Bird sound recognition
ECE 443 Term Project

Ning Zhang, Department of Data Science
Chen Wang, Department of Chemical Engineering

University of Rochester

December 16, 2016

ABSTRACT

In this article a simple method is put forward for bird species recognition. In this method we first give steps for fre-processing of bird sound recordings data, then 3 kinds of different feature extraction methods can be applied on the pre-processed data to generate a series of features as observation sequences. An HMM(Hidden Markov Model) can thus be trained with the observation sequences and be used for predicting unknown bird species.

Index Terms : bird species recognition, spectrogram, hidden Markov Model

1 Introduction

The authors are curious about the ability of HMM in the speech recognition field, yet this topic is too complicated and is out of our range. Thus a more simple yet similar topic was considered, and we decide to perform HMM on bird species recognition based on their recordings. Bird sound recordings give important information of environment quality, while the technology of bird species automatic recognition remains to be immature. Therefore this topic also interests lots of researchers.

2 Data Collection

Our experiment data are fetched from xeno-canto.org[2], which is a bird sound database website supported by Naturalis Biodiversity Center. The recordings on it are uploaded, shared and thus used by ornithologists all over the world under a CC BY-NC-SA 4.0 license[1].
Among the 333,600 recordings of 9676 bird species on xeno-canto, we pick several species for the study, with choosing principles listed as below:

(a) First of all, the recordings should be interesting and should have a unique characteristic that distinguishes themselves from other recordings.
(b) Most recordings chosen are calls and ralatively simple songs. A bird can make various kinds of vocalizations, for example, call, song, flight call, begging call ,alarm call and sometimes,drumming, which can only be made by some woodpecker species. A call is a simple word of bird chirp that usually contains only one syllable, while a song can contain many syllables and can be very complicated and flexibale in frequency and length.
(c) The vocalizations in recordings are usually accompanied by vocalizations from other birds of the same species, other bird species, or even insects. From this sense the training data are picked mostly among the solo recordings, in which there is only vocalization of one single bird and in which the signal reaching a significant signal/noise ratio level.
(d) Ripped from old-style tape equipments, the old-aged recordings, which usually were uploaded in the late 1990s, are avoided because of their coarse qualities.
(e) Some species have rare population, and thus their vocalizations are less likely to be collected. For training a HMM, a relatively abundant dataset is essensial, therefore the rare species are also avoided.
Totally we have 78 recordings, 8 species and totally about one hour of recording length. The species are listed as Table 1 :

Table 1: Bird species involved

Code	scientific name	common name
BirdA	Anas platyrhynchos	Mallard
BirdB	Ortalis vetula	Plain chachalaca
BirdC	Cariama cristata	Red-legged seriema
BirdD	Agelaius phoeniceus	Red-winged blackbird
BirdE	Zenaida macroura	Mourning dove
BirdF	Sylvia communis	Common whitethroat
BirdG	Spizaetus tyrannus	Black hawk-eagle
BirdH	Cistothorus platensis	Sedge wren

3 Data Pre-processing

3.1 Digital Format Conversion

The raw data fetched from Xeno-Canto are stored in *.mp*3 format, hence we first transfrom the vocalizations recordings into *.wav* format. The recordings in stereo mode are merged into a mono channel mode, maintaining a sampling rate of 44100Hz, and the bit depth is set to 8bit in order to gain a faster processing speed. A tentative filter was applied on the vocalizations recordings for noise reduction, but it came out that the processing speed was unbearablely slow, as well as that some bird's call can reach

a scope from 50Hz to as high as 7000Hz, which covers most frequency bands of natural noises occured in the recordings, and removing those noises will inevitably lose large amount of concerned information of the vocalization signal. Therefore it is not possible to reduce the noise with traditional filters like high-pass or low-pass filters.

3.2 Spectrogram

As a common technology to visualize the result of a short-time Fourier Transform(STFT), spectrogram is a stack of multiple spectrums from a time series[5]. The x-axis is usually time and the y-axis is usually frequency (Figure 1). When we intercept a vertical frame from a spectrogram, a column of amplitude values will be generated, which is usually represented by a clormap or grayscale in a spectrogram figure. However, it is not necessarily needed to visualize a spectrogram for HMM training.

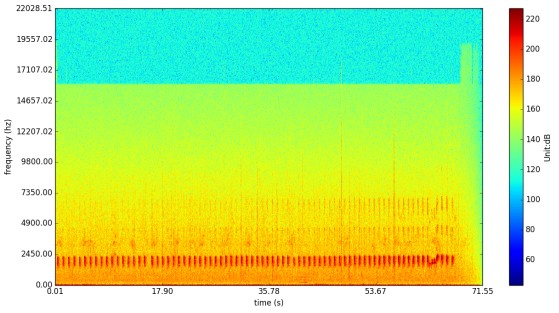

Figure 1: Spectrogram of *Nothocercus julius*

The process of spectrogram generation is performed in a conventional way[5]:
(a) First cut the signal data into several segments with a fixed length(usually a integer power of 2 for FFT efficiency). The segment length in our study is 4096(2^{12}).
(b) Apply a window on the segment by taking convolution of the segment and a windowing function. The window function applied is Hamming Window.
(c) Take FFT of the segment, thus gaining a spectrum of the segment.
(d) Arrange all spectrum together in time order.

Although the hearing range of birds can vary from 50Hz to 12kHz, their most sensitive band is between 1kHz and 5kHz[4]. Therefore in our research only the part with frequency lower than 7kHz is kept, while the spectrogram itself can reach a upper frequency bound of 22.05kHz(half the sampling rate)(Figure 1).

3.3 Mute Area Skipping

Mute area skipping is a widely-used method in after effect of sound data. The certeria of skipping is that when a part of sound signal has a amplitude that is lower than some threshold decibel number, and when this lower signal lasts for longer than some threshold

time. Here the thresholds are 500ms and