MUSICAL SEGMENTATION TECHNIQUES FOR BIRD SONG CLASSIFICATION

Regis Verdin, Avrosh Kumar

December 2015

Department of Music Technology, Georgia Institute of Technology

ABSTRACT

Bird identification using audio recordings is a well-known problem in the field of bio-acoustics. The basic approach is to extract a set of features from the audio, and to then use a machine-learning classifier to determine the species of the bird. Here, we approach the problem by denoising the audio, and then segmenting it at the approximate locations of repetitions in the recordings. The repetitions presumably correspond with phrases of the bird vocalizations. The segmentation uses a self-similarity lag matrix to find repeated sections in the audio, which are assumed to be repetitions of the bird songs or calls. Spectral features and MFCCs are extracted. Finally, we compare the classification results obtained from segmented and un-segmented data.

Index Terms— bird song, machine learning, segmentation, classification, self-similarity lag matrix

1. INTRODUCTION

Audio content analysis techniques have been applied to the identification of bird species. There are several uses for automating bird identification. It can help recreational birders identify birds, especially in situations where visual identification is limited. It can also be a tool for ecological research, where a researcher may be interested in learning about the presence, absence, or population of a bird species in a given location or ecosystem. Fully or partially automating these tasks significantly reduces the amount of labor required.

Bird vocalizations can be roughly grouped into two basic types: songs and calls. Songs are typically longer and more complex, while calls are short and relatively simple. The contexts for their uses are also different: songs are often used in courtship, while calls are used as alarms. For simplicity, we treat songs and calls as equivalent in our classification system. With a sufficiently large data set, it is safe to assume that both the calls and songs of a species will be represented in the classifier. There are variations between individual birds in a single species singing the same song or call, so the task of identifying them is comparable to the automatic recognition of covers in music.

1.1. BirdClef Data Set

We use the BirdClef 2014 dataset to train and test our classifier. The dataset is created yearly for a classification competition, and draws all of its files from the larger Xeno-Canto set. Xeno-Canto is a website featuring user-uploaded and tagged field recordings of birds. The BirdClef set contains 9,688 audio files of birds from Brazil, with 428 unique species. For the BirdClef set, the files are normalized and downmixed to mono audio at 44kHz. Each recording has an accompanying XML file with user-provided metadata. The metadata tag of interest here is “species”. We treat the user-identified species as the ground truth in our work, using it to train the classifier and later test our accuracy.

The BirdClef set contains recordings of varying quality, because they are recorded in a real world setting. A typical file contains many repetitions of one bird call or song, along with varying amounts of noise and irrelevant information: insects, wind, microphone handling, other bird species, wind, human speech, and so on. In our system, we take measures to reduce the effects of the noise. The goal is to be able to use the system in real world setting, so a noisy dataset was desirable for our evaluation process.

2. RELATED WORK

The research presented here is based in part on *Audio-based Bird Species Identification with Music Processing Approaches* [1]. We follow the approach outlined in that paper, while introducing our own segmentation process. The basic approach to audio content analysis is as follows: preprocessing, feature extraction, and classification. For bird song identification, the data sets consist of either field recordings or studio recordings. This difference informs the algorithms used, as field recordings require more sophisticated preprocessing to obtain good results.

2.1. Preprocessing

Segmentation of the audio recordings is used for several main reasons: for example, to break long bird songs into

smaller "syllables", or to extract single phrases of bird songs or calls. Manual segmentation can be used for this task, but it becomes unfeasible when the data set becomes adequately large to train the classifier well. Several approaches have been taken to implement automatic segmentation. One approach is used with the songs of Finches, breaking them into component syllables [2]. The order of the syllables can then be determined, giving the researcher a sense of the form of an individual song. Vartaki and Lerch [1] use visual segmentation of the spectrogram to produce a 26 dimension feature vector, where the features themselves are obtained through image processing of the visual segments. Ruiz-Munoz et al. [3] implement unsupervised segmentation to a similar end.

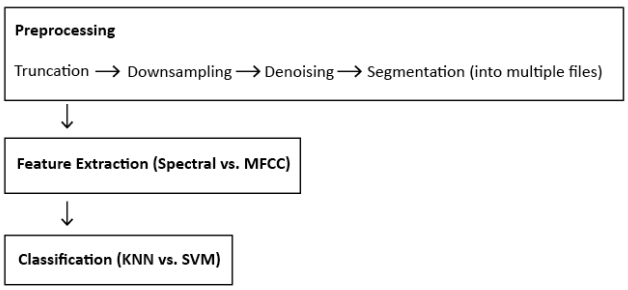
To increase the accuracy of the segmentation, denoising is usually applied before segmenting. Boll's spectral subtraction technique [4] is used here: the FFT of a noisy segment in the audio is subtracted from all other FFT blocks in the file. Various methods can been used to obtain the initial noisy segment to be subtracted.

