DOMAIN TUNING METHODS FOR BIRD AUDIO DETECTION

Sidrah Liaqat, Narjes Bozorg, Neenu Jose, Patrick Conrey, Antony Tamasi and Michael T. Johnson

University of Kentucky Speech and Signal Processing Lab
Electrical Engineering Dept.
Lexington, KY 40506
mike.johnson@uky.edu

ABSTRACT

This paper presents several feature extraction and normalization methods implemented for the DCASE 2018 Bird Audio Detection challenge, a binary audio classification task, to identify whether a ten second audio segment from a specified dataset contains one or more bird vocalizations. Our baseline system is adapted from the Convolutional Neural Network system of last year's challenge winner *bulbul* [1]. We introduce one feature modification, an increase in temporal resolution of the Mel-spectrogram feature matrix, tailored to the fast-changing temporal structure of many song-bird vocalizations. Additionally, we introduce two feature normalization approaches, a front-end signal enhancement method to reduce differences in dataset noise characteristics and an explicit domain adaptation method based on covariance normalization. Results show that none of these approaches gave significant benefit individually, but that combining the methods lead to overall improvement. Despite the modest improvement, this system won the award for "Highest-scoring open-source/reproducible method" for this task.

Index Terms— audio classification, convolutional neural network, bioacoustic vocalization analysis, domain adaptation

1. INTRODUCTION

The DCASE 2018 Bird Audio Detection Challenge (BADC, DCASE 2018 Challenge Task 3) [2] is a binary audio classification task to determine whether a fixed-length ten second audio segment contains one or more bird vocalizations across a wide variety of bird species and background noise environments. This focuses on the challenging problem of domain adaptation, with the evaluation audio segments to be classified identified as coming from one of three different evaluation datasets, one of which is represented in the training data and two of which are not. This problem of dataset adaptation, also referred to as domain adaptation, domain shift, domain transfer, or dataset bias, is of great interest in a number of domains such as image and audio classification. Recently, the success of deep-learning based approaches requiring large amounts of training data have led to an interest in how to adapt existing well-trained models to new, smaller datasets.

The particular domain of the BADC is that of bioacoustics signal processing and analysis. Currently bioacoustics research often requires extensive amounts of manual labor for segmentation, detection and labeling of voice activity from hours of field recordings [3], and because of this automated analysis of bioacoustics data can be a powerful noninvasive and economical tool for monitoring the diversity, migration patterns [4] and ecosystem health [5] of vocally active animal species. In recent years, speech processing and machine learning techniques for human speech have begun to be used to study animal communication for detection and classification, with applications to censusing [6], understanding the effect of noise on animal communication [7], and other areas of acoustic ecology and ethology. The emphasis of the BADC is to develop a highly generalizable and robust bird classification task that is robust across species and acoustic environments. Although it is presented as a detection problem, it is not "detection" in the sense of typical bioacoustics terminology because it does not involve locating the start and end points of the individual vocalizations.

Our team's submission for the BADC is based on a Convolutional Neural Network (CNN) structure adapted from the baseline architecture of last year's challenge *bulbul* [1]. Nearly all of the top performers of last year's challenge were based on a similar structure, using time-frequency features such as Mel-frequency spectrogram or cepstral features as input to a CNN architecture. Using this baseline, we have introduced three specific modifications to the front-end feature processing methods.

The first of these is adjustment of the time and frequency resolution, which was fairly consistent across many of last year's challenge submissions. The idea behind this change, described in more detail in Section 4.1, is that the variations in the vocalizations of many bird species, especially passerines (songbirds), have a much finer spectral and/or temporal structure than human speech.

The second modification, described in Section 4.2, is the introduction of acoustic signal enhancement, specifically Log-Spectral Amplitude (LSA) estimation combined with Iterative Minimal Controlled Recursive Averaging (IMCRA). The idea behind this is not simply for signal enhancement, which does not typically give improvement

to neural-network based speech or audio classification systems, but instead as a type of dataset normalization intended to decrease the differences between the background noise characteristics of the different datasets.

