“A SoundHound for the Sounds of Hounds”
Weakly Supervised Modeling of Animal Sounds

Robert Colcord[1], Ethan Geller[1], Matthew Horton[1]
[1]Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA

Abstract

We propose a hybrid approach to generating acoustic models for recognizing specific species of animals based on their vocalizations using unsupervised analysis and recognition of similar sounds in a larger recording.

In many applications of machine learning to audio analysis and processing, such as voice recognition, the primary challenge is finding and annotating training data. This makes unsupervised clustering of similar sounds attractive as a means for creating acoustic models. The Cornell Lab of Ornithology hosts a large quantity of field recordings of a large variety of species of animals. However, many of these recordings are exceptionally long, and the animal in question sometimes will not be audible for several minutes at a time within these larger recordings. Analyzing audio, however, on the basis of similar repetitive sounds could form a foundation for a potential solution to this problem. That is what this project explored.

Dataset and Features

We used ten field recordings of animal sounds as our training dataset. The species represented in the recordings include chimpanzee, northern elephant seal, white-nosed coati, red deer, brown howler monkey, and red squirrel. Each of these field recordings lasts anywhere from twenty minutes to an hour, and the represented animal vocalizes several times. In all, this comprised fifty-five minutes and fifteen seconds of audio data. The field recordings were acquired with permission from the Macaulay Library at the Cornell Lab of Ornithology[2]. All processing and feature extraction of the data was done in Matlab.

The key feature of the audio data used in the generation of models is the mel-frequency cepstrum (MFC) of the sound.The MFC is a tool used in audio analysis that represents short-term power spectrum of a sound. Mel-frequency cepstrum coefficients (MFCCs) are the individual amplitude values of an MFC. The process of finding the MFCCs at a certain point in time t is as follows:

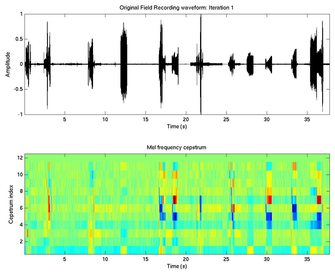
1. Extract signal x(t:t+h), wherein h is the window size.
2. Smooth the cepstrum of the extracted signal:
   Xcep= FFT(win*IFFT(log(FFT(x))))
   where win is a boxcar window.
3. Measure energy at specific frequencies mapped from the Mel scale.

Examples of MFC plots can be seen in Methods

Methods

1. Mel-Frequency Cepstral Coefficients

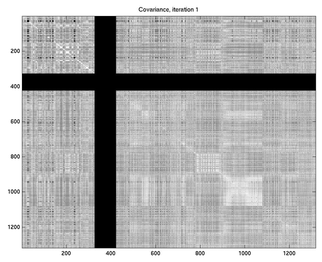
For each of the sounds we analyzed, we found the mel-frequency cepstrum of the signal. The figure below shows the original waveform plotted over the MFC.



2. Similarity Matrix

A similarity matrix, K, was found by comparing the mel-frequency cepstrum of the sound in question with itself according to the formula:

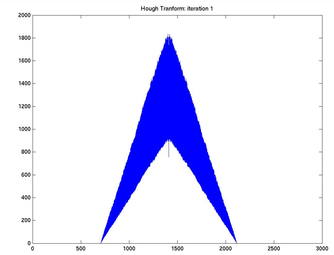
$$K_{ij} = 0.5 * (1 + (\frac{<x_i, x_j>}{\| x_i, x_j \|})^2)$$



3. Line Segment Detection using a Hough Transform

Next, we look for consecutive points in our matrix that exhibit high similarity. To do this, we need to find line segments that run parallel to the diagonal of our matrix.

We do this by running a Hough Transform with a theta at 315 degrees. Then, after finding peaks in our Hough transform, we look for consecutive high similarities in the corresponding diagonal.

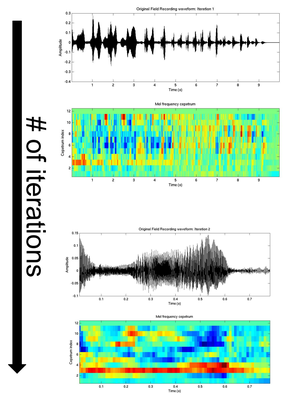


4. Removal of insignificant data

After finding locations of line segments, we remove audio from locations in our recording that do not correspond to those line segments.

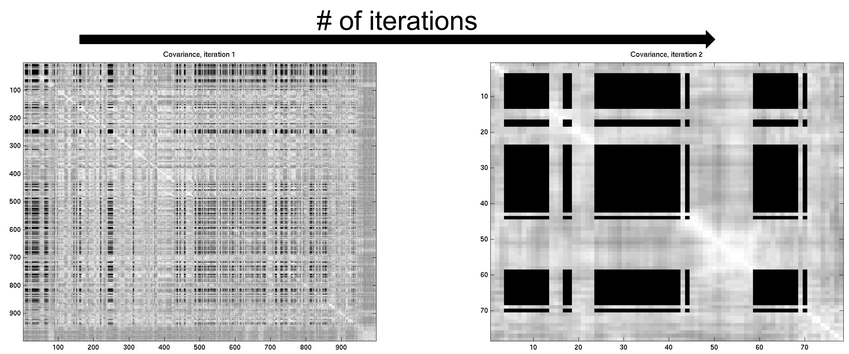
5. Iterate until convergence

Using the new audio file composed of concatenated similar sounds, we begin anew.



time and spectral representations of signals after successive iterations of the process described here.

similarity matrices after iterative removal of noise and unique sounds. Because we are comparing a recording to itself, this matrix will always be symmetric, and any point on the diagonal will always be 1.



References

A. Jansen, S. Thomas, and H. Hermansky, “Weak top-down constraints for unsupervised acoustic model training ,” in *Proceedings of ICASSP,* 2013.

A. Norouzian, R. Rose, S.H. Ghalehjegh, and A. Jansen, “Zero Resource Graph-Based Confidence Estimation for Open Vocabulary Spoken Term Detection,” in *Proceedings of ICASSP,* 2013.

Acknowledgements

Special thanks to the Cornell Lab of Ornithology for providing the animal sound field recordings used in the project.

Special thanks to Junjie, Andrew Ng, and the entire CS229 teaching staff for their guidance.