

RESEARCH ARTICLE

Towards using virtual acoustics for evaluating spatial ecoacoustic monitoring technologies

Neel P. Le Penru^{1,2}  | Becky E. Heath³  | Jamie Dunning⁴  | Lorenzo Picinali¹  | Robert M. Ewers⁵ | Sarab S. Sethi⁶ 

¹Audio Experience Design, Dyson School of Design Engineering, Imperial College London, London, UK; ²Science and Solutions for a Changing Planet DTP, Imperial College London, London, UK; ³Department of Zoology, University of Cambridge, Cambridge, UK; ⁴Faculty of Biological Sciences, University of Leeds, Leeds, UK; ⁵Georgina Mace Centre for the Living Planet, Department of Life Sciences, Imperial College London, London, UK and ⁶Department of Life Sciences, Imperial College London, London, UK

Correspondence

Neel P. Le Penru

Email: neel.le-penru16@imperial.ac.uk

Funding information

Natural Environment Research Council,
Grant/Award Number: NE/S007415/1

Handling Editor: Nicolas Lecomte

Abstract

1. Small microphone arrays and sound-source localisation algorithms are increasingly prevalent in the passive acoustic monitoring (PAM) of ecosystems. These technologies enable analysis of natural soundscapes' spatial features, yielding additional insights into biodiversity and ecosystem health. While many of these technologies have been evaluated in the field, there is a lack of controlled, repeatable methods to test them.
2. We developed an ambisonic virtual sound environment (VSE) for simulating real natural soundscapes to evaluate spatial PAM technologies. We validated this novel approach using a PAM recorder with a six-microphone array, from which we extracted a typical suite of ecoacoustic metrics, including acoustic indices and avian species predictions and localisations from the software BirdNET and HARKBird, respectively. We first verified whether the VSE could replicate natural soundscapes well enough to test PAM technologies by comparing these metrics between field and VSE-based recordings. To pilot the VSE as an environment for testing PAM hardware, we assessed how orientation impacts the six-microphone array's performance by using the same suite of metrics to compare VSE recordings made with the array at various pitch angles. Finally, we piloted the VSE as a test platform for PAM software by investigating how BirdNET and HARKBird perform on bird calls added to the VSE-replicated soundscapes.
3. While the VSE and field recordings had similarities in some metrics, including spectral composition and BirdNET predictions, ambisonics' perceptual bias and susceptibility to spatial aliasing limited the spatial analyses that could be undertaken. Our trials nonetheless revealed that device orientation impacts the performance of HARKBird and certain ecoacoustic indices, and that BirdNET and HARKBird perform best on louder, more directional bird calls.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

4. Our results demonstrate the potential for this approach, but highlight limitations to using an ambisonics-based VSE. We thus provide guidelines for the use and refinement of such systems towards more standardised, controlled benchmarking of PAM technologies, empowering practitioners to make more informed decisions on using these vital tools.

KEYWORDS

virtual sound environment, spatial ecoacoustics, Ambisonics, microphone array, BirdNET, HARKBird, bioacoustics, passive acoustic monitoring

1 | INTRODUCTION

Spatial audio technologies are increasingly widespread for the passive acoustic monitoring (PAM) of natural environments. Regarding hardware, this primarily consists of multi-microphone recording approaches. These encompass both distributed microphone arrays, which consist of several recording devices spread across space (Blumstein et al., 2011; Collier et al., 2010; Mennill et al., 2012; Mennill & Vehrencamp, 2008), and, increasingly, integrated microphone arrays, in which two microphones (Hobson et al., 2002; Payo et al., 2021) or more (e.g. Celis-Murillo et al., 2009; Heath et al., 2024; Suzuki et al., 2017; Wijers et al., 2019) are contained within a single, multichannel recording device. Regarding software, sound-source localisation algorithms continue to improve, partly thanks to deep learning methods (Grumiaux et al., 2022) and the development of localisation applications for animal vocalisations, such as the avian call localisation tool HARKBird (Sumitani et al., 2019; Suzuki et al., 2017). Together these spatial PAM hardware and software developments are providing superior metrics of ecosystem health and biodiversity (Wijers et al., 2019), including species abundance estimates (Blumstein et al., 2011; Heath, 2022).

While many of these technologies have been tested in the field, real natural soundscapes are inherently and continually variable, and somewhat unpredictable, limiting the control and replicability of these test methods. This can partly be avoided with playback experiments, in which one or more loudspeakers play back pre-recorded sounds in the field (Cretois et al., 2022; De Rosa et al., 2022; Znidersic & Watson, 2022). Further control and replicability can be achieved by simulating more of the field recording process, but few methods for this have been proposed—especially for testing spatial PAM technologies. Kaneko and Gamper (2022) recently developed an approach to simulate distributed microphone array field recordings of a large forest and test their robustness to noise, reverberation and measurement errors, such as imperfect microphone placement. However, as their approach involves simulating field recordings, it cannot be used to test spatial PAM hardware. Moreover, their approach is not intended for testing multichannel recorders with integrated microphone arrays.

Other fields, such as Audiology, have leveraged various sound-field reconstruction techniques to create virtual sound environments (VSEs) (Mansour et al., 2019; Oreinos & Buchholz, 2016;

Simon et al., 2021) that replicate real spatial sound-scenes in a laboratory setting, typically to evaluate human hearing and test multi-microphone hearing aids. VSEs have been employed in audiology to overcome the limited control and reproducibility of real-world tests (Grimm et al., 2019; Simon et al., 2020)—challenges which also arise for field testing PAM technologies. Several studies have used ambisonics, a technique for creating, recording, modifying and reproducing spatial sound fields in an efficient and flexible manner, to create VSEs for evaluating hearing aids (Favrot & Buchholz, 2010; Oreinos & Buchholz, 2016; Simon et al., 2020, 2021) and measuring certain features of speech intelligibility and perception (Hui et al., 2021; Mansour et al., 2019).

In ambisonics, sound fields are decomposed into spherical harmonic components (akin to a Fourier series for 3-D space (Müller, 2006)), starting with an omnidirectional component, plus a number of directional components, each corresponding to a specific directional pattern, for example bi-directional (front-back, left-right, up-down) for the first order, and more complex ones for higher ambisonics orders. Monophonic signals can also be encoded to the spherical harmonic domain, allowing additional sources to be added to any location within a sound field. This spherical harmonic representation can then be decoded to any suitable loudspeaker array, provided a sufficient number of regularly spaced loudspeakers (at least $(n+1)^2$, where n is the ambisonic order) (Daniel, 2001; Moreau et al., 2006; Zotter & Frank, 2019). First order ambisonics was primarily developed by Michael Gerzon (1973, 1975). It is often used today for entertainment, such as rendering spatial audio over headphones (Zotter & Frank, 2019), but has a small 'sweet spot' in which the sound field is most accurately reproduced (Frank, 2014; Simon et al., 2020, 2021). Higher-order ambisonics has a larger sweet spot, but requires more microphones, loudspeakers and processing for recording and reproduction (Moreau et al., 2006; Zotter & Frank, 2019).

Even for higher orders, the sweet spot's size is limited, in part due to spatial aliasing during recording and playback (Moreau et al., 2006; Simon et al., 2020). This amounts to distortions between sound waves coming from different directions (Epain & Daniel, 2008), which can affect their expected arrival time and level differences within the reconstructed sound field (Simon et al., 2021), leading to inaccurate directionality. Higher-order ambisonic playback can suffer from aliasing as it always uses multiple loudspeakers to reproduce signals, even for single point sources (Simon et al., 2020). Furthermore, ambisonics

is primarily intended to reconstruct sound fields in a perceptually accurate manner for human listeners (Frank, 2014; Gerzon, 1992), but may not produce a precise physical replica of the sound field, even in the sweet spot. This underpins some of the above inaccuracies in ambisonic reconstruction as well as other spectral, spatial and temporal artefacts it can create (Oreinos & Buchholz, 2016).

Nonetheless, ambisonics is a powerful tool for capturing, altering and reproducing real, complex spatial sound fields (including their reverberation and other acoustic properties) instead of the more labour-intensive method of reconstructing them with room acoustics models, for example (Oreinos & Buchholz, 2016). This has driven ambisonics' uptake for applications requiring VSEs to recreate dynamic, noisy environments, and should be suited to reproducing natural soundscapes, whose acoustics are challenging to model.

We therefore developed a higher- (specifically, third-) order ambisonic VSE, consisting of a spherical 25 loudspeaker array, for reproducing real spatial natural soundscapes to evaluate spatial PAM recorders and software in a controlled, replicable manner. There has been some work on multi-loudspeaker playback in ecoacoustics, namely for studying avian behaviour (Celis-Murillo et al., 2009; Mennill & Vehrencamp, 2008; Quirós-Guerrero et al., 2017). However, there is little work involving full-sphere sound field recording and reconstruction in ecoacoustics, beyond more artistic and experiential applications (Monacchi, 2013). Moreover, we demonstrate that full-sphere reconstruction can be achieved in a comparatively resource-efficient way, through repurposed or increasingly accessible equipment.

Here, we detail this VSE test platform and work to validate it (Figure 1). This followed three objectives, to investigate whether the VSE could:

1. *Soundscape reproduction efficacy*: Effectively simulate real, spatial natural soundscapes for testing PAM technologies. We verified this by comparing a standard suite of ecoacoustic metrics between recordings of real and VSE-replicated soundscapes.
2. *Impact of device orientation*: Test PAM hardware and factors important to spatial PAM deployments. As a proof-of-concept for using the VSE as a test environment, we re-recorded VSE soundscapes with MAARU (multichannel acoustic autonomous recording unit; Heath et al., 2024), a six-microphone PAM recorder, oriented at various pitch angles, to study how device orientation impacts the same standard ecoacoustic metrics. These include the popular avian classification algorithm BirdNET (Kahl et al., 2021) and the avian localisation tool HARKBird (Sumitani et al., 2019). Together these applications can provide lower-bound avian species abundance estimates (Heath, 2022) by revealing whether there are several birds of a particular species calling simultaneously from different locations around a microphone array. However, HARKBird is expected to be susceptible to device orientation. Hence, we explore whether the VSE could reveal the degree to which HARKBird localisations are altered by MAARU's orientation.
3. *BirdNET and HARKBird performance on additional avian calls*: Test PAM analysis software. To demonstrate this, we used the platform

to explore the performance of BirdNET and HARKBird on additional avian calls embedded in the VSE soundscapes. We first added calls by ambisonic encoding; however, given the limitations of ambisonics' spatial accuracy, we also reproduced calls' direct paths from individual loudspeakers within the VSE and compared both embedding methods.

2 | MATERIALS AND METHODS

2.1 | Platform description

The concept of this test platform is to take high quality, spatial recordings of natural soundscapes and use a spatial sound-field reconstruction method to accurately reproduce them in a controlled environment (Figure 1a,c). One can then use the PAM hardware and software being tested to repeatedly record and analyse these 'virtual' soundscapes under various conditions (Figure 1d,l). For example, one can alter the set-up of the PAM recorder or modify the reconstructed soundscape to simulate effects like the presence of additional species (Figure 1l). This also enables PAM technologies to be tested on compatible spatial recordings of soundscapes from diverse and challenging environments (e.g. hard to reach or stormy areas), all within a single laboratory setting.