The third modification, described in Section 4.3, is an explicit domain adaptation technique that applies a source-target covariance transform to the underlying features for an input vocalization based on which dataset it is from.

This paper is organized as follows: in the next section a brief description of data used for training and testing the neural network is provided. Section 3 gives an overview of the baseline system, and section 4 introduces each of the improvements that were implemented to the baseline system in further detail. Section 5 gives results and discussion followed by conclusion in Section 6.

2. DATA

The data provided for the challenge consists of audio recordings from three development datasets and three evaluation datasets which are normalized in amplitude, saved as a 16-bit single channel PCM at a 44.1kHz sampling frequency [2]. Each development dataset has a metadata file associated with it, with a binary label to mark bird presence or absence. The labels are manually annotated by visual analysis of the spectrograms and listening to the audio clips, resulting in a small number of mislabeled files.

The development datasets include Birdvox-DCASE-20k, Warblr10k and Freefield1010. The Birdvox-DCASE-20k dataset was recorded during autumn 2015 in Ithaca, NY, USA as part of a bioacoustics monitoring project. About half of the 20000 files contain at least one bird vocalization [8]. The Birdvox-DCASE-20k dataset was originally recorded at a 24kHz sampling rate and was resampled to 44.1kHz to match the other challenge datasets, and therefore contains no content above 12kHz.

The Warblr10k dataset consists of 8000 audio clips recorded using smartphones, crowdsourced by users of Warblr app in the United Kingdom, with 75.6% of the recordings labeled for bird presence. The Freefield1010 dataset [9] consists of 7690 audio segments derived from files with the field-recording tag in the Freesound crowdsourced global audio archive, with about 25% of dataset labeled as having one or more birds present.

The evaluation data includes 2000 files from Warblr10k, 6620 files from Chernobyl, and 4000 files from PolandNFC. The Chernobyl dataset was collected from the Chernobyl Exclusion Zone as a part of the Transfer Exposure Effects (TREE) project to study the long-term effects of the Chernobyl accident on ecology. The PolandNFC dataset was collected along the Baltic coast of Poland during autumn of 2016. In addition, there was a randomly selected smaller subset of the Warblr10k and Chernobyl (but not PolandNFC) evaluation datasets consisting of approximately 1000 files used for posting ongoing results on the challenge leaderboard. We refer to this as the Leaderboard Evaluation dataset and all testing results in this paper are on this dataset unless otherwise specified. Results are given as Area under the Curve (AUC) of the Receiver Operating Characteristic curve of the submitted prediction probabilities.

Each dataset has unique characteristics in terms of ambient background noise, species present in the recordings, and variety of non-avian interfering sound sources. Warblr, the only dataset present in both training and evaluation sets, and Freefield are crowd-sourced and therefore represent a much wider range of background sound sources, but both the datasets are from UK and therefore have similar species. All three other datasets are remote monitoring data with internal consistency across species and conditions, but vary widely in location and habitat.

It should also be noted that all five of these datasets contain a large number and wide range of bird species that includes both passerine and non-passerine species. Passerine vocalizations tend to have distinct song-like patterns moving around a single or dual frequency (due to the dual-frequency action of the syrinx sound production mechanism), while non-passerine vocalizations are often broadband with unique spectral characteristics. The binary task of classifying whether one or more bird calls is present or non-present inherently represents recognition and classification of many different sound event characteristics.

3. BASELINE SYSTEM

The baseline CNN system was modeled after the baseline architecture of last year's challenge *bulbul* [1] architecture as distributed by the challenge organizers. This is a feed forward network with four 2D CNN layers followed by three dense layers, as shown in detail in Figure 1. The neural network was trained using the log Mel filter bank energy features extracted from small frames of each audio signal. Vocalizations are resampled to a 22.05kHz sampling frequency and divided into 46ms frames using a Hamming window function, with a step size of 14ms, yielding an overlap of 70% across frames. A Fast Fourier Transform is computed, and then Mel filter banks with 80 bands are calculated across a frequency range from 50Hz to 12kHz. The logarithm of the normalized sum magnitude of the filter bank energies is computed for each window. These features were normalized to range between 0 and 1 before feeding to the network input.

