DOI: 10.61782/fa.2023.0897

forumacusticum 2023

CONSISTENT BIRDSONG SYLLABLE SEGMENTATION USING DEEP SEMI-SUPERVISED LEARNING

Houtan Ghaffari[1*] **Paul Devos**[1]

[1] Department of Information Technology, WAVES Research Group, Ghent University, Belgium

ABSTRACT

Birdsong analysis requires a segmentation step to isolate syllables. It's a laborious task requiring expertise while subject to human bias and error. Automated methods for birdsong analysis are valuable in biology and linguistics. However, current models need large labeled datasets or at least isolated syllables after human-aided segmentation. Although some models are automatic during the deployment phase, there is no method for avoiding costly annotation during the development phase. Also, this issue underlies a significant weakness of current models, the lack of consistency and generalization across datasets and birds since there are no globally known rules for birdsong syntax. We argue that an automated method is necessary to achieve this feat, where human intervention in the annotation process should be avoided as much as possible. We leverage a semi-supervised model to get consistent segmentation free of human bias. The model achieved near-expert results using a few labeled songs to segment hours of recordings. Moreover, we show the possibility of a species-specific model instead of the commonly used individual-specific ones. Such a method opens the venue for merging clustering and segmentation methods to propose a fully automated framework and accelerate research in fields that study birdsong.

Keywords: *birdsong analysis, syllable segmentation, semi-supervised learning, deep learning*

1. INTRODUCTION

The song acquisition in singing bird species resembles behavioral, neural, and genetic similarities to humans' speech acquisition [1,2]. These characteristics make them a proper model for gaining insights into neural mechanisms of sensory-motor learning, plasticity, and neurogenesis [1, 3]. Moreover, studying vocal communication in animal models improves our understanding of the neurogenetic basis for speech and communication disorders [4].

The utility of statistical models of the birds' vocal behavior spans a diverse spectrum of biological and linguistic studies [5, 6]. However, creating such models requires annotating days of recordings in the controlled lab environment. The data should be clean to prevent propagating error to the downstream tasks, unlike open-field experiments for conservation that are more tolerant to noisy predictions [7].

The canonical procedure to annotate birdsong consists of two steps, segmentation and labeling [8]. Most analyses require a segmentation at the level of vocal units called syllables [2, 5]. First, the song is segmented by an amplitude thresholding algorithm into syllables [8–10]. Afterward, the expert goes over the result to add the missed syllables, remove noise or unwanted patterns taken as a syllable, and adjust the boundaries of each detected syllable in time. The next step is an arbitrary label assignment to each isolated syllable to form coherent and mutually exclusive groups by subjective auditory and visual assessment of the expert(s). It's a time-consuming and labor-intensive procedure that requires expertise.

Despite their prevalence in bioacoustics, thresholding segmentation algorithms are often explained without enough rigor or hard to reproduce due to lengthy heuristics [9, 10]. Also, each bird and recording condition needs different parameters found manually by trial and error, even

*[*Corresponding author*: houtan.ghaffari@ugent.be.]*

for individuals of the same species. Moreover, for birds with complex songs (e.g., canaries), a single threshold parameter can't find all the syllables due to their song variability [5]. Such segmentation methods are not automated enough to save time in practice. It's also noteworthy that the experts need to manually discard some syllable-like vocal units for not belonging to a song (e.g., bird calls) or being badly vocalized or ambiguous [11]. Also, a vocal pattern might be broken into two syllables or taken as one based on the decision of the annotator(s) [5, 11]. All of these hard decisions contribute to inconsistent annotation across trials.

The problem with subjective annotation exacerbates in the labeling stage. The syllable categories are defined separately for each individual since there are no globally known rules for animal vocal analysis, unlike human languages. Hence, the labeling procedure is not generalizable across datasets, creating extra labor and ambiguity in cross-laboratory and cross-species comparisons [4, 12]. Prior work has emphasized the necessity to formalize the birdsong for proper use as a model system for human speech and cognition evolution [2]. These observations call for methods free of human bias in segmentation and labeling, which can generalize across birds and datasets. Moreover, it's hard to collect sufficient data for training separate models for each individual, especially with deep neural networks that have shown superior performance to prior birdsong models [5, 6, 13, 14]. Hence, having a unified segmentation method and a shared set of syllable categories (at least within a species) can provide the models with more data on top of consistency [6].