We adopted a flexible, resource-efficient approach to implementing the above concept. We used the 19-channel ZYLIA ZM-1 microphone array (ZYLIA, n.d.; Poznań, Poland) with ZYLIA's recording software (version 2.1.2, ZYLIA, 2022b) to take high quality recordings of various natural soundscapes (Figures 1a and 2a). The ZM-1 is considerably more affordable than other higher-order ambisonic recorders such as *MH Acoustics*' 'Eigenmike' devices.^{1,2} We used ZYLIA's software to convert the recordings to third-order ambisonics (SN3D normalisation, Furse-Malham channel ordering) (version 1.6.1; ZYLIA, 2022a) and reproduced these ambisonic recordings through an array of 25 pre-owned *Genelec* 8010A loudspeakers³ (*Genelec* Oy, Iisalmi, Finland; Figure 1c), mounted to a sphere constructed from two commercially available, steel hemispherical frames (Lifetime Products Inc., Clearfield, UT, USA; Figure 2).⁴ The third-order ZM-1 recordings were decoded to the loudspeaker array using the free ICST Ambisonics externals for *Max/MSP* (Institute for Computer Music and Sound Technology, 2021).

For most of this work the VSE was housed in a non-acoustically treated room (Figures 1 and 2), mirroring conditions in many ecoacoustics laboratories, given their focus on field recording and the high cost of creating anechoic/soundproof spaces. However, this room likely introduced artefacts from reverberation, room modes and so on.

¹<https://mhacoustics.com/products>.

²[https://www.creativefieldrecording.com/2017/03/30/ambisonic-field-recording-resources/\#:~:text=MH\%20Acoustics\%20EigenMike\%20em32\%20\(\%24, via\%20the\%20EigenMike\%20Interface\%20Box](https://www.creativefieldrecording.com/2017/03/30/ambisonic-field-recording-resources/\#:~:text=MH\%20Acoustics\%20EigenMike\%20em32\%20(\%24, via\%20the\%20EigenMike\%20Interface\%20Box).

³<https://www.genelec.com/8010a>.

⁴<https://www.lifetime.com/lifetime-90136-dome-climber-earthtone>.

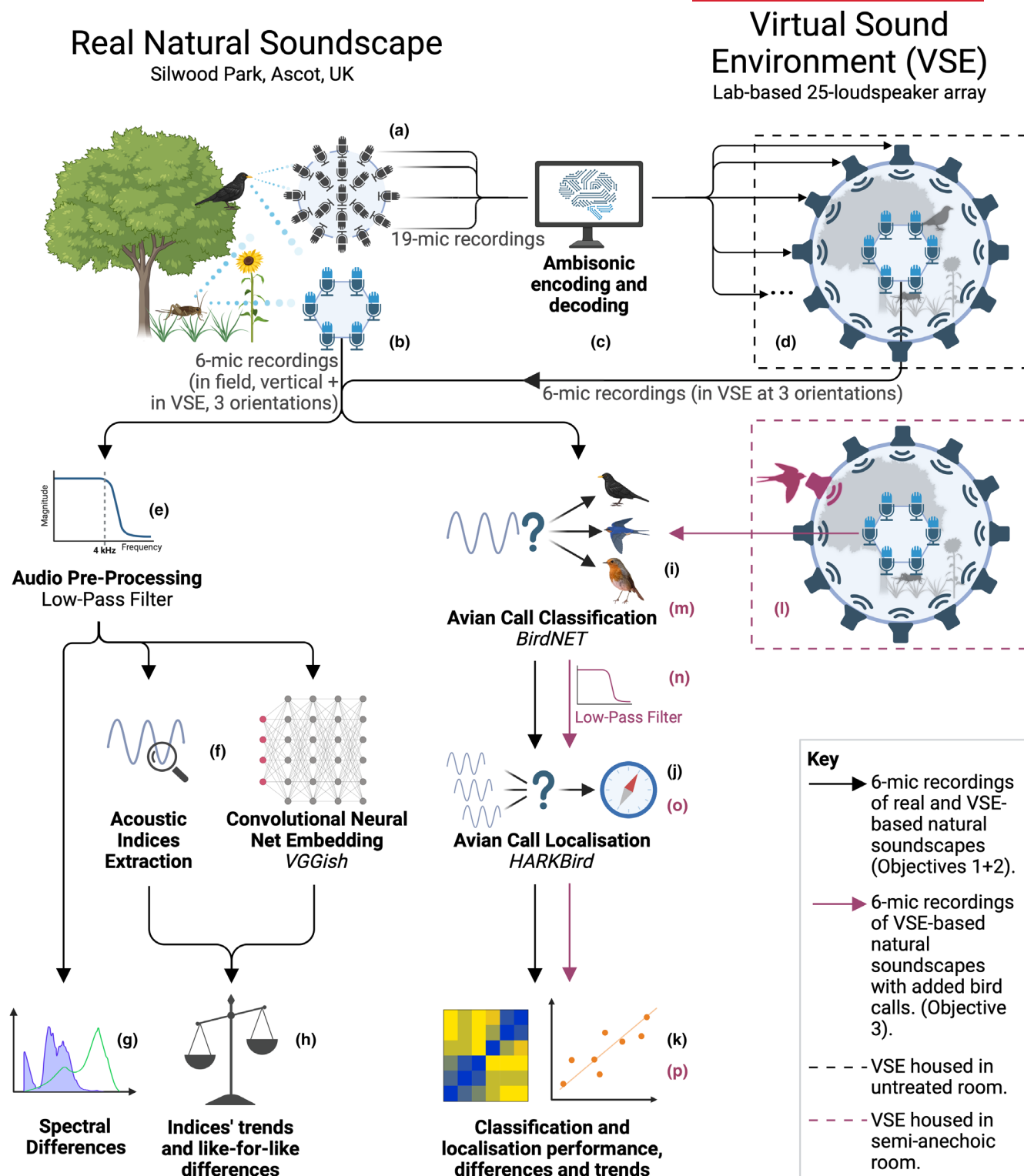


FIGURE 1 System overview and experimental structure. We captured real, spatial natural soundscapes with a high-quality 19-microphone array (a) and a multichannel acoustic autonomous recording unit (MAARU), the six-microphone PAM recorder being tested (b). The 19-mic recordings were encoded to the ambisonics domain and decoded to a 25-loudspeaker virtual sound environment (VSE) (c), allowing the soundscapes to be re-recorded by MAARU (d) with adjustments to relevant recording parameters, namely, device orientation. The six-microphone recordings from the field (b) and VSE (d) were passed through a suite of typical ecoacoustic metrics (f, i, j). Various analyses (g, h, k) were performed to compare the field and VSE recordings and explore the impact of device orientation. The process was repeated for VSE recordings with additional bird calls (l), to further explore the performance of software tools BirdNET and HARKBird (m–p).

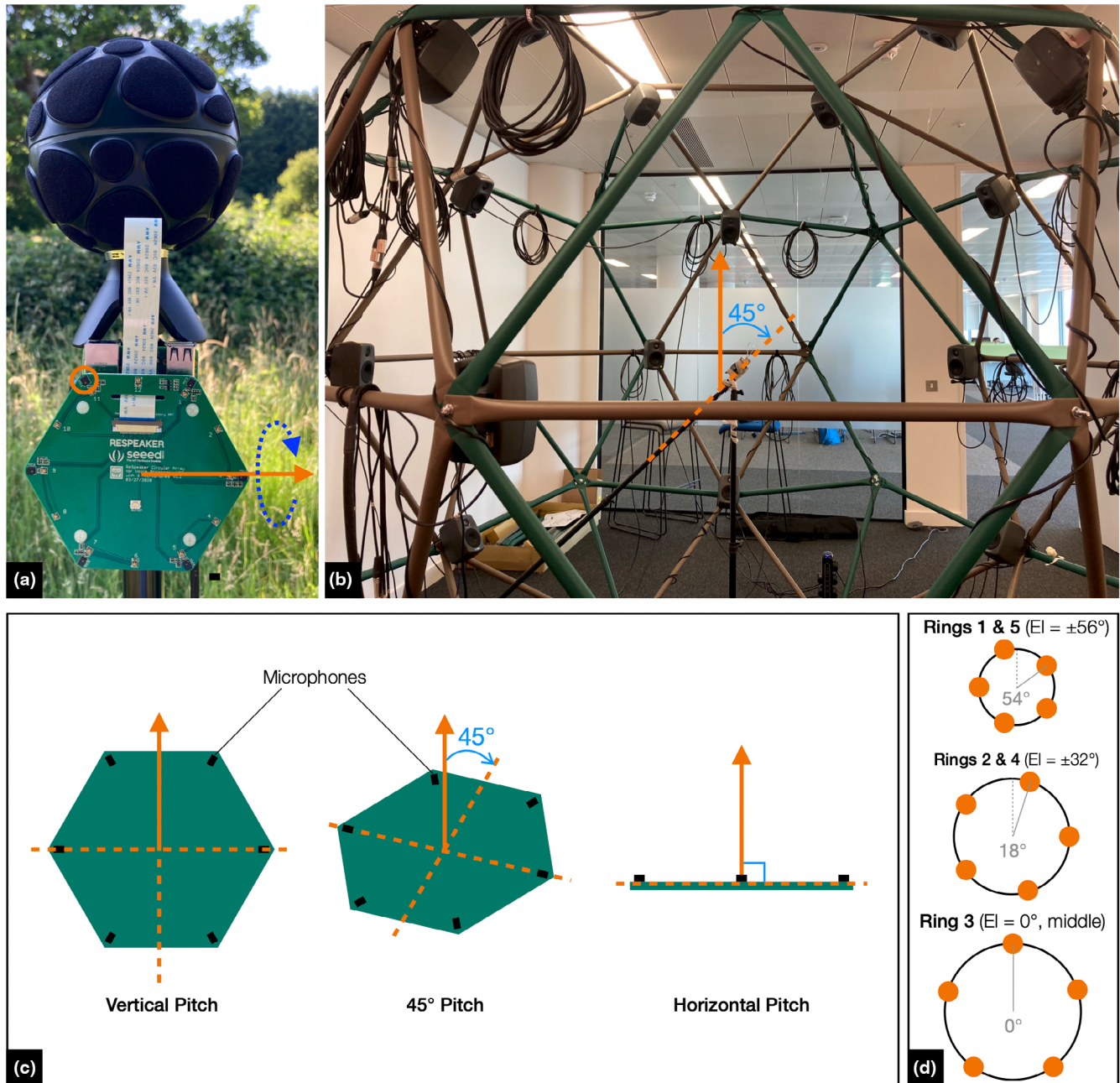


FIGURE 2 (a) Field recording setup with multichannel acoustic autonomous recording unit (MAARU), the six-microphone PAM recorder being tested, beneath the spherical 19-mic ZM-1 used for ambisonic recording. Recordings for all single-channel analyses were obtained from Mic 1 of MAARU, circled. The arrows show the rotation axis and direction to change MAARU's pitch angle in the virtual sound environment (VSE) recordings for Objective 2. (b) VSE in the laboratory space it was first housed in, with MAARU oriented at 45° pitch. (c) MAARU's pitch angles for the VSE recordings. (d) Arrangement of loudspeakers on the VSE. The VSE consists of five rings at the elevations indicated (EI = Elevation). Each ring contains five evenly spaced loudspeakers, the first of which is at the azimuth indicated inside the rings.