Batch normalization layers along with dropout layers were employed in the neural network for improved regularization. The dropout layer has a dropout rate of 0.5. The network also uses L2 layer at the end of CNN layers with a regularization parameter of 0.01. For training, ADAM optimizer is used with an initial learning rate of 0.001. The learning rate was reduced by a factor of 0.2 if there was no improvement in validation accuracy over five consecutive

epochs. The network is trained on binary cross entropy loss using accuracy as a metric. Intermediate activation layers were leaky RELU with a final prediction probability output computed using a sigmoid activation function. Training was done on batches of 16 audio samples over 30 to 40 epochs. Optional data augmentation was built into the system, using a simple cyclic pitch and time shifting approach. For this, the pitch shift was limited to 5% but cyclic time shift could be as much as 90%.

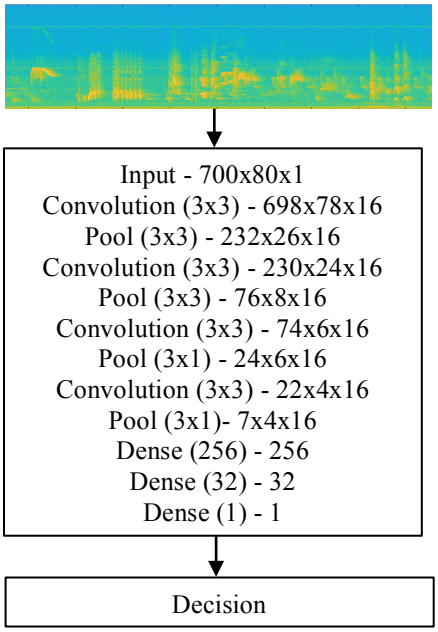

Figure 1: Baseline network architecture

Experiments were conducted both within individual datasets and across datasets using the developmental datasets. Since the number of positive examples varies within each dataset, the selection of positive and negative examples was equalized using class weights to avoid a mismatch in class representations. The dataset organization used for the experiments is given in Table 1 and Table 2.

For testing on the final evaluation dataset, the network was trained on all three development datasets, Birdvox-DCASE-20k, Warblr10k and Freefiled1010, combined. The datasets were shuffled to keep the training of the network impervious to the sequence in which examples are presented to it. Since the number of examples in Freefield1010 and Warblr10k datasets is almost equal, the class imbalance evens out, so equal weights were assigned to positive and negative examples. The development dataset was split into 33905 training examples (95%) and 1784 validation examples (0.05%). The test dataset consisted of the complete evaluation dataset having 12620 examples; 6620 from Chernobyl, 2000 from warblr10k-eval and 4000 from PolandNFC dataset.

Table 1: Within dataset experiments

Dataset	Train (80%)	Test (0.15%)	Validation (.05%)
Birdvox_20k	16000	3000	1000
Freefiled1010	6152	1153	385
Warblr10k	6400	1200	400

Table 2: Cross dataset experiments

Training datasets (84% training and 16% validation)	Test dataset	Class weights	
		-ve	+ve
BirdVox+ freefield	Warblr	43%	57%
Freefield + warblr	BirdVox	50%	50%
BirdVox + Warblr	Freefield	57%	43%

4. PROPOSED IMPROVEMENTS AND RESULTS

4.1. Temporal and frequency resolution

Most bioacoustics signals are nonstationary, like human speech, with changing frequency content over time. Choosing the frame length is a tradeoff between spectral and temporal resolution, with a long frame yielding better spectral resolution but poorer temporal resolution, and vice versa. Bird vocalizations, especially passerine songs, typically have higher frequencies and very fast temporal patterns compared to human speech, with modulations as fast as a few milliseconds [10]. Most of the previous challenge systems, including the baseline bulbul system, have a window size that is relatively long for typical song-bird vocalizations, which would prevent feature representation of small time-scale modulations and transients. Prior work for the BirdClef2017 challenge has also considered resolution issues for bird call recognition [11].

