Robust detection of overlapping bioacoustic sound events

Anonymous Author(s)
Affiliation
Address
email

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 the foundation of bioacoustics research. In practice, these sounds often overlap, but identifying each individual acoustic unit is necessary for a diversity of tasks, including species recognition and population estimation, which can be critical for ecology and conversation (1). 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). To understand these communication systems, large-scale identification of individual vocalizations, including accurate classification of overlapping sounds, is crucial.

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 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 recordings of eight female zebra finches (ZFs) spontaneously interacting in a laboratory environment, annotated with onset and offset of each vocalization, and featuring a high degree of overlap. We find that `Voxaboxen` consistently outperforms alternatives, even in the presence of a high degree of overlap, on our new dataset as well as seven previously-published bioacoustics datasets. Taken together, our results demonstrate the general effectiveness of `Voxaboxen` for bioacoustic SED, includ-

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

ing for situations with overlapping vocalizations. To democratize putting boxes around vocalizations, we open source the code for our model and new dataset. To summarize, the contributions of this paper are as follows: (1) introducing `Voxaboxen`, and SED model leveraging pretrained audio encoders, which can predict overlapping vocalizations; (2) releasing a new dataset, Overlapping Zebra Finch (OZF), specifically focused on overlapping vocalizations; (3) experimental evaluation on a diverse set of eight 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 accommodate overlapping objects, and have occasionally 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 differs in that it matches boxes based on feature similarity, which can be inaccurate for animal vocalizations, where highly stereotyped events mean that different events can share very similar features. Our approach accounts for this by matching based on intersection over union (IoU) instead.

Given an audio recording with a mixture of sound sources, source separation is the task of predicting the audio of the pre-mixture sounds. Prior work in bioacoustics (19) has demonstrated the effectiveness of source separation for improving accuracy in downstream classification tasks. 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. Approaches typically assume a maximum number of speakers (e.g., two or four), and assume that speakers can be re-identified by their vocal characteristics across multiple 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 of the audio clip, and α, β are hyperparameters. In (1), the model's predicted detection probability at time t is $\hat{p}_t$, and the target p_t is obtained by smoothing each event onset with a Gaussian kernel 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, and for $x \in \text{Events}$, $\text{Onset}(x)$ and $\text{Dur}(x)$ denote the onset time and duration of x, and s is a hyperparameter. The regression term L_{reg} is L1 loss, applied only to frames in $\{\text{Onset}(x) \mid x \in \text{Events}\}$, i.e. frames where an event begins. Similarly, the classification term L_{cls} is a categorical cross-entropy loss, again applied only when an event begins.

At inference time, we apply a peak-finding algorithm to the time-series of detection probabilities. Detection peaks above a threshold become boxes, with duration and class prediction determined by the value of the regression and classification predictions at the peak. The detection threshold is swept (for computing metrics), or fixed as a hyperparameter; see Section 5. Finally, we apply soft non-maximal suppression (21) to remove duplicate boxes.

3.2 Bidirectional Predictions

One drawback of using these predicted boxes directly is the difficulty for the model in making accurate regression predictions. In preliminary experiments, we observed that both onset and duration predictions can be slightly inaccurate, meaning that the model sometimes correctly detects an event but the edges of the bounding box are slightly off where they should be. To reduce error in bounding box edges, 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 then compute an optimal way to fuse the forward and backward predictions into a single set of predictions, by casting the problem as a maximal bipartite graph matching problem. The bipartite graph has all boxes as vertices. 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 (divided into 60-second files) of 8 adult (> 1 year) female ZFs housed in a large group cage in a sound attenuating chamber (TRA Acoustics, Ontario, Canada). We continuously recorded using Audacity (3.3.3) through two omnidirectional microphones (Countryman, Menlo Park, CA) positioned above and to the side of the cage. Food and water were provided *ad libitum* and all procedures were approved by the McGill University Animal Care and Use Committee in accordance with Canadian Council on Animal Care guidelines. Female ZFs make short, discrete vocalizations of about 100ms, consisting of a flat or downward sweeping harmonic stack, with most energy located between 0.5 and 8 kHz. The recordings were divided among three annotators, who marked the onset and offset time of each vocalization using Raven Pro (Cornell Lab of Ornithology, v.1.6.5). Annotators covered 25 minutes each. One 5 minute section was annotated by all three, where the mean pairwise inter-annotator F1@0.5IoU of 93.5, and 78.1 on the subset that overlaps.