2.1.1 | Calibration

The VSE was calibrated in two steps to reproduce same approximate sound pressure level (SPL) as the original soundscapes. First, the 8010A loudspeakers were calibrated relative to one another. A calibrated SPL meter was placed in the centre of the VSE and each loudspeaker's gain was adjusted so it reproduced a 1 kHz sine wave to around 75 ± 1 dB. Second, the entire array was calibrated

relative to the ZM-1's built-in gain. White noise was encoded to third-order ambisonics at zero amplitude and elevation, decoded through the VSE and monitored with the SPL meter. It was then recorded with the ZM-1 positioned in the sphere's centre, and aligned at 0° azimuth and elevation. Next, the ZM-1-recorded white noise was reproduced through the VSE. Using the SPL meter, the gains of all loudspeakers were adjusted by around -12 dB so that the ZM-1's recording of the white noise was reproduced within around

± 1 dB of the original white noise. This was repeated for a 1 kHz sine wave, which also yielded an approximately -12 dB adjustment for all loudspeakers.

2.2 | Recording devices, locations and schedule

We piloted the VSE on MAARU (Heath et al., 2024), a six-microphone spatial PAM recorder (Figure 1b,d), with its weatherproofing enclosure removed. MAARU advances Sethi et al.'s (2018) open-source, monophonic PAM device with Sseed Studio's (n.d.) affordable 'ReSpeaker 6-Mic Circular Array Kit' which features six omnidirectional Micro-ElectroMechanical Systems (MEMS) microphones on the vertices of a flat hexagonal PCB (Sseed Technology Co., Ltd., Shenzhen, China). To restrict file size while still allowing for a breadth of analyses, we used MAARU's intended 16 kHz sampling rate.

We conducted field recordings in July 2022 at Imperial College London's Silwood Park campus (Ascot, Berks., UK), using MAARU oriented vertically and the ZM-1 (sampling rate 48 kHz) mounted directly above it (Figures 1a,b and 2). We took 10 min recordings at each of six sites spanning various levels of forest density and proximity to a stream or human activity. Government-issued fieldwork licences were not required for this work.

We re-recorded the simulated soundscapes using MAARU at the centre of the VSE (Figure 1d), initially oriented vertically (as in the field), then at 45° pitch, and finally horizontally (90° pitch; Figure 2).

2.3 | Soundscape reproduction efficacy and impact of device orientation

We extracted a standard suite of ecoacoustic metrics from MAARU's field and VSE recordings (240 mins total, Table S1) to compare MAARU's recordings towards Objectives 1 and 2 (Figure 1f-k). This comprised spectral power differences, seven acoustic indices, feature embeddings from an audio-specific convolutional neural network (CNN), and avian species predictions and localisations from BirdNET and HARKBird, respectively. Apart from HARKBird, these metrics use monophonic recordings. Such metrics are generally consistent across MAARU's microphones due to MEMS' microphones reliability and their proximity on MAARU's array. Indeed, the median RMS of recordings across all microphones for the 240 mins was 0.366 dB (range 0.062–8.892 dB). We thus extracted all monophonic metrics from MAARU's upper left microphone.

2.3.1 | Spectral differences, acoustic indices and VGGish features

We calculated power spectrograms for all the recordings with 'pspectrum' in MATLAB (version 2022b; The MathWorks, Inc.,

Natick, MA, USA), using 0.5 s Kaiser windows with 10% overlap, 0.7 leakage and 1024 Discrete Fourier Transform points. We then subtracted the field recordings' power spectrograms from those for the VSE recordings. In around two thirds of all recordings, we initially noticed a substantial increase in power around 7–8 kHz. This was likely introduced by background noise in the non-anechoic laboratory environment that may have been exacerbated by the frequency responses of the ZM-1 and Genelec loudspeakers. To remove this, we low-passed all recordings with a 4 kHz cutoff and 12 dB roll-off (Figure 1e). The low-passed data were used for most subsequent analyses except for BirdNET and HARKBird's use for Objective 2, as most real bird calls in these recordings were too attenuated by low-passing to be detected. This, however, was different for the HARKBird's use for Objective 3 (Figure 1o,p) where low-passing was beneficial (see Section 2.4).

We computed the following canonical acoustic indices (Bradfer-Lawrence et al., 2019; Heath et al., 2021; Machado et al., 2017) on 30 s windows.⁵ using the R packages *seewave* (version 2.2.0; Sueur et al., 2022) and *soundecology* (version 1.3.3; Pijanowski & Villanueva-Rivera, 2018): Acoustic Complexity Index (ACI), Acoustic Diversity Index (ADI), Acoustic Evenness (AEve), Bioacoustic Index (Bio), Acoustic Entropy (H), Median of the Amplitude Envelope (M), and Normalised Difference Soundscape Index (NDSI). We also used the 'vggishPreprocess' and 'predict' functions from MATLAB's *Deep Learning Toolbox* (version 14.5; Mathworks, 2022) with default parameters to compute the 128 dimension embedding of VGGish, a CNN pre-trained on a broad corpus of labelled audio samples from YouTube (Hershey et al., 2017). This embedding has been highly effective for numerous ecological analyses and classification tasks (Heath et al., 2021; Sethi et al., 2020).

We performed an adapted Bland–Altman analysis (Giavarina, 2015) to examine scaled differences in acoustic indices and VGGish features between field and VSE recordings. Bland–Altman analysis assesses the agreement between two methods of measurement and involves evaluating the difference between both measures to identify bias. This bias can be measured relative to the average of both measures, or one of the measures if it can be considered a reference. 'Limits of agreement' defining upper and lower bounds for acceptable bias can then be set based on context or differences' 95% confidence interval if they are normally distributed (Giavarina, 2015).

We used the field recordings as reference. However, the indices/features studied vary considerably in magnitude and range, lack clear contextual limits of agreement and were not normally distributed. Therefore, per Heath et al. (2021) and Araya-Salas et al. (2019), we calculated indices/features' differences as a percentage of their range in the field recordings. We zeroed differences for VGGish features with zero range in field recordings to avoid singularities, and averaged the VGGish differences across all 128 features to obtain a univariate difference set akin to other indices. As an ideal VSE should

⁵60 s or more is a more common window size (Bradfer-Lawrence et al., 2019) but would lead to too few data points given the brevity of our recordings.

Frequency band	Vertical		45°		Horizontal	
	Mean (dB)	Variance (dB)	Mean (dB)	Variance (dB)	Mean (dB)	Variance (dB)
0–1 kHz	12.19	2.37	12.78	2.3	12.96	2.28
1–2 kHz	9.02	0.76	11.45	0.53	13.09	0.58
2–3 kHz	5.54	1.82	10.67	1.1	13.54	1.23
3–4 kHz	4.49	2.4	7.62	1.64	12.41	1.73
4–5 kHz	4.7	1.86	6.35	1.95	12.32	1.86
5–6 kHz	1.67	2.61	3.34	2.3	7.85	1.86
6–7 kHz	0.73	1.26	1.31	1.35	2.61	1.68
7–8 kHz	0.05	0.63	0.06	0.78	0.09	0.97
Average across all frequency bands	4.8	1.71	6.7	1.49	9.36	1.52

TABLE 1 Average differences in spectral power (dB) per 1 kHz band (up to the 8 kHz Nyquist frequency) between all virtual sound environment re-recordings and the reference field recordings.

show minimal differences, we followed Heath et al. (2021) in using $\pm 5\%$ limits of agreement and considering data strongly altered if their interquartile range omitted zero. We also explored $\pm 2.5\%$, $\pm 7.5\%$ and $\pm 10\%$ limits of agreement, as a $\pm 5\%$ difference is challenging to interpret ecologically, given the extracted indices/features' correlations with biodiversity are highly context-dependent and may vary by recording/playback equipment (Sethi et al., 2023). Further, while differences were expected to be consistent across VSE recordings, we also assessed for correlations between MAARU's VSE pitch angle and indices' scaled differences.

2.3.2 | BirdNET and HARKBird

We ran unfiltered recordings from MAARU's upper left microphone through BirdNET (version 2.1; Kahl et al., 2021) and from its six microphones through HARKBird (Figure 1d). We passed the coordinates of each field site to the BirdNET GUI to restrict it to birds from this region, and used default values for the other parameters. While BirdNET results would be similar across recording conditions with an ideal VSE, BirdNET's confidence score—which often positively correlates with prediction accuracy—can vary greatly by species and recording/playback hardware (Wood & Kahl, 2024). We thus restricted all BirdNET predictions to those with a greater than or equal confidence to the average for their species in the corresponding field recording. Species only predicted in the re-recordings were filtered using the corresponding site's overall average field recording confidence. After filtering, recordings had 20 predictions on average (range=0–54). We filtered the HARKBird localisations to those whose timestamps overlapped with the filtered BirdNET predictions, and then found the percentage 'overlap' by calculating the proportion of field recordings' predictions and localisations that appeared in the VSE re-recordings. We used inferential statistics to explore differences in BirdNET predictions' confidence values and HARKBird's localisations between all recording conditions. We also evaluated correlations between device orientation and confidence values and localisations.

To compare BirdNET's precision and recall across all recording conditions, a professional ornithologist, well-experienced with conducting

UK bird surveys, manually labelled the omnidirectional channel of the ZM-1's field recordings using non-overlapping 3 s windows to match BirdNET. As desired precision and recall vary by application, we calculated them for a range of confidence thresholds: 0.25 (average predictions per recording=14, range=0–35), 0.5 (average predictions per recording=9, range=0–29) and 0.75 (average predictions per recording=5, range=0–20; Figure 6), and 0.01 and 0.9 (Table S2).

2.4 | BirdNET and HARKBird performance on additional avian calls

We performed 10 further VSE re-recordings (100mins total, Table S1), with MAARU oriented horizontally and additional avian calls embedded at certain spatial locations in the first five sites' ambisonic field recordings (Figure 1e; Table 4) during quieter passages. We then verified how accurately these were identified by BirdNET and localised by HARKBird.

We added six common English bird vocalisations (Table 4) from audio recordings freely available on Xeno-Canto (Ålberg, 2006, 2008; Bot, 2009; Krabbe, 1997; Matusiak, 2010; Poelstra, 2011). We first added the calls' direct paths to the soundscapes by encoding them to the ambisonics domain using the ICST Ambisonics externals in Max/MSP (Institute for Computer Music and Sound Technology, 2021), enabling them to originate from any direction. To avoid spatial aliasing in ambisonic reconstruction—which can be particularly problematic for HARKBird's underlying localisation algorithm, MUSIC (MULTiple Signal Classification) (Schmidt, 1986; Suzuki et al., 2017)—we also reproduced the direct paths from individual loudspeakers closest to each ambisonic encoding position, providing superior directionality.