For the labeling problems discussed here, unsupervised classification and representation of the syllables have shown promising results [4, 6, 12, 13, 15]. However, clustering models require isolated syllables. Thus, the need to address the segmentation challenge first. Cohen et al. [5] proposed to train a convolutional-recurrent neural network on frame-level labeled spectrograms of the songs. They achieved good results using small to moderate annotations, even for birds with complex songs. They intelligently circumvented the segmentation step by classifying the non-syllable frames as background. However, the experts' subjective labeling challenge remains open since it wasn't a binary segmentation task but categorization. The fully supervised methods are individual-specific and don't scale well with the number of birds unless adequate annotation is available for all birds [5, 16].

Leveraging unlabeled data has shown success in Bioacoustics [14, 15, 17, 18]. One strategy to reduce the reliance on human annotation and mine the information from uncurated data is semi-supervised learning [19]. The general goal of this paradigm is to learn the high-level structures from unlabeled data and only rely on the annotations for learning the fine-grained details of a given task [20].

Many semi-supervised algorithms utilize the cluster assumption, which states that the decision boundary should lie in low-density regions of the feature space (i.e., to reduce the mistakes by small perturbations) [21]. Grandvalet and Bengio [22] argued that unlabeled data are not necessarily helpful for discriminators unless the classes are well separated. They used the entropy of the model's predictions as a measure of class overlap and proposed to minimize it for unlabeled data as a regularization. Pseudo-labeling [23] is another effective method equivalent to entropy regularization [22] where the model first predicts unlabeled data and then gets trained simultaneously on labeled and pseudo-labeled samples (a.k.a. self-training). The 'pseudo' means the labels come from the model rather than a human. However, such methods suffer from reinforcing their errors during training (a.k.a. confirmation bias). The soft pseudo-labels are more informative than hard ones since the models are prone to propagating mistakes from overconfident wrong predictions [24]. Also, recent semi-supervised methods only use unlabeled examples that the model can predict confidently [25]. Hence, incorporating soft and confident pseudo-labels can alleviate the confirmation bias.

Our contribution is providing an automated method for the time-consuming segmentation step in the birdsong analysis using small annotation. It serves as the first step in a fully automated framework for a data-efficient, generalizable, and consistent birdsong annotation. We leveraged the Mean Teacher [26] semi-supervised algorithm with only soft [24] and confident [25] predictions to train a neural network for syllable segmentation. The model can cleanly segment hours of recordings while requiring only a few seconds of labeled songs. We also provide experimental evidence on the generalization of this method to multiple individuals, which showed the possibility of a global model for a species.

2. METHODOLOGY

2.1 Data

The four Bengalese finches dataset is from a public repository kindly published by [27]. The recordings have a sampling rate of 32 kHz, annotated by temporal position and category of the syllables. The unannotated files and ones longer than 30 seconds were removed, resulting in 148 recordings for the bird with id "bl26lb16", 767 for "gr41rd51", 780 for "gy6or6", and 337 for "or60yw70".

The three canaries dataset is generously open-sourced by [5] (please see the reference for related links). The recordings have a sampling rate of 44.1 kHz and are fully annotated. Since this dataset was large, 300 recordings were sampled randomly from each canary.

For data preparation, the waveforms were high-pass filtered at 500 Hz and transformed to a power spectrogram using a centered fft window of 512 samples and a hop length of 64. The spectrograms were compressed to the decibel scale and further normalized to $[0, 1]$ range by min-max normalization. Finally, the first row of the spectrograms (DC) was discarded.

