FUSING SHALLOW AND DEEP LEARNING FOR BIOACOUSTIC BIRD SPECIES CLASSIFICATION

Justin Salamon[1,2,], Juan Pablo Bello[1], Andrew Farnsworth[3] and Steve Kelling[3]*

[1]Music and Audio Research Laboratory, New York University, New York, NY, USA
[2]Center for Urban Science and Progress, New York University, New York, NY, USA
[3]Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA

ABSTRACT

Automated classification of organisms to species based on their vocalizations would contribute tremendously to abilities to monitor biodiversity, with a wide range of applications in the field of ecology. In particular, automated classification of migrating birds' flight calls could yield new biological insights and conservation applications for birds that vocalize during migration. In this paper we explore state-of-the-art classification techniques for large-vocabulary bird species classification from flight calls. In particular, we contrast a "shallow learning" approach based on unsupervised dictionary learning with a deep convolutional neural network combined with data augmentation. We show that the two models perform comparably on a dataset of 5428 flight calls spanning 43 different species, with both significantly outperforming an MFCC baseline. Finally, we show that by combining the models using a simple late-fusion approach we can further improve the results, obtaining a state-of-the-art classification accuracy of 0.96.

Index Terms— Convolutional neural networks, bioacoustics, flight calls, deep learning, data augmentation.

1. INTRODUCTION

Automatic classification of animal vocalizations has great potential to enhance the monitoring of species movements and behaviors. This is particularly true for monitoring nocturnal bird migration, where automated classification of migrants' flight calls could yield new biological insights and conservation applications for birds that vocalize during migration. Among an increasingly important array of bioacoustic tools for conservation science [1] that describe presence, abundance, and behavior of vocal species, there is a significant body of research on automatic species classification from audio (e.g. [2, 3, 4, 5, 6, 7, 8, 9]). See [10] for a detailed survey of automatic birdsong recognition. Recently, a number of approaches have been proposed that employ generalizable machine learning techniques that can be easily adapted to multiple species [7, 11, 12]. However, these studies were focused on bird song (and marine mammals), not flight calls. Flight calls are species-specific vocalizations produced primarily during periods of sustained flight (i.e. nocturnal migration). Among other differences from vocalizations analyzed in the aforementioned studies, flight calls are primarily single note vocalizations that are less than 200 ms long, whereas most songs contain several types of notes and may vary from seconds to minutes in duration. Studies focusing specifically on automatic flight call classification include [13, 14, 15, 16].

For fully automated bioacoustic migration monitoring based on flight calls, several challenges must be addressed: distinguishing between flight calls and confounding factors such as geophony (e.g. wind, water), biophony (e.g. insects, frogs) and anthrophony (e.g. speech, transportation); distinguishing between the flight calls of a large number of related target species; and potentially having to deal with temporally overlapping calls. Here we focus on one of these challenges, namely distinguishing between a large number of target species, i.e. large-vocabulary classification.

In this study we explore different state-of-the-art classification techniques for large-vocabulary bioacoustic classification. In particular, we contrast the unsupervised dictionary learning approach presented in [16] with a deep convolutional neural network architecture which, in combination with data augmentation techniques, has been shown to outperform the dictionary learning approach for environmental (not bioacoustic) sound classification [17]. To the best of our knowledge this is the first application of a deep convolutional neural network to flight call classification, and one of its first applications to bioacoustic classification in general. It is also the first application, as far as we know, of audio data augmentation (beyond simple time shifts) to bioacoustic classification. Furthermore, we examine whether the dictionary learning method (which we consider a "shallow learning" technique because it learns a single representational layer from the input data) and the deep learning architecture are complementary to each other and whether combining their output in a late-fusion fashion can yield improved classification accuracy.

2. METHOD

2.1. Unsupervised dictionary learning

Our "shallow learning" approach [16] is based on dictionary learning. We use the *spherical k-means* algorithm [18] to learn a dictionary of representative code words, and then encode our data against the learned dictionary. In this variant of the k-means clustering algorithm [19] the centroids are constrained to have unit L2 norm (preventing them from becoming arbitrarily large or small), and represent the distribution of meaningful directions in the data. The algorithm is efficient and scalable, competitive with slower and more complex techniques such as sparse coding, and it has been shown that its resulting set of centroids can be used as bases (a dictionary) for mapping new data into a feature space which reflects the discovered regularities [20, 18, 8]. Spherical k-means (SKM) has been exploited for classifying music [21], birdsong [8], urban (environmental) sounds [22, 23] and most recently flight calls [16].

