SUPERVISED HIERARCHICAL SEGMENTATION FOR BIRD SONG RECORDING

Teresa V. Tjahja, Xiaoli Z. Fern, Raviv Raich, Anh T. Pham

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

ABSTRACT

A common framework of identifying bird species from audio recordings involves detecting bird song segments, which will be subsequently input to a classifier. In-field recordings are contaminated with various environmental noise. For such recordings, supervised segmentation has been observed to outperform unsupervised energy-based approaches. Prior supervised segmentation work considers only pixel-level predictions and ignores the supervision provided at the segment-level. We propose a hierarchical approach that learns to isolate bird song syllables based on both pixel-level and segment-level information. Experimental results suggest that our method outperforms an existing supervised method that learns only from pixel-level supervision.

Index Terms— Audio segmentation, supervised segmentation, bird species classification

1. INTRODUCTION

The field of bioacoustics studies animal vocalization to assess biodiversity, which serves as an indicator of environmental health. Due to its various potential benefits to society and science, interest in bioacoustics has been steadily increasing, mainly to help monitor and mitigate environmental impacts resulting from human activities [1]. Among the species that are observable in their natural habitat, birds are the most commonly chosen as their behavior reflects critical environmental changes in both global and local scales [2]. Bird species classification and detection are two of the most common goals of bird song analysis. Several studies have achieved promising results [3–6], with each of them utilizing syllables due to their important role as the basic building blocks of bird songs [7].

In previous work, a supervised method for extracting individual bird song segments from noisy audio recordings has been introduced [8]. Each recording is first converted to a spectrogram and a whitening filter is applied to normalize the noise level. A Random Forest classifier is then trained with a set of human-annotated spectrograms to assign a probability for each pixel in the spectrogram belonging to bird song segments. The probability map is then smoothed with a Gaussian filter and thresholded with a global value to produce a binary mask. Segments of individual syllables are then extracted as connected components from the spectrogram based on the binary mask. While this approach has been shown to perform well on noisy field recordings, it has two inherent limitations. First, using a single global threshold is generally suboptimal because different spectrograms or even different syllables within a single spectrogram may require different thresholds to be extracted properly. Second, human-annotated spectrograms may contain valuable information about what bird song syllables look like. Such information cannot be captured at the pixel level, and thus not utilized by the aforementioned method, which only learns at the pixel level.

In this paper, we propose to address these two limitations. Specifically, we consider multiple thresholds to build a hierarchy of candidate segments for each spectrogram. Learning from syllable-level supervision, we train a quality predictor to assess the quality of each candidate segment. The final segments are then identified by applying an efficient bottom-up selection procedure to the hierarchy.

We evaluate our method both at the pixel level and segment level using 200 field recordings. The results suggest that, compared to the baseline method, our proposed method achieves significantly better segment-level quality while maintaining comparable pixel-level quality.

2. PROPOSED METHOD

Our proposed method takes as input the probability map generated by the Random Forest classifier in the previous work [8]. Fig. 1 displays an example probability map generated from a time-frequency spectrogram by the Random Forest classifier. Our method consists of three main steps. In the first step, we apply multiple thresholds to the input probability map, producing a hierarchy of multiple levels of segments. In the second step, a quality predictor is trained to assign a score to each candidate segment in the hierarchy. The last step selects the final segments from the hierarchy based on the quality scores. The output of our method is a binary mask for each spectrogram, where each connected component represents a bird song segment.

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

Fig. 1. A probability map generated by the Random Forest classifier.

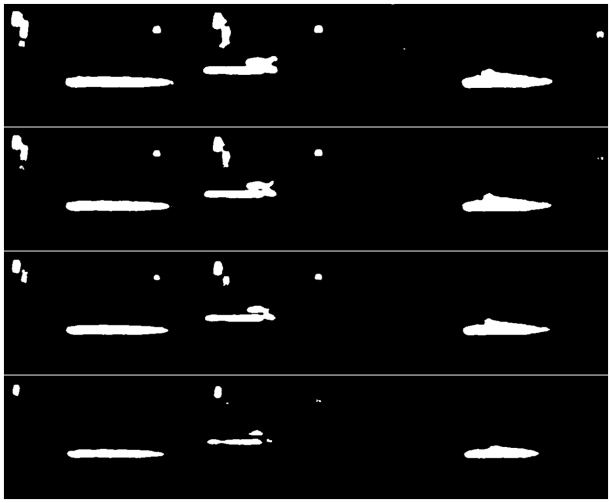

Fig. 2. Several levels of segmentation hierarchy built by thresholding the probability map in Fig. 1. Each white blob is considered a segment. From top to bottom, the thresholds are 0.20, 0.30, 0.40, and 0.60.

