L. Agapito, M. Bronstein, and C. Rother (Springer, Cham, Switzerland, 2014), pp. 255–270.
[6]L. P. Koh and S. A. Wich, "Dawn of drone ecology: Low-cost autonomous aerial vehicles for conservation," Trop. Conserv. Sci. **5**(2), 121–132 (2012).
[7]N. Pettorelli, W. F. Laurance, T. G. O'Brien, M. Wegmann, H. Nagendra, and W. Turner, "Satellite remote sensing for applied ecologists: Opportunities and challenges," J. Appl. Ecol. **51**(4), 839–848 (2014).
[8]F. Trolliet, C. Vermeulen, M.-C. Huynen, and A. Hambuckers, "Use of camera traps for wildlife studies: A review," Biotechnol. Agron. Soc. Environ. **18**(3), 446–454 (2014).
[9]A. Digby, M. Towsey, B. D. Bell, and P. D. Teal, "A practical comparison of manual and autonomous methods for acoustic monitoring," Methods Ecol. Evol. **4**(7), 675–683 (2013).
[10]P. Laiolo, "The emerging significance of bioacoustics in animal species conservation," Biol. Conserv. **143**(7), 1635–1645 (2010).
[11]E. Browning, R. Gibb, P. Glover-Kapfer, and K. E. Jones, *Passive Acoustic Monitoring in Ecology and Conservation* (WWF-UK, Woking, UK, 2017).
[12]L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro, Jr., and D. Llusia, "Terrestrial passive acoustic monitoring: Review and perspectives," BioScience **69**(1), 15–25 (2019).
[13]D. T. Blumstein, D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, A. H. Krakauer, C. Clark, K. A. Cortopassi, S. F. Hanser, B. McCowan, A. M. Ali, and A. N. G. Kirschel, "Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus," J. Appl. Ecol. **48**(3), 758–767 (2011).
[14]S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," Ecol. Inf. **61**, 101236 (2021).
[15]T.-H. Lin and Y. Tsao, "Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval," Remote Sens. Ecol. Conserv. **6**(3), 236–247 (2020).
[16]A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," Neural Netw. **13**(4), 411–430 (2000).
[17]D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of Advances in Neural Information Processing Systems 13 (NIPS 2000)*, Denver, CO (November 28–30, 2000), pp. 556–562.
[18]M. R. Izadi, R. Stevenson, and L. N. Kloepper, "Separation of overlapping sources in bioacoustic mixtures," J. Acoust. Soc. Am. **147**(3), 1688–1696 (2020).
[19]T. Denton, S. Wisdom, and J. R. Hershey, "Improving bird classification with unsupervised sound separation," arXiv:2110.03209 (2021).
[20]I. R. Urazghildiiev and C. W. Clark, "Comparative analysis of localization algorithms with application to passive acoustic monitoring," J. Acoust. Soc. Am. **134**(6), 4418–4426 (2013).
[21]T. A. Rhinehart, L. M. Chronister, T. Devlin, and J. Kitzes, "Acoustic localization of terrestrial wildlife: Current practices and future opportunities," Ecol. Evol. **10**(13), 6794–6818 (2020).
[22]A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," IEEE Trans. Signal Process. **42**(9), 2481–2491 (1994).
[23]M. Wajid, A. Kumar, and R. Bahl, "Design and analysis of air acoustic vector-sensor configurations for two-dimensional geometry," J. Acoust. Soc. Am. **139**(5), 2815–2832 (2016).
[24]R. A. Holler, *The Evolution of the Sonobuoy from World War II to the Cold War*, Navmar Applied Sciences Corp., Warminster, PA, 2014).
[25]M. T. Silvia and R. T. Richards, "A theoretical and experimental investigation of low-frequency acoustic vector sensors," in *Proceedings of OCEANS '02 MTS/IEEE*, Biloxi, MS (October 29–31, 2002),Vol. 3, pp. 1886–1897.
[26]C. R. Greene, Jr., M. W. McLennan, R. G. Norman, T. L. McDonald, R. S. Jakubczak, and W. J. Richardson, "Directional frequency and recording (difar) sensors in seafloor recorders to locate calling bowhead whales during their fall migration," J. Acoust. Soc. Am. **116**(2), 799–813 (2004).
[27]M. A. McDonald, "Difar hydrophone usage in whale research," Can. Acoust. **32**, 155–160 (2004).
[28]A. M. Thode, T. Sakai, J. Michalec, S. Rankin, M. S. Soldevilla, B. Martin, and K. H. Kim, "Displaying bioacoustic directional information from sonobuoys using 'azigrams'," J. Acoust. Soc. Am. **146**(1), 95–102 (2019).
[29]A. M. Thode, A. S. Conrad, E. Ozanich, R. King, S. E. Freeman, L. A. Freeman, B. Zgliczynski, P. Gerstoft, and K. H. Kim, "Automated two-dimensional localization of underwater acoustic transient impulses using vector sensor image processing (vector sensor localization)," J. Acoust. Soc. Am. **149**(2), 770–787 (2021).
[30]A. M. Thode, A. Conrad, M. Lammers, and K. Kim, "Tracking multiple humpback whales simultaneously using time-frequency representations of active intensity on DIFAR acoustic vector sensors," J. Acoust. Soc. Am. **149**(4), A56 (2021).
[31]L. Tenorio-Hallé, A. M. Thode, M. O. Lammers, A. S. Conrad, and K. H. Kim, "Multi-target 2D tracking method for singing humpback whales using vector sensors," J. Acoust. Soc. Am. **151**(1), 126–137 (2022).
[32]H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," IEEE Trans. Acoust. Speech Signal Process. **35**(10), 1365–1376 (1987).
[33]G. L. D'Spain, J. C. Luby, G. R. Wilson, and R. A. Gramann, "Vector sensors and vector sensor line arrays: Comments on optimal array gain and detection," J. Acoust. Soc. Am. **120**(1), 171–185 (2006).
[34]R. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986).
[35]M. H. Hayes, *Statistical Digital Signal Processing and Modeling* (Wiley, New York, 2009).
[36]G. D'Spain, W. Hodgkiss, and G. Edmonds, "Energetics of the deep ocean's infrasonic sound field," J. Acoust. Soc. Am. **89**(3), 1134–1158 (1991).
[37]G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," Annu. Rev. Stat. Appl. **6**, 355–378 (2019).
[38]A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: Background, derivation, and applications," Wiley Interdiscip. Rev. Comput. Stat. **4**(2), 199–203 (2012).
[39]R. J. Steele and A. E. Raftery, "Performance of Bayesian model selection criteria for Gaussian mixture models," in *Frontiers of Statistical Decision Making and Bayesian Analysis*, edited by M.-H. Chen, D. K. Dey, P. Müller, D. Sun, and K. Ye (Springer, New York, 2010), pp. 113–130.
[40]M. Hawkes and A. Nehorai, "Acoustic vector-sensor processing in the presence of a reflecting boundary," IEEE Trans. Signal Process. **48**(11), 2981–2993 (2000).
[41]F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (Springer, New York, 2019).
[42]M. A. Gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound," in *Proceedings of Audio Engineering Society Convention 50*, London, UK (March 4–7, 1975).
[43]C. Faller and M. Kolundzija, "Design and limitations of non-coincidence correction filters for soundfield microphones," in *Proceedings of Audio Engineering Society Convention 126*, Munich, Germany (May 7–10, 2009).