Please direct correspondence to: justin.salamon@nyu.edu.
This work was partially supported by National Science Foundation award NSF IIS-1633259 and a Google Faculty Award.

We learn the dictionary from time-frequency patches (TF-patches) extracted from the log-scaled mel-spectrogram representation of each audio clip. The mel-spectrogram consists of 40 bands between 2000-11025 Hz, and is computed using the Essentia library [24] with a Hann analysis window of 11.6 ms and a hop size of 1.45 ms. Each TF-patch spans the entire frequency range and has a duration of 46 ms. Prior to learning the dictionary the data is PCA-whitened (this improves the discriminative power of the learned features [18]) keeping components to explain 99% of the variance in the data as in [21]. We then learn a dictionary with 256 codewords (centroids), and encode data against it by taking the matrix product of a datum's input representation (mel-spectogram) with the dictionary matrix, and summarizing the result over the time-axis using three summary statistics (mean, standard deviation and maximum). The resulting features are used to train (and test) a Support Vector Machine (SVM) classifier with a radial basis function kernel implemented in Python using Scikit-learn [25]. For further details about the algorithm and the choice of parameters the reader is referred to [16].

2.2. Deep convolutional neural network

We use the deep convolutional neural network (CNN) architecture proposed for environmental sound classification in [17]. The model is comprised of 3 convolutional layers interleaved with 2 pooling operations, followed by 2 fully connected (dense) layers. We use the same log-scaled mel-spectrograms described in the previous section as the input to the network, the only exception being that we increase the number of bands to 128 given the model's increased capacity. Since the clips in our evaluation dataset (described below) are of varying duration, we trim each clip to the middle 150 ms. To make the network invariant to small time-shifts, each clip is shifted in time by up to ± 10 ms to generate a total of 13 clips. Following [17], the network is parameterized as follows:

- ℓ_1: 24 filters with a receptive field of (5,5), followed by (4,2) strided (non-overlapping) max-pooling over the time and frequency dimensions respectively, and a rectified linear unit (ReLU) activation function $h(x) = \max(x, 0)$.
- ℓ_2: 48 filters with a receptive field of (5,5), followed by (4,2) strided max-pooling and a ReLU activation function.
- ℓ_3: 48 filters with a receptive field of (5,5), followed by a ReLU activation function (no pooling).
- ℓ_4: 64 hidden units, followed by a ReLU activation function.
- ℓ_5: 43 output units, followed by a softmax activation function.

During training the model optimizes cross-entropy loss via mini-batch stochastic gradient descent [26]. We use a constant learning rate of 0.01 and apply dropout [27] to the input of the last two layers with probability 0.5. L2-regularization is applied to the weights of the last two layers with a penalty factor of 0.001. The model is trained for 100 epochs and is checkpointed after each epoch. A validation set is used to identify the parameter setting (epoch) achieving the highest classification accuracy. The CNN is implemented in Python using Lasagne [28], and data stream multiplexing (for training) is implemented using Pescador [29].

In [17] we also showed that the performance of the CNN model can be improved by increasing the size of the training set using *data augmentation*, that is, the application of audio deformations that modify the audio signal while maintaining the semantic validity of the recording's label. Following this, we apply the following augmentations: adding background noise (from 4 different field recordings captured at night containing geophony), dynamic range compression (using 4 parameterizations: music, film, speech, radio), pitch shifting (by 4 conservative values of -0.5, -0.25, 0.25, 0.5 semitones, and 4 less conservative values of -2, -1, 1, 2 semitones), and time stretching (by 4 ratios: 0.81, 0.93, 1.07, 1.23). The augmentations are applied using the MUDA library [30], to which the reader is referred for further details about the implementation of each audio deformation.

An important point for this study is that the deformations applied should maintain the semantic, and in our case biological, validity of the labels. That is, the resulting flight calls must still resemble plausible calls by each species after augmentation. Variation in these signals is typically significantly greater among species than within species, in particular with differences in syllabic structure, duration, and frequency [31, 32]. However, individual variation in flight calls of nocturnally migrating birds has been documented by a small number of studies, primarily in call frequencies and duration [33, 34, 31, 35] as well as in call structure [36]. For example, American Redstart (Setophaga ruticilla) typically exhibits variation of ± 15 ms in duration, ± 2.3 kHz in frequency, and of five discrete call structure variants [34]. Of the augmentations applied in this study, only pitch shifting and time stretching have the potential to invalidate a recording's label. To avoid this, the shifting/stretching parameter ranges were chosen to be within the naturally occurring ranges for the majority of the species in the evaluation dataset, which is described further down.