To formalize the notation, let's denote the input space by $\mathcal{X} \in \mathbb{R}^{F \times T}$, where F stands for the frequency bins and T indicates the number of frames. The output space is shown by $\mathcal{Y} \in \{0, 1\}^T$, which is a sequence of labels for each input frame. Therefore, the labeled dataset is shown by $D_L = \{(x_i, y_i)|(x_i, y_i) \in (\mathcal{X}, \mathcal{Y})\}_{i=1}^{N_L}$ and the unlabeled one as $D_U = \{x_i|x_i \in \mathcal{X}\}_{i=1}^{N_U}$ where N_L and N_U denote the size of the datasets. Since this is a presence/absence segmentation task, only the temporal position of the syllables was used where 1 means part of a syllable and 0 means irrelevant frame.

2.2 Thresholding Algorithm

There are three baseline thresholding algorithms, named thresholding-A, B, and C, in descending order of quality. The thresholding-C is a general algorithm deployed by the authors following standard steps: (i) high-pass filtering at 500 Hz to reduce noise, (ii) Hilbert envelope extraction, (iii) smoothing using a Hann window, and (iv) thresholding to extract the uninterrupted sequence of samples as syllables. The threshold was adaptively set to the envelope's average for each recording. This algorithm reflects the case of blindly applying a thresholding method without laborious tuning by an expert. The algorithms A and B were adapted from [28], which are specifically tailored for the Bengalese finches of [27]. The thresholding-B uses a sophisticated and manually tuned algorithm that uses experimentally hard-coded threshold value, species-tailored band-pass filtering, ignoring short gaps to attach neighboring segments, and discarding segments shorter than a minimum expected syllable duration for the selected birds. This is the segmentation method with default parameters in [28]. The thresholding-A further uses the tuned values of these parameters for each individual Bengalese finch (minimum silent gap, minimum syllable duration, and threshold value). In terms of automation, algorithms A and B require expertise and labor, but C is fully automated. However, it is applied blindly without using prior knowledge about the bird's vocal behavior.

2.3 Mean Teacher Algorithm

Many semi-supervised algorithms are formulated with consistency regularization loss function [26, 29–31]. Mean Teacher [26] is one such method that uses two identical models to play the teacher and student roles. Let's denote the student's predictive function and output by:

$$\hat{y}_i = p(y_i|x_i, \eta; \theta) \tag{1}$$

where $p(y_i|.) \in [0, 1]$ represents the model's predictive function, the θ represents its parameters, and η represents noisy operations such as data augmentation and dropout [32]. Similarly, for the teacher:

$$\hat{y}'_i = p(y_i|x_i, \eta'; \theta') \tag{2}$$

There are two loss functions for the student model in this framework. One is the cross-entropy of ground truth and student predictions for the labeled examples:

$$\mathcal{L}_s(\theta) = -\frac{1}{N_L} \sum_{(x_i, y_i) \in D_L} y_i \log \hat{y} + (1 - y_i) \log (1 - \hat{y}) \tag{3}$$

The second loss is the consistency between the student and teacher in predicting two differently augmented views of the unlabeled examples. This loss function is weighted using a sigmoid ramp-up function to decrease its importance in the first few epochs of training since the teacher is no better than the student at first [26, 33]. The consistency loss is the difference between the teacher and student predictions:

$$\mathcal{L}_c(\theta) = -\frac{1}{N_U} \sum_{(x_i) \in D_U} w_e * d(\hat{y}_i, \hat{y}'_i) \tag{4}$$

where w_e is a coefficient from sigmoid ramp-up at the current epoch [26]. The cross-entropy was used for the difference function d, but mean-squared-error is also effective with better theoretical properties [26, 33]. A random amplitude gain modulation and Gaussian noise were applied to the waveform for data augmentation. The time and frequency masking [34] were not beneficial for this task. Also, the student's input was augmented heavier than the ones for the teacher. It's important to use noise in the model's forward pass for Mean Teacher to work properly. A strong dropout in the student can provide it, but it is not necessary for the teacher [26]. The teacher is not optimized with gradient descent, but with an exponential moving average of its parameters and the ones from the student. Hence, the gradient of Eqn. (4) is used for training the student, but the teacher's parameters at step t gets updated by the following formula:

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t \tag{5}$$

where α determines how quickly the teacher should incorporate the student's information at the current step (lower α means faster change for the teacher). Following the [26] ablations and our observations, $\alpha = 0.99$ was used for the first 200 epochs and increased to $\alpha = 0.999$ for another 100 epochs of training.

