Robust detection of overlapping bioacoustic sound events

Louis Mahon[*]
University of Edinburgh
Edinburgh, UK
louis.mahon@ed.ac.uk

Benjamin Hoffman[*]
Earth Species Project
Berkeley, CA, USA
benjamin@earthspecies.org

Logan James
McGill University
Montreal, Quebec, Canada

Maddie Cusimano
Earth Species Project
Berkeley, CA, USA

Masato Hagiwara
Earth Species Project
Berkeley, CA, USA

Sarah C Woolley
McGill University
Montreal, Quebec, Canada

Felix Effenberger
Earth Species Project
Berkeley, CA, USA

Sara Keen
Earth Species Project
Berkeley, CA, USA

Jen-Yu Liu
Earth Species Project
Berkeley, CA, USA

Olivier Pietquin
Earth Species Project
Berkeley, CA, USA

[*] Equal contribution

Abstract

We propose a method for accurately detecting bioacoustic sound events that is robust to overlapping events, a common issue in domains such as ethology, ecology and conservation. While standard methods employ a frame-based, multi-label approach, we introduce an onset-based detection method which we name `Voxaboxen`. For each time window, `Voxaboxen` predicts whether it contains the start of a vocalization and how long the vocalization is. It also does the same in reverse, predicting whether each window contains the end of a vocalization, and how long ago it started, and fuses the two sets of bounding boxes with a graph-matching algorithm. We also release a new dataset of temporally-strong labels of zebra finch vocalizations designed to have high overlap. Experiments on eight datasets, including our new dataset, show `Voxaboxen` outperforms natural baselines and existing methods, and is robust to vocalization overlap.

1 Introduction

Detecting animal sounds is a foundational task in bioacoustics research (1). Temporally fine-grained detection, i.e. identifying the time boundaries of each acoustic unit, is critical for answering questions arising in animal communication. When multiple individuals from a single species co-occur, the sounds they produce can overlap with each other, often with important functional consequences, e.g. in bats (2), zebra finches (3), frogs (4), and elephants (5). Therefore, to understand these and other animal communication systems, it is vital that we can accurately detect overlapping animal sounds.

Motivated by this, we desire a sound event detection (SED) method that can predict the onset time, offset time, and class label (e.g., species label) for overlapping sound events. Commonly, SED methods adopt a frame-based approach: for each time frame, for each class, predicting whether a

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

sound of that class occurs in that frame (6; 7; 8; 9), and merging consecutive frames with the same class into a single event. This does not accommodate overlaps from the same class. To address this limitation, we propose a method we name `Voxaboxen`, For each frame, `Voxaboxen` makes a binary prediction as to whether it contains an event onset, plus a regression prediction for how long that event will last, and a class prediction (e.g. species label). This design choice means the duration of one predicted event can extend past the onset of a second event, thus allowing the model to predict overlapping vocalizations without them being merged.

To investigate how well `Voxaboxen` deals with overlapping vocalizations, we introduce a new dataset of eight female zebra finches (ZFs) spontaneously interacting in a laboratory environment, annotated with onset and offset of each vocalization and featuring a high overlap. We find that `Voxaboxen` consistently outperforms alternatives on our new dataset and seven previously-published bioacoustics datasets, even with high levels of overlap. We open source the code for our model and dataset. Our contributions are: (1) `Voxaboxen`, an SED model leveraging pretrained audio encoders to predict overlapping vocalizations; (2) the Overlapping Zebra Finch (OZF) dataset; (3) experimental evaluation on eight diverse datasets, showing SotA performance for `Voxaboxen`.

2 Related Work

In bioacoustics applications, SED has typically been framed as a multi-label classification problem (1), with temporal resolution ranging from tens of milliseconds (10; 8), to multiple seconds (11; 12). Recent post-processing techniques decouple event durations and detections (9; 13); but still use frame-based predictions and cannot handle within-class overlaps. Other approaches include matrix factorization algorithms (14) or probabilistic models (15). Visual object detection methods such as Faster-RCNN (16) can handle overlaps and have been applied to bioacoustic SED (17). CornerNet (18) is an object detection method that, similar to `Voxaboxen`, matches predicted boundaries into a single event, but uses feature similarity, which can be unreliable for stereotyped animal vocalizations. Our approach accounts for this by matching based on intersection over union (IoU) instead.

Source separation methods decompose mixed audio into individual sources and have shown promise for bioacoustic classification ((19). In our context, a source separation model could theoretically separate vocalizations from multiple individuals into different audio tracks, thus reducing the complexity of the audio passed to a downstream detection model. We investigate this approach as an alternative to `Voxaboxen`. A related task is speaker diarization, which segments multi-speaker recordings and assigns each segment to a speaker. This typically requires assumptions about maximum speaker count and re-identification across segments (20); in contrast, we assume no maximum number of speakers, and do not expect to re-identify individuals within a recording.

