TIME-FREQUENCY SEGMENTATION OF BIRD SONG IN NOISY ACOUSTIC ENVIRONMENTS

*Lawrence Neal, Forrest Briggs, Raviv Raich, and Xiaoli Z. Fern**

School of EECS, Oregon State University, Corvallis, OR 97331-5501

ABSTRACT

Recent work in machine learning considers the problem of identifying bird species from an audio recording. Most methods require segmentation to isolate each syllable of bird call in input audio. Energy-based time-domain segmentation has been successfully applied to low-noise, single-bird recordings. However, audio from automated field recorders contains too much noise for such methods, so a more robust segmentation method is required. We propose a supervised time-frequency audio segmentation method using a Random Forest classifier, to extract syllables of bird call from a noisy signal. When applied to a test data set of 625 field-collected audio segments, our method isolates 93.6% of the acoustic energy of bird song with a false positive rate of 8.6%, outperforming energy thresholding.

***Index Terms*—** Audio segmentation, bird species identification, time-frequency segmentation

1. INTRODUCTION

Classification of bird species from audio is a recent application of machine learning. Several methods have proven successful in correctly labeling the species of single birds in low-noise environments [4, 1, 2]. We propose a method of pre-processing and segmenting noisy field recordings of bird song, to isolate each bird syllable from the rest of the signal.

Bird song has a structure consisting of single-vocalization syllables, many of which make up a song. The structure of songs varies within bird species, but the structure of individual syllables remains relatively constant. Thus, many methods of species recognition are based on classification at the level of syllables [3]. In ideal conditions, an audio recording of bird song consists of sequential bird call syllables separated by silence. Current methods of classification use the characteristics of this audio signal, such as Mel-frequency cepstral coefficients, to label each syllable in the audio signal as a bird species [1].

The process of segmenting, or extracting syllables from an audio signal is simple in ideal conditions. If bird call is the only source in a signal, then increased audio energy will denote a syllable. In field conditions, this assumption does not always hold. Rather than bird song alone, there may be many sources of sound in a recording. Wind, streams, and other background noise degrades the signal, and noises from other animals and surrounding events interrupt bird song. Complicating the issue, vocalizations from two or more individual birds may occur simultaneously during a recording. However, accurate segmentation is essential for successful species classification [4].

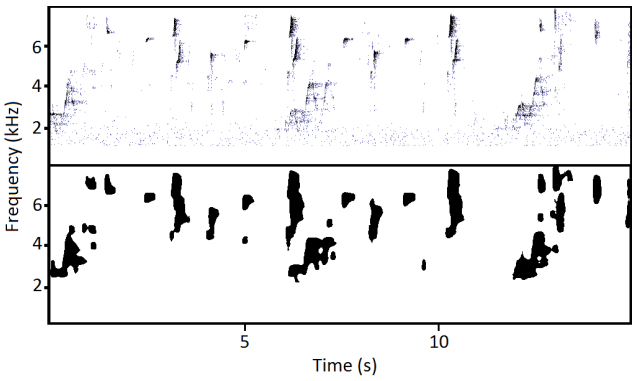

Fig. 1. Above: A noise-reduced spectrogram of a Swainson's Thrush and a Pacific-Slope Flycatcher. Below: The binary mask generated by the proposed method. Each darkened region corresponds to a detected syllable of bird song.

In noisy environments, classification attempts face two challenges. First, time-domain segmentation based purely on audio energy will introduce false syllables, corresponding to non-bird noises. Second, whenever multiple birds sing at once, time-domain segmentation will group mixtures of syllables together, degrading classification accuracy. We have a 4 terabyte dataset of audio from automated recorders at sites in the H.J. Andrews Experimental Forest. Recordings in this dataset contain high noise and simultaneous bird syllables.

We propose a method of segmentation that addresses these problems, using supervised machine learning in the time-frequency domain. We transform input signals into a spectrogram representation, then apply a supervised classifier to create a binary mask labeling each time-frequency unit as

*The authors thank Matthew Betts, Sarah Frey, Adam Hadley, Rachael Chung and Vivian Brown for their contributions.

either bird sound or background. This process allows us to extract individual syllables of bird song, even when syllables overlap in time. Fig. 1 shows such a binary mask. We evaluate this binary mask against a manually-labeled ground truth using a metric of true positive rate (TPR) versus false positive rate (FPR), and a metric of energy-weighted TPR versus FPR. Applied to our set of field recordings, our method achieves 90.5% TPR with 9.3% FPR ($\theta = 0.08$) for the first metric, and 93.6% TPR to 8.6% FPR ($\theta = 0.12$) for the second. By both metrics, the proposed method outperforms segmentation by energy thresholding.