The total cost function is the sum of the Eqn. (3) and Eqn. (4) while the latter being weighted by a sigmoid ramp-up that reaches the value of 1 in 100 epochs. To alleviate the confirmation bias, soft pseudo-labels were used while incorporating only the confident predictions for the consistency loss [24, 25]. Any prediction below 0.2 and above 0.8 was considered as confident. During training, the labeled songs were fed to the model in full length, but the unlabeled ones were randomly cropped into 5-seconds chunks at each iteration to speed up the training by batching. After the training, the teacher's predictions on the full-length recordings were taken as the segmentation.

2.4 Model

A mixture of convolution and recurrent layers is a popular structure for time-series processing, especially when the data is scarce [5, 35, 36]. The model consists of three 2d-convolution blocks, each one having a LeakyReLU activation, a max-pooling layer that preserves the temporal dimension, and a 2d-dropout. After the convolution blocks, it has a bi-directional LSTM layer followed by a 1d-dropout, ending with a linear projection for the binary task of syllable detection. The channel-wise dropout [37] alleviates the overfitting while injecting the noise into the model required by the semi-supervised frameworks [25, 26]. Please see the Fig. 1 for a graphical illustration.

Table 1. Results of the thresholding methods on Bengalese finches. Each section annotated by the method's name shows the metrics defined on the top for each individual on the rows. For ease of read, pay attention to the Jaccard or f1 columns.

animal id	f1	precision	recall	Jaccard
Thresholding-C				
bl26lb16	69.02	96.69	53.66	52.69
gr41rd51	85.40	98.11	75.60	74.52
gy6or6	71.68	99.53	56.00	55.86
or60yw70	81.67	99.15	69.43	69.02
Thresholding-B				
bl26lb16	91.67	97.84	86.23	84.62
gr41rd51	95.91	98.89	93.10	92.14
gy6or6	91.75	99.60	85.04	84.75
or60yw70	96.54	98.79	94.40	93.32
Thresholding-A				
bl26lb16	97.73	95.66	99.88	95.56
gr41rd51	98.25	96.63	99.93	96.56
gy6or6	99.17	98.46	99.90	98.36
or60yw70	99.04	98.16	99.94	98.10

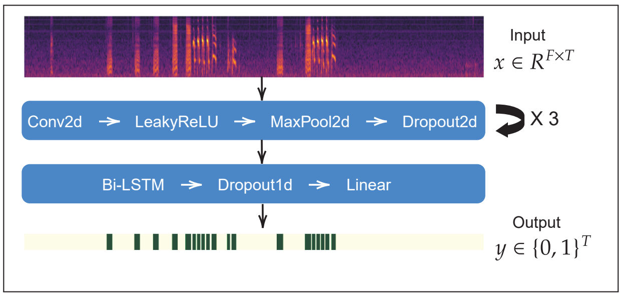
Input
$x \in R^{F\times T}$
Conv2d → LeakyReLU → MaxPool2d → Dropout2d
X 3
Bi-LSTM → Dropout1d → Linear
Output
$y \in \{0,1\}^T$

Figure 1. The neural network architecture with an input-output example.

The optimization algorithm was Adam [38] with the default parameters. We used a learning rate scheduler consisting of a linear warmup for 10 epochs, a constant rate of $1\text{e}{-3}$ for 140 epochs, and a cosine decay for 150 epochs while keeping the minimum learning rate at $1\text{e}{-6}$.

Table 2. Results of the thresholding methods for the canaries. Thresholding-A can't be applied here, please see the section 3.

animal id	f1	precision	recall	jaccard
Thresholding-C				
llb3	70.36	98.98	54.58	54.28
llb11	68.07	97.61	52.25	51.59
llb16	77.04	99.32	62.92	62.65
Thresholding-B				
llb3	91.56	95.96	87.54	84.43
llb11	85.80	92.79	79.79	75.13
llb16	89.62	95.61	84.35	81.20

3. EXPERIMENTS AND RESULTS

All the results in the following tables are in percentage, and higher is better. The results of thresholding methods for the Bengalese finches are shown in Tab. 1, and for canaries in Tab. 2. The thresholding-A is not perfect due to the additional cleanups by experts during annotation. Also, it's not applicable for the canaries since manually tuned parameters by experts are not available. Cohen et al. [5] demonstrated that thresholding doesn't extract all the syllables for complex songs as seen in canaries. It's due to the variation in the amplitude and structure of their songs. The following sections present our methods.