3 Method

3.1 Bounding Box Regression

Our method, which is architecture-agnostic, uses a frame-based audio encoder $\phi\colon \mathbb{R}^T \to \mathbb{R}^{T' \times F}$ to produce a sequence of latent vectors. Here T is the original number of samples, T' is the final number of frames, and F is the feature dimension. A final linear layer $h\colon \mathbb{R}^F \to \mathbb{R}^{2+C}$ makes three types of predictions, for each time frame: a prediction of the probability that an event starts in that frame, a prediction of the duration of the event (should it start in that frame), and a prediction of a class label (logits across C classes). Using gradient descent, we minimize the loss function $L = L_{det} + \lambda L_{reg} + \rho L_{cls}$, $\lambda, \rho \geq 0$, which includes a detection term L_{det}, a regression term L_{reg}, and a classification term L_{cls}. The detection term is inspired by the penalty-reduced focal loss in (18):

$$L_{det} = -\frac{1}{T}\sum_{t=1}^{T} \begin{cases} (1-\hat{p}_t)^{\alpha} \log \hat{p}_t & p_t = 1 \\ (1-p_t)^{\beta} \hat{p}_t^{\alpha} \log(1-\hat{p}_t) & p_t < 1. \end{cases} \tag{1}$$

Here, T is the duration in frames, α, β are hyperparameters, $\hat{p}_t$ is the predicted detection probability at time t, and p_t is obtained by applying Gaussian smoothing to event onsets and taking the maximum value at each frame, across all events (following (18)):

$$p_t = \max_{x \in \text{Events}} \exp\left(-\frac{(t - \text{Onset}(x))^2}{\text{Dur}(x)^2/s}\right). \tag{2}$$

In (2), Events is the set of events in an audio clip, $\mathrm{Onset}(x)$ and $\mathrm{Dur}(x)$ denote the onset time and duration of x, and s is a hyperparameter. L_{reg} (L1) and L_{cls} (cross-entropy) are applied only at event onset frames. At inference, detection probability peaks above threshold become boxes, with duration and class prediction determined by peak predictions. We apply soft non-maximal suppression (21) to remove duplicates.

3.2 Bidirectional Predictions

To improve bounding box accuracy, we make a second set of *backward* predictions which are the mirror image of the first (*forward*) set. The backward predictions are a binary prediction for each frame as to whether it contains an offset, plus a regression for how long the event lasted. We fuse these by casting the problem as maximal bipartite graph matching, where forward and backward boxes are linked by an edge if their IoU exceeds a threshold. The Hopcroft-Karp-Karzonov algorithm (22) computes the maximal matching sub-graph, and edge-linked box pairs are fused. The onset of the fused box is defined to be the midpoint of the onset of the forward box, with the offset minus duration of the backward box (and similarly for the offset of the fused box).

4 Overlapping Zebra Finch Dataset (OZF)

We recorded 65 minutes (in 60-second files) of 8 adult (> 1 year) female ZFs in a sound attenuating chamber using two omnidirectional microphones. All procedures were approved by the McGill University Animal Care and Use Committee. Female ZFs make short, discrete vocalizations of about 100ms, with most energy between 0.5 and 8 kHz. Three annotators marked onset and offset times of each vocalization using Raven Pro, covering 25 minutes each, with 5 minutes reviewed by all annotators (mean pairwise inter-annotator agreement: 93.5%F1@0.5IoU overall and 78.1% for the subset overlaps). The dataset contains 8504 vocalizations, with 1449 (17.04%) overlapping at least one other (1463 total overlaps). Vocalizations per 60s file range from 19 to 245, with 0-73 overlapping. Vocalization durations range from 3-350 ms (mean 109 ms). The 17.04% overlap rate is consistent with prior work showing evidence for turn-taking in female ZFs (23).

