SIMULTANEOUS SEGMENTATION AND CLASSIFICATION OF BIRD SONG USING CNN

Revathy Narasimhan, Xiaoli Z. Fern, Raviv Raich

School of EECS, Oregon State University, Corvallis, OR 97331-5501, USA
{narasimr, xfern, raich}@oregonstate.edu

ABSTRACT

In bioacoustics, automatic animal voice detection and recognition from audio recordings is an emerging topic for animal preservation. Our research focuses on bird bioacoustics, where the goal is to segment bird syllables from the recording and predict the bird species for the syllables. Traditional methods for this task addresses the segmentation and species prediction separately, leading to propagated errors. This work presents a new approach that performs simultaneous segmentation and classification of bird species using a Convolutional Neural Network (CNN) with encoder-decoder architecture. Experimental results on bird recordings show significant improvement compared to recent state-of-the-art methods for both segmentation and species classification.

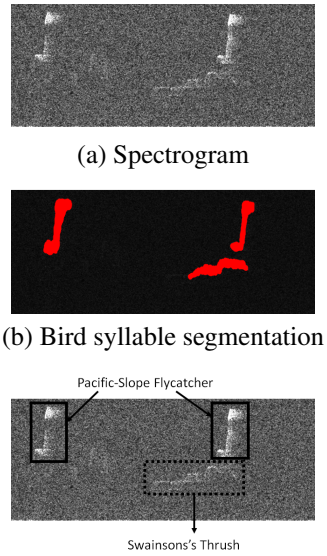
Index Terms— Convolutional Neural Network, encoder-decoder architecture, bioacoustic species classification

1. INTRODUCTION

Machine learning for bioacoustic is an emerging field of research. The recent depletion in animal species has mandated the need to preserve ecosystem biodiversity. An effective tool of tracking animals and understanding biodiversity is via sensors such as microphones, which leads to data in audio signal form. In this paper, we focus on the problem of detecting and recognizing bird species from in-situ recordings of bird songs obtained from their natural habitats.

Fig. 1 shows the typical stages of a bioacoustic species recognition system. Given an audio recording, it is first processed to create a spectrogram as shown in Fig. 1(a) (enhanced to increase readability), which represents the intensity of sound at different frequencies as a function of time. The spectrogram then goes through segmentation to extract bird syllables, as in Fig. 1(b), where each syllable is shown as a red segment representing a single distinct utterance by a bird [1]. Finally, the syllables are analyzed to identify the species of the vocalizing birds, as shown in Fig. 1(c).

Detecting and recognizing bird syllables from natural in-situ recordings is challenging for several reasons. First, there are often multiple birds singing simultaneously in the recording, leading to bird syllables overlapping in time and potentially in frequency. Second, complex environmental noise such as rain or car sound often overlap with bird syllables and occlude the pattern, making the recognition task more challenging.



(a) Spectrogram

(b) Bird syllable segmentation

(c) Bird species recognition

Fig. 1. Stages of a bioacoustic species recognition system.

As suggested by Fig. 1, traditional approaches for bird species recognition decouples the segmentation and classification stages. In segmentation, the goal is to separate the foreground bird syllable from the background, as in [2]. In classification, we focus on the extracted foreground segments and identify their bird species based on the characteristics of the segments, as in [3]. For example, in Fig. 1(c), the bird syllables in the solid boxes are Pacific-Slope Flycatcher and the bird syllable in the dashed box is Swainson's Thrush.

Decoupling the segmentation and classification steps has a critical limitation that the classification performance will heavily depend on the segmentation results. Intuitively, the two tasks are strongly interrelated and can benefit one another mutually. For example, when comparing two different segmentations, the one that leads to higher classification

This work is partially supported by the National Science Foundation grants CCF-1254218, DBI-1356792 and IIS-1055113.

confidence will likely be a better segmentation. In this paper, we propose to simultaneously segment and classify the bird syllables. We build on a deep learning framework that has been successfully used in many computer vision tasks. Specifically, our approach uses a Convolutional Neural Network (CNN) with encoder-decoder architecture [4], which was originally designed for semantic segmentation of images. We evaluate the method on real recordings collected from a forest and the results show significant improvement compared to current state-of-the-art methods.

2. RELATED WORKS

Some of the early methods proposed for bird syllable segmentation in [5, 6, 7] are mainly unsupervised. The boundary of each bird syllable is determined by analyzing signal information such as frequency, energy and amplitude.

There has also been several efforts in supervised learning of segmentation models based on annotated spectrograms. Some methods [8, 9, 10], assume annotations come in the form of time-frequency boxes that contain bird vocalizations. In this work, we are interested in pixel level segmentation, which provides more detailed information for downstream analysis of the bird syllables. In [2], a random forest classifier is trained from spectrograms that are manually annotated at the pixel-level to generate a probability for each pixel of the spectrogram, which indicates the likelihood of it belonging to a foreground bird syllable. The probability map is then Gaussian smoothed and a global threshold is applied to produce a binary mask for bird syllable segments. The main limitation of this method was the use of a single global threshold for segmentation, because different spectrograms may require multiple different thresholds for the bird syllable segments to be extracted precisely.

