CHIRPING UP THE RIGHT TREE: INCORPORATING BIOLOGICAL TAXONOMIES INTO DEEP BIOACOUSTIC CLASSIFIERS

Jason Cramer,[1] *Vincent Lostanlen,*[1,2] *Andrew Farnsworth,*[2] *Justin Salamon,*[3] *Juan Pablo Bello*[1]

[1] New York University, Music and Audio Research Laboratory, New York, NY, USA
[2] Cornell Lab of Ornithology, Ithaca, NY, USA
[3] Adobe Research, San Francisco, CA, USA

ABSTRACT

Class imbalance in the training data hinders the generalization ability of machine listening systems. In the context of bioacoustics, this issue may be circumvented by aggregating species labels into super-groups of higher taxonomic rank: genus, family, order, and so forth. However, different applications of machine listening to wildlife monitoring may require different levels of granularity. This paper introduces TaxoNet, a deep neural network for structured classification of signals from living organisms. TaxoNet is trained as a multitask and multilabel model, following a new architectural principle in end-to-end learning named "hierarchical composition": shallow layers extract a shared representation to predict a root taxon, while deeper layers specialize recursively to lower-rank taxa. In this way, TaxoNet is capable of handling taxonomic uncertainty, out-of-vocabulary labels, and open-set deployment settings. An experimental benchmark on two new bioacoustic datasets (ANAFCC and BirdVox-14SD) leads to state-of-the-art results in bird species classification. Furthermore, on a task of coarse-grained classification, TaxoNet also outperforms a flat single-task model trained on aggregate labels.

Index Terms— Acoustic signal detection, audio databases, classification algorithms, multilayer neural network, phylogeny

1. INTRODUCTION

The range of vocalizations of an animal often bears the "acoustic fingerprint" of its species [1]. Consequently, the deployment of autonomous recording units allows to sample these vocalizations in their natural environment, with numerous applications in ecology and conservation biology [2, 3]. However, analyzing large volumes of bioacoustic data requires to automate species classification and resort to machine listening techniques [4].

Some vocalizations from migratory birds, known as flight calls, are particularly identifiable in terms of species [5]. Thus, a potential solution for mapping bird migration in real time would be to record these flight calls from an acoustic sensor network on the ground [6]. In this context, two open-access data science challenges named LifeCLEF [7] and DCASE Bird Audio Detection [8] have concluded that deep convolutional networks (convnets) achieve state-of-the-art results in bird species classification [9].

Although convnets have recently proven successful in avian bioacoustics, they depend on the availability of a large training set. In the case of a flight call, collecting reliable human annotations is costly and time-consuming, because it requires expert knowledge.

Please direct correspondence to jtcramer@nyu.edu.
This work is partially supported by National Science Foundation award 1633259 (BIRDVOX).

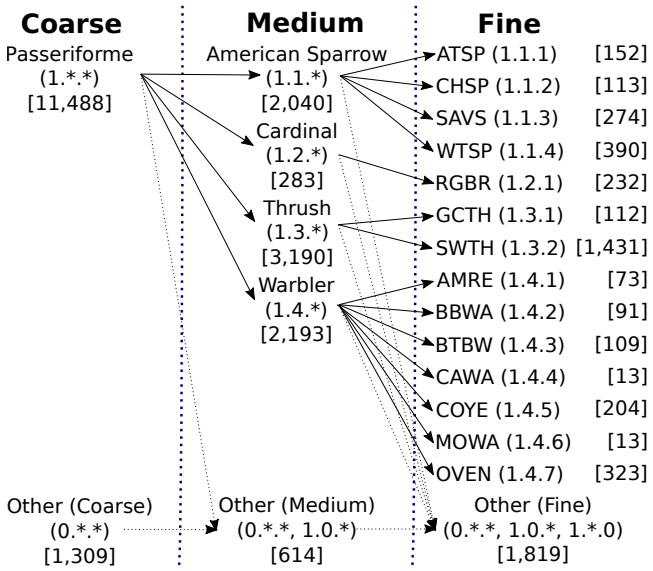


Fig. 1: Taxonomy of flight calls by animal order (coarse), family (medium), and species (fine). The numbers in brackets denote the size of quantity of annotated training samples in the BirdVox-14SD dataset. The suffix "0" and the wildcard "*" denote "other" and "unknown" respectively.