2. BACKGROUND

The proposed segmentation method is part of a system that classifies bird audio recordings by species. Such classification of bird audio data, from multiple sites in the field, can provide presence/absence data to map bird species to locations and times. This data in turn can be used for the modelling of bird species distributions and conservation planning. A system based on such methods has the capability to provide higher resolution data at less cost than manual surveys [1].

Numerous methods have been proposed to classify segments of bird call by species [2, 5, 6, 7]. In each example, classifiers are applied to segments of audio data, each representing a vocalization from a single bird. These segments are obtained from the source audio through time-domain segmentation. Graciarena, et al. use a simple voice activity detection system to segment audio [5]. Lakshminarayanan, et al. compute Kullback-Liebler divergence between the power spectral density of each audio frame and the uniform distribution. Local minima of the KL divergence are used to identify the boundaries of segments, and segments with the greatest energy are classified [2]. Somervuo et al. apply an iterative thresholding algorithm to find high-energy audio segments [7]. These methods work well when input audio consists of sequential single-species calls with minimal noise, but cannot accurately segment noisy field recordings.

2.1. Random Forest

Our segmentation is based on supervised classification using a Random Forest classifier. Random Forest (RF) is an ensemble classifier consisting of a collection of decision trees [8]. Given a set of training examples $\mathcal{T}$, each tree h_i in the RF classifier is independently built from a bootstrap sample selected randomly with replacement from $\mathcal{T}$. Trees are constructed by recursively applying the following procedure:

- Take as input a set of examples T, where each $T_i = (x, y)$, x is a feature vector, and y is the corresponding class label.
- If all labels y are the same, create a leaf node with the value y.
- Select a random subset $\mathcal{F}$ of $log_2(k)+1$ features, where k is the number of features in x.
- For each feature $d \in \mathcal{F}$, sort T on d and find the threshold value θ_d that splits T into two sets T_{left} and T_{right}, such that the Gini index $G(T_{left}, T_{right})$ is maximized.
- Choose the feature and threshold (d, θ_d) such that G is maximized. If all possible values of G are equal, then make a leaf node with the majority label. Otherwise, create two child nodes by recursively applying the procedure using T_{left} and T_{right} as input.

Each interior node of an RF tree corresponds to a test of the form $x_d < \theta$. Traversing the tree with any input vector x will lead to a leaf node, which contains a single class label y. When classifying an input x, each decision tree in the RF classifier casts a vote. The output label for x is equal to the proportion of trees that voted for y.

3. PROBLEM FORMULATION

Our segmentation method begins by decomposing a time-domain audio signal $A(t)$ into a two dimensional time-frequency spectrogram S. This involves separating A into a set of overlapping frames $\{F_0, F_1, ...\}$. Each frame F_t corresponds to a set of values $\{A(t), A(t+1), ...A(t+s-1)\}$, where s is the number of samples in each frame. For each frame F_t, a short-time Fourier transform is applied to generate coefficients $\{f_0(t), f_1(t), ...f_{\frac{s}{2}}(t)\}$. The spectrogram S is composed from the magnitude of the coefficients of each frame.

In segmenting a spectrogram S, we use a supervised classifier to assign a binary label to each element in S, corresponding to whether the audio signal at that time and frequency is bird sound or noise. The output of our method is a binary mask overlaying the spectrograms. We consider each contiguous positive-labeled region in this mask to correspond to a bird syllable in the spectrogram. This approach allows us to isolate the desired bird call signal from background noise, and to separate distinct syllables that overlap in time but not frequency.

4. PROPOSED METHOD

4.1. Preprocessing

In each input audio file, a Hamming window is first applied to each frame. A short-time FFT is then applied with a frame size of 512 samples, and an overlap of 256 samples between each subsequent frame, transforming the signal into a time-frequency spectrogram. A whitening filter is subsequently applied to the spectrogram, to normalize the level of environmental noise at each frequency. Frequency ranges below 1kHz contain little or no bird call [3], so a band-pass filter removes frequencies under 1kHz.