5 Experimental Evaluation

Implementation Details We first extract features from the raw audio using a backbone encoder, and then make the predictions described in Section 3 from the extracted features. The encoder converts input audio (mono, 16 kHz) to a frame-based representation, which is a sequence of latent vectors produced at 50 Hz (window size 10s, hop size 5s). For the main experiments, we use BEATs (24) as a backbone encoder. BEATs is a 12-layer encoder-only transformer (hidden size 768, 8 attention heads), pretrained on Audioset (25). We explore alternative backbones in Section 5.2. The detection, regression and classification predictions are made via linear layers. The loss function hyperparameters were fixed at $\alpha = 2$, $\beta = 4$, and $s = 6$ following (18). These values showed stable training behavior in preliminary experiments and were found to transfer well from the visual to temporal domain. During training and inference, audio is divided into 10-second windows, with 5s step size between windows. Training lasts for 50 epochs, with the encoder frozen for the first 3 epochs. This protocol is identical for all encoder backbones. We use Adam with ams-grad, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a cosine annealing scheduler. For all models, we select a learning rate from {1e-4, 3e-5, 1e-5}, based on mean average precision @0.5IoU on the validation set. We apply soft non-maximal suppression (21) with $\sigma = 0.5$. Because BirdAVES (based on HuBERT) has a known batch artifact in which padding affects the embeddings, we used batch_size=1 to ensure a fair comparison. All experiments were conducted on a single 40GB A100 GPU, with training time less than one day per dataset.

Datasets In addition to the OZF dataset, we evaluated `Voxaboxen` using seven existing datasets (Table 1), selected for their taxonomic diversity: amphibians (AnuraSet), insects (Katydid), birds (BirdVox-10h, Hawaiian Birds, Powdermill), and mammals (Humpback, Meerkat). Dataset preprocessing is described on the project github. For Katy, BV10 and OZF, the events of interest were brief and, for Katy and BV10, often above the 8kHz Nyquist frequency assumed by several of the models we evaluated. For all models, we used a half-time version of BV10 and OZF, and a sixth-time version of Katy. This effectively increased the output frame rate to 100 Hz for BV10 and OZF, and 300 Hz for Katy, which was shown to dramatically improve performance in initial experiments.

Evaluation As a metric, we first match predicted events to true events as in (26), only counting matches that exceed a certain IoU threshold. Then, we compute mean average precision (mAP) using 1001 equally-sized intervals. We report results for an IoU threshold of 0.5.

Comparison Models We compare `Voxaboxen` to several frame-based methods. Three of these pretrained transformers with linear classification heads: Frame-ATST (7) (25 Hz output frame rate, pretrained on AudioSet), BEATs (24) (50 Hz, AudioSet) and BirdAVES (27)[1] (50 Hz, animal sound datasets). Outputs are median filtered, with kernel size (ks) 1/3/7/11, selected based on mAP @0.5IoU on the validation set. We also compare to a convolutional-recurrent neural network (CRNN) (6; 10; 8) that operates on log-mel spectrograms and uses a 2d conv layer (ks=7, hidden size 64), mean-pooling in the frequency dimension (ks=2), two 2d residual blocks (ks=3), mean pooling in both directions (ks=2), and a bi-LSTM, with hidden size 1024. Lastly, we compare to two computer vision object detection models, Faster-RCNN (16) (X-101 model checkpoint pretrained on MS COCO)[2] and SEDT (28), an encoder-decoder transformer, adapted to detect 1d events from a spectrogram[3].

5.1 Main Results

Metric	Method	AnSet	BV10	HawB	HbW	Katy	MT	Pow	OZF
mAP@0.5IoU	CRNN	9.89	35.59	22.72	21.03	17.24	82.97	35.45	71.80
	Faster-RCNN	8.06	55.49	7.39	21.66	25.93	84.22	14.08	90.20
	SEDT	0.18	3.79	2.79	3.95	2.30	18.58	2.71	2.26
	Frame-ATST	14.87	40.62	32.19	33.62	17.88	87.58	45.42	73.48
	BEATs	15.71	48.01	35.37	37.13	20.12	86.08	50.32	77.94
	BirdAVES	14.21	42.09	32.67	26.54	19.11	86.11	43.52	78.33
	`Voxaboxen`	**27.08**	**77.32**	**53.87**	**59.92**	**36.04**	**90.96**	**56.77**	**97.92**

Table 1: Mean average precision scores at 0.5 IoU. Best results in **bold**. With one exception, `Voxaboxen` outperforms existing methods, sometimes by far, such as on BV10, HawB, and OZF.

Table 1 shows that `Voxaboxen` outperforms other methods in almost all cases, often by large margins (e.g. 10+ points on BV10, HawB, HbW, and Katy). Faster-RCNN performs well on single-class datasets (OZF, MT) but struggles with multi-class tasks. Among frame-level SED models, BEATs is the strongest, consistent with our findings for the backbone choice (Table 2). SEDT, which was pretrained on mostly ambient city noises, transfers poorly to animal vocalizations.

Metric	Method	AnSet	BV10	HawB	HbW	Katy	MT	Pow	OZF
mAP@0.5IoU	`Voxaboxen`	**27.08**	**77.32**	**53.87**	**59.92**	**36.04**	**90.96**	**56.77**	**97.92**
	with BirdAVES encoder	22.86	46.33	49.22	48.04	26.59	88.78	50.21	96.36
	no fwd-bck matching	25.04	75.97	52.10	56.99	34.97	89.39	50.02	95.77

Table 2: Ablations on encoder choice and forward-backward matching method. The main model uses the BEATs encoder. Performance drops validate BEATs encoder and bidirectional matching method.