Furthermore, flight call datasets are typically imbalanced in terms of species labels, due to disparities in abundance and vocal activity [10]. Lastly, the presence of background noise may hamper the identifiability of some vocalizations, even to an expert ear, thereby resulting in label uncertainty [11]. For these reasons, flight call classification remains an open research problem. Beyond the example of flight calls, these three issues issues recur throughout many applications of convnets to the life sciences [12].

In this article, we propose a new supervised learning framework for hierarchical classification in deep neural networks. The key idea is to decompose each species label into a sequence of taxa of decreasing rank, typically: order, family, genus, and species itself. In this way, species classification becomes a multitask problem, wherein each task corresponds to retrieving the taxon at a different rank. The companion website of this paper presents a quantitative benchmark of design choices in the formulation of hierarchical multitask learning, in the practical setting of flight call classification.[1]

[1] Companion website: https://github.com/BirdVox/cramer2020icassp

2. RELATED WORK

Hierarchical classification is an instance of structured classification [13] in which classes are akin to nodes on a tree [14]. It has found applications in multiple domains, including natural language [15], environmental sounds [16], and music [17]. One recent publication proposes a hierarchical classification approach for analyzing frog calls [18]; but this approach does not rely on neural networks, and does not accommodate out-of-vocabulary labels.

The appeal behind formulating bird species recognition as a hierarchical classification problem consists in drawing insights from evolutionary biology [19]. The theoretical framework of Darwinian classification, perfected in the 20th century thanks to the technology of genomic sequencing, has led to a comprehensive phylogeny of the class of birds (*Aves*) based on the notion of last common ancestor [20]. Although this phylogeny does not always reflect acoustic similarity between vocalizations, it has the advantage of being systematic and robust to intraspecific variations in phenotype. In contrast, hierarchical taxonomies for urban sounds requires to collect similarity judgments between acoustic events [21], which may vary across individuals, cultures, and use cases [22].

Identifying an animal vocalization at multiple levels of the taxonomy has a strong potential in applied bioacoustics. Indeed, in many conservation science initiatives, the most relevant level of the taxonomy may not be known ahead of time. To address this problem, the simplest methodology would be to train a separate classifier at each level of the hierarchy of interest. However, this approach incurs a significant software development cost, i.e. to train and maintain multiple "flat" classification models. On the contrary, a hierarchical model requires a single pass of training; in addition, it may discover informative correlations between related taxa, thereby mitigating the effect of class imbalance. In this sense, hierarchical classifiers follow a paradigm of multitask learning [23, 24]. A recent publication combines multitask hierarchical classification with metric learning to detect acoustic events [25]; but this hierarchy-aware training loss is not reflected is the architecture of the proposed neural network.

3. NEW DATASETS: ANAFCC AND BIRDVOX-14SD

3.1. Taxonomy

Figure 1 illustrates our proposed taxonomy. We focus on a set of 14 species of migratory birds, all belonging to the *Passeriformes* order, and gathered into four families: American sparrows (*Passeridae*), cardinals (*Cardinalidae*), thrushes (*Turdidae*), and New World Warblers (*Parulidae*). In the following, we refer to the three taxonomic ranks at hand (order, family, species) as coarse, medium, and fine level respectively. Each of the levels contains a catch-all category for sounds which fall in none of the taxa of interest.

3.2. American Northeast Avian Flight Call Classification

To train our model, we aggregate flight calls from different sources: BirdVox-70k [26], CLO-43SD, CLO-SWTH, CLO-WTSP [4], the Macaulay Library [27], Xeno-Canto [28], and Old Bird [29][2]. An expert ornithologist (AF of the authors) verified or re-annotated those clips and aligned each flight call precisely at the center of its corresponding clip. We release this set of recordings and annotations as the American Northeast Avian Flight Call Classification dataset; henceforth ANAFCC[3].

[2]Official website of Old Bird, Inc.: http://www.oldbird.org

[3]Download ANAFCC: https://doi.org/10.5281/zenodo.3666782

3.3. BirdVox 14 Species Dataset

For evaluation, we collect 6600 hours of audio from ten autonomous recording units in Ithaca, NY, USA, following the same protocol as BirdVox-full-night [26]. In order to maximize diversity across sensor locations, time of day, week in the season, and background noise characteristics (as represented by vector quantizations of median MFCCs), we select 150 two-hour recordings by means of the Entrofy library[4]. Then, an expert ornithologist (AF of the authors) annotated each of these recordings by pinpointing flight calls and labeling species. We refer to this dataset as the BirdVox 14 Species Dataset; henceforth BirdVox-14SD[5]. BirdVox-14SD aims to provide a broader snapshot of avian activity in a full migratory season than BirdVox-full-night, which is based on a single night of migration.