3.1 Bengalese finches Individual-Specific Model

Five labeled songs were picked randomly for each bird, and the rest were used for the semi-supervised objective (without annotations) and testing. The semi-supervised model was compared to a supervised model trained solely on the five labeled songs. To reflect a real few-shot scenario, no validation set was used for early stopping. However, a small portion of the recordings were used in preliminary experiments for model design. The models showed fast convergence while being consistent across the epochs with negligible variation in performance. All models were trained for 300 epochs, and the reported metrics are from the last training epoch, see the Tab. 3.

Table 3. Results of training a separate model for each Bengalese finch. Both methods are strong, compare the Jaccard score to thresholding-A in Tab. 1.

animal id	f1	precision	recall	Jaccard
Supervised				
bl26lb16	96.96	95.29	98.69	94.09
gr41rd51	97.81	97.43	98.20	95.72
gy6or6	97.87	97.57	98.17	95.83
or60yw70	97.59	97.57	97.62	95.30
Semi-Supervised				
bl26lb16	97.86	97.70	98.03	95.81
gr41rd51	98.25	98.59	97.92	96.57
gy6or6	98.36	98.01	98.72	96.78
or60yw70	98.50	98.50	98.49	97.03

Table 4. Results of training a model for all Bengalese finches simultaneously. The performance improved compared to individual modeling in Tab. 3.

animal id	f1	precision	recall	Jaccard
Supervised				
bl26lb16	98.09	98.16	98.02	96.25
gr41rd51	98.59	98.89	98.29	97.22
gy6or6	98.03	96.94	99.16	96.14
or60yw70	98.27	97.55	99.00	96.60
Semi-Supervised				
bl26lb16	98.18	98.00	98.36	96.42
gr41rd51	99.04	98.87	99.21	98.10
gy6or6	98.78	98.75	98.81	97.59
or60yw70	98.96	98.92	99.00	97.94

3.2 Bengalese finches Species-Specific Model

The setup and data are identical to the section 3.1. However, a unified model was trained for all birds, totaling 20 labeled songs from four birds and their unlabeled songs for the semi-supervised objective and testing. The results are in Tab. 4, and a sample output of the model in Fig. 2.

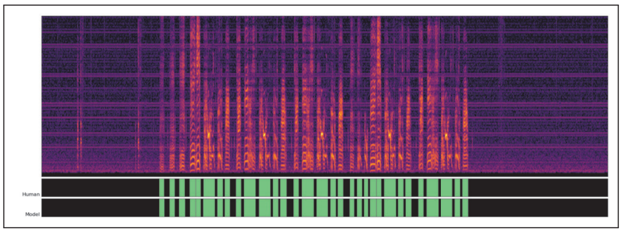

Figure 2. A sample output of the semi-supervised species-level model of Bengalese finch. Top ribbon is Human annotation and the bottom one from model.

Table 5. Results of the canary species-specific model. Both models show reasonable performance given low data and complexity of canary song. This is not achievable by a thresholding method.

animal id	f1	precision	recall	Jaccard
		Supervised		
llb3	97.00	95.66	98.39	94.18
llb11	97.41	96.50	98.33	94.94
llb16	97.37	96.81	97.94	94.87
		Semi-Supervised		
llb3	97.11	95.82	98.43	94.38
llb11	97.60	97.18	98.01	95.31
llb16	97.57	97.36	97.79	95.26

3.3 Canaries Species-Specific Model

We limit the canaries to species-specific modeling since it's stronger, computationally cheaper, and should be the method of choice in applications. A set of 29 recordings were picked randomly while ensuring each syllable type was present at least once. The other recordings were used for semi-supervised objective and testing. See the results in Tab. 5 and Fig. 3

3.4 Supervised vs. Semi-Supervised method

The supervised model performed close to the semi-supervised while being easier to implement. The Bengalese finch with id ”gy6or6” was used to compare the two methods as training size increases, see the Fig. 4. The result suggests a trade-off between the computation and annotation costs.