In both cases, we added a simulation of each field site's reverberation, following a method outlined by Picinali (2006). The reverberation is produced by convolving the direct path of each added sound with the zero-padded Impulse Response of the acoustic reflections from the corresponding site. We found the impulse response at a distance of 2 m at 0°, 90°, 180° and 270° about the ZM-1 microphone

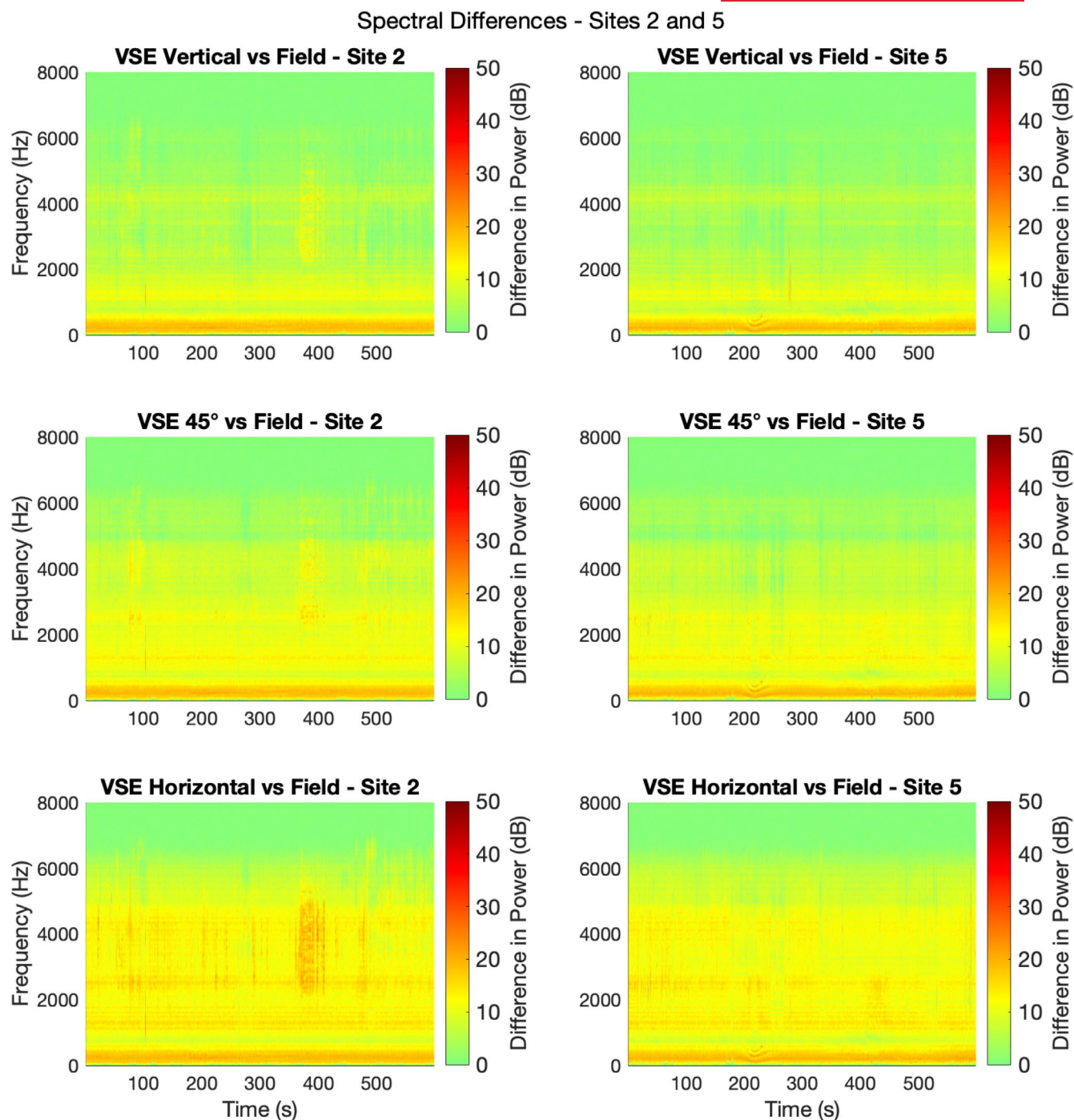


FIGURE 3 Differences in spectral power between each virtual sound environment (VSE) re-recording and the field recording (audio captured from top left microphone of multichannel acoustic autonomous recording unit and low-passed above 4 kHz with a 12 dB roll-off) for Sites 2 and 5, which exemplify patterns seen across all sites.

for all but the final site, which had insufficient space about the microphone to do so. Each added bird vocalisation was convolved with the impulse response measured closest to its azimuth.

We initially noticed inaccurate HARKBird localisations for both encoding methods, suggesting spatial aliasing between MAARU's microphones due to its size and geometry (Epain & Daniel, 2008). MAARU's closest microphones are approximately 4.5 cm apart, hence aliasing could occur for sounds with a wavelength of 9 cm or

less—roughly 3800 Hz and above. We therefore reused the 4 kHz low-pass filter from previous analyses on these recordings before passing to HARKBird (Figure 1n),⁶ notably improving localisation

⁶We did not apply the 4 kHz low-pass filter to recordings passed to BirdNET and HARKBird for Objectives 1 and 2, as the real bird calls in these recordings were too quiet for detection after low-passing, even with a 6–12 dB gain boost (beyond which clipping would occur).

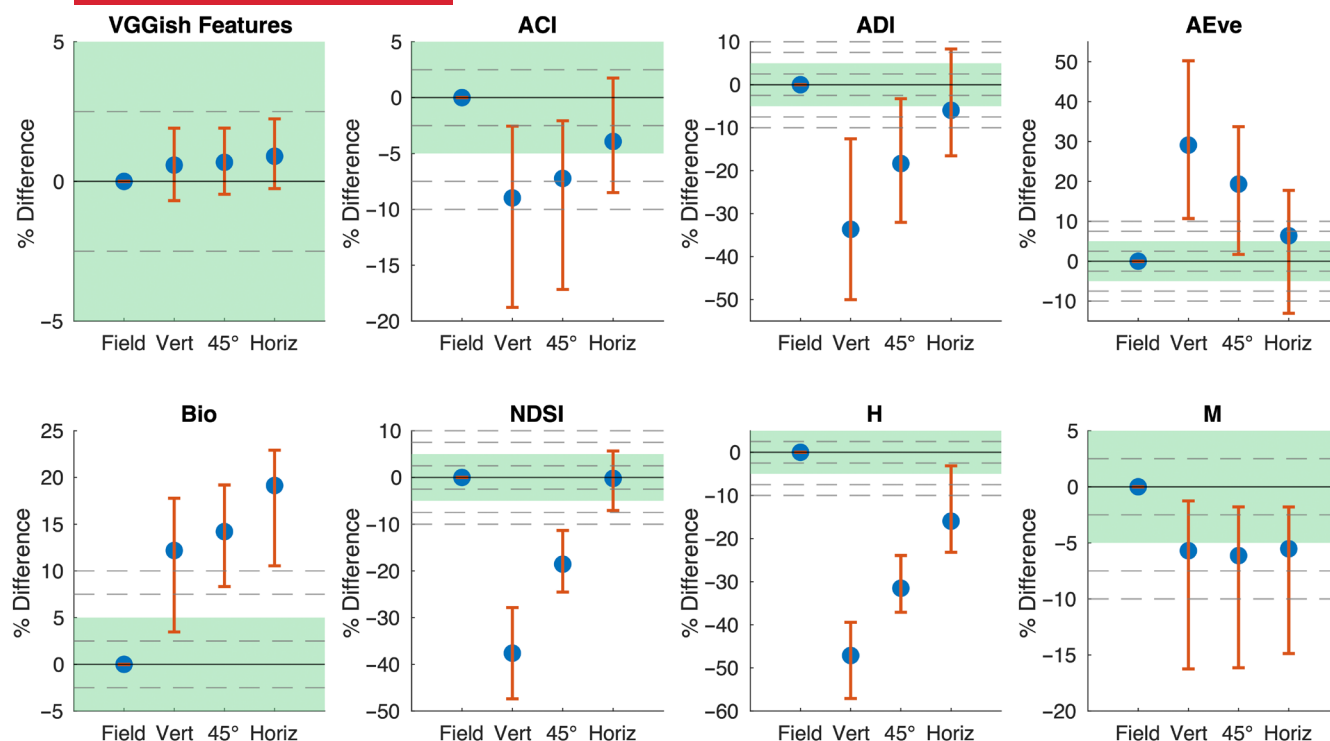


FIGURE 4 Plots of indices' average scaled like-for-like differences across all sites. Blue dots indicate the median percentage differences, whiskers show the upper and lower quartiles, and the shaded area denotes the $\pm 5\%$ limits of agreement. Where appropriate, the grey dashed lines represent $\pm 2.5\%$, $\pm 7.5\%$ and $\pm 10\%$ as alternative limits of agreement.

accuracy. To verify other factors did not cause the inaccuracies, we ran HARKBird on unfiltered recordings of bird calls added from individual loudspeakers without reverberation, and with neither reverberation nor the background ambisonic soundscapes (Figure S2). However, performance was best on low-passed, single loudspeaker recordings (with reverberation and the background soundscape), so we ultimately compared this approach to the ambisonic encoding.

After passing the recordings through BirdNET and HARKBird, we investigated the proportion of added calls that were accurately classified and localised and how the prediction confidence and localisation error varied by calls' elevation and embedding method.

3 | RESULTS

3.1 | Soundscape reproduction efficacy and impact of device orientation

3.1.1 | Spectral differences

There were progressively greater overall level increases for the vertical, 45° and horizontal re-recordings (Table 1), particularly in the 1–5 kHz bands. However, the average variance in spectral difference across all frequency bands for all re-recordings was only around 1.6 dB (range 1.49–1.71 dB). To exemplify these patterns, data from Sites 2 and 5 are shown in Figure 3.

TABLE 2 Spearman's rank correlation coefficients and p -values (with a significance level of 5% [bold]) assessing the relationship between MAARU's orientation and: (i) acoustic indices' and VGGish features' scaled like-for-like differences, (ii) HARKBird localisations and (iii) BirdNET confidence values (for all detected birds and each of the four most predicted species).

	Rho	p -value
Mean VGGish features	0.077	<0.001
ACI (Acoustic Complexity Index)	0.238	<0.001
ADI (Acoustic Diversity Index)	0.428	<0.001
AEve (Acoustic Evenness)	-0.392	<0.001
Bio (Bioacoustic Index)	0.238	<0.001
NDSI (Normalised Difference Soundscape Index)	0.772	<0.001
H (Acoustic Entropy)	0.586	<0.001
M (Median of the Amplitude Envelope)	-0.013	0.812
HARKBird localisations	-0.583	0.001
BirdNET confidence—all (19 species, 205 predictions)	-0.005	0.944
BirdNET confidence—Eurasian Jackdaw (68 predictions)	-0.131	0.286
BirdNET confidence—Parakeet (55 predictions)	0.008	0.951
BirdNET confidence—European Goldfinch (40 predictions)	-0.012	0.94
BirdNET confidence—Coal Tit (15 predictions)	0.103	0.715

3.1.2 | Acoustic indices and VGGish features

In our adapted Bland–Altman Analysis, only the VGGish features remained within $\pm 5\%$ difference to the field recordings across all VSE recordings, while the acoustic indices varied considerably (Figure 4; Figure S1). These trends were similar for other limits of agreement, with only ACI and M appearing acceptably altered under the more generous alternative limits. Interestingly, absolute percentage differences rarely increased as re-recording orientation progressively changed. Rather, most decreased or showed no trend. Significant

Spearman's rank correlations were found for all indices but M, the strongest being for ADI, AEve, NDSI and H (Table 2).