2.1. Generating Segment Hierarchy

Our system uses 10 thresholds (0.20 to 0.65 with 0.05 step) to build a 10-level segmentation hierarchy for each spectrogram. Based on empirical observations, thresholds below 0.20 produce severe under-segmentation that creates large blobs of segments, in addition to noise segments. Meanwhile, thresholds above 0.65 results in highly fragmented segments due to over-segmentation. Fig. 2 shows a part of the hierarchy constructed for the probability map shown in Fig. 1.

2.2. Segment-level Quality Predictor

Once the hierarchy is constructed, the next step is to assess the quality of each segment in the hierarchy. To achieve this, the human-annotated spectrograms are utilized to learn a regression model for predicting segment quality. To build such a regression model, we first generate segmentation hierarchies for the training spectrograms. For each resulting segment, we compute a quality value (between 0 and 1) based on its overlap ratio with ground-truth segments. Consider a segment I_s in the hierarchy of a particular spectrogram and a corresponding ground truth segment I_G, both represented as a binary mask. If the segment does not overlap with any ground-truth segment for that spectrogram, its quality is 0. Otherwise, if

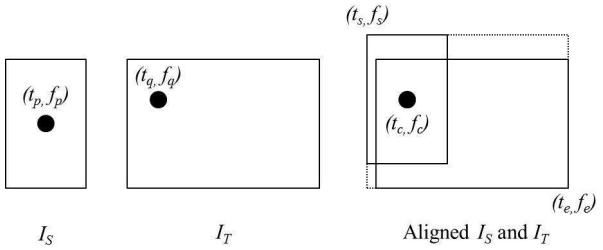

Fig. 3. Alignment of a segment I_S and a template I_T based on energy peaks.

I_s overlaps with a ground-truth segment I_G, the overlap ratio between I_S and I_G is defined as:

$$R(I_S, I_G) = \frac{\sum_{t=t_s}^{t_e} \sum_{f=f_s}^{f_e} I\{I_S(t,f) = 1 \wedge I_G(t,f) = 1\}}{\sum_{t=t_s}^{t_e} \sum_{f=f_s}^{f_e} I\{I_S(t,f) = 1 \vee I_G(t,f) = 1\}} \quad (1)$$

where (t, f) represents the time-frequency coordinate in the original spectrogram. If I_S overlaps with more than one ground truth segment, the quality is defined to be the maximum overlap ratio.

Next, we define a set of template-based features to represent the segments. We consider each ground-truth segment in the training spectrograms as a template. Given a set of templates, to represent a segment I_S, we compute the similarity between I_S and each template I_T. The idea is that the ground truth segments provide us with good examples of what bird song segments should resemble, and how similar a segment is to these "good examples" can be indicative of its quality. To compute the similarity between I_S and I_T we first align them based on their energy peaks, and compute their overlap ratio as defined in Equation 1. Note that we only allow segments to be aligned if their peaks' frequency values differ no more than 30 pixels. For such cases, the alignment process is illustrated in Figure 3. Since there can be more than one peak in a segment, 3 peaks are randomly selected from the segment and 3 peaks from the template, resulting in 9 pairs of peaks. The maximum overlap ratio between the segment and the template aligned at those peak pairs is then used.

We build our segment quality predictor using Support Vector Regression (SVR) [9] with a linear kernel. Each segment is represented by a N_T-dimensional feature vector, with N_T as the number of templates. Depending on the number of example spectrograms, there can be hundreds of ground-truth segments. In our implementation, to speed up calculation and reduce the dimension of the feature vector, the ground-truth segments are grouped into 50 clusters using the K-means algorithm. Each cluster medoid is then used as a template.

To select the optimal regularization parameter C for the SVR, 10-fold cross-validation is performed at the spectrogram-level. For instance, the training spectrograms, instead of the training segments, are split into training and validation sets. A

model is trained on the candidate segments obtained from the spectrograms in the training set and used to predict the quality of each candidate segment obtained from the spectrograms in the validation set. A final segmentation result is obtained for each validation spectrogram using the selection algorithm described in Section 2.3. The segment-level mapping score obtained by the evaluation metric as explained in Section 3.1 is used as the score for each model. The C parameter of the model that achieves the maximum score is then selected.

2.3. Segment Selection

After all segments in the hierarchy are assigned quality scores, we need to select a set of high-quality segments to form the final segmentation result. We follow the convention that the top level of the hierarchy contains segments obtained with the lowest threshold. As we move down the hierarchy, the new segments at level h will necessarily each be enclosed within one of the segments in level $h-1$, which we refer to as its parent. Given such a hierarchy, our goal is to select a subset of high quality segments such that no two segments in the subset should have ancestor-descendent relationship (to avoid redundant segments). Note that the prior approach that uses a global threshold corresponds to selecting a fixed level in the hierarchy for all spectrograms. Our approach not only allows different levels to be selected for different spectrograms, but is also capable of selecting segments from different levels within the same spectrogram.