3.4. Cross-dataset evaluation

In the following, we train and validate our models on ANAFCC and evaluate them on BirdVox-14SD. We train all of our models up to 1.152M examples by means of the Adam optimizer with an initial learning rate of 10^{-4}. Because the class distributions differ between the two datasets, we create a training-validation split by formulating the allocation of examples to the validation set as a knapsack problem [30] treating individual data sources containing a particular fine-level class as items, with weights corresponding to the number of examples contained within. We consider validation knapsack sizes between 15-30% of the total number of examples in ANAFCC, and for each find the optimal validation knapsack with Google OR-Tools. Among these candidate splits, we choose the split with the lowest average of Jensen-Shannon divergence between the fine-level class distributions of the split subsets and BirdVox-14SD.

3.5. Per-Channel Energy Normalization (PCEN)

For the sake of cross-library compatibility, we use the same preprocessing frontend as BirdVoxDetect [31]: we resample each audio clip to a 22,050 Hz, apply a log-scale mel-frequency spectrogram with a window size 256 (12 ms), hop size of 32 (1.5 ms), and 128 mel-frequency subbands ranging from 2000 Hz to 11,025 Hz. We then apply per-channel energy normalization (PCEN) [32] with librosa v0.7.0 [33] with parameters chosen identically to those used in prior work on flight call detection [34], that is, $\varepsilon = 10^{-6}$, $\alpha = 0.8$, $\delta = 10$, $r = 0.25$, and $T_{\text{PCEN}} = 60$ ms. Lastly, we extract the center 104 frames (150 ms) of the PCEN representation as 2-D input to the convnet. We apply random digital audio effects as data augmentation: pitch shifting, time stretching, and addition of background noise [35]. We sample training data uniformly with respect to fine-level classes by means of the pescador library [36].

3.6. Baseline

The base architecture of our models mirrors the deep convolutional networks described by [10], shown to obtain state-of-the-art accuracy on the CLO-43SD dataset when combined with shallow learning approaches. This model consists of three convolutional layers (ℓ_1, ℓ_2, ℓ_3), a flattening layer (henceforth $\ell_{3,\text{flat}}$), a fully connected hidden layer (ℓ_4), and a fully connected output layer (ℓ_5). All hidden layers use rectified linear units (ReLU) as activation functions. Furthermore, note that we disable bias weights on all layers.

[4]Source code of Entrofy: https://github.com/dhuppenkothen/entrofy

[5]Download BirdVox-14SD: https://doi.org/10.5281/zenodo.3667094

4. MULTILABEL HIERARCHICAL COMPOSITION IN THE TAXONET CLASSIFIER

4.1. Partition of nodes at every layer

In the single-task baseline, only the deepest layer ℓ_5 contains output nodes. Conversely, TaxoNet associates $\ell_{3,\text{flat}}$ to the coarse-level predictions, ℓ_4 to the medium-level predictions, and ℓ_5 to fine-level predictions. We partition each layer ℓ_k such that each taxon is allocated a number of nodes that is proportional to its number of subtaxa at the level below (ℓ_{k+1}). Such a partition implies that the linear span of the layer at hand gets decomposed into a direct sum of orthogonal subspaces, each corresponding to a different taxon.

Furthermore, we induce a sparsity constraint on the synaptic weights: each of the subspaces in ℓ_k only activates the subspaces in the ℓ_{k+1} partition which correspond to its children nodes in the taxonomy. In this coarse-to-fine scheme, the subspace decomposition progresses recursively as depth increases, until becoming an orthogonal basis at the deepest layer.