4.2. Random Forest Training

Our method requires as input a training set of audio files with corresponding binary masks. Each audio file is converted to a spectrogram, using the same parameters as the input data. Time-frequency units covered by the mask are used as positive examples of bird sound when training the classifier. All other elements, including silence, static noise, and non-bird sound, are used as negative examples. The masks in this set were created manually by visual and auditory examination of each spectrogram and its corresponding audio. For the purposes of training and evaluation, these masks are assumed to be ideal binary masks corresponding to audible bird call. Our implementation uses 40 RF decision trees, built from a set of 500,000 randomly-sampled training examples, with a class balance of 10% bird call and 90% noise examples.

4.3. Classification

In training and classification, a feature vector $x_{t,f}$ is extracted for each time-frequency unit $S(t, f)$ in the spectrogram. The vector $x_{t,f}$ describes the spectral characteristics of a rectangular window surrounding (t, f), and is defined by the following:

- The frequency value f
- The values within a rectangular window surrounding (t, f)

  $$S(i,j), i \in [t - t_w, t + t_w],\ j \in [f - f_w, f + f_w]$$

  centered at (t, f), where $2t_w + 1$ is the size of the window in the time dimension and $2f_w + 1$ is its size in the frequency dimension.
- The variance σ^2 of the units in this window

  $$\sigma^2 = \frac{1}{(2t_w+1)(2f_w+1)} \sum_{i=t-t_w}^{t+t_w} \sum_{j=f-f_w}^{f+f_w} (S_{i,j} - \mu)^2$$

  where μ is the mean value in this window.

We use a t_w value of 6 T-F units and a f_w value of 12 units, yielding a window spanning 192ms by 750hz in the T-F domain. In the classification process, a probability mask M_p is generated by the outputs of the RF classifier, in which each value $M_p(i, j)$ corresponds to the fraction of RF trees that labeled $S_{i,j}$ as bird call.

4.4. Output and Analysis

After classification, a Gaussian convolution is applied to create a smoothed probability mask M_s.

$$M_s = M_p \star g, \text{ where } g(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

This convolution is applied with a square kernel of 17x17 time-frequency units, and $\sigma = 3.0$. After smoothing, the probability mask is converted to a binary mask M_b by applying a threshold

$$M_b(x, y) = 1 \text{ if } M_s \geq \theta, \text{ or 0 otherwise}$$

where $0 \leq \theta \leq 1$. The value θ controls the trade-off between false positive and false negative. A larger value leads to lower false positive rate but higher false negative rate. The smallest time-frequency regions identifiable as bird syllables had a duration of approximately 160ms and a frequency range of approximately 300hz. Any regions in the binary mask less than 90% of this size are discarded from the final segmentation.

5. EVALUATION

The output of our method is a binary mask M_b over the time-frequency representation of the input audio signal. Ideally, each contiguous positive region in the mask represents one syllable of bird call. However, it is often ambiguous whether a region consists of one syllable or several. Thus, we evaluate the binary mask itself, by comparing it directly to a human-provided ideal mask. We evaluate the output binary mask M_b against the manually-labeled ideal binary mask M_i with the following two metrics:

- **Time-frequency area metric:** Find all labeled units (i, j) such that $M_b(i, j) = 1$ and $M_i(i, j) = 1$. Define the true positive count TP to be the number of units found. The false positive value FP is similarly defined for $M_b(i, j) = 1,\ M_i(i, j) = 0$, the true negative TN for $M_b(i, j) = 0,\ M_i(i, j) = 0$, and the false negative FN for $M_b(i, j) = 0,\ M_i(i, j) = 1$.
- **Acoustic energy metric:** Find all labeled units (i, j) such that $M_b(i, j) = 1$ and $M_i(i, j) = 1$, as in the previous metric. Define the true positive value TP to be the sum $\sum S(i, j),\ (i, j) \in TP$ of all spectrogram energy values in the true positive set.

For each of the two metrics, a true positive rate $TPR = TP/(TP+FN)$ and a false positive rate $FPR = FP/(FP+TN)$ are defined. We plot coordinate pairs $(FPR, TPR)_\theta$ for a varying threshold θ, to display the recall/precision trade-off of the method.

Two energy thresholding methods are evaluated along with the Random Forest method. The first is an energy thresholding applied directly to the spectrogram S. Each unit (i, j) in the output binary mask has a value 1 if $S(i, j) > \theta$, where $0 < \theta < 1$. The second method applies the same Gaussian blur described in 4.4 to the spectrogram. The energy thresholding is then applied to the blurred spectrogram.