3.1.3 | BirdNET and HARKBird

Device orientation impacted the performance of BirdNET and HARKBird. For BirdNET, laboratory re-recordings overlapped with the field recordings by 51.2% (vertical), 40.4% (45°) and 55.4% (horizontal) on average. Figure 5 shows spectrograms of vocalisations

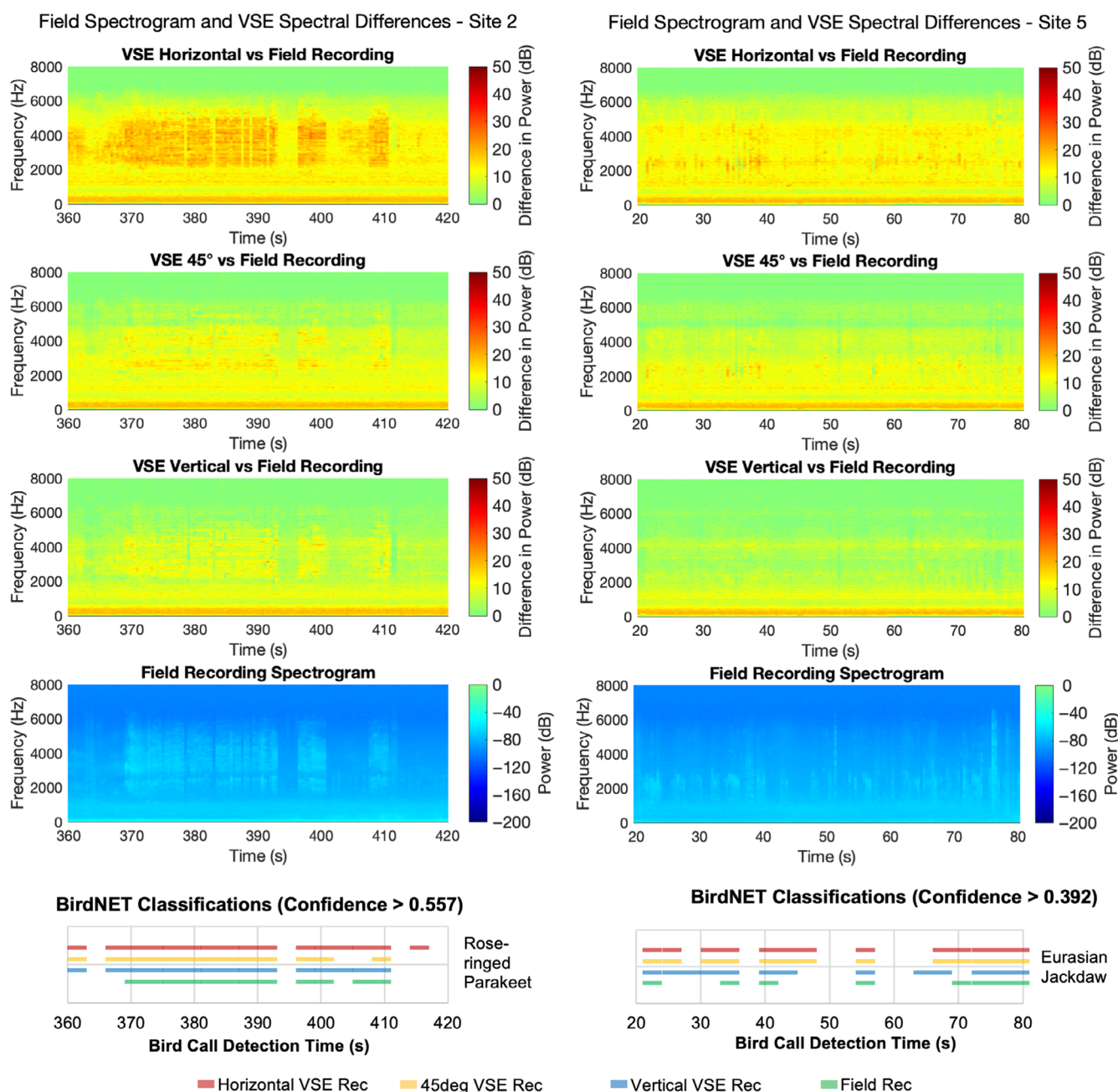


FIGURE 5 Field recording spectrogram, virtual sound environment recordings' spectral differences to the field recording, and BirdNET predictions over time for a 60s window in the recordings for Site 2 and Site 5. Both recording segments contain many avian calls, reflected in the increases in spectral power in the spectrograms. The BirdNET predictions mostly correctly align with the spectral patterns but there is some inconsistency in predictions between recording conditions.

that were repeatedly correctly predicted by BirdNET in each recording condition, but the predictions did not always occur at the same times (i.e. different overlap). For HARKBird, allowing $\pm 45^\circ$ in localisation error, average overlaps were just 2.8% (vertical), 4.7% (45°) and 12.3% (horizontal).

BirdNET confidence values and HARKBird localisations were not normally distributed. Friedman tests showed significant differences between the recording conditions for confidence values ($\chi^2(3) = 61.3$, $p < 0.001$) and localisations ($\chi^2(3) = 24.0$, $p < 0.001$). Dunn post-hoc tests revealed significant differences in confidence between the field recording and each VSE recording, and in localisations between the field and VSE vertical recordings, and the VSE vertical and 45° recordings (Table 3).

There was a significant Spearman's rank correlation between device orientation and HARKBird localisations, but not between orientation and BirdNET confidence (Table 2).

BirdNET's precision and recall generally fell within similar ranges across recording conditions though there was some variation between conditions and confidence thresholds (Figure 6). The metrics were limited or not calculable for Sites 3 and 4 due to a lack of avian activity. Precision was frequently over 0.8, sometimes limited by false positive predictions. Recall was always relatively low (below 0.3), but more consistent between recording conditions and confidence thresholds. Field recordings often had high precision but the lowest recall due to fewer BirdNET predictions.

3.2 | BirdNET and HARKBird performance on additional avian calls

BirdNET correctly identified almost all added bird calls, with most over 0.7 confidence (Table 4). HARKBird localised only 30% of calls added

by ambisonic encoding to within 45° , compared to 70% of those added from individual loudspeakers. Performance was worse for some species and the Blue Tit and Robin were sometimes not detected by BirdNET or HARKBird, respectively (Table 4). Calls added at higher elevations seem correlated with larger BirdNET confidence and HARKBird error (Figure 7), though this is likely due to other factors for BirdNET.

We compared the BirdNET confidence values and HARKBird errors from both embedding methods using the Mann-Whitney U test, as the data were not normally distributed.

While the difference in BirdNET confidence was not significant ($p = 0.453$, $U = 26,787$, $Z = -0.751$), HARKBird errors were significantly lower for the direct-from-loudspeaker bird calls ($p < 0.001$, $U = 25,233$, $Z = -17.839$).

4 | DISCUSSION

Results from the spectral differences, VGGish, and, to a lesser degree, BirdNET, show potential for VSEs to simulate soundscapes well enough to test key aspects of PAM technology. However, our results also highlight important caveats to this approach, including the limitations of ambisonics and the space housing our VSE.

4.1 | Soundscape reproduction efficacy and impact of device orientation

4.1.1 | Spectral differences

Though not a spatial consideration, it was fundamental to verify that overall spectral power patterns were preserved in the VSE (Figure 3), as confirmed by the relatively low variance in spectral

		Rec condition			
		Field—Vertical	VSE—Vertical	VSE— 45°	VSE—Horizontal
Rec condition	Field—Vertical	—	BN: <0.001 HB: <0.001	BN: <0.001 HB: 0.670	BN: <0.001 HB: 0.035
	VSE—Vertical	—	—	BN: 0.972 HB: 0.006	BN: 0.300 HB: 0.283
	VSE— 45°	—	—	—	BN: 0.829 HB: 0.670
	VSE—Horizontal	—	—	—	—
Median of BN conf. values		0.823	0.928	0.887	0.942
SD of BN conf. values		0.219	0.171	0.156	0.185
BN No. species predicted		4			
BN No. predictions		75			
Median of HB localisations		185	331	260	265
SD of HB localisations		40.7	2.23	4.41	23.6
HB No. localisations		9			

TABLE 3 Results from Dunn post-hoc tests following Friedman tests assessing for significant differences in BirdNET (BN) confidence and HARKBird (HB) localisations between all recording conditions (field and virtual sound environment (VSE)). The significance level is set at 0.83% following Bonferroni correction (bold).