More recently, the Supervised Hierarchical Segmentation (SHS) method in [11] was proposed to overcome this limitation. SHS applies multiple thresholds to the probability map to generate a hierarchy of candidate segments for each spectrogram. The candidate segments are then evaluated using a learned quality predictor and finally a selection criterion is optimized to identify a set of non-overlapping segments from the hierarchy.

Once the bird syllable segments are extracted from the spectrograms, Multi-instance Multi-label Learning (MIML) has been proposed as an effective bird species classification framework because it naturally models the phenomenon of multiple birds singing in the same recording [12]. The current state-of-the-art for MIML species classification is the Multi-instance Multi-label Learning (MLR) model presented in [13]. MLR learns a model from spectrograms that are labeled with the species present to predict the species of individual bird syllable segment. The recording level prediction can be then obtained by taking the union of the segment level predictions.

A deep autoencoder neural network [14] was proposed to generate a binary mask of the bird syllables in the spectrogram. In [15] and [16] segmentation of bird syllables is achieved by using unsupervised methods such as template matching and energy-based algorithms, then segments are classified using CNN.

All of the efforts mentioned above perform bird syllable segmentation and bird species classification individually, this decoupling leads to propagated errors. But in our approach segmentation and classification of bird syllables are performed simultaneously using the encoder-decoder CNN [4].

3. PROBLEM FORMULATION

We consider a supervised learning approach. The core of our approach is to learn a function that given an input spectrogram, annotates each pixel of the spectrogram as either belonging to a particular bird species or background noise. Below we describe the set up for supervised learning of the annotation function.

Given a set of training spectrograms in the form of $\{X, Y\}$ pairs where X represents the training spectrogram, and Y is the corresponding ground truth pixel-wise annotation. That is, each pixel of X has a corresponding pixel in Y annotated with a label $\in \{0, ..., n+1\}$, where 0 means background, and $1, \ldots, n$ are the species labels and $n+1$ is used to signal the learner to ignore this pixel because its label is unknown/uncertain. The goal is then to learn a function that given a new spectrogram as an input, accurately annotates each pixel of the spectrogram with a label in {0,...,n}.

4. THE DEEP LEARNING METHOD

In this work, we apply the convolution neural network proposed in [4]. The input to the network is a spectrogram, and the output is an image of the same size with pixels labeled as either background or one of the n species.

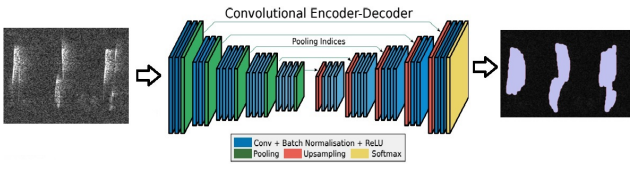


Fig. 2. Convolutional encoder-decoder network in [4] implemented for bird species classification.

Fig. 2 shows the architecture of the network, which has a two-part encoder-decoder structure. Each encoder in the encoder network (the left half) performs convolution with a filter bank to obtain rich feature hierarchies from the input spectrogram. We then batch normalize them and apply element-wise ReLU ($max(0, x)$). Then, 2×2 maxpooling with a stride 2

(non-overlapping window) is applied. The encoder architecture has 5 levels with different number of convolutional layers at each level, the input image is downsampled by a factor of 2 after each layer to provide a more global context for each pixel to identify complex syllable patterns. Due to downsampling of the input image, the boundary details are lost. To preserve them, for every max-pooling layer, the max intensity pixel locations computed by max-pooling are saved and provided to the decoder network. The decoder architecture also has 5 levels, which are structured similar to the encoder network. The decoder network maps the downsampled features from the encoder to input size features using the saved pixel locations from the encoder, which enables the network to perform simultaneous segmentation and classification. The output of the decoder network is a multi-channel feature map that is given to a multi-class soft-max layer to compute class probabilities for each pixel independently. For more detailed description of the network architecture please refer to [4].

The training of the network is achieved by minimizing the cross-entropy loss [17] computed by

$$L(f) = -\sum_{i}\sum_{c} y_{ic} * \ln(f_{ic})$$

where i indexes over all pixels, and c goes over all class labels. y_{ic} is a binary indicator that takes value 1 if pixel i has label c, and f_{ic} is the soft-max probability of class c for pixel i. The filters in the encoder-decoder architecture are trained using backpropagation to minimize the cross-entropy loss.