2.3. Baseline

To benchmark the shallow and deep learning approaches, we also provide the results from the baseline method implemented in [16]. The method extracts Mel-Frequency Cepstral Coefficients (MFCC) [37] using 40 mel bands and keeps the first 25 coefficients, which are summarized over time using 11 summary statistics as in [38]. The resulting feature vectors are used to train (and test) an SVM classifier with a radial basis function kernel.

3. EVALUATION

For evaluation we use the publicly available CLO-43SD dataset [16]. The dataset is comprised of 5428 audio clips of flight calls from 43 different species of North American wood-warblers (in the family Parulidae). The clips come from a variety of recording conditions, including clean recordings obtained using highly-directional shotgun microphones, recordings obtained from noisier field recordings using omnidirectional microphones, and recordings obtained from birds in captivity using the method described in [35]. Every clip contains a single flight call from one of the 43 target species. A list of the species included in this dataset is provided in supplementary document "S1 Table" of [16].

The methods compared in this study are evaluated in terms of classification accuracy. We divide the CLO-43SD dataset into 5 folds and perform 5-fold cross validation, and report the results as a box-plot generated from the per-fold accuracies. For identfying the best training epoch for the CNN model we use 1 of the 4 training folds as a validation set, and train on the remaining 3 folds.

4. RESULTS AND DISCUSSION

The results for the MFCC baseline, SKM and CNN models are provided in Figure 1 (left of the dashed line). Mean accuracies are indicated by the red squares. We see that both models based on

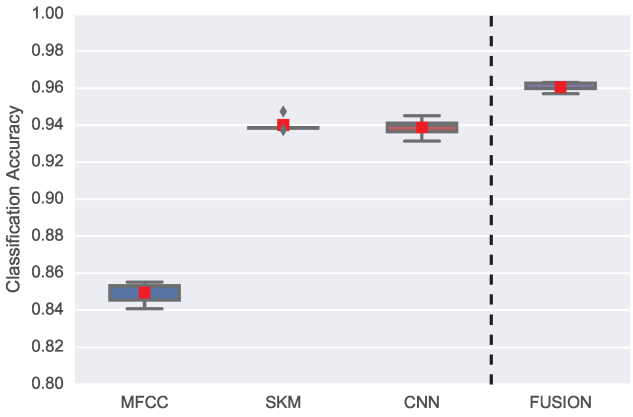

Fig. 1. Classification accuracy: baseline (MFCC), dictionary (shallow) learning (SKM), deep convolutional neural network (CNN), and late fusion of the SKM and CNN models (FUSION).

feature learning significantly outperform the MFCC baseline, obtaining mean classification accuracies of 0.94 (compared to 0.85 for the baseline). Interestingly, the CNN does not outperform the SKM model, unlike the results observed in [17]. It might be possible to achieve some improvement through further exploration of the CNN architecture space. Alternatively it might be the case that, despite the use of data augmentation, there are still some species for which there aren't enough samples in the dataset to reflect the range of natural variation in the species' flight call. To assess the influence of data augmentation on the CNN model, we also trained it on the original dataset without augmentation. This yielded a lower classification accuracy of 0.92 (0.916), confirming the beneficial influence of augmentation on the performance of the CNN model.

Given the results, we wanted to see whether the SKM and CNN models were making the same predictions (and mistakes) or whether they were behaving differently. To answer this, in Figure 2 we plot the *difference* between the confusion matrix yielded by the CNN model and the confusion matrix yielded by the SKM model. Along the diagonal, positive (red) values indicate the CNN makes more correct predictions and negative (blue) values indicate the SKM model makes more correct predictions for the corresponding species. Off the diagonal, positive (red) values indicate greater confusion by the CNN model for the corresponding pairs of species, while negative (blue) values indicate greater confusion by the SKM model for the corresponding pairs of species.