Our selection algorithm starts from the lowest level (i.e. segments obtained with the highest threshold). At each level h, the average quality of segments belonging to the same parent is compared to the quality of the parent segment. If the parent's quality is higher than the children's average quality, the children are removed from the tree. Otherwise, the parent is removed and the children are assigned to the parent's parent. The intuition behind our algorithm is that if the average quality of the children is lower than the parent, then any solution containing the children will likely to be sub-optimal since replacing them with the parent will likely improve the overall quality. Similarly, if the parent scores lower than the average of its children, any solution containing the parent will likely be improved by substituting it with its children. The pseudocode for the algorithm is presented in Fig. 4.

The final result of applying our proposed method is displayed in Fig. 5. The top spectrogram contains the bird song segments extracted by our method. Each segment is labeled with its predicted quality and its level in the segmentation hierarchy in parentheses. The second spectrogram shows the ground-truth mask, while the third and last spectrograms show the binary segmentation masks generated by our method and the algorithm presented in [8] with 0.40 threshold, respectively.

```
SELECTSEGMENTS(Tree)
H ← MAXLEVEL(Tree)
FOR h ← H-1, …, 1
     S ← {s : s.level = h, s ∈ Tree}
     FOR ALL s ∈ S
          R ← {r : r.parent = s, r ∈ Tree}
          IF s.quality > AVERAGEQUALITY(R)
               REMOVEFROMTREE(R)
          ELSE
               FOR ALL r ∈ R
                    r.parent ← s.parent
               REMOVEFROMTREE(s)
RETURN Tree
```

Fig. 4. Pseudocode for segment selection.

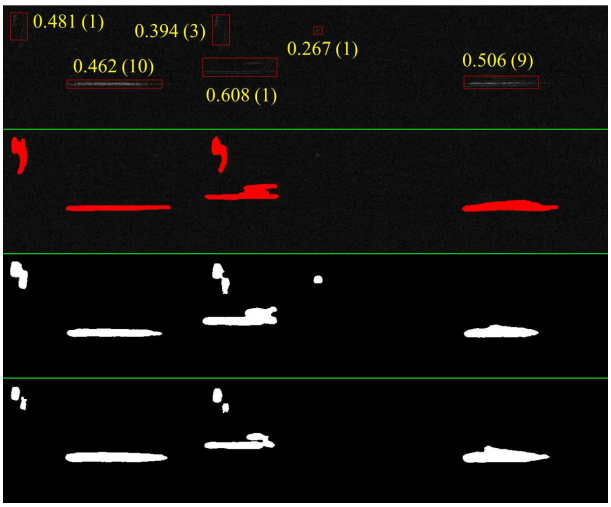

Fig. 5. The final result of applying the proposed method to the probability map in Fig. 1.

3. EVALUATION

In this section, we empirically evaluate our proposed method and compare it to prior work on supervised segmentation.

3.1. Evaluation Metrics

For each spectrogram, the system-generated binary mask is evaluated against the ground-truth mask annotated manually by human. Two types of metrics are used in the evaluation.

Pixel-level measure. We assess the pixel-level quality of the system-generated binary mask by computing its True Positive Rate (TPR) and False Positive Rate (FPR) compared to the ground-truth. This is a standard measure that has been widely used to evaluate the quality of segmentation results at the pixel level. One limitation of this measure is that it fails to capture how well the system-generated segmentation matches the ground-truth at the segment level. For example, consider two different segmentation results. In the first one, a

ground truth segment is split into two closely-spaced smaller segments. In the second result, the segment is slightly shrunk in size, but remains to be a single segment. The pixel-level measure will produce very similar TPR and FPR. However, conceptually, the second result is much preferable since it better maintains the integrity of the segment.

Segment-level measure. To address the limitation of the pixel-level measure, we design a novel segment-level quality measure, called *segment mapping score*, which forces a one-to-one mapping between the system-generated and ground-truth segments. Specifically, for each spectrogram, a complete bipartite-graph $G = (V_S, V_G, E)$ is constructed, where each element in V_S represents a system-generated segment and each element in V_G represents a ground-truth segment, and $|V_S| = |V_G| = N_V = \max(N_S, N_G)$, with N_S and N_G the number of system-generated and ground-truth segments, respectively. If $N_S < N_V$, then dummy nodes are appended to V_S so that $|V_S| = N_V$, and similarly for V_G. The value for each edge $e_{ij} \in E$ is the overlap ratio between $v_i \in V_S$ and $v_j \in V_G$ in the spectrogram. If one of the nodes e_i or e_j is a dummy node, $e_{ij} = 0$. Then, we find a maximum bipartite matching between V_S and V_G. The *segment mapping score* is calculated as the average overlap ratio across all matched pairs. Consider the aforementioned example. The segmentation result that splits the ground truth segment into two will achieve significantly lower score compared to the single segment alternative, correctly reflecting our preference.