To investigate this, we experimented with changing both the temporal resolution by varying the step and window sizes, as well as changing the frequency resolution by varying the number of filter banks. The high temporal resolution condition used a window size of 12ms with 80 Mel-spaced filter banks (dimension 1669x80), while the high spectral resolution condition used a window size of 32ms along with 160 Mel-spaced filter banks (dimension 624x160). Leaderboard Evaluation results, shown in Table 3 below, indicate that the increased temporal resolution has little impact while the increased spectral resolution has a negative impact.

Table 3: Temporal and spectral resolution results

	AUC	Acc	Val acc
Baseline (B)	86.83	0.89	0.88
High-res temporal (HT)	86.43	0.90	0.89
High-res frequency (HF)	83.54	0.89	0.87

4.2. Signal enhancement

In noisy environments, anthropogenic noise and adverse causes may mask bird song, especially the notes occurring at lower frequencies. In urban environments, birds may modify their songs to low frequency regions to minimize masking effect by anthropogenic noise [12]. Each of the datasets in the BADC has a unique set of background noise characteristics. Our hypothesis for the cross-dataset conditions of the BADC is that applying a front-end signal enhancement may increase similarity across datasets and allow the network to generalize to new noise conditions.

To investigate this, we used the Improved Controlled Recursive Algorithm (IMCRA) noise tracking approach with a log-spectral amplitude estimation technique as proposed by Cohen [13], to implement signal enhancement on all datasets. Noise estimation is updated by averaging the past spectral power values using smoothing parameters that are adjusted with the probability of target signal presence within sub bands. IMCRA includes two iterations of smoothing and minimum tracking. During the first iteration the signal presence probability is detected in each frequency band, and in the second iteration the minimum tracking will be updated by smoothing parameter both in time and frequency domains. This was used with High Temporal features. Results, shown in Table 4 below, show a small degradation to the results from this approach.

Table 4: IMCRA-LSA Signal enhancement results

	AUC	Acc
HT	86.43	0.90
HT enhanced (HTE)	84.47	0.88

4.3. Domain adaptation

One of the primary issues with this challenge problem is the training/test mismatch. There have been a number of different methods suggested for domain adaptation in the image processing literature, to allow well-trained models to be quickly used on new smaller datasets.

In this work we have implemented the CORAL domain adaptation method described in [14] which aligns the second order statistics of source dataset to the target dataset before training the network. This amounts to whitening the training input and then re-coloring it with the covariance characteristic of a chosen target dataset. For this task, we do normalization frame-wise based on the 80 dimensional frequency features. For a baseline 700x80 feature matrix, the normalized feature matrix is

$$\mathbf{A}' = \mathbf{A} \times \left(\mathbf{C}_{\text{source}} + \mathbf{I}\right)^{-\frac{1}{2}} \times \left(\mathbf{C}_{\text{target}} + \mathbf{I}\right)^{\frac{1}{2}}$$

where $\mathbf{A}$ is the original feature matrix, $\mathbf{C}_{\text{source}}$ and $\mathbf{C}_{\text{target}}$ are the 80x80 source and target covariance matrices, and $\mathbf{I}$ is an identity matrix added for regularization.

Although the choice of a target is arbitrary, since the Warblr10k dataset was present in both development and evaluation dataset, we selected it as the target. In the domain adaptation experiments, feature matrices from all other datasets were transformed to the covariance characteristic of the Warblr10k dataset before being applied to the network. Results, shown in Table 5 below, again show little change due to the domain adaptation method.