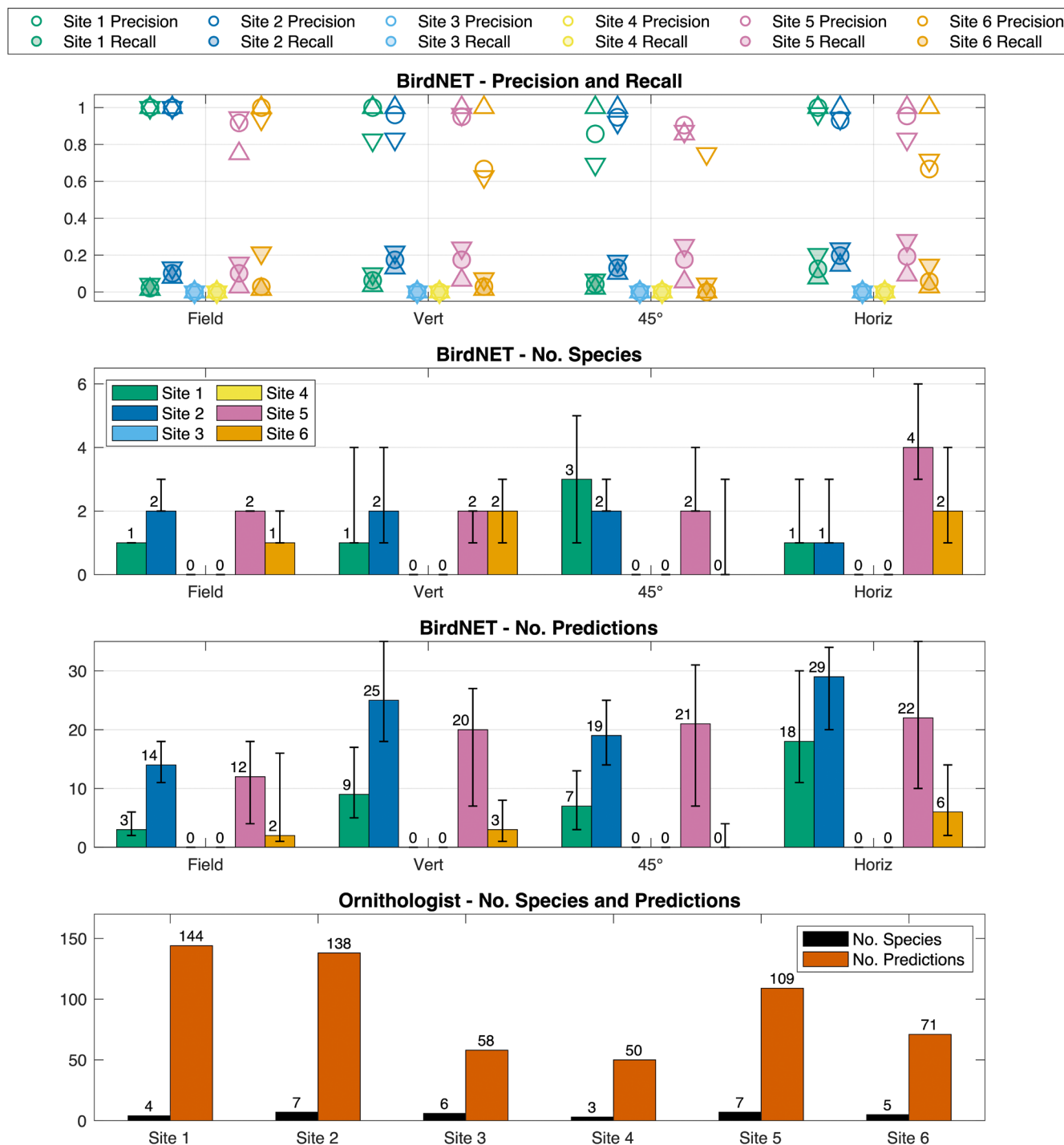


FIGURE 6 Precision and recall of BirdNET predictions on field recordings and virtual sound environment recordings at various orientations, filtered using minimum confidence thresholds of 0.25, 0.5 and 0.75. Precision and recall were calculated based on manual labelling of the omnidirectional channel of the ambisonic field recordings by a professional ornithologist. Values for precision when BirdNET made no predictions (sites 3 and 4) and recall when no bird calls were actually present are omitted to avoid division by zero. Downward facing triangles denote a confidence threshold of 0.25, circles 0.5 and upward facing triangles 0.75. Numbers denote the number of species or predictions (total number of 3 s windows with calls from each species) detected by the ornithologist and BirdNET (0.5 confidence threshold). Error bars indicate the number of species or predictions for the 0.25 (lower) and 0.75 (upper) confidence thresholds. Raw data and those for 0.01 and 0.9 confidence thresholds are presented in Table S2.

differences across all re-recordings (Table 1). This also suggests that level increases—though unexpected—were consistent within all re-recordings (and not, for instance, specific to transient sounds like

bird calls). As such, level increases likely stemmed from factors that impacted the recordings overall, such as reverberation and standing waves from the untreated room initially housing the system.

Site	Common name	Ambisonic encoding			Individual loudspeaker		
		Az, El (deg)	BN conf	HB Az Err (deg)	Az, El (deg)	BN conf	HB Az Err (deg)
1	Robin	102, 56	0.97	166	126, 56	0.93	-51
1	Sparrow	182, 67	0.56	-160	198, 56	0.44	-118
2	Blue Tit	30, 45	0.4	10	18, 32	N/A	-16
2	Wood Pigeon	180, 45	0.73	-145	198, 56	0.9	10
3	Greenfinch	63, 35	0.93	-151	90, 56	0.99	1
3	Blackbird	260, 71	0.57	-145	270, 56	0.62	23
4	Blackbird	0, 0	0.71	35	0, 0	0.53	0
4	Blue Tit	150, 25	0.4	-35	162, 32	N/A	-22
5	Greenfinch	13, 68	0.9	92	342, 56	0.99	-18
5	Robin	243, 83	0.81	-133	270, 56	0.95	N/A
Mean (abs. values)		—	0.7	107	—	0.8	29
SD (abs. values)			0.21	59		0.22	37

Note: The added calls were embedded by either encoding to the ambisonics domain or playback from individual loudspeakers in the VSE. Positive azimuth errors are clockwise from the true value, and negative error values anticlockwise.

Abbreviations: Abs, absolute; Az, azimuth; BN, BirdNET; Conf, confidence; deg, degrees; El, elevation; Err, error; HB, HARKBird; VSE, virtual sound environment.

In addition, the increases in spectral power as device orientation changed are most likely due to the larger number of loudspeakers in the horizontal plane of our ambisonic VSE. This stems from ambisonics' intended purpose of reproducing spatial soundscapes in a perceptually accurate manner, as humans are particularly sensitive to localising sounds horizontally.

4.1.2 | VGGish features and acoustic indices

We reiterate that while it is challenging to ecologically interpret the extracted features/indices (Sethi et al., 2023), differences for an ideal VSE should be near zero. Acoustic indices' large and variable differences could stem from the limited recording duration and/or window length. These patterns may also reflect acoustic indices' inherent limitations: compared to VGGish features, acoustic indices have been found to be less robust to audio compression (Heath et al., 2021) and worse ecological descriptors when used as the input to various ecosystem classification tasks (Sethi et al., 2020). Conversely, VGGish features could also be more robust to ambisonic reproduction as they originate from a CNN trained on human-labelled audio (Gemmeke et al., 2017; Hershey et al., 2017) and may thus be biased towards human perception.

4.1.3 | BirdNET and HARKBird

As BirdNET uses monophonic recordings, we anticipated most field recordings' species predictions to be detected in the VSE recordings.

TABLE 4 Summary of BirdNET predictions and HARKBird localisations on bird calls added to various spatial soundscape recordings from Silwood Park reproduced through our ambisonic VSE.

However, we expected higher and more consistent overlap. The 40.4%–55.4% overlap range may relate to BirdNET's relatively low recall on our recordings, with the large number of missed calls sometimes resulting in species predictions at different time-stamps in the field and VSE recordings.

BirdNET also tends to return more predictions and higher confidence values on louder bird calls. This may explain the higher percentage overlap for the horizontal recordings, which were loudest due to the concentration of loudspeakers in the VSE's horizontal plane (Figure 3). This increase in level also likely underpins the significantly higher confidence values for the VSE recordings, while BirdNET's lack of spatial considerations is reflected in the absence of both significant differences in confidence between re-recordings themselves (Table 3) and significant correlations between confidence and device orientation (Table 2).

For HARKBird, we expected the significant correlation between localisations and device orientation (Table 2), but it was surprising that so few HARKBird localisations in the laboratory recordings matched the field recordings, and that the largest overlaps were observed for the 45° and horizontal laboratory recordings. Given this and the post-hoc results (Table 3), it seems sound source direction can be inaccurate in our ambisonic playback system and/or MAARU can suffer from considerable spatial aliasing between microphones. Indeed, on average 28.4% of HARKBird field localisations appeared in the vertical re-recordings, but most had >45° error. Moreover, the underlying MUSIC algorithm HARKBird uses to localise calls typically performs poorly on correlated signals (Schmidt, 1986). Ambisonic playback's spreading of sound source position through the activation of multiple loudspeakers (Simon

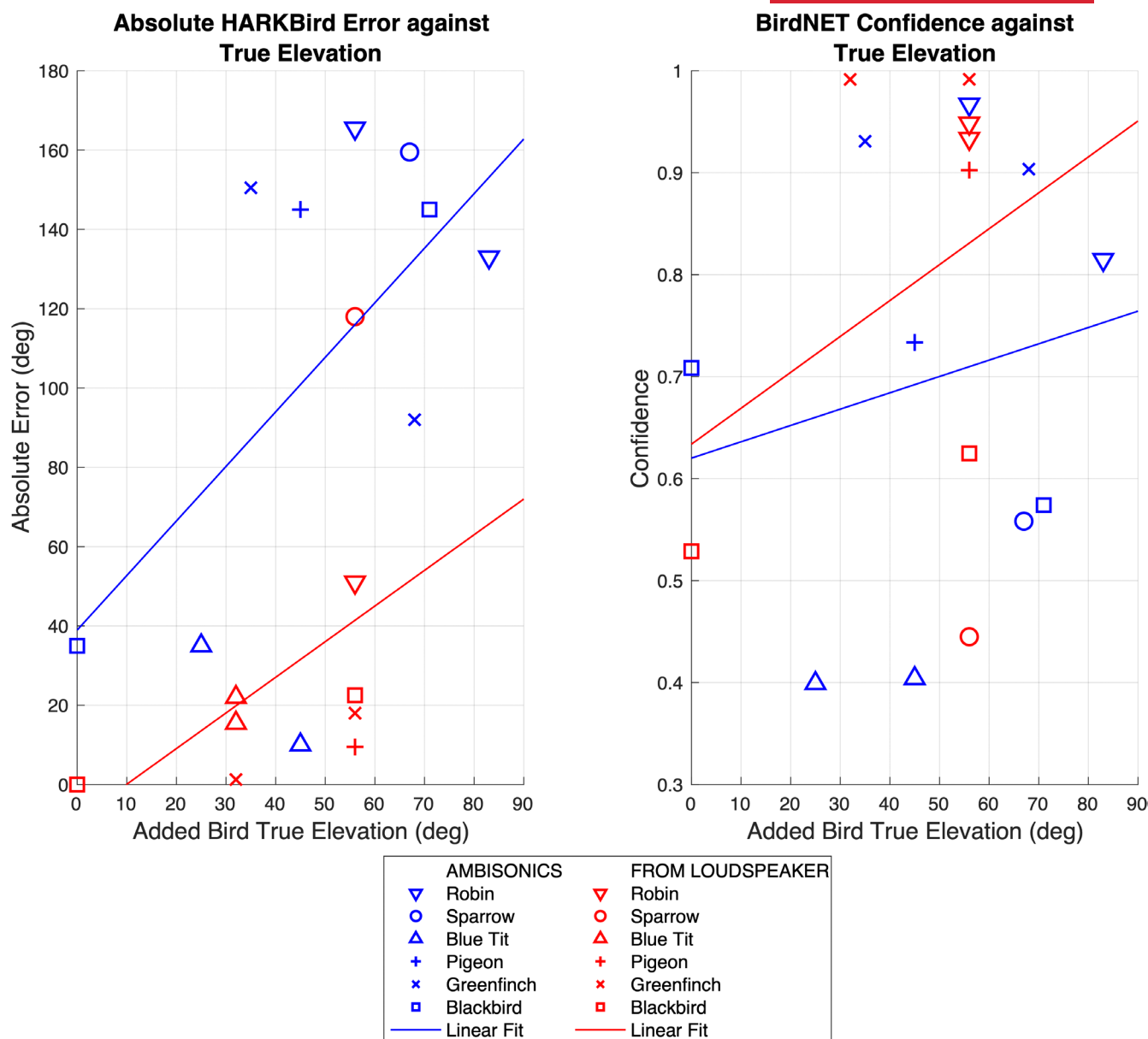


FIGURE 7 Graphs of HARKBird error and BirdNET confidence against the true elevation values of bird calls added to virtual sound environment-based spatial natural soundscapes via ambisonic encoding or reproduction from individual loudspeakers. Calls added from individual loudspeakers generally had both higher BirdNET confidence and lower HARKBird errors.

et al., 2020) results in correlated, coherent signals originating from multiple locations, which likely further degraded HARKBird's performance.

Overall, we met aspects of Objectives 1 (Efficacy of Soundscape Reproduction) and 2 (Impact of Device Orientation), given the VSE's preservation of spectral patterns (despite overall level increases), VGGish features' minimal differences between the field and VSE, and the ability to use the VSE to draw baseline conclusions on device orientation's impact on BirdNET and HARKBird. However, we also identified limitations to both HARKBird and our VSE. Future work should thus explore other, more accurate spatial audio reproduction techniques, such as those based on sound field reconstruction rather than human perception.