3.2. Experiment Results

We applied our method to a set of 200 manually-annotated 10-second recordings collected with omni-directional microphones in natural environments. The 200 recordings contain various levels of difficulty, from a relatively clean recording to noisy ones with overlapping bird song segments. We also compare our method with the previous work in [8] as the baseline. Since we do not know the optimal global threshold for the baseline method, our experiments considered a total of five threshold values. Table 1 shows the final evaluation results.

From the results, we can see that the pixel-level quality of the baseline method is very sensitive to the global threshold. The value 0.4 as recommended by the original work achieves a good trade-off between TPR and FPR, and the top segment-level quality score amongst the different choices of θ. When evaluated at the pixel level, our method is comparable with the baseline methods, consistent with the general trend of trading off TPR and FPR. This is not surprising, because fundamentally we use the same pixel-level predictor. At the segment level, however, our method achieves far superior performance than the baseline methods, regardless of the θ value. This suggests that by learning from the segment level, the proposed approach was able to produce better overall segments despite having no gain at the pixel level.

Table 1. Evaluation of the baseline and proposed method.

	Pixel Level		Segment Level
	TPR	FPR	
Baseline ($\theta = 0.2$)	0.889	0.051	0.211
Baseline ($\theta = 0.3$)	0.829	0.035	0.226
Baseline ($\theta = 0.4$)	0.761	0.024	0.226
Baseline ($\theta = 0.5$)	0.693	0.017	0.217
Baseline ($\theta = 0.6$)	0.619	0.012	0.203
Proposed Method	0.784	0.028	0.309

4. CONCLUSION AND FUTURE WORK

We proposed a supervised hierarchical segmentation method to extract bird song segments from noisy recordings. The novel contributions of our work are:

- We introduced a hierarchical segmentation method that allows bird song segments to be extracted with different thresholds. This is particularly useful for in-situ recordings where different bird song signals may present itself at different strength levels.
- We introduced a novel supervised approach for predicting the quality of a segment as a whole. This is the first effort, to the best of our knowledge, that learns to perform segmentation based on both segment- and pixel-level supervision.
- We introduced an efficient bottom-up algorithm for selecting segments given a hierarchy of segments and their predicted quality scores, which has been empirically observed to work well.

Our method is most appropriate for analyzing bird song recordings obtained within natural environments with varying level of noise and signal strength. One potential limitation of our method is its sensitivity to the annotation quality of the training annotation. If the annotations are incomplete or inconsistent, our approach may not learn effectively. Future work will investigate whether the improved segment-level quality can lead to improved species classification accuracy.

5. REFERENCES

[1] NIPS Int. Conf., *Proc. Neural Information Processing Scaled for Bioacoustics, from Neurons to Big Data*, USA, 2013, http://sabiod.org/nips4b.

[2] Forrest Briggs, Balaji Lakshminarayanan, Lawrence Neal, Xiaoli Z. Fern, Raviv Raich, Sarah J. K. Hadley, Adam S. Hadley, and Matthew G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *The Journal of the Acous-*

tical Society of America, vol. 131, no. 6, pp. 4640–4650, 2012.

[3] A Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, April 2003, vol. 5, pp. V–545–8 vol.5.

[4] P. Somervuo and A Harma, "Bird song recognition based on syllable pair histograms," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, May 2004, vol. 5, pp. V–825–8 vol.5.

[5] Chang-Hsing Lee, Chin-Chuan Han, and Ching-Chien Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1541–1550, Nov 2008.

[6] Wei Chu and D.T. Blumstein, "Noise robust bird song detection using syllable pattern-based hidden markov models," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 345–348.

[7] Chih-Hsun Chou, Pang-Hsin Liu, and Bingjing Cai, "On the studies of syllable segmentation and improving mfccs for automatic birdsong recognition," in *Asia-Pacific Services Computing Conference, 2008. APSCC '08. IEEE*, Dec 2008, pp. 745–750.

[8] L. Neal, F. Briggs, R. Raich, and X.Z. Fern, "Time-frequency segmentation of bird song in noisy acoustic environments," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 2012–2015.

[9] Harris Drucker, Chris J.C. Burges, Linda Kaufman, Chris J. C, Burges* Linda Kaufman, Alex Smola, and Vladimir Vapnik, "Support vector regression machines," 1996.