Table 5: Covariance normalizations results

	AUC	Acc	Val acc
Baseline (B)	86.83	0.89	0.88
Covariance normalized (CN)	86.61	0.87	0.88

4.4. Combined systems

In addition to the individual modifications, several combined systems were implemented. This includes combining high-temporal resolution features with enhancement and covariance normalization, a score fusion system that consisted of a fully connected three-layer neural net using the second from the last layer of each of three individual networks (concatenated 3 32x1 outputs) followed by two dense layers, and several different combinations of simple averaging. Results are shown in Table 6, and indicate that both direct and weighted score fusion methods lead to significant improvement.

Table 6: Composite system results

	AUC
Baseline (B)	86.83
Sequence HT→SE→CV	70.05
Score fusion –Parallel B/HT/HF → 3 FC layers	89.54
Boosting (prediction averaging) (B, HT, HTE, CN, HF)	89.94
Boosting (weighted prediction averaging) (Score Fusion, B, HT, HF)	90.25

5. DISCUSSION AND CONCLUSION

This paper has presented CNN-based methods for the DCASE 2018 Bird Audio Detection challenge, including experiments adjusting the temporal and frequency resolution, signal enhancement for the purpose of dataset normalization, and a method for explicit domain adaptation based on covariance normalizations. Overall results on the leaderboard evaluation dataset show that although none of these approaches gave significant improvements on overall AUC or accuracy metrics, but that combining them together using score fusion approaches were beneficial, improving AUC from a baseline of 86.83 to 90.25 on the leaderboard dataset. The system scored 83.9% on the challenge evaluation dataset, winning the award for Highest-scoring open-source/reproducible method on this challenge task.

6. REFERENCES

[1] Grill, T. and J. Schlüter. *Two convolutional neural networks for bird detection in audio signals*. in *2017 25th European Signal Processing Conference (EUSIPCO)*. 2017.

[2] Stowell, D., et al., *Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge.* 2018.

[3] Christine Erbe, M.L.D., *Animal Bioacoustics.* 2017.

[4] Stepanian, P.M., et al., *Extending bioacoustic monitoring of birds aloft through flight call localization with a three-dimensional microphone array.* Ecol Evol, 2016. **6**(19): p. 7039-7046.

[5] Ross, S.R.P.J., et al., *Listening to ecosystems: data-rich acoustic monitoring through landscape-scale sensor networks.* Ecological Research, 2018. **33**(1): p. 135-147.

[6] Adi, K., M.T. Johnson, and T.S. Osiejuk, *Acoustic censusing using automatic vocalization classification and identity recognition.* J Acoust Soc Am, 2010. **127**(2): p. 874-83.

[7] Catherine, P.O., *Chapter 2: Effects of noise pollution on birds: A brief review of our knowledge.* Ornithological Monographs, 2012. **74**(1): p. 6-22.

[8] Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., & Bello, J.P., *Birdvox-full-night : a Dataset and Benchmark for Avian Flight Call Detection.* 2018.

[9] Stowell, D. and M.D. Plumbley, *An open dataset for research on audio field recording archives: freefield1010.* 2013.

[10] Clemins, M.T.J.a.P.J., *Hidden Markov Model Signal Classification*, in *Comparative Bioacoustics: An Overview*. 2017, Bentham Science. p. 358-414.

[11] Sevilla, A. and H. Glotin. *Audio Bird Classification with Inception-v4 extended with Time and Time-Frequency Attention Mechanisms*. in *CLEF*. 2017.

[12] Wood, W.E. and S.M. Yezerinac, *Song Sparrow (Melospiza melodia) Song Varies with Urban Noise (Le Chant de Melospiza melodia Varie avec le Bruit Urbain).* The Auk, 2006. **123**(3): p. 650-659.

[13] Cohen, I., *Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging.* IEEE Transactions on Speech and Audio Processing, 2003. **11**(5): p. 466-475.

[14] Sun, B., J. Feng, and K. Saenko, *Return of Frustratingly Easy Domain Adaptation.* 2015.