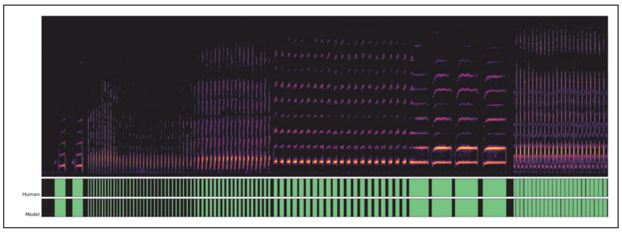

Figure 3. A sample output of the semi-supervised species-level model of canary. Top ribbon is Human annotation and the bottom one from model.

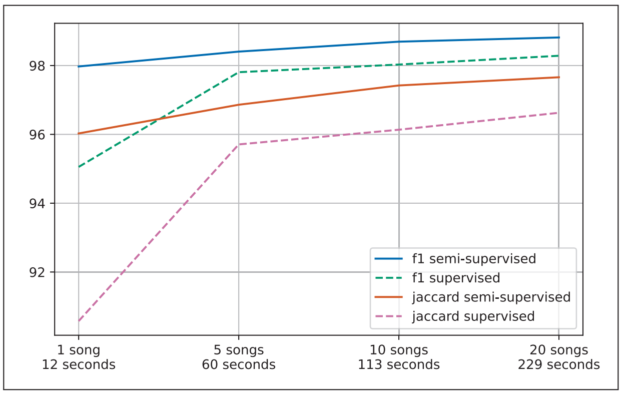

Figure 4. Supervised vs. semi-supervised for a Bengalese finch. The duration is more background than song.

4. CONCLUSION

Many studies in birdsong analysis require accurate segmentation and categorization of the syllables. Current models for curating birdsong datasets lack consistency and reusability across laboratories due to heavy reliance on experts' subjective decisions and heuristics. However, mistakes in annotation propagate to the final result, leading to inconsistencies or non-reproducible findings. We proposed a data-efficient semi-supervised model to alleviate the segmentation stage problems and labor.

The model showed near-expert performance while using very few labeled songs. Additionally, it can handle all individuals within a species in one model without any change to hyper-parameters. To our knowledge, this is impossible by the thresholding algorithms. The main difference is that learning models are pattern-seeking instead of being sensitive to the magnitude. The semi-supervised model outperformed its supervised counterpart, which reveals the benefits of leveraging unlabeled data in Bioacoustics. However, the supervised model had good performance with reasonable training data. One might be interested in extending this work to field recordings, but that requires addressing challenges such as overlapping sounds and diverse types of noise found outdoors. Providing such an assessment is beyond the scope of the current paper, and requires precisely labeled filed recordings.

Finally, the neural networks were not sensitive to the initial small training sets. However, it is not expected from any learning method to correctly evaluate a pattern that was absent during the training. We suggest two solutions to gain consistent results: (i) provide at least one example of each vocal pattern in the training set, (ii) segment all vocalization and clean them at the clustering stage. The former solution might inject inconsistency, but the latter makes it possible to think about unsupervised segmentation and full automation. A potential future direction is to merge the deep segmentation with a deep clustering method to provide consistent and effortless annotation on demand for large datasets.

5. REFERENCES

[1] M. S. Brainard and A. J. Doupe, "What songbirds teach us about learning," *Nature*, vol. 417, no. 6886, pp. 351–358, 2002.

[2] R. C. Berwick, K. Okanoya, G. J. Beckers, and J. J. Bolhuis, "Songs to syntax: the linguistics of birdsong," *Trends in cognitive sciences*, vol. 15, no. 3, pp. 113–121, 2011.

[3] D. G. Mets and M. S. Brainard, "Learning is enhanced by tailoring instruction to individual genetic differences," *eLife*, vol. 8, p. e47216, sep 2019.