5.2 Ablation Studies

Table 2 shows ablations on encoder backbone and forward-backward matching. BirdAVES underperformed BEATs despite being designed specifically for animal sounds, likely due differences in pre-training data volume and training regimes. Removing forward-backward matching (i.e. using forward only) consistently reduces mAP by 1-2 points (larger on HbW, Pow).

Limitations: Performance may vary between species or recording conditions. Short or high-frequency vocalizations require time-stretching. Results shown here are from single runs and therefore do not have error bars.

Data Availability

Data used in this study is available at `https://zenodo.org/records/15507508`. The code used in this study is available at `https://github.com/earthspecies/voxaboxen`.

[1] `https://github.com/earthspecies/aves`

[2] `https://github.com/facebookresearch/detectron2`

[3] `https://github.com/Anaesthesiaye/sound_event_detection_transformer`

References

[1] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, p. e13152, 2022.

[2] E. Gillam and M. B. Fenton, *Bat Bioacoustics*. New York, NY: Springer, 2016, ch. Roles of Acoustic Social Communication in the Lives of Bats, pp. 117–139.

[3] J. E. Elie, H. A. Soula, N. Mathevon, and C. Vignal, "Dynamics of communal vocalizations in a social songbird, the zebra finch (Taeniopygia guttata)," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. 4037–4046, 2011.

[4] S. Clulow, M. Mahony, L. Elliott, S. Humfeld, and H. C. Gerhardt, "Near-synchronous calling in the hip-pocket frog Assa darlingtoni," *Bioacoustics*, vol. 26, no. 3, pp. 249–258, 2017.

[5] J. Soltis, K. Leong, and A. Savage, "African elephant vocal communication I: antiphonal calling behaviour among affiliated females," *Animal Behaviour*, vol. 70, no. 3, pp. 579–587, 2005.

[6] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[7] N. Shao, X. Li, and X. Li, "Fine-tune the pretrained ATST model for sound event detection," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 911–915.

[8] K. Martin, O. Adam, N. Obin, and V. Dufour, "Rookognise: Acoustic detection and identification of individual rooks in field recordings using multi-task neural networks," *Ecological Informatics*, vol. 72, p. 101818, 2022.

[9] J. Ebbers, F. G. Germain, G. Wichern, and J. L. Roux, "Sound event bounding boxes," in *Interspeech 2024 Proceedings*, 2024, pp. 562–566.

[10] Y. Cohen, D. A. Nicholson, A. Sanchioni, E. K. Mallaber, V. Skidanova, and T. J. Gardner, "Automated annotation of birdsong with a neural network that segments spectrograms," *Elife*, vol. 11, p. e63853, 2022.

[11] B. Ghani, T. Denton, S. Kahl, and H. Klinck, "Global birdsong embeddings enable superior transfer learning for bioacoustic classification," *Scientific Reports*, vol. 13, no. 1, p. 22876, 2023.

[12] D. Robinson, M. Miron, M. Hagiwara, and O. Pietquin, "Naturelm-audio: An audio-language foundation model for bioacoustics," in *Proceedings of the International Conference on Learning Representations (ICML)*, 2025.

[13] T. Yoshinaga, K. Tanaka, Y. Bando, K. Imoto, and S. Morishima, "Onset-and-offset-aware sound event detection via differentiable frame-to-event mapping," *IEEE Signal Processing Letters*, pp. 186–190, 2024.

[14] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*. Springer, 2012, pp. 341–371.

[15] D. Stowell and D. Clayton, "Acoustic event detection for multiple overlapping similar sources," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[17] S. Zsebők, M. F. Nagy-Egri, G. G. Barnaföldi, M. Laczi, G. Nagy, É. Vaskuti, and L. Z. Garamszegi, "Automatic bird song and syllable segmentation with an open-source deep-learning object detection method–a case study in the collared flycatcher (Ficedula albicollis)," *Ornis Hungarica*, vol. 27, no. 2, pp. 59–66, 2019.

[18] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[19] T. Denton, S. Wisdom, and J. R. Hershey, "Improving bird classification with unsupervised sound separation," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 636–640.

[20] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Interspeech 2021 Proceedings*, 2021, pp. 3111–5.

[21] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–improving object detection with one line of code," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5561–5569.

[22] J. E. Hopcroft and R. M. Karp, "An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs," *SIAM Journal on computing*, vol. 2, no. 4, pp. 225–231, 1973.