The plot indicates that the two models are in fact making quite a number of different predictions. Interestingly, not only do they make different types of confusions (i.e. between different pairs of species) as indicated by the non-zero values off the diagonal, but also each model performs better for a specific set of species, as indicated by the non-zero values along the diagonal of the matrix. For example, the CNN correctly classifies 8 more BTBW (Black-throated Blue Warbler) clips compared to the SKM model, whereas the latter correctly classifies 10 more BTNW (Black-throated Green Warbler) clips. This indicates that the output of the two models is potentially complementary. To verify this, we evaluated an "oracle" model which, given the predictions made by the two models, always chooses the correct prediction if available (by comparing them to the reference labels). The oracle yielded a classification accuracy of 0.97 (0.974), suggesting that given the right fusion approach we could surpass the 0.94 accuracy obtained by either model on its own.

To this end, we experimented with a number of late-fusion techniques, in which we combine the output of the models post-prediction. To do so, we require the "confidence" of each model in its predictions. For the CNN we simply use the softmax activation values returned by the last layer of the model. The 43 values (one for each class) sum to 1 and can be treated as probabilities. For the SKM model, which uses an SVM classifier, we obtain the confidence value for each clip by applying Platt scaling to the distance of the clip from the SVM's separation hyper-plane [39]. This results in 43 values for each clip (one per class), which also sum to 1. To fuse the confidence values returned by the two models, we experimented with two approaches: the first involved a simple combination of the confidence values by taking their arithmetic or geometric mean, and then making a new prediction for each clip by taking the argmax over the resulting confidence values. The second approach involved learning the fusion by treating the confidence values of the two models as features and training a third model (a discriminative classifier) to predict the label. We experimented with a number of models including SVM with different kernels, Random Forest, Logistic Regression and Naive Bayes. Curiously, simply taking the geometric mean followed by the argmax produced as good results as the best performing learned fusion (yielded by the SVM) – a mean classification accuracy of 0.96. The result is displayed to the right of the dashed line in Figure 1, and the improvement over the individual SKM and CNN models is statistically significant ($p = 0.0003$ according to a paired two-sided t-test).

5. SUMMARY

The automated classification of migrating birds' flight calls has the potential to yield new biological insights and conservation applications for birds that vocalize during migration. In this paper we explored two state-of-the-art classification techniques for large-vocabulary bird species classification from flight calls: a "shallow learning" unsupervised dictionary learning method and a deep convolutional neural network combined with data augmentation. The models were evaluated on a dataset of 5428 flight calls from 43 different species, and were compared against a baseline model based on MFCCs. We showed that the two models perform comparably, yielding a mean classification accuracy of 0.94 and significantly outperforming the MFCC baseline (0.85). We also compared the performance of the CNN model with and without augmentation and noted that the augmentation contributes to the performance of the model. By examining the difference between the confusion matrices yielded by the CNN and SKM models we noted that they make different types of mistakes and tend to perform better on specific sets of species. This led us to experiment with late-fusion techniques, ultimately resulting in a state-of-the-art classification accuracy of 0.96 using a simple geometric mean fusion approach.

In the future we intend to experiment further with the CNN architecture to study its influence on performance. We also intend to apply the insights gained in this study, such as the utility of data augmentation and model fusion, to the task of bird species recognition in continuous audio streams such as the ones captured by remote acoustic sensors.

6. REFERENCES

[1] Paola Laiolo, "The emerging significance of bioacoustics in animal species conservation," *Biological Conservation*, vol.

