Detection and Segmentation of Bird Song in Noisy Environments

Lawrence Neal (nealla@onid.oregonstate.edu) Mentor: Xiaoli Fern (xfern@eecs.oregonstate.edu)

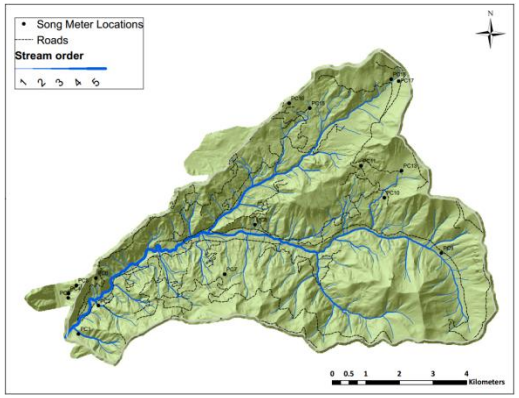
Abstract

By studying the activity and population distribution of bird species, ecologists can gain valuable insights into factors that affect the whole ecosystem. However, current methods of tracking bird activity involve manual field surveys, which are slow and low-resolution. The OSU Bioacoustics Group is building a system to automatically gather per-species bird activity data using machine learning analysis of audio gathered from remote automatic recorders.

After audio is recorded in the field, in order to classify bird species, each 'syllable' of bird song must be extracted from the source audio, in a process we call segmentation.

H.J. Andrews Forest

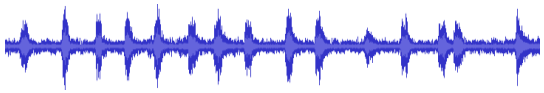
Automated audio recorders have been placed at sites in the H.J. Andrews Long-Term Ecological Research Forest in the western Cascade mountains. Each spring and summer since 2009, audio data has been gathered daily from 16 sites placed across a range of elevations and terrain. Analyzing this data can reveal changes in bird activity due to climate change and other factors.



Time-Frequency Analysis

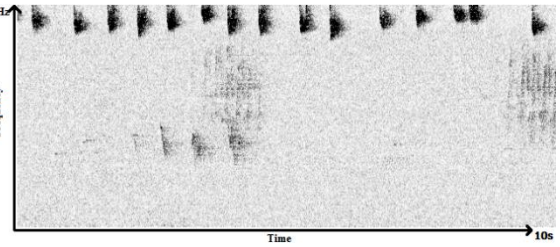
The objective of the Bioacoustics project is to identify the species of birds singing in each audio recording. To do this, each song must be separated from background noise and from other songs. Because background noise varies and multiple birds sing at once, this cannot be done by simply taking cuts in time from the original audio.

Instead, a discrete Fourier transform is applied to generate a spectrogram of each input audio cut. The spectrogram separates sound into its component frequencies, which allows us to segment individual regions of sound in the time-frequency domain, even when they overlap in time.



Left: Audio waveform of three calling birds over 10 seconds

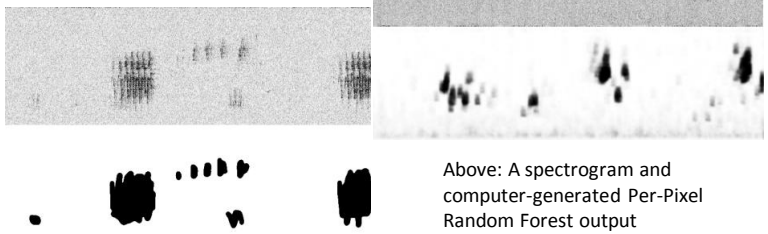
Right: Spectrogram of the audio, after noise filter (Darkened pixels indicate high amplitude)



Machine Learning for Segmentation

Supervised machine learning allows software to mimic human output for tasks such as classifying an object. The segmenter uses a Random Forest decision tree ensemble algorithm to learn the difference between areas of the spectrogram containing bird call and areas containing only noise. The algorithm is trained on a set of 625 hand-drawn spectrogram masks.

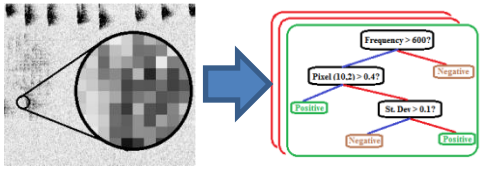
Below: A spectrogram and hand-drawn training label of bird song (each black spot is one syllable)



Above: A spectrogram and computer-generated Per-Pixel Random Forest output

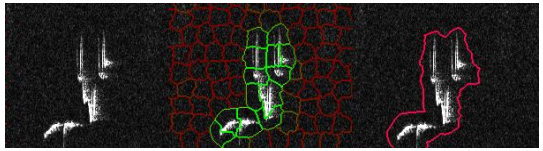
Two Learning Methods

The "Per-Pixel Random Forest method" trains a classifier to output true/false labels for each pixel of a spectrogram, based on the pixel's frequency and the pixel values of a 'window' around it.



Spectrogram pixel values are sampled and compared to training masks to learn sets of decision tree rules

The "Superpixel Merger method" first breaks input spectrograms into small regions of a few hundred pixels, then trains two classifiers to first separate foreground from background, then merge or split adjacent superpixels.



Left: Example spectrogram syllables. Center: Superpixel foreground/background labels. Right: Merged component of superpixels

Results

Applied to the training data using cross-fold validation, the Per-Pixel Random Forest method classifies spectrogram pixels with a true positive rate of 90.5% and a false negative rate of 9.3%, compared to hand-drawn masks[1].

The Superpixel Merger method achieves similar per-pixel accuracy, while successfully separating some syllables that overlap in both time and frequency.

[1] Lawrence Neal, Forrest Briggs, Raviv Raich, and Xiaoli Z.Fern. "Time-Frequency Segmentation of Bird Song in Noisy Acoustic Environments." *Proc. International Conference on Acoustics, Speech and Signal Processing, 2011.*