2.2. Feature Extraction and Classification

Many features have been used in classifying bird vocalizations. Some rely on connected component features from the spectrogram image, while others use spectral features in the audio, or a combination of both image and audio-based features. The classifiers typically used are SVM and Random Forest.

3. ALGORITHM OVERVIEW & DESCRIPTION

3.1. Overview of training and classification



The training process for the classifier is as follows. The BirdClef audio files are first read into Matlab. Many file numbers were missing from the set (in terms of the original numbering scheme), so a try-catch was used to ensure accurate index labeling. Accompanying XML files are also loaded into a matrix, with each "species" label converted to an integer. This resulted in 428 unique class labels. The audio files exceeding 40 seconds in length (about 20% of the data set) were then truncated to 40 seconds, to speed up the training and testing time. The classifier could be trained on the entire length of each file, but given the number of repetitions of each vocalization in the files it was deemed acceptable to truncate them. Onset detection was considered to find an ideal starting point for truncation, by starting the recordings at the time of the first bird call. However, we decided against it given the uncertainty that a chosen onset would be the desired bird versus some background sound.

Following the truncation, the recordings were downsampled to 22kHz, and denoised using spectral subtraction. The denoising algorithm used is based on a denoiser designed to remove vuvuzela sounds from soccer game audio [5]. The algorithm selects the segment between 0.4 and 1 seconds (0.6 total length), and assumes this to be a noisy segment. A potential issue here is that the segment might include the bird vocalization. In a commercial product this might be a good place for a user to manually select a bird-less segment. However, the birds were not often in the first second of our recordings, so it was relatively successful. In figure 1, we show the denoising process with file id 6021.

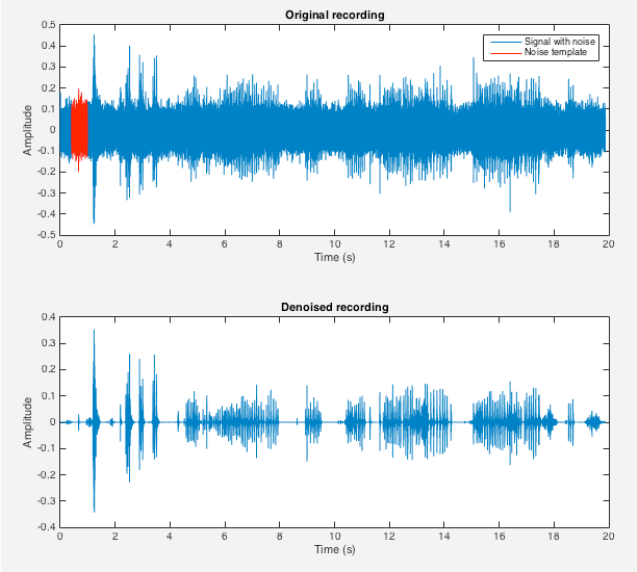


Figure 1

Segmentation (described below in more detail) was applied, with the aim of removing much of the irrelevant information and resulting in several recording segments per file. A variety of features are extracted from the audio segments, and are used to train a SVM and K-nearest-neighbor classifier. A similar overall process was followed for testing data in the classifier, using 10-fold cross validation.

3.2. Segmentation

There were several motivations for segmenting the audio in our situation. First, the BirdClef database contains 9,688 recordings, with 428 unique species, giving an average of 22.6 recordings per species. The low number of samples per class is not ideal for training the classifier. Our segmentation tries to address this issue by extracting each repetition of the phrase from the audio file, and using each extracted segment as a separate training sample. This results in many more training samples per class. Unlike the segmentation in much related work, however, we do not extract features from the image itself.

Taking inspiration from the musicality of bird songs, we use a technique typically used to analyze musical structure [6]. This is used to find the repeated sections of our audio files, which presumably correspond with the primary bird song phrase. MFCCs are extracted from the audio, and a self-distance matrix (SDM) from the MFCC is constructed. Here, strong diagonal lines (other than the main diagonal) indicate repetitions, where a series of times on one axis have high similarity values to a different series of times on the other axis. To extract the locations of these repetitions, we create a self-similarity lag matrix using the SDM. An adaptive threshold is used to convert the image to binary values. A combination of erosion and dilation is used on the lag matrix to smooth and emphasize horizontal lines.