Out of a total of 8504 vocalizations in the dataset, 1449 (17.04%) overlap with at least one other. The total number of overlaps is slightly higher at 1463, as some can overlap more than one other. The number of vocalizations per 60 s file ranges from 19 to 245, with between 0 and 73 overlapping. The duration of silence per 60 s file ranges from 35.5 to 58.1 seconds. We observe a roughly linear relationship between the two. The duration of each vocalization ranges from 3 ms to 350 ms and is strongly peaked around the mean of 109 ms. It is possible to show that, assuming independent vocalizations from each bird that can be modeled with a Poisson distribution, the expected number of pairwise overlaps is $d(n-1)(1-1/B-1/n)$, where d is the ratio of total call durations to time window size, n is the number of vocalizations and B is the number of birds (details provided at project github). In our case, $B = 8$, $n \approx 120$ and $d \approx 0.1$. Plugging in the values for n and d from each 60s file, the average ratio of overlap to number of vocalizations should be 20.46%, significantly above the 17.04% we observe. This 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 an encoder-only transformer, with 12 layers, hidden size 768 and 8 attention heads, pretrained on Audioset (25). In Section 5.2, we explore different choices of backbone. The detection, regression and classification predictions are then each made using a linear layer. The loss function hyperparameters were fixed at $\alpha = 2$, $\beta = 4$, and $s = 6$ following (18). During training and inference, audio is divided into 10-second windows, with 5-second step size between windows. Training lasts for 50 epochs, with the encoder frozen for the first 3 epochs. 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 val set. We apply soft non-maximal suppression (21) with $\sigma = 0.5$.

Datasets In addition to our newly released OZF dataset, we evaluated `Voxaboxen` using seven existing datasets (Table **??**), selected for their taxonomic diversity: amphibians (AnuraSet), insects (Katydid), birds (BirdVox-10h, Hawaiian Birds, Powdermill), and mammals (Humpback, Meerkat). The preprocessing steps we performed on these datasets is described at 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 use a half-time version of BV10 and OZF, and a sixth-time version of Katy. This effectively increases the output frame rate to 100 Hz for BV10 and OZF, and 300 Hz for Katy. Initial experiments indicated that using these slowed-down versions dramatically improved performance.

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 the performance of `Voxaboxen` to several frame-based methods. Three of these consist of a linear layer on top of a encoder-only transformer, initialized with pre-trained weights. The encoders are Frame-ATST (7) (25 Hz output frame rate, pretrained on AudioSet), BEATs (24) (50 Hz, pretrained on AudioSet) and BirdAVES (27)[1] (50 Hz, pre-trained on animal sound datasets). Outputs are median filtered, with kernel size (ks) 1, 3, 7, or 11, selected based on mean average precision @0.5IoU on the val set.

As an additional frame-based method, we compare to a convolutional-recurrent neural network (CRNN) (6; 10; 8). Model inputs are log-mel spectrograms (256 mel bands), and the model consists of 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 finally a bi-LSTM, with hidden size 1024. The weights are randomly initialized. Finally, we compare to two existing 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.

As shown in Table 1, `Voxaboxen` outperforms other methods in almost all cases, and in is far ahead of all other models in several cases, e.g. 10+ points on mAP@0.5 on BV10, HawB, HbW, and Katy. The diversity of animal sounds in the datasets especially highlights the general effectiveness of our method. Faster-RCNN generally performs well on OZF and MT; however, it struggles with datasets with more than one class (AnSet, HawB, and Pow), as well as HbW. Of the frame-level SED models, Frame-ATST, BEATs and BirdAVES, BEATs is generally the strongest, which is consistent with our findings for the backbone choice in `Voxaboxen` (see Table 2). SEDT is poor. Pretrained on datasets mostly of ambient city noises, it transfers badly to animal vocalizations.

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

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

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

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: Ablation studies on the backbone encoder and the forward-backward matching method. The main model uses the BEATs encoder. Best results in **bold**. Both ablation settings give a moderate, consistent drop in performance, showing the superiority of the BEATs encoder over BirdAVES, and the effectiveness of the `Voxaboxen` forward-backward matching method.

5.2 Ablation Studies

Table 2 shows the effect of changing the encoder backbone of `Voxaboxen`, and of removing the forward-backward matching procedure. We found that using BirdAVES as a backbone for `Voxaboxen` reduced performance compared with the version of that used the BEATs encoder. This was surprising considering BirdAVES was designed specifically for animal sounds; however differences in pre-training data volume and training regimes may explain the performance difference. Removing forward-backward matching (i.e. only using forward predictions) also consistently lowers the mAP scores. Mostly the difference is 1-2 points but larger for some datasets, e.g. HbW and Pow.

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`.

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.