[23] J. I. Benichov, S. E. Benezra, D. Vallentin, E. Globerson, M. A. Long, and O. Tchernichovski, "The forebrain song system mediates predictive call timing in female and male zebra finches," *Current Biology*, vol. 26, no. 3, pp. 309–318, 2016.

[24] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023, pp. 5178–5193.

[25] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[26] I. Nolasco, S. Singh, V. Morfi, V. Lostanlen, A. Strandburg-Peshkin, E. Vidaña-Vila, L. Gill, H. Pamuła, H. Whitehead, I. Kiskin *et al.*, "Learning to detect an animal sound from five examples," *Ecological Informatics*, vol. 77, p. 102258, 2023.

[27] M. Hagiwara, "AVES: Animal vocalization encoder based on self-supervision," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[28] Z. Ye, X. Wang, H. Liu, Y. Qian *et al.*, "Sound event detection transformer: An event-based end-to-end model for sound event detection," *arXiv preprint arXiv:2110.02011*, 2021.

[29] J. S. Cañas, M. P. Toro-Gómez, L. S. M. Sugai, H. D. Benítez Restrepo, J. Rudas, B. Posso Bautista, L. F. Toledo, S. Dena, A. H. R. Domingos, F. L. de Souza *et al.*, "A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring," *Scientific Data*, vol. 10, no. 1, p. 771, 2023.

[30] A. Navine, S. Kahl, A. Tanimoto-Johnson, H. Klinck, and P. Hart, "A collection of fully-annotated soundscape recordings from the island of Hawai'i," 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7078499

[31] A. N. Allen, M. Harvey, L. Harrell, A. Jansen, K. P. Merkens, C. C. Wall, J. Cattiau, and E. M. Oleson, "A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset," *Frontiers in Marine Science*, vol. 8, p. 607321, 2021.

[32] S. Madhusudhana, H. Klinck, and L. B. Symes, "Extensive data engineering to the rescue: building a multi-species katydid detector from unbalanced, atypical training datasets," *Philosophical Transactions of the Royal Society B*, vol. 379, no. 1904, p. 20230444, 2024.

[33] L. M. Chronister, T. A. Rhinehart, A. Place, and J. Kitzes, "An annotated set of audio recordings of eastern north american birds containing frequency, time, and species information," *Ecology*, vol. 102, no. 6, p. e03329, 2021.

A Technical Appendices and Supplementary Material

B OZF Further Details

B.1 Segmentwise Statistics of the Real-World Portion

Figure 1 reports the distribution of vocalizations and overlaps across each 60-second audio file. Figure 2 reports these per 10-second segment.

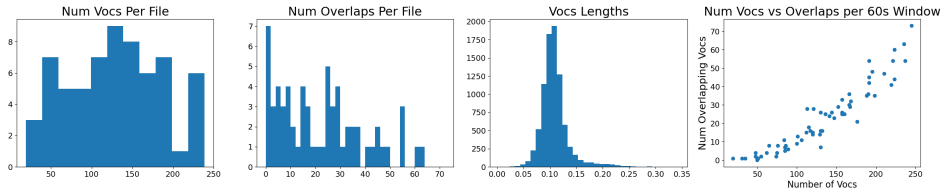

Figure 1: Summary of the live portion of the dataset we release. Left: number of vocalizations per 60 s file. Second: number of overlapping pairs per 60 s file. Third: distribution of the lengths of vocalizations. Right: number of vocalizations per 60 s file vs number overlapping pairs.

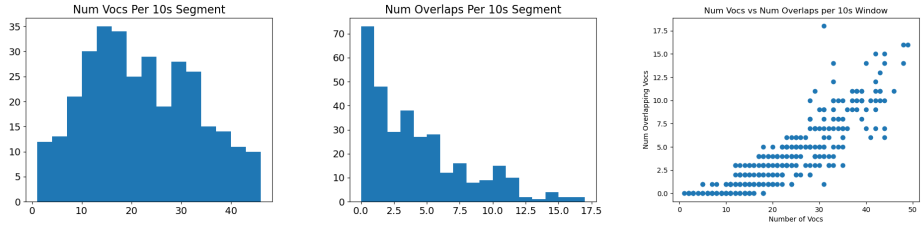

Figure 2: Statistics from the real-world portion of OZF, across 10 s segments. Top: the distribution of the number of vocalizations entirely contained within each 10 s segment, across all audio files. Middle: the distribution of the number of pairwise overlaps of these vocalizations. Bottom: the number of vocalizations vs the number of pairwise overlaps across each 10 s segment.