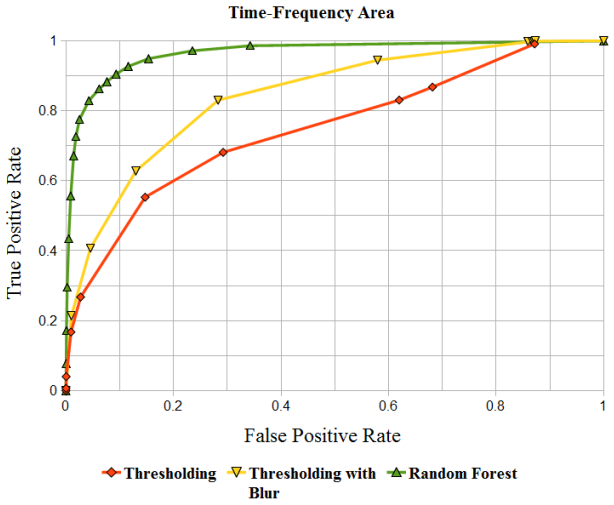

Fig. 2. ROC curves of the time-frequency area metric.

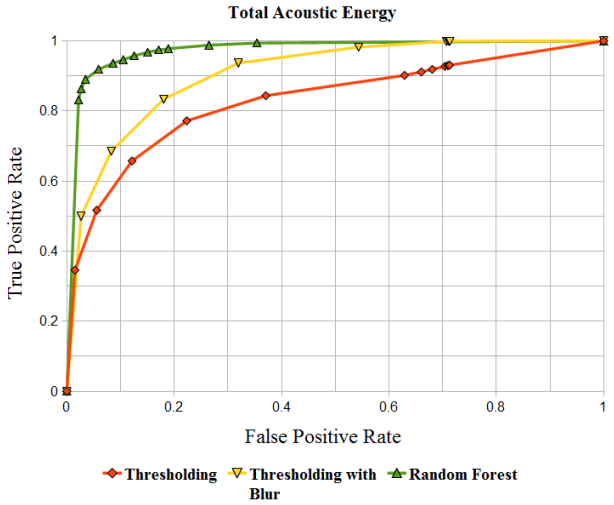

Fig. 3. ROC curves of the acoustic energy metric.

The dataset used for all three methods is an annotated set of 625 audio segments, each 15 seconds, in 16kHz PCM format. The audio segments are selected, two per hour, from a 24 hour recording at each of 13 sites across the H.J. Andrews Experimental Forest. These data were recorded between May and July 2009. Two-fold cross validation is used in evaluation of the RF classifier.

In Fig. 2 and Fig. 3, the ROC curves of the RF method are plotted against the energy thresholding methods, using the area-based and energy-based metrics, respectively. The thresholding method with Gaussian blur outperforms the direct thresholding in both metrics. By both metrics, RF outperforms spectrogram energy thresholding.

6. CONCLUSION AND FUTURE WORK

We proposed a method to segment syllables of bird call from a noisy audio signal, based on a time-frequency representation and a Random Forest supervised classifier. Our method produced a binary mask covering 90.5% of the time-frequency area of bird syllables with a FPR of 9.3% ($\theta = 0.08$), and 93.6% of the spectral energy of bird vocalizations with a FPR of 8.6% ($\theta = .12$), out-performing energy-based methods. Future improvements to the method may include replacing thresholding with a more complex region-growing algorithm to more effectively separate concurrent syllables from separate individuals. To apply the method to larger datasets, runtime could be decreased by using a smaller set of features.

7. REFERENCES

[1] Forrest Briggs, Raviv Raich, and Xiaoli Z. Fern, "Audio classification of bird species: a statistical manifold approach," in *IEEE International Conference on Data Mining*, December 2009, pp. 51–60.

[2] B. Lakshminarayanan, R Raich, and X Fern, "Audio classification of bird species: a statistical manifold approach," in *International Conference on Machine Learning and Applications*, December 2009, pp. 53 – 59.

[3] A. Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *IEEE International Conference on Acoustics Speech and Signal Processing*, April 2003, pp. 545–548.

[4] Seppo Fagerlund, "Bird species recognition using support vector machines," *EURASIP Journal on Applied Signal Processing*, pp. 64–64, January 2007.

[5] Martin Graciarena, Michelle Delplanche, Elizabeth Shriberg, Andreas Stolcke, and Luciana Ferrer, "Acoustic front-end optimization for bird species recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing*, March 2010, pp. 293 – 296.

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

[7] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 2252 – 2263, November 2006.

[8] Leo Breiman, "Random forests," *Machine Learning*, pp. 5–32, January 2001.