4.2. Prediction of out-of-vocabulary taxa

Let C_k the number of in-vocabulary taxa at rank k , and $c \leq C_k$ some in-vocabulary taxon among them. We denote by $\mathbf{x}_{k,c} \in \mathbb{R}^{n_{k,c}}$ the layer partition of $n_{k,c}$ nodes associated with class c. Let $\mathbf{a}_{k,c|\theta} \in \mathbb{R}^{n_{k,c}}$ be the trainable vector of neural network weights associated to taxon c. TaxoNet predicts the probability of presence of c as

$$y_{k,c} = \sigma\left(\sum_{m=1}^{n_{k,c}} \mathbf{a}_{k,c|\theta}(m)\mathbf{x}_{k,c}(m)\right), \quad (1)$$

where σ denotes the sigmoid function. Moreover, TaxoNet predicts the probability of presence of the catch-all taxon ("other") as the complement of the most likely in-vocabulary taxon:

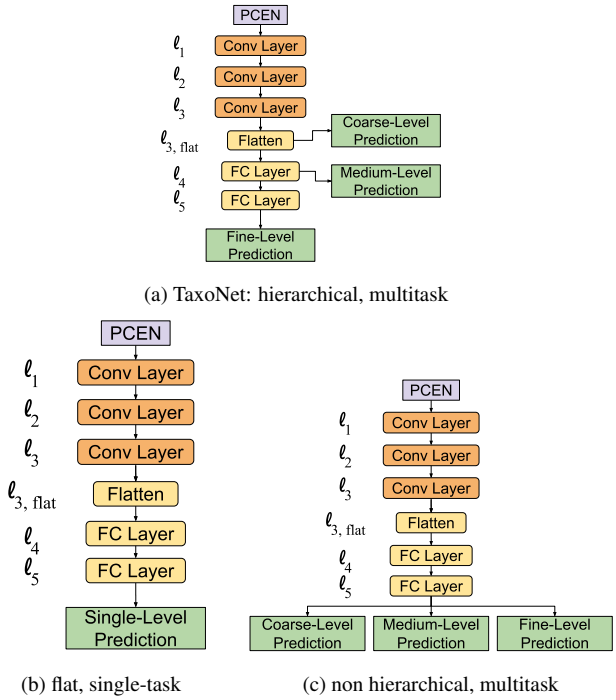
$$y_{k,\text{other}} = 1 - \left(\max_{1\leq c\leq C_k} y_{k,c}\right). \quad (2)$$

Note that the probabilities y_k, c and $y_k,$ other range between zero and one, but do not necessarily sum to one. Therefore, we train each level k in TaxoNet by means of a separate binary cross-entropy loss for each y_k, c and for $y_k,$ other. In this sense, each TaxoNet layer resembles a multilabel classifier. The rationale behind this design choice is that we intend to focus the representational power of TaxoNet on in-vocabulary taxa, while dedicating no trainable parameters to potential out-of-vocabulary taxa. Given Equation 2, an out-of-vocabulary sample may be correctly classified at the shallower level (ℓ_{k-1}) while being correctly excluded from layer ℓ_k and deeper.

4.3. Results

We evaluate all proposed models according to two metrics: micro-averaged and macro-averaged accuracy. The former assigns an identical penalty to every misclassified sample, regardless of its class. The latter operates in two stages: it first measures the probability of correct classification on a classwise basis, and then averages these probabilities uniformly across classes. Because our the ANAFCC and BirdVox-14SD datasets are imbalanced, classifiers tend to fare worse in terms of macro-averaged accuracy than micro-averaged accuracy.

Table 1 summarizes our benchmark; it also includes results from our ablation study (see Section 5). In terms of species classification, we find that TaxoNet outperforms the baseline, a single-task classifier, both in micro-averaged (66.33% vs. 61.13%) and macro-averaged



(a) TaxoNet: hierarchical, multitask

(b) flat, single-task

(c) non hierarchical, multitask

Fig. 2: Computational graphs of: (a) TaxoNet model (hierarchical multitask); (b) flat single-task model; (c) non-hierarchical multitask models. See Section 5 for details.

(55.69% vs. 54.80%) accuracies. This result suggests that TaxoNet has potential for improving the state of the art in bioacoustic species classification.

We also observe that TaxoNet outperforms single-task baselines on coarser tasks. In terms of four-way family classification, for example, TaxoNet reaches a micro-averaged accuracy of 76.50%, compared with 73.80% for the baseline. This result suggests that TaxoNet is indeed capable of multitasking, and adapt the taxonomic rank of its output to user preferences at runtime.