B.2 Call Database for the Synthetic Portion

To construct the synthetic datasets, we created a database of female ZF calls. To construct this, female ZFs were recorded using the same setup as with the live portion of OZF. Calls were detected using an initial version of `Voxaboxen`, and cropped versions of the calls were saved. We then performed a denoising procedure: First, using BirdMixit (19) each of these cropped calls was separated into four stems. Then, our trained model was again run across each of these four stems, and we retained a stem when `Voxaboxen` both 1) detected a call and 2) the model detection confidence was higher on this stem than the other three stems. Finally, we observed that even after these steps there remained stems that contained no zebra finch vocalizations. To remove these, we performed a quality-filtering step: for each stem, we predicted the species of the call using BirdNET (Kahl et a., 2021). We retained only stems where BirdNET predicted a species with British English common name containing the word "Finch". The stems that passed this quality filter became the database of denoised female zebra finch calls.

C Expected Number of Overlaps from Independent Memoryless Sources

Given a window of time and some set V of vocalizations whose onsets occur during this window, we are interested in the expected value of the number of pairs that overlap, assuming that the probability density function for the point of each onset is uniform and independent. Let L be the length of the time window, 60s in our case, so that the pdf equals $\frac{1}{L}$. Then, for any two vocalizations with onsets

as v_1 and v_2 of respective durations l_1 and l_2, the probability of overlap is

$$\frac{l_1 + l_2}{L}\,, \tag{3}$$

because they will overlap if and only if v_1 falls in the interval $(v_2 - l_1, v_2 + l_1)$, which is of length $l_1 + l_2$.

The expected number of overlaps, $\mathbb{E}[X]$, is the sum, across all ordered pairs of vocalizations, of the the indicator random variable for the event that they overlap, which equals the probability as given in (3). Let $|V| = n$, and let l_i be the duration of the i vocalization, then

$$\begin{aligned}
\mathbb{E}[X] =& \sum_{i=1}^{n}\sum_{j=1}^{i-1}\frac{l_i + l_j}{L} = \frac{1}{L}\left(\sum_{i=1}^{n}\sum_{j=1}^{i-1} l_i + l_j\right) \\
=& \frac{1}{L}\left(\sum_{i=1}^{n}(i-1)l_i + \sum_{i=1}^{n}\sum_{j=1}^{i} l_j\right) \\
=& \frac{1}{L}\left(\sum_{i=1}^{n}(i-1)l_i + \sum_{j=1}^{n-1}\sum_{i=j}^{n} l_j\right) \\
=& \frac{1}{L}\left(\sum_{i=1}^{n}(i-1)l_i + \sum_{j=1}^{n-1}(n-j)l_j\right) \\
=& \frac{1}{L}\left(\sum_{i=1}^{n}(i-1)l_i + \sum_{i=1}^{n-1}(n-i)l_i\right) \\
=& \frac{1}{L}\left(\sum_{i=1}^{n}(i-1)l_i + \sum_{i=1}^{n}(n-i)l_i\right) \\
=& \frac{1}{L}\left(\sum_{i=1}^{n}(i-1)l_i + (n-i)l_i\right) \\
=& \frac{1}{L}\left(\sum_{i=1}^{n}(n-1)l_i\right) \\
=& (n-1)\frac{\sum_{i=1}^{n} l_i}{L}\,.
\end{aligned}$$

Thus, we see that the expected number of overlaps for uniform independent vocalizations is the product of two factors. The first is the vocalization density

$$d = \frac{\sum_{i=1}^{n} l_i}{L}\,, \tag{4}$$

which is the ratio between the length of all vocalizations played back to back and the duration of the time window in which they occur, or equivalently, the expected number of vocalizations occurring at any one point. The second factor is the number of vocalizations (minus 1).

$$\mathbb{E}[X] = d(n-1)\,, \tag{5}$$

where d is as in (4).

In the case of our released dataset, we must also account for the fact that there is a finite number of birds (eight), and overlaps can only occur between vocalizations from two different birds. For two given vocalizations, let S be the event that they come from different birds, which has probability

$$P(S) = 1 - \frac{\sum_{j=1}^{B}(\sum_{i=1}^{n}\mathbb{1}(b_i = j))^2}{n^2}\,, \tag{6}$$

where B is the number of birds and b_i is the bird that produced the ith vocalization. Let $Z_j = \sum_{i=1}^{n} \mathbb{1}(b_i = j)$ be the random variable counting the number of times the jth bird vocalises in a given time window (60s for our dataset). Assuming this distribution is the same for all birds, we can drop the subscript and just write Z. The expression in (6) is linear apart from the square on Z, so we have

$$P(S) = 1 - \frac{\sum_{j=1}^{B} \mathbb{E}[Z^2]}{n^2} = 1 - \frac{B\mathbb{E}[Z^2]}{n^2}$$
$$= 1 - \frac{B(\mathbb{E}[Z]^2 + \text{Var}(Z))}{n^2}\,.$$