While testing the soft-max layer predicts a label in $\{0, ..., n\}$ for each pixel. All non-zero pixels are considered as foreground and the connected components are extracted as syllables and a single species label is obtained for each syllable using majority voting. Finally, a recording level prediction can be achieved by taking the union of all species present in the spectrogram.

5. EXPERIMENTAL RESULTS

We evaluate our approach on two real world bird acoustics datasets, the HJA [13] and MLSP [18] datasets for both segmentation and classification of the bird syllables.

5.1. Data description

The MLSP dataset has a total 645 recordings of 19 different bird species. All 645 recordings are labeled at the recording level, i.e., with a list of bird species present in the recording. Of these 645 recordings, 250 recordings are annotated pixel-wise by bird experts with their species. This is achieved by manually marking out the syllables and labeling each syllable with its species. Regions that are uncertain or contain overlapping syllables of different species are marked as the "unknown" class. The HJA dataset contains 550 recordings of 14 different bird species that are a subset of the bird species in MLSP. The HJA recordings have only recording-level labels.

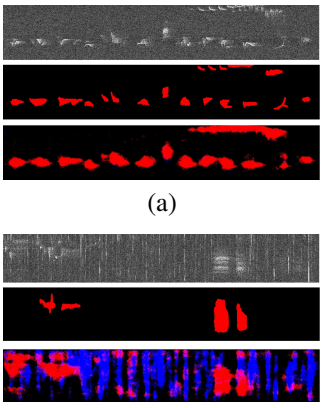
5.2. Segmentation

While our method is designed to simultaneously segment and classify the bird species, it can be easily applied to perform segmentation only by restricting the pixel labels to foreground and background. In the first set of experiments, we compare the segmentation performance achieved by our deep learning method with previous methods on bird song segmentation [11], which uses the same training information as our method.

We use the same experimental set up as used in [11] and train on 50 spectrograms annotated at pixel level with foreground/background labels and evaluated the segmentation performance on the remaining 200 MLSP spectrograms.

5.2.1. Experimental examples

We visualize some of our segmentation results in Fig. 3. In Fig. 3(b), there are many instances where bird syllable is overlapped with rain, but the CNN network was able to successfully identify the bird syllables.



(a)

(b)

Fig. 3. Segmentation results (best viewed in color). In (a) and (b) the first image is the enhanced input spectrogram (for better visualization), second is the ground truth labeled by human experts and third is the result from our approach (red label represents bird syllable and blue label represents rain).

5.2.2. Quantitative evaluation

We measure the segmentation accuracy by the number of correctly predicted pixels as background and foreground over the total number of pixels. We compare our method with the Supervised Hierarchical Segmentation (SHS) method [11] and Neal's method [2] on the MLSP dataset. The results are presented in Table 1. Note that by increasing the threshold in the probability map, the False Positive Rate (FPR) of Neal's method decreases, since all background pixels are predicted

Algorithm	TPR	FPR
SegNet	**0.988**	0.02
SHS	0.78	0.03
Neal-0.2	0.89	0.05
Neal-0.3	0.83	0.04
Neal-0.5	0.76	0.02
Neal-0.5	0.69	0.02
Neal-0.6	0.62	**0.01**

Table 1. Performance comparison of different segmentation methods on the MLSP dataset, in terms of True positive rate (TPR) and False positive rate (FPR).

as background due to high threshold. Our method achieves the significantly higher True Positive Rate (TPR) while maintaining a low false positive rate, suggesting far superior segmentation results. Additionally, note that our method does not require any Gaussian smoothing as a post processing step as in SHS [11] and Neal's method [2].

5.3. Species Classification

Our method is the first to perform simultaneous bird syllable segmentation and classification at the pixel level. We were thus unable to compare to any prior method for pixel-level species prediction. Instead, we compare the recording-level prediction accuracy with prior state-of-the-art. Toward this, we use the 250 pixelwise annotated spectrograms from MLSP for training and test the learned model on 550 spectrograms from the HJA dataset such that we can compare our results with previously reported state-of-the-art on the same dataset. As testing in done on the HJA dataset, during training we consider only the 14 bird classes present in the HJA dataset the additional 5 bird species in the MLSP dataset are labeled as unknown.

5.3.1. Experimental examples

We visualize some of our predictions in Fig. 4, where unique color codes are used to represent bird syllables of different species. Our approach successfully identifies the unique bird syllable patterns in Fig. 4(a) and (b). Also Fig. 4(a) shows that our method successfully eliminates noise such as rain marked as blue in the ground truth annotation of the spectrogram.

5.3.2. Quantitative results

We compare the performance of our method with the previous state-of-the-art result on the HJA dataset achieved by the MLR method [13]. The results are shown in Table 2, using hamming loss proposed in [19] as the evaluation metric. The hamming loss computes the number of object-label pair misclassified i.e, a proper label is missed or a wrong label is predicted. When the performance is ideal, the hamming loss is zero. Thus smaller the value of hamming loss, better is the performance.