Thirdly, we compare TaxoNet against a single-task species classifier on a coarse task. In the case of flight calls, the coarse task consists in deciding whether a given sound arises from a passerine (order *Passeriformes*) or not. We observe a wide gap in accuracy: $\ell_{3,\text{flat}}$ in TaxoNet fares at 94.69% while the coarsened ℓ_5 in the baseline fares at 77.72%. This result that TaxoNet is not only better than the baseline on average; it also makes less glaring mistakes. Specifically, TaxoNet may occasionally confuse flight calls across species of the same family, or across families of the same order; but it would rarely label a non-bird sound as a flight call or vice versa.

4.4. Open-source distribution under MIT license

We distribute TaxoNet as a pretrained model as part of a Python library named BirdVoxClassify[6]. BirdVoxClassify is a dependency of BirdVoxDetect as of v0.2, thus allowing joint detection and classification of flight calls in a single pass of audio streaming, mel-spectrogram analysis, and PCEN.

[6]Repository: https://github.com/BirdVox/birdvoxclassify

Model	# Trained Params	FIN Micro Acc.	FIN Macro Acc.	MED Micro Acc.	MED Macro Acc.	COA Acc.
Flat ST [FIN]	641K	61.13	54.80	64.61	50.40	77.72
Flat ST [MED]	640K	-	-	73.80	56.04	**94.75**
Flat ST [COA]	640K	-	-	-	-	93.85
TaxoNet	649K	**66.33**	55.69	76.50	61.60	94.69
Non-H. MT	641K	61.82	55.83	75.10	55.87	94.39
H. Baseline	650K	58.74	**58.06**	75.83	60.04	94.54
H. Cont.	640K	63.47	41.46	**79.36**	**65.08**	**94.75**
H. Comp. MC	649K	60.39	52.30	75.94	56.96	94.67

Table 1: Model comparison. ST: single task, MT: multitask, H: Hierarchical, Cont: Containment, Comp: Composition, ML: multilabel, MC: multiclass. For single-task models, predictions at coarser taxonomic levels are obtained by coarsening the prediction, and "other" is mapped to "other" at the corresponding level.

5. ABLATION STUDY

This section briefly discusses some alternative design choices to TaxoNet for multitask species classification. The full set of experimental results is presented in Table 1.

5.1. Non-hierarchical multitask model (Non-H. MT)

This model, depicted in Figure 2 (c), predicts taxa at all levels from the deepest layer, ℓ_5. Therefore, the output of ℓ_4 therefore serves as a shared representation for all tasks.

5.2. Hierarchical baseline model (H. Baseline)

This model is hierarchical, in the sense that $\ell_{3,\text{flat}}$ and ℓ_4 respectively predict coarse-level and medium-level taxa. However, this model is not compositional: $\ell_{3,\text{flat}}$ is densely connected to ℓ_4 and ℓ_4 is densely connected to ℓ_5. This is unlike TaxoNet, in which connections are sparse and nested recursively according to hierarchical composition.

5.3. Hierarchical containment model (H. Cont.)

This model is similar to TaxoNet, but replaces all inner products with trainable vectors $\mathbf{a}_{k,c|\theta}$ by unweighted averages. It uses $\tanh$ in lieu of sigmoid for σ so that zero maps to zero: consequently, if the activation for a class is zero, all subclasses in the taxonomy will, by definition, also have activations of zero.

The hierarchical containment model seems to outperform all other models with respect to medium-level accuracy, but achieves the lowest medium-level macro-accuracy. The reason for this is not entirely clear, but could be due to difficulties in training due to the model zeroing out activations for all subtaxa of an inactive node.

5.4. Hierarchical composition multiclass (H. Comp. MC)

This model is identical to TaxoNet, except that it devotes a specific trainable vector $\mathbf{a}_{k,\text{other}|\theta}$ to represent the probability of the "other" class $y_{k,\text{other}}$. Consequently, the output probabilities of each rank k now sum to one, and hierarchical classification can be formulated as a multiclass problem with a softmax nonlinearity; whereas TaxoNet formulates it as a multilabel problem with sigmoid nonlinearities.