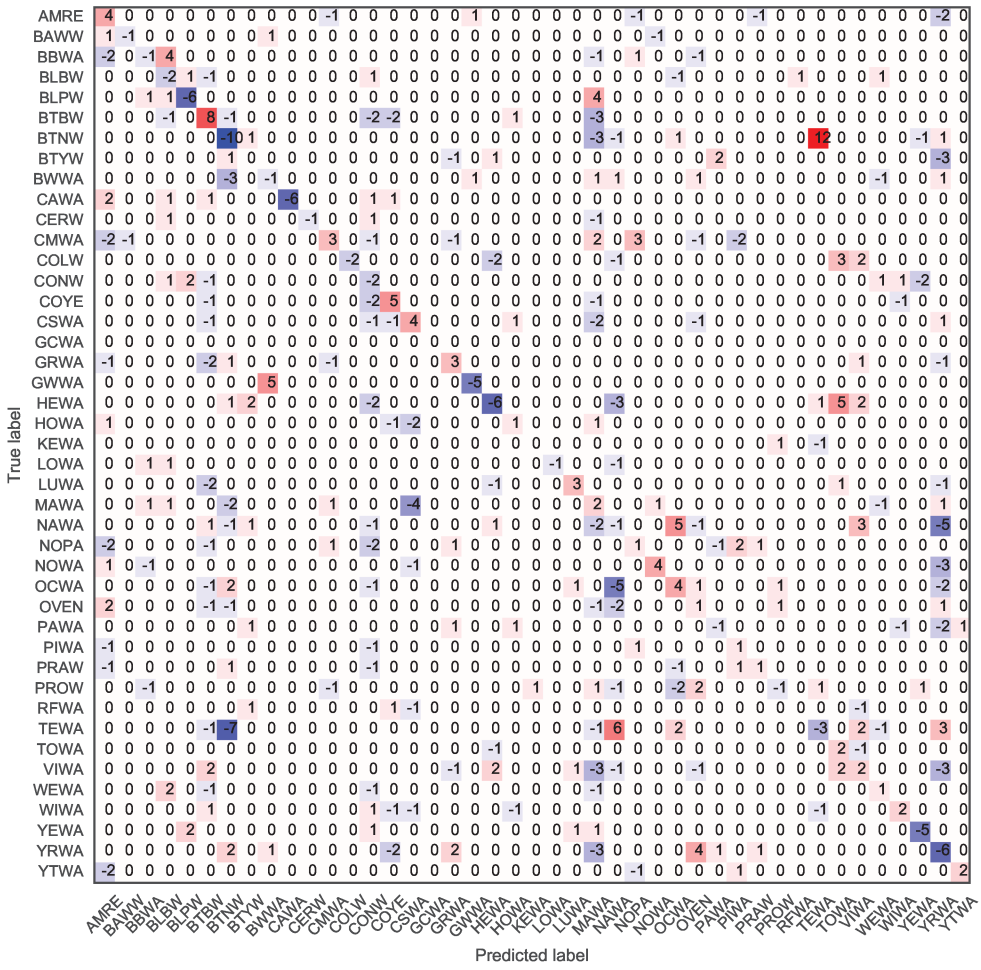

Fig. 2. Difference between the confusion matrix yielded by the CNN model and the confusion matrix yielded by the SKM model.

143, no. 7, pp. 1635–1645, 2010, Conservation planning within emerging global climate and economic realities.

[2] Miguel A Acevedo, Carlos J Corrada-Bravo, Héctor Corrada-Bravo, Luis J Villanueva-Rivera, and T Mitchell Aide, "Automated classification of bird and amphibian calls using machine learning: A comparison of methods," *Ecological Informatics*, vol. 4, no. 4, pp. 206–214, 2009.

[3] T. Damoulas, S. Henry, A. Farnsworth, M. Lanzone, and C. Gomes, "Bayesian classification of flight calls with a novel dynamic time warping kernel," in *9th Int. Conf. on Machine Learning and Applications (ICMLA)*, Dec 2010, pp. 424–429.

[4] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1524–1534, 2010, Pattern Recognition of Non-Speech Audio.

[5] S. Bastas, M. Wadood Majid, G. Mirzaei, J. Ross, M.M. Jamali, P.V. Gorsevski, J. Frizado, and V.P. Bingman, "A novel feature extraction algorithm for classification of bird flight calls," in *IEEE Int. Symposium on Circuits and Systems (ISCAS),*, Seoul, South Korea, May 2012, pp. 1676–1679.

[6] T. Mitchell Aide, Carlos Corrada-Bravo, Marconi Campos-Cerqueira, Carlos Milan, Giovany Vega, and Rafael Alvarez, "Real-time bioacoustics monitoring and automated species identification," *PeerJ*, vol. 1, pp. e103, 7 2013.

[7] O. Dufour, T. Artieres, H. Glotin, and P. Giraudet, "Clusterized mel filter cepstral coefficients and support vector machines for bird song identification," in *Soundscape Semiotics - Localization and Categorization*, number 2013, pp. 89–93. INTECH, 2013.

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

[9] T. Ganchev, O. Jahn, M. Isaac Marques, J. Maimone de Figueiredo, and K.-L. Schuchmann, "Automated acoustic detection of vanellus chilensis lampronotus," *Expert Systems with Applications*, vol. 42, no. 15–16, pp. 6098–6111, 2015.

[10] D. Stowell and M. Plumbley, "Birdsong and c4dm: a survey of

uk birdsong and machine recognition for music researchers," Tech. Rep., Centre for Digital Music, Queen Mary, University of London, 2011.