4.2 | BirdNET and HARKBird performance on additional avian calls

Our tests towards Objective 3 revealed strong performance from BirdNET and potential for HARKBird on clear, highly directional signals.

We note that all added bird calls were relatively loud compared to others in the soundscapes. Since a large range of SPLs could occur in the field depending on a bird's proximity to the microphone, we set added calls' gains to a level deemed perceptually appropriate for a nearby bird. This, however, is likely the main reason for BirdNET's high confidence values. Further work could explore adding calls

at progressively lower levels. Nonetheless, these tests reiterated BirdNET's robustness to spatial variation given predictions' confidence values were not significantly different when added via ambisonic encoding or single loudspeaker playback, though it is unclear why the added Blue Tit vocalisations were not detected when reproduced directly from the loudspeaker.

The superior accuracy of HARKBird localisations on bird calls added from single loudspeakers further highlights the issue of spatial aliasing in ambisonic playback, with multiple loudspeakers activating for directional calls in our third-order VSE, and the difficulties that the MUSIC algorithm underlying HARKBird faces under such conditions (Schmidt, 1986; Suzuki et al., 2017). However, even with single loudspeaker playback, most errors were larger than in other testing of MAARU, in which calls have been localised to within $\pm 10^\circ$ (Heath et al., 2024). This is most likely due to the addition of reverberation in both embedding methods blurring the source position of the bird calls, as it was generated from the impulse response at the closest of four azimuths (0° , 90° , 180° and 270°) to each added call's direct path, and could therefore differ by up to $\pm 45^\circ$. Moreover, as the reverberation was generated in the ambisonic domain, it will have been reproduced from multiple loudspeakers, resulting in lower directionality.

Further, in previous experiments with MAARU, sounds were all reproduced in the horizontal plane. However, as reflected in Figure 7, some error is expected as elevation increases, since sound sources are no longer in the same plane as MAARU's microphones and may thus appear to originate from different positions.

While further work could interrogate the above findings, the ability to generate these insights indicates promise for using a VSE to evaluate the performance of PAM software. Further, in totality our results suggest potential for using VSEs in other ecological applications requiring soundscape reproduction. For example, VSEs could provide a more realistic and immersive sonic environment for laboratory-based playback experiments to measure animal behaviour (De Rosa et al., 2022).

5 | GUIDELINES AND RECOMMENDATIONS

In this proof-of-concept, we have showcased the potential for VSEs to be used as test platforms for spatial and single-channel PAM hardware and software. Further, we have highlighted the benefits and pitfalls of ambisonic sound field reconstruction in such VSEs. Based on this, we propose the following Guidelines and Recommendations for the use and further development of such platforms:

1. Guideline 1. With careful experimental design, VSEs can be used to test monophonic PAM hardware and software tools, and show potential for testing spatial PAM technologies.
 - Recommendation 1. We hope a standard suite of analyses may be adopted for testing different classes of PAM technologies, for example spectral differences and, as in (Heath et al., 2021),

a set of typical acoustic indices alongside embedding into an audio-specific CNN.

2. Guideline 2. Ambisonics is a useful and increasingly accessible method for reproducing real, spatial natural soundscapes in a VSE, thanks to its ability to capture actual sound fields and its setup-agnostic nature. Care should be taken, however, if using an ambisonic VSE to test spatial PAM technologies, due to the potential for inaccurate sound field reconstruction arising from ambisonics' use of multiple loudspeakers and its bias to human perception.
 - Recommendation 2. We recommend considering ambisonics' limitations, using higher-order ambisonics (third order and above), and/or conducting analyses focussed on lower frequencies so as to have a larger sweet spot of more accurate sound field reconstruction. Alternatively, we suggest exploring other methods of sound field reconstruction (e.g. vector-based amplitude panning (Pulkki, 1997)).
3. Guideline 3. VSEs should ideally be housed in an acoustically treated, anechoic space. However, ambisonic VSEs can be designed and built in a scalable, resource-efficient way.
 - Recommendation 3. We recommend using affordable or existing spaces and equipment for the creation of further, diverse VSEs, providing this permits for sufficiently accurate sound field reconstruction (e.g. sufficient number and spacing of loudspeakers for ambisonic systems) and the VSE is carefully calibrated. Doing so, VSEs can be constructed in days (weeks if sourcing most parts), while hours of field recording can allow for hours to days of laboratory re-recordings, depending on experimental conditions. Therefore, ecoacoustic VSEs can be built and piloted in a month. We hope this facilitates the development of more VSE test platforms for PAM technologies across manifold spaces and budgets, while reducing the potential environmental impact of this work.

AUTHOR CONTRIBUTIONS

Neel P. Le Penru, Lorenzo Picinali, Robert M. Ewers, and Sarab S. Sethi conceived the ideas and designed methodology. Jamie Dunning performed the manual labelling of avian calls; Neel P. Le Penru collected all other data. Becky E. Heath ran recordings through HARKBird; Neel P. Le Penru undertook all other analysis. Neel P. Le Penru led the manuscript writing. All authors contributed critically to drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

The authors thank Dr. David Orme for analysis advice, Alessia Borrelli for data collection assistance and Dr. Connor Wood for advice on analysing BirdNET outputs. Figure 1 was created with BioRender. NPLP was supported by the Natural Environment Research Council (grant number NE/S007415/1).

CONFLICT OF INTEREST STATEMENT

None.

PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14405>.

DATA AVAILABILITY STATEMENT

Data and code are available via <https://doi.org/10.5281/zenodo.13157933> (Le Penru et al., 2024).

ORCID

Neel P. Le Penru  <https://orcid.org/0009-0003-5407-3723>

Becky E. Heath  <https://orcid.org/0000-0002-9111-6994>

Jamie Dunning  <https://orcid.org/0000-0001-8234-8526>

Lorenzo Picinali  <https://orcid.org/0000-0001-9297-2613>

Sarab S. Sethi  <https://orcid.org/0000-0002-5939-0432>

REFERENCES

- Ålberg, P. (2006). XC28208 Eurasian blue tit (*Cyanistes caeruleus*). <https://xeno-canto.org/28208>
- Ålberg, P. (2008). XC27554 European robin (*Erithacus rubecula*). <https://xeno-canto.org/27554>
- Araya-Salas, M., Smith-Vidaurre, G., & Webster, M. (2019). Assessing the effect of sound file compression and background noise on measures of acoustic signal structure. *Bioacoustics*, 28(1), 57–73. <https://doi.org/10.1080/09524622.2017.1396498>
- Blumstein, D. T., Mennill, D. J., Clemens, P., Girod, L., Yao, K., Patricelli, G., Deppe, J. L., Krakauer, A. H., Clark, C., Cortopassi, K. A., Hanser, S. F., McCowan, B., Ali, A. M., & Kirschel, A. N. G. (2011). Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus. *Journal of Applied Ecology*, 48(3), 758–767. <https://doi.org/10.1111/j.1365-2664.2011.01993.x>
- Bot, S. (2009). XC35048 European greenfinch (*Chloris chloris*). <https://xeno-canto.org/35048>
- Bradfer-Lawrence, T., Gardner, N., Bunnefeld, L., Bunnefeld, N., Willis, S. G., & Dent, D. H. (2019). Guidelines for the use of acoustic indices in environmental research. *Methods in Ecology and Evolution*, 10(10), 1796–1807. <https://doi.org/10.1111/2041-210X.13254>
- Celis-Murillo, A., Deppe, J. L., & Allen, M. F. (2009). Using soundscape recordings to estimate bird species abundance, richness, and composition. *Journal of Field Ornithology*, 80(1), 64–78. <https://doi.org/10.1111/j.1557-9263.2009.00206.x>
- Collier, T. C., Kirschel, A. N. G., & Taylor, C. E. (2010). Acoustic localization of antbirds in a Mexican rainforest using a wireless sensor network. *The Journal of the Acoustical Society of America*, 128(1), 182–189. <https://doi.org/10.1121/1.3425729>
- Cretois, B., Rosten, C. M., & Sethi, S. S. (2022). Voice activity detection in eco-acoustic data enables privacy protection and is a proxy for human disturbance. *Methods in Ecology and Evolution*, 13(12), 2865–2874. <https://doi.org/10.1111/2041-210X.14005>
- Daniel, J. (2001). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia* [PhD, Université Paris VI]. http://gyronymo.free.fr/audio3D/download_Thesis_PwPt.html
- De Rosa, A., Castro, I., & Marsland, S. (2022). The acoustic playback technique in avian fieldwork contexts: A systematic review and recommendations for best practice. *Ibis*, 164(2), 371–387. <https://doi.org/10.1111/ibi.13033>
- Epain, N., & Daniel, J. (2008). Improving spherical microphone arrays. *Audio Engineering Society Convention*, 124. <http://www.aes.org/e-lib/browse.cfm?elib=14609>
- Favrot, S., & Buchholz, J. M. (2010). LoRA: A loudspeaker-based room Auralization system. *Acta Acustica united with Acustica*, 96(2), 364–375. <https://doi.org/10.3813/AAA.918285>
- Frank, M. (2014). How to make Ambisonics sound good. *Forum Acusticum 2014*. European Acoustics Association, Krakow, Poland.
- Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. <https://doi.org/10.1109/ICASSP.2017.7952261>
- Gerzon, M. A. (1973). Periphery: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1), 2–10.
- Gerzon, M. A. (1975). The design of precisely coincident microphone arrays for stereo and surround sound. *Audio Engineering Society Convention*, 50. <http://www.aes.org/e-lib/browse.cfm?elib=2466>
- Gerzon, M. A. (1992). General metatheory of auditory localisation. *Audio Engineering Society Convention*, 92. <http://www.aes.org/e-lib/browse.cfm?elib=6827>
- Giavarina, D. (2015). Understanding bland Altman analysis. *Biochemia Medica*, 25(2), 141–151. <https://doi.org/10.11613/BM.2015.015>
- Grimm, G., Luberadzka, J., & Hohmann, V. (2019). A toolbox for rendering virtual acoustic environments in the context of audiology. *Acta Acustica united with Acustica*, 105(3), 566–578. <https://doi.org/10.3813/AAA.919337>
- Grumiaux, P.-A., Kitić, S., Girin, L., & Guérin, A. (2022). A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America*, 152(1), 107–151. <https://doi.org/10.1121/10.0011809>
- Heath, B. E. (2022). *The assessment and development of methods in (spatial) sound ecology* [PhD Thesis, Imperial College London]. <http://spiral.imperial.ac.uk/handle/10044/1/107090>
- Heath, B. E., Sethi, S. S., Orme, C. D. L., Ewers, R. M., & Picinali, L. (2021). How index selection, compression, and recording schedule impact the description of ecological soundscapes. *Ecology and Evolution*, 11(19), 13206–13217. <https://doi.org/10.1002/ece3.8042>
- Heath, B. E., Suzuki, R., Le Penru, N. P., Skinner, J., Orme, C. D. L., Ewers, R. M., Sethi, S. S., & Picinali, L. (2024). Spatial ecosystem monitoring with a Multichannel Acoustic Autonomous Recording Unit (MAARU). *Methods in Ecology and Evolution*, 1–12. <https://doi.org/10.1111/2041-210X.14390>
- Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). CNN architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 131–135. <https://doi.org/10.1109/ICASSP.2017.7952132>
- Hobson, K. A., Rempel, R. S., Greenwood, H., Turnbull, B., & Van Wilgenburg, S. L. (2002). Acoustic surveys of birds using electronic recordings: New potential from an omnidirectional microphone system. *Wildlife Society Bulletin (1973–2006)*, 30(3), 709–720.
- Hui, C. T. J., Au, E., Xiao, S., Hioka, Y., Masuda, H., & Watson, C. I. (2021). Differences in speech intelligibility in noise between native and non-native listeners under ambisonics-based sound reproduction system. *Applied Acoustics*, 184, 108368. <https://doi.org/10.1016/j.apacoust.2021.108368>
- Institute for Computer Music and Sound Technology. (2021). *Ambisonics externals for MaxMSP (3.0.2)* [Max/MSP]. Zurich University of the Arts. <https://www.zhdk.ch/forschung/icst/software-downloads-5379/downloads-ambisonics-externals-for-maxmsp-5381>
- Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61, 101236. <https://doi.org/10.1016/j.ecoinf.2021.101236>
- Kaneko, S., & Gamper, H. (2022). Large-scale simulation of bird localization systems in forests with distributed microphone arrays.