The results show that our method achieves significantly lower hamming loss on the HJA dataset. It is important to note that MLR is trained with spectrograms labeled at the recording level, thus is at a disadvantage in this comparison. However, because MLR decouples the segmentation step and species classification step, it is challenging for it to consider additional segment-level training information due to significant misalignment between human annotated segments and the automatically extracted segments. This presents a major limitation for such decoupled methods. Our method, in contrast, does not suffer from this issue and can take advantage of the fine grained annotation to achieve better segmentation and classification performance.

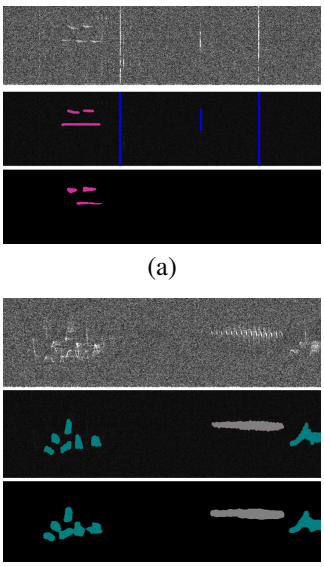


Fig. 4. Classification results (best viewed in color). In (a) and (b) the first image is the enhanced input spectrogram (for better visualization), second is the ground truth labeled by human experts and third is the result from our approach.

Algorithm	hamming loss
SegNet	**9.24**
MLR	11.1

Table 2. Classification performance comparison.

6. CONCLUSION

We present a novel deep learning based method for simultaneous bird syllable segmentation and species prediction from noisy in-situ recordings of bird songs. Our method applies a popular encoder-decoder deep network structure that has been successfully applied to computer vision tasks. Our results show that this method achieves improved performance for both syllable segmentation and species classification in comparison with the current state-of-the-art methods. Future work includes developing recurrent neural network (RNN) model which take into account the temporal relation between bird syllables.

7. REFERENCES

[1] C.K. Catchpole and P.J.B Slater, "Bird song: biological themes and variations," *Cambridge Univ Press*, 2008.

[2] L. Neal, F. Briggs, R. Raich, and X. Z. Fern, "Time-frequency segmentation of bird song in noisy acoustic environments," *ICASSP*, pp. 2012–2015, 2011.

[3] F. Briggs, X. Z. Fern, and R. Raich, "Rank-loss support instance machines for miml instance annotation," *KDD*, pp. 534–542, 2012.

[4] B. Vijay, K. Alex, and C. Roberto, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv:1511.00561v2 [cs.CV]*, 2015.

[5] Anderson S. E., A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *J Acoustical Soc America*, 1996.

[6] Jancovi P. and Kokuer M. K., "Detection of sinusoidal signals in noise by probabilistic modelling of the spectral magnitude shape and phase continuity," *ICASSP*, 2011.

[7] P. Jancovic, M. Kokuer, and Russell M., "Bird species recognition from field recordings using hmm-based modelling of frequency tracks," *ICASSP*, pp. 8252–8256, 2014.

[8] Dan S., Mike W., Yannis S., and Herv G., "Bird detection in audio: A survey and a challenge," *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 13–16, 2016.

[9] Christopher W. C. and Kurt M. F., "Advanced technologies for acoustic monitoring of bird populations," *SERDP Project*, 2009.

[10] "Bioacoustics research program. (2014). raven pro: Interactive sound analysis software (version 1.5) [computer software]. ithaca, ny: The cornell lab of ornithology. available from http://www.birds.cornell.edu/raven," .

[11] T. V. Tjahja, X. Z. Fern, R. Raich, and A. T. Pham, "Supervised hierarchical segmentation for bird song recording," *ICASSP*, pp. 763–767, 2016.

[12] F. Briggs, X. Z. Fern, and R. Raich, "Context-aware miml instance annotation," *IEEE International Conference on Data Mining*, pp. 41–50, 2013.

[13] A. T. Pham, X. Z. Fern, and R. Raich, "Dynamic programming for instance annotation in multi-instance multi-label learning," *IEEE Trans. on PAMI*, 2014.

[14] Ilyas P., "Deep learning for detection of bird vocalisations," *arXiv:1609.08408v1 [cs.SD]*, 2016.

[15] Hendrik V. K., Jan V. B., and Frans W., "A deep neural network approach to the lifeclef 2014 bird task," *CLEF*, 2014.

[16] Herve G., Herve Gl., Willem-Pier V., Robert P., and Alexis J., "Lifeclef bird identification task 2016 the arrival of deep learning," *CLEF*, 2016.

[17] E. Shelhamer J. Long and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[18] F. Briggs, Y. Huang, and R. Raich, "The 9th annual mlsp competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment," *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–8, 2013.

[19] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, pp. 2291–2320, 2012.