If we model the vocalizations of each individual bird as a Poisson distribution, then we have

$$\mathbb{E}[Z] = \text{Var}(Z) = \lambda = \frac{n}{B}\,,$$

giving

$$P(S) = 1 - \frac{B((\frac{n}{B})^2 + \frac{n}{B})}{n^2} = 1 - \frac{\frac{n^2}{B} + n}{n^2} = 1 - (\frac{1}{B} + \frac{1}{n})\,.$$

The value from (5) is then the probability of overlap between two vocalizations given they come from separate birds: $\mathbb{E}[X|S] = d(n-1)$, and the total probability of overlap is then

$$\mathbb{E}[X] = \mathbb{E}[X|S]P(S) = d(n-1)(1 - \frac{1}{B} - \frac{1}{n})\,. \tag{7}$$

C.1 Difference Between Expected and Observed Overlaps

Table 3 shows the observed number of pairwise overlaps per file, compared with the expected number from (7). The former is consistently lower than the latter. Indeed, looking at the 'difference' column, we see it has mean 9.73, and standard deviation 9.05. We can model this difference as a normal distribution by the central limit theorem, as it is the sum of 8 independent samples from the distribution of a single bird. With 65 files, the estimated population standard deviation of this normal distribution is

$$\frac{9.05}{\sqrt{65-1}} = \frac{9.05}{8} = 1.13\,,$$

so the t-value is $\frac{9.73}{1.13} = 8.61$. This is highly significant, as the significance threshold for 64 degrees of freedom is 3.23 at 99.9% confidence.

C.2 Evaluation Dataset Preprocessing

AnuraSet We used the portion of the frog call dataset presented in (29) that includes strong temporal annotations (onset, offset, and species label). We randomly assigned files into train, validation, and test sets with ratios $60\%/20\%/20\%$. For our purpose, we retained only annotations corresponding to the ten most commonly occurring species in the dataset.

BirdVox-10h We used the version of the BirdVox dataset presented in (26). We divided each recording into three segments: the first 60% was assigned to the train set, the next 20% was assigned to the validation set, and the final 20% was assigned to the test set. For our purpose, we merged all annotations (species labels for multiple passerine species) into a single class (passerine vocalization). This was done to avoid having many classes with few example vocalizations.

Hawaiian Birds We used the dataset of Hawaiian soundscapes presented in (30). We randomly assigned files into train, validation, and test sets with ratios $60\%/20\%/20\%$. For our purpose, we retained only annotations corresponding to the nine most commonly occurring bird species in the dataset.

Humpback We used the "initial" audit portion of the dataset of humpback whale vocalizations presented in (31), retaining only the 75-second clips containing at least one annotation. We randomly assigned these clips into train, validation, and test sets with ratios $60\%/20\%/20\%$. Finally, we retained only annotations corresponding to humpback whales, and discarded other annotations (e.g. ship noise).

file	n	d	B	expected overlaps	observed overlaps	difference
0	106	0.19	8	16.94	11	5.94
1	117	0.22	8	21.80	16	5.80
2	157	0.29	8	39.17	25	14.17
3	191	0.36	8	59.25	42	17.25
4	195	0.36	8	60.76	48	12.76
5	221	0.38	8	73.36	54	19.36
6	51	0.08	8	3.59	1	2.59
7	160	0.30	8	41.91	25	16.91
8	223	0.37	8	71.97	44	27.97
9	19	0.03	8	0.48	1	-0.52
10	48	0.09	8	3.80	2	1.80
11	31	0.06	8	1.40	1	0.40
12	50	0.08	8	3.43	0	3.43
13	147	0.28	8	34.87	23	11.87
14	210	0.39	8	71.42	47	24.42
15	191	0.36	8	59.05	45	14.05
16	219	0.40	8	75.20	41	34.20
17	237	0.41	8	85.24	54	31.24
18	235	0.44	8	90.18	63	27.18
19	51	0.08	8	3.60	1	2.60
20	50	0.09	8	3.57	1	2.57
21	85	0.15	8	11.14	7	4.14
22	136	0.24	8	28.27	25	3.27
23	141	0.23	8	28.35	24	4.35
24	74	0.13	8	8.32	2	6.32
25	166	0.33	8	47.87	30	17.87
26	65	0.13	8	6.90	8	-1.10
27	223	0.39	8	74.50	60	14.50
28	168	0.31	8	45.60	32	13.60
29	158	0.29	8	39.44	25	14.44
30	120	0.20	8	20.36	15	5.36
31	84	0.16	8	11.16	11	0.16
32	245	0.46	8	97.64	73	24.64
33	191	0.37	8	60.43	54	6.43
34	75	0.11	8	7.21	4	3.21
35	130	0.26	8	28.77	7	21.77
36	130	0.23	8	25.50	26	-0.50
37	86	0.14	8	10.15	8	2.15
38	152	0.27	8	35.18	29	6.18
39	55	0.10	8	4.40	2	2.40
40	176	0.29	8	44.43	21	23.43
41	115	0.18	8	18.08	18	0.08
42	157	0.28	8	37.95	33	4.95
43	35	0.04	8	1.29	1	0.29
44	85	0.16	8	11.32	5	6.32
45	89	0.16	8	12.18	6	6.18
46	198	0.35	8	60.38	35	25.38
47	113	0.25	8	24.49	28	-3.51
48	112	0.22	8	20.79	15	5.79
49	101	0.18	8	15.48	13	2.48
50	157	0.29	8	38.71	26	12.71
51	144	0.26	8	31.71	26	5.71
52	62	0.13	8	6.76	4	2.76
53	167	0.32	8	46.38	29	17.38
54	120	0.23	8	23.86	14	9.86
55	76	0.15	8	9.99	8	1.99
56	190	0.33	8	53.87	36	17.87
57	130	0.20	8	22.82	16	6.82
58	166	0.36	8	51.14	36	15.14
59	121	0.25	8	26.36	28	-1.64
60	132	0.23	8	25.98	16	9.98
61	48	0.08	8	3.26	4	-0.74
62	100	0.19	8	16.00	9	7.00
63	129	0.23	8	25.27	14	11.27
64	188	0.34	8	55.06	35	20.06