Paulus et al. discuss using similarity lag matrices to music structure detection. We apply this concept to find the start and end times of bird phrases. We find the longest segments in our lag matrix using an algorithm which scans horizontally for a series of 1s, which correspond to repeating sections. A tolerance for gaps of up to 2 consecutive zeros is also allowed. For example, 1-1-1-0-0-1-0-1 would be counted as a section of length 8. Scanning the entire lag matrix (left to right, then bottom to top), we obtain the longest section and its corresponding starting and ending times.

The recording is repeatedly cut using the time stamps from the algorithm above: we extract the features from the longest repeated section detected in the song. The segment is cut from the original, and the next longest (or equal length) repeated section is found, added to the new data set, and cut from the original recording. This process repeats until either 8 passes have been made, or the original recording has been cut to less than 5 seconds. We are not differentiating whether the detected repetition is a silence.

Figures 2-6 demonstrate the iterative process of phrase segmentation. The red boxes show the detected phrase. Note the decreasing length of the sample. We have extracted 5 features sets from this sample, all labeled with the same class.

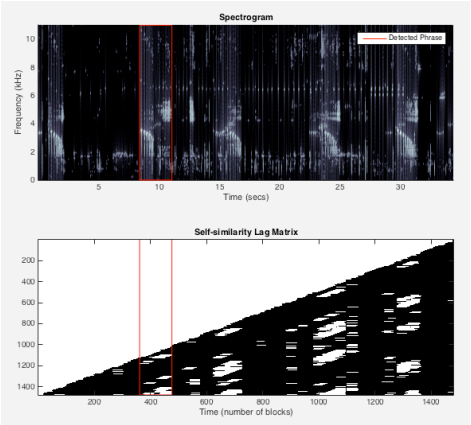


Figure 2

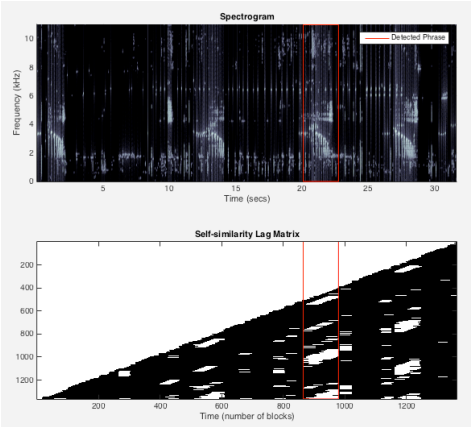


Figure 3

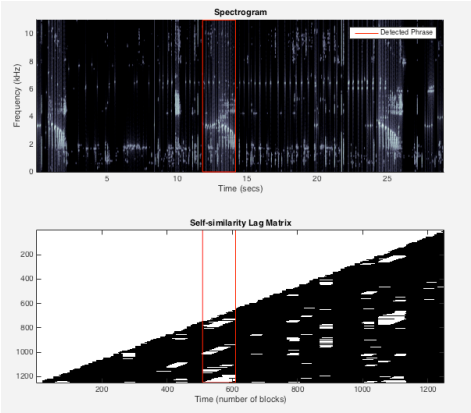


Figure 4

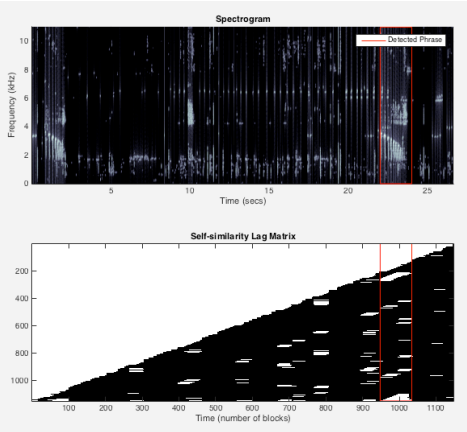


Figure 5

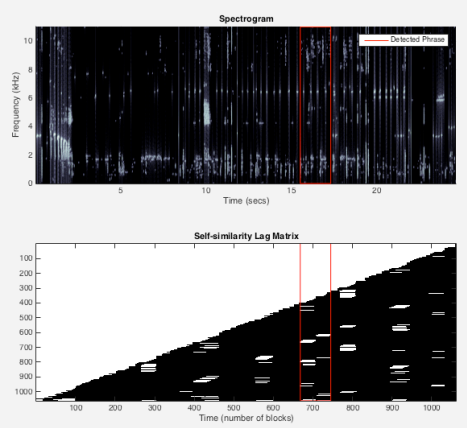


Figure 6

3.2. Feature Extraction and Classification

After pre-processing of the training set, we extract features and train a classifier. The following normalized features were extracted together: spectral flux, spectral centroid, spectral roll-off, spectral flatness, spectral crest, and (time-domain) zero crossings. Mel-frequency cepstral coefficients (MFCCs) were also extracted. The features (either MFCC or spectral + zero-crossings) were used to train an SVM classifier, with gamma=0.1 and cost=1. For comparison, a K-nearest-neighbor classifier was also tested, with K=3.

4. EVALUATION AND DISCUSSION