6. CONCLUSION AND FUTURE PERSPECTIVES

In this paper, we have presented a neural network architecture, named TaxoNet, which performs a joint prediction at multiple levels of a known taxonomy. On a task of bird species classification from flight calls, we showed that TaxoNet performs comparably to a specialized model trained to predict at any particular taxonomic level. Moreover, TaxoNet outperforms the strategy of coarsening predictions produced by models trained to predict at finer taxonomic levels. Through this work, we have also released the ANAFCC and BirdVox-14SD datasets as well as the BirdVoxClassify Python package.

We must acknowledge that the taxonomy with which we have worked with is relatively small: 14 nodes at the finest level. It is of interest to inquire about the effect of the depth and width of taxonomy on the accuracy of TaxoNet. Besides, we have only discussed a few methods for incorporating taxonomy into the training methods and architecture design of deep hierarchical models, but there are a number of other ways that applying inductive bias inspired by biological taxonomy could be explored. In particular, hierarchical multitask training of classifiers have the potential to be useful in cases where due to annotator skill or interest, annotations for flight calls may vary in specificity.

There are a number of opportunities for future work in the TaxoNet framework of hierarchical classification. Class balancing techniques can to improve the balance at every level of the taxonomy, especially the coarser levels that are currently heavily imbalanced. The multiclass models can be extended to produce joint probabilities that ensure hierarchical consistency in the model's predictions, which is crucial for robust hierarchical classification [14]. Lastly, the complex optimization landscape of multitask training alone can make model fitting difficult; but methods such as using a hierarchical form of curriculum learning [37] might help to alleviate these issues in the near future.

7. ACKNOWLEDGMENT

We thank Jessie Barry, Ian Davies, Tom Fredericks, Jeff Gerbracht, Sara Keen, Holger Klinck, Anne Klingensmith, Ray Mack, Peter Marchetto, Ed Moore, Matt Robbins, Ken Rosenberg, and Chris Tessaglia-Hymes for designing autonomous recording units and collecting data. We thank Bill Evans from Old Bird, Inc. for contributing to the ANAFCC dataset. We acknowledge that the land on which the BirdVox-14SD dataset was collected is the unceded territory of the Cayuga nation, which is part of the Haudenosaunee (Iroquois) confederacy.

8. REFERENCES

[1] P. Laiolo, "The emerging significance of bioacoustics in animal species conservation," *Biological conservation*, vol. 143, no. 7, pp. 1635–1645, 2010.

[2] F. Bairlein, "Migratory birds under threat," *Science*, vol. 354, no. 6312, pp. 547–548, 2016.

[3] S. R. Loss, T. Will, and P. P. Marra, "Direct mortality of birds from anthropogenic causes," *Annual Review of Ecology, Evolution, and Systematics*, vol. 46, 2015.

[4] J. Salamon, J. P. Bello, A. Farnsworth, M. Robbins, S. Keen, H. Klinck, and S. Kelling, "Towards the automatic classification of avian flight calls for bioacoustic monitoring," *PLOS ONE*, vol. 11, no. 11, pp. 1–26, 11 2016.

[5] A. Farnsworth, "Flight calls and their value for future ornithological studies and conservation research," *The Auk*, vol. 122, no. 3, pp. 733–746, 2005.

[6] H. Pamula, A. Pocha, and M. Klaczynski, "Towards the acoustic monitoring of birds migrating at night," *Biodiversity Information Science and Standards*, vol. 3, pp. e36589, 2019.

[7] A. Joly, H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W. P. Vellinga, J. C. Lombardo, R. Planque, S. Palazzo, and H. Müller, "LifeCLEF 2017 lab overview: multimedia species identification challenges," in *CLEF*, 2017, pp. 255–274.

[8] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[9] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *Proc. EUSIPCO*. IEEE, 2017, pp. 1764–1768.

[10] J. Salamon, J.P. Bello, A. Farnsworth, and S. Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," in *Proc. ICASSP '17*, March 2017, pp. 141–145.

[11] V. Lostanlen, K. Palmer, E. Knight, Ch. Clark, H. Klinck, A. Farnsworth, J. Cramer, T. Wong, and J. P. Bello, "Long-distance detection of bioacoustic events with per-channel energy normalization," in *Proc. DCASE*, 2019.

[12] S. Belongie and P. Perona, "Visipedia circa 2015," *Pattern Recognition Letters*, vol. 72, pp. 15 – 24, 2016, Special Issue on ICPR 2014 Awarded Papers.