Table 3: Comparison of the expected number of overlaps by equation 7 and the observed number of overlaps, by 60s file.

Katydid We used the dataset of katydid calls presented in (32). We randomly assign files into train, validation, and test sets with ratios $60\%/20\%/20\%$. For our purpose, we merged all annotations (species labels) into a single class (katydid call). This was done to avoid having many classes with few example calls.

Meerkat We used the dataset of on-body Meerkat recordings presented in (26) (abbreviated as MT in *loc. cit.*). We divided each recording into three segments: the first 60% was assigned to the train set, the next 20% was assigned to the validation set, and the final 20% was assigned to the test set. For our purpose, we merged all annotations (vocalization type labels) into a single class (meerkat vocalization). This was done to avoid having many classes with few example vocalizations.

Powdermill We used the dataset of Northeastern United States soundscapes presented in (33). We randomly assigned files into train, validation, and test sets with ratios $60\%/20\%/20\%$. For our purpose, we retained only annotations corresponding to the six most commonly occurring bird species in the dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect our paper's contributions and are validated in sections 3-5. The claims include introducing Voxaboxen for detecting overlapping bioacoustic events, releasing the OZF dataset, and demonstrating high performance across eight datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations at the end of Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This is an empirical methods paper without formal theoretical results that would require proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide implementation details in Section 5, including architecture specifications, hyperparameters, training protocol, and steps for dataset preprocessing. Code and data availability are stated at the end of Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide links to open access data and code repositories in the "Data Availability" section at the end of Section 5.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 5 we provide specifications for model training (number of epochs, learning rates tested, optimizer parameters, batch size, and window sizes) and steps for hyperparameter selection.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report single-run results without error bars. Given the consistently superior performance of Voxaboxen relative to baselines we believe these results demonstrate the robustness of our method. We note this limitation at the end of Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the compute resources needed to reproduce experiments the "Implementation Details" paragraph in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This study involves data collection from animal subject; in Section 4 we state that all data collection procedures received ethical approval from the McGill University Animal Care and Use Committee.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: In the introduction we discuss the impacts of this work for ecology and conservation and note that our work can be particularly helpful for applications such as species recognition and population estimation. We expect the work has minimal risk of misuse and therefore do not explicitly discuss potential negative societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA] .

Justification: This work presents a method for detecting overlapping animal vocalizations and does not pose a high risk for misuse. Although the improved detection capabilities presented here might theoretically help humans locate endangered species, this is a general methodological contribution that improves on temporal precision in existing detection tools rather than enhancing detection ability for detecting or identifying species beyond what was already possible with existing detectors.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all existing datasets and models used and provide GitHub links for existing codebases that were used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The OZF dataset is released with documentation. We also describe recording and annotation protocol, inter-annotator agreement, and dataset statistics in Section 4.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: The workdid not involve crowdsourcing or human subjects beyond the members of the research team that annotated vocalizations.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes] .

Justification: We note that all procedures were approved by the McGill University Animal Care and Use Committee in Section 4.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: LLMs were not used for this work.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.