[4] Z. D. Burkett, N. F. Day, O. Peñagarikano, D. H. Geschwind, and S. A. White, "Voice: A semi-automated pipeline for standardizing vocal analysis across models," *Scientific reports*, vol. 5, no. 1, pp. 1–15, 2015.

[5] 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, jan 2022.

[6] T. Morita, H. Koda, K. Okanoya, and R. O. Tachibana, "Measuring context dependency in birdsong using artificial neural networks," *PLoS computational biology*, vol. 17, no. 12, p. e1009707, 2021.

[7] S. Zsebők, M. F. Nagy-Egri, G. G. Barnaföldi, M. Laczi, G. Nagy, Éva 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," *Ornis Hungarica*, vol. 27, no. 2, pp. 59–66, 2019.

[8] A. Kershenbaum, D. T. Blumstein, M. A. Roch, Ç. Akçay, G. Backus, M. A. Bee, K. Bohn, Y. Cao, G. Carter, C. Cäsar, *et al.*, "Acoustic sequences in nonhuman animals: a tutorial review and prospectus," *Biological Reviews*, vol. 91, no. 1, pp. 13–52, 2016.

[9] P. Du and T. W. Troyer, "A segmentation algorithm for zebra finch song at the note level," *Neurocomputing*, vol. 69, no. 10, pp. 1375–1379, 2006. Computational Neuroscience: Trends in Research 2006.

[10] S. Fagerlund, "Automatic recognition of bird species by their sounds," *Finlandia: Helsinki University Of Technology*, 2004.

[11] D. Stowell and M. D. Plumbley, "Birdsong and c4dm: A survey of uk birdsong and machine recognition for music researchers," *Centre for Digital Music, Queen Mary University of London, Tech. Rep. C4DM-TR-09-12*, 2010.

[12] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.

[13] J. Goffinet, S. Brudner, R. Mooney, and J. Pearson, "Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires," *eLife*, vol. 10, p. e67855, may 2021.

[14] K. R. Coffey, R. G. Marx, and J. F. Neumaier, "Deepsqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations," *Neuropsychopharmacology*, vol. 44, no. 5, pp. 859–868, 2019.

[15] S. C. Keen, K. J. Odom, M. S. Webster, G. M. Kohn, T. F. Wright, and M. Araya-Salas, "A machine learning approach for classifying and quantifying acoustic diversity," *Methods in ecology and evolution*, vol. 12, no. 7, pp. 1213–1225, 2021.

[16] R. O. Tachibana, N. Oosugi, and K. Okanoya, "Semi-automatic classification of birdsong elements using a

linear support vector machine," *PloS one*, vol. 9, no. 3, p. e92584, 2014.

[17] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, p. e488, 2014.

[18] V. Morfi, R. F. Lachlan, and D. Stowell, "Deep perceptual embeddings for unlabelled animal sound events," *The Journal of the Acoustical Society of America*, vol. 150, no. 1, pp. 2–11, 2021.

[19] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2006.

[20] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[21] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *International workshop on artificial intelligence and statistics*, pp. 57–64, PMLR, 2005.

[22] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances in Neural Information Processing Systems* (L. Saul, Y. Weiss, and L. Bottou, eds.), vol. 17, MIT Press, 2004.

[23] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, p. 896, 2013.

[24] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5982–5991, 2019.

[25] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[26] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[27] D. Nicholson, J. E. Queen, and S. J. Sober, "Bengalese finch song repository," Sept. 2022.

[28] D. Nicholson, "evfuncs," 3 2021.

[29] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," *Advances in neural information processing systems*, vol. 27, 2014.

[30] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[31] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[33] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *International Conference on Learning Representations*, 2017.

[34] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, ISCA, sep 2019.

[35] D. Quang and X. Xie, "Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences," *Nucleic acids research*, vol. 44, no. 11, pp. e107–e107, 2016.

[36] H. Ghaffari, "An efficient method for the classification of croplands in scarce-label regions," *arXiv preprint arXiv:2103.09588*, 2021.

[37] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 648–656, 2015.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.