- JASA Express Letters, 2(10), 101201. <https://doi.org/10.1121/10.0014809>
- Krabbe, N. (1997). XC25554 Common Blackbird (*Turdus merula*). <https://xeno-canto.org/25554>
- Le Penru, N. P., Borrelli, A., Heath, B. E., Dunning, J., Picinali, L., Ewers, R. M., & Sethi, S. S. (2024). Towards using virtual acoustics for evaluating spatial ecoacoustic monitoring technologies—Code and data (Version 2) [dataset]. Zenodo. <https://zenodo.org/records/13157933>
- Machado, R. B., Aguiar, L., & Jones, G. (2017). Do acoustic indices reflect the characteristics of bird communities in the savannas of Central Brazil? *Landscape and Urban Planning*, 162, 36–43. <https://doi.org/10.1016/j.landurbplan.2017.01.014>
- Mansour, N., Marschall, M., Westermann, A., May, T., & Dau, T. (2019). Speech intelligibility in a realistic virtual sound environment: 23rd International Congress on Acoustics. *Proceedings of the 23rd International Congress on Acoustics*, 7658–7665.
- Mathworks. (2022). MATLAB deep learning toolbox (14.5) [computer software]. <https://uk.mathworks.com/help/deeplearning/>
- Matusiak, J. (2010). XC44727 house sparrow (*Passer domesticus*). <https://xeno-canto.org/44727>
- Mennill, D. J., Battiston, M., Wilson, D. R., Foote, J. R., & Doucet, S. M. (2012). Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour. *Methods in Ecology and Evolution*, 3(4), 704–712. <https://doi.org/10.1111/j.2041-210X.2012.00209.x>
- Mennill, D. J., & Vehrencamp, S. L. (2008). Context-dependent functions of avian duets revealed by microphone-Array recordings and multispeaker playback. *Current Biology*, 18(17), 1314–1319. <https://doi.org/10.1016/j.cub.2008.07.073>
- Monacchi, D. (2013). Fragments of extinction: Acoustic biodiversity of primary rainforest ecosystems. *Leonardo Music Journal*, 23, 23–25. https://doi.org/10.1162/LMJ_a_00148
- Moreau, S., Daniel, J., & Bertet, S. (2006). 3D sound field recording with higher order ambisonics—Objective measurements and validation of spherical microphone. *Audio Engineering Society Convention 120*. Audio Engineering Society, Paris, France.
- Müller, C. (2006). *Spherical harmonics* (Vol. 17). Springer.
- Oreinos, C., & Buchholz, J. M. (2016). Evaluation of loudspeaker-based virtual sound environments for testing directional hearing aids. *Journal of the American Academy of Audiology*, 27(7), 541–556. <https://doi.org/10.3766/jaaa.15094>
- Payo, D., González, L., Kogan, P., & Eguia, M. (2021). Sentinel: Versatile real-time acoustic autonomous monitoring system for studying natural soundscapes. *Euronoise 2021*. Euronoise, Madeira, Portugal.
- Picinali, L. (2006). Techniques for the extraction of the impulse response of a linear and time-invariant system. *Proceedings of the DMRN Doctoral Research Conference*, Goldsmiths College, University of London, UK, 22–23 July 2006. DMRN Doctoral Researchers Conference, London, UK. <https://dora.dmu.ac.uk/handle/2086/5401>
- Pijanowski, B. C., & Villanueva-Rivera, L. J. (2018). *soundecology: Soundscape ecology* (1.3.3) [computer software]. <https://cran.r-project.org/web/packages/soundecology/>
- Poelstra, J. (2011). XC76918 common wood pigeon (*Columba palumbus*). <https://xeno-canto.org/76918>
- Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6), 456–466.
- Quirós-Guerrero, E., Janeiro, M. J., Lopez-Morales, M., Cresswell, W., & Templeton, C. N. (2017). Riverside wren pairs jointly defend their territories against simulated intruders. *Ethology*, 123(12), 949–956. <https://doi.org/10.1111/eth.12694>
- Schmidt, R. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3), 276–280. <https://doi.org/10.1109/TAP.1986.1143830>
- Seed Studio. (n.d.). *ReSpeaker 6-Mic circular array kit for Raspberry Pi* [Wiki]. Seed Wiki. https://wiki.seedstudio.com/ReSpeaker_6-Mic_Circular_Array_kit_for_Raspberry_Pi/
- Sethi, S. S., Bick, A., Ewers, R. M., Klinck, H., Ramesh, V., Tuanmu, M.-N., & Coomes, D. A. (2023). Limits to the accurate and generalizable use of soundscapes to monitor biodiversity. *Nature Ecology & Evolution*, 7(9), 1373–1378. <https://doi.org/10.1038/s41559-023-02148-z>
- Sethi, S. S., Ewers, R. M., Jones, N. S., Orme, C. D. L., & Picinali, L. (2018). Robust, real-time and autonomous monitoring of ecosystems with an open, low-cost, networked device. *Methods in Ecology and Evolution*, 9(12), 2383–2387. <https://doi.org/10.1111/2041-210X.13089>
- Sethi, S. S., Jones, N. S., Fulcher, B. D., Picinali, L., Clink, D. J., Klinck, H., Orme, C. D. L., Wrege, P. H., & Ewers, R. M. (2020). Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. *Proceedings of the National Academy of Sciences of the United States of America*, 117(29), 17049–17055. <https://doi.org/10.1073/pnas.2004702117>
- Simon, L. S. R., Dillier, N., & Wüthrich, H. (2020). Comparison of 3D audio reproduction methods using hearing devices. *Journal of the Audio Engineering Society*, 68(12), 899–909.
- Simon, L. S. R., Wüthrich, H., & Dillier, N. (2021). *Investigating Higher order spatial impulse response rendering for the evaluation of hearing devices*. DAGA 2021. DAGA, Vienna, Austria. https://pub.dega-akustik.de/DAGA_2021/data/articles/000152.pdf
- Sueur, J., Aubin, T., Simonis, C., Lellouch, L., Aumond, P., de Baudouin, A., Brown, E. C., Corbeau, G., Depaetere, M., Desjonquères, C., Fabianek, F., Gasc, A., Hauptert, S., Kasten, E., Lees, J., Marchal, J., Mikulec, A., Pavoine, S., Pinaud, D., ... Zhivomirov, H. (2022). *seewave: Sound analysis and synthesis* (2.2.0) [computer software]. <https://cran.r-project.org/web/packages/seewave/index.html>
- Sumitani, S., Suzuki, R., Chiba, N., Matsubayashi, S., Arita, T., Nakadai, K., & Okuno, H. G. (2019). An integrated framework for field recording, localization, classification and annotation of birdsongs using robot audition techniques—Harkbird 2.0. *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8246–8250. <https://doi.org/10.1109/ICASSP.2019.8683743>
- Suzuki, R., Matsubayashi, S., Hedley, R. W., Nakadai, K., & Okuno, H. G. (2017). HARKBird: Exploring acoustic interactions in bird communities using a microphone array. *Journal of Robotics and Mechatronics*, 29(1), 213–223. <https://doi.org/10.20965/jrm.2017.p0213>
- Wijers, M., Loveridge, A., Macdonald, D. W., & Markham, A. (2019). CARACAL: A versatile passive acoustic monitoring tool for wildlife research and conservation. *Bioacoustics*, 30(1), 41–57. <https://doi.org/10.1080/09524622.2019.1685408>
- Wood, C. M., & Kahl, S. (2024). Guidelines for appropriate use of BirdNET scores and other detector outputs. *Journal of Ornithology*, 165, 777–782. <https://doi.org/10.1007/s10336-024-02144-5>
- Znidersic, E., & Watson, D. M. (2022). Acoustic restoration: Using soundscapes to benchmark and fast-track recovery of ecological communities. *Ecology Letters*, 25(7), 1597–1603. <https://doi.org/10.1111/ele.14015>
- Zotter, F., & Frank, M. (2019). *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality* (Vol. 19). Springer International Publishing. <https://doi.org/10.1007/978-3-030-17207-7>
- ZYLIA. (2022a). *ZYLIA ambisonics converter* (1.6.1) [computer software]. Zylia Sp. z o.o. <https://www.zylia.co/zylia-ambisonics-converter-download.html>
- ZYLIA. (2022b). *ZYLIA studio* (2.1.2) [computer software]. Zylia Sp. z o.o. <https://www.zylia.co/zylia-studio.html>
- ZYLIA. (n.d.). *ZYLIA ZM-1 microphone*. ZYLIA—3D AUDIO RECORDING & POST-PROCESSING. <https://www.zylia.co/zylia-zm-1-microphone.html>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Table S1: Breakdown of Zylia ZM-1 and MAARU recordings used for each Objective of this study, as well as the ecoacoustic metrics extracted from MAARU's recordings.

Table S2: BirdNET's Precision (Prec.), recall (Rec.), number of predicted species (Spec.) and number of predictions (Pred.) across all recording sites and conditions (field and virtual sound environment (VSE) at various orientations) towards Objectives 1 (efficacy of soundscape reproduction) and 2 (Impact of Device Orientation) for the following minimum confidence thresholds (leftmost column): 0, 0.25, 0.5, 0.75 and 0.9.

Figure S1: Plots of scaled differences in indices' values between VSE and field recordings as a percentage of each index's field recording range, for all sites.

Figure S2: Absolute error in HARKBird's estimates of the azimuth of avian calls embedded in VSE-replicated soundscapes against calls'

true elevation values, for four methods of embedding: ambisonic encoding, playback from individual loudspeakers (recordings low-passed with a 4 kHz cutoff and 12 dB roll-off prior to passing to HARKBird), individual loudspeaker playback with no reverb, and soloed playback from individual loudspeakers.

How to cite this article: Le Penru, N. P., Heath, B. E., Dunning, J., Picinali, L., Ewers, R. M., & Sethi, S. S. (2025). Towards using virtual acoustics for evaluating spatial ecoacoustic monitoring technologies. *Methods in Ecology and Evolution*, 16, 108–125.

<https://doi.org/10.1111/2041-210X.14405>