[11] R. Hernandez Murcia, "Bird identification from continuous audio recordings," in *1st Int. Workshop of Machine Learning for Bioacoustics ICML4B joint to ICML 2013*, Atlanta, GA, USA, Jun. 2013.

[12] H. Glotin, Y. Lecun, P. Dugan, C. Clark, X. Halkias, and J. Sueur, Eds., *Proceedings of the 1st Int. Workshop of Machine Learning for Bioacoustics ICML4B joint to ICML 2013*, Atlanta, GA, USA, Jun. 2013.

[13] A. Taylor, "Bird flight call discrimination using machine learning," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3370–3370, 1995.

[14] T. Schrama, M. Poot, M. Robb, and H. Slabbekoorn, "Automated monitoring of avian flight calls during nocturnal migration," in *International Expert meeting on IT-based detection of bioacoustical patterns*, Dec. 2007, pp. 131–134.

[15] M. Marcarini, G.A. Williamson, and L. de Sisternes Garcia, "Comparison of methods for automated recognition of avian nocturnal flight calls," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, Mar. 2008, pp. 2029–2032.

[16] 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. e0166866, Nov. 2016.

[17] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, In Press.

[18] A. Coates and A. Y. Ng, "Learning feature representations with K-means," in *Neural Networks: Tricks of the Trade*, pp. 561–580. Springer, 2012.

[19] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[20] I.S. Dhillon and D.M. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1, pp. 143–175, 2001.

[21] S. Dieleman and B. Schrauwen, "Multiscale approaches to music audio feature learning," in *14th Int. Soc. for Music Info. Retrieval Conf.*, Curitiba, Brazil, Nov. 2013.

[22] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 171–175.

[23] J. Salamon and J. P. Bello, "Feature learning with deep scattering for urban sound analysis," in *2015 European Signal Processing Conference*, Nice, France, Aug. 2015.

[24] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "ESSENTIA: an audio analysis library for music information retrieval," in *14th Int. Soc. for Music Info. Retrieval Conf.*, Brazil, Nov. 2013, pp. 493–498.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[26] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *19th International Conference on Computational Statistics (COMPSTAT)*, Paris, France, Aug. 2010, pp. 177–186.

[27] N. Srivastava, G. E. 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. 1, pp. 1929–1958, 2014.

[28] S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S.K. Sønderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, and J. Kelly, "Lasagne: First release," `https://github.com/Lasagne/Lasagne`, 2015.

[29] B. McFee and E. J. Humphrey, "pescador: 0.1.0," `https://github.com/bmcfee/pescador`, 2015.

[30] B. McFee, E.J. Humphrey, and J.P. Bello, "A software framework for musical data augmentation," in *16th Int. Soc. for Music Info. Retrieval Conf.*, Malaga, Spain, Oct. 2015, pp. 248–254.

[31] A. Farnsworth, *Ecological and evolutionary characteristics of flight-calls of the wood-warblers (Parulidae)*, Ph.D. thesis, Cornell University, Ithaca, NY, USA, 2007.

[32] A. Farnsworth and I. J. Lovette, "Phylogenetic and ecological effects on interspecific variation in structurally simple avian vocalizations," *Biological Journal of the Linnean Society*, vol. 94, no. 1, pp. 155–173, 2008.

[33] W. R. Evans, "Nocturnal flight call of bicknell's thrush," *The Wilson Bulletin*, pp. 55–61, 1994.

[34] W. R. Evans and M. O'Brien, "Flight calls of migratory birds: Eastern north american landbirds," Old Bird Incorporated, 2002.

[35] M. Lanzone, E. Deleon, L. Grove, and A. Farnsworth, "Revealing undocumented or poorly known flight calls of warblers (parulidae) using a novel method of recording birds in captivity," *The Auk*, vol. 126, no. 3, pp. 511–519, Jun. 2009.

[36] E. T. Griffiths, S. Keen, M. Lanzone, and A. Farnsworth, "Can nocturnal flight calls of the migrating songbird, american redstart, encode sexual dimorphism and individual identity?," *PloS one*, vol. 11, no. 6, pp. e0156578, 2016.

[37] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *International Symposium on Music Information Retrieval (ISMIR)*, Plymouth, Mass, USA, Oct. 2000.

[38] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[39] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, pp. 61–74. MIT Press, Cambridge, MA, 1999.