The segmentation algorithm described above was tested with using MFCCs. We generated results from the dataset, both with and without segmentation. The segmented version resulted in about 3 times as much training data. In figure 7, we see an improvement of 13.1% for SVM and 4.97% for K-NN when segmentation is used.

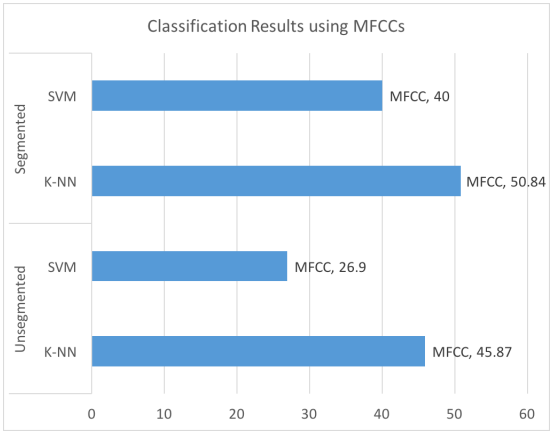


Figure 7

For spectral features, there is a large reduction in the quality of classification when segmentation is used. A possible reason for this is that the samples are too short for spectral features to be effective.

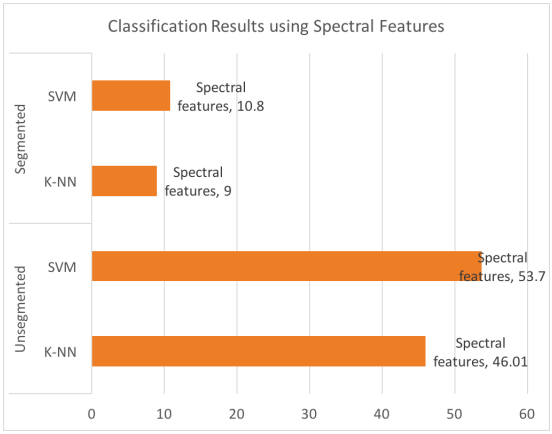


Figure 8

5. CONCLUSIONS AND FUTURE WORK

A novel segmentation technique was applied to bird song classification. We successfully extracted features from segmented phrases of recordings of varying sounds qualities. We achieved better results in classification using segmentation of MFCCs. However, there is room for future improvement in our approach. We have not accounted for the possibility of repeated detection of silences in the segmentation process. The presence of background species in the recordings also introduces the possibility of mis-training the classifier.

6. REFERENCES

[1] A. Vartaki, A. Lerch, “Augmenting Audio-based Bird Species Identification with Music Processing Approaches”, Georgia Tech Center for Music Technology, Georgia Institute of Technology, Atlanta, Georgia, July 2014.

[2] R.O. Tachibana, N. Oosugi, K. Okanoya, “Semi-Automatic Classification of Birdsong Elements Using a Linear Support Vector Machine”, *PLoS ONE 9(3),* Tokyo, Japan, March 2014

[3] J.F.Ruiz-Munoz. M. Orozco-Alzate. G. Castellanos-Dominguez, “Multiple Instance Learning-Based Birdsong Classification Using Unsupervised Recording Segmentation”, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, Bogota, Colombia, July 2015

[4] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction.", *IEEE Transactions 27.2 (1979): 113-120*, April 1979

[5] C.Vincent, “Vuvuzela Sound Denoising Algorithm”, http://www.mathworks.com/matlabcentral/fileexchange /27912-vuvuzela-sound-denoising-algorithm. Last accessed Dec 7 2015, Jun 2010.

[6] J. Paulus, M.Müller, and A. Klapuri. "State of the Art Report: Audio-Based Music Structure Analysis." *ISMIR*. 2010.