[13] D. Charte, F. Charte, S. García, and F. Herrera, "A snapshot on nonstandard supervised learning problems: taxonomy, relationships, problem transformations and algorithm adaptations," *Progress in Artificial Intelligence*, vol. 8, no. 1, pp. 1–14, 2019.

[14] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *DMKD*, vol. 22, no. 1-2, pp. 31–72, 2011.

[15] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings," *Information Sciences*, vol. 471, pp. 216 – 232, 2019.

[16] F. Saki and N. Kehtarnavaz, "Real-time hierarchical classification of sound signals for hearing improvement devices," *Applied Acoustics*, vol. 132, pp. 26 – 32, 2018.

[17] J. J. Burred and A. Lerch, "Hierarchical automatic audio signal classification," *JAES*, vol. 52, no. 7/8, pp. 724–739, 2004.

[18] J. G. Colonna, J. Gama, and E. F. Nakamura, "A comparison of hierarchical multi-output recognition approaches for anuran classification," *Machine Learning*, vol. 107, no. 11, pp. 1651–1671, Nov 2018.

[19] A. Farnsworth and I. J. Lovette, "Phylogenetic and ecological effects on interspecific variation in structurally simple avian vocalizations," *BJLS*, vol. 94, no. 1, pp. 155–173, 04 2008.

[20] R. O. Prum, J. S Berv, A. Dornburg, D. J. Field, J. P. Townsend, E. M. Lemmon, and A. R Lemmon, "A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing," *Nature*, vol. 526, no. 7574, pp. 569–573, 2015.

[21] J. F. Gemmeke, D. P. W. 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 *Proc. ICASSP '17*. IEEE, 2017, pp. 776–780.

[22] Mark Cartwright, Ana Elisa Mendez Mendez, Jason Cramer, Vincent Lostanlen, Graham Dove, Ho-Hsiang Wu, Justin Salamon, Oded Nov, and Juan Bello, "Sonyc urban sound tagging (sonyc-ust): A multilabel dataset from an urban acoustic sensor network," in *Proc. DCASE*, New York University, NY, USA, October 2019, pp. 35–39.

[23] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, 09 2017.

[24] W. Waegeman, K. Dembczyński, and E. Hüllermeier, "Multi-target prediction: a unifying view on problems and methods," *DMKD*, vol. 33, no. 2, pp. 293–324, Mar 2019.

[25] A. Jati, N. Kumar, R. Chen, and P. Georgiou, "Hierarchy-aware loss function on a tree structured label space for audio event detection," in *Proc. ICASSP*. IEEE, 2019, pp. 6–10.

[26] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Birdvox-full-night: a dataset and benchmark for avian flight call detection," in *Proc. ICASSP '18*, April 2018.

[27] I. Betancourt and C. M. McLinn, "Teaching with the Macaulay library: an online archive of animal behavior recordings," *JMBE*, vol. 13, no. 1, pp. 86, 2012.

[28] W. P. Vellinga and R. Planqué, "The Xeno-Canto collection and its relation to sound recognition and classification.," in *CLEF*, 2015.

[29] W. R. Evans and D. K. Mellinger, "Monitoring grassland birds in nocturnal migration," *Studies in Avian Biology*, vol. 19, pp. 219–229, 1999.

[30] S. Martello, "Knapsack problems: algorithms and computer implementations," *Wiley-Interscience Series in Discrete Mathematics and Optimization*, 1990.

[31] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Robust sound event detection in bioacoustic sensor networks," *PLOS ONE*, vol. 14, no. 10, pp. e0214168, 2019.

[32] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *Proc. ICASSP '17*. IEEE, 2017, pp. 5670–5674.

[33] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, et al., "librosa/librosa: 0.7.1," Oct. 2019.

[34] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello, "Per-channel energy normalization: Why and how," *IEEE SPL*, vol. 26, no. 1, pp. 39–43, 2018.

[35] B. McFee, E. J. Humphrey, and J. P. Bello, "A software framework for musical data augmentation.," in *ISMIR*, 2015, pp. 248–254.

[36] B. McFee, C. Jacoby, E. J. Humphrey, V. Lostanlen, H. van Kemenade, W. Pimenta, and H. Schreiber, "pescadores/pescador: 2.1.0," Aug. 2019.

[37] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML 2009*, New York, NY, USA, 2009, ICML '09, p. 41–48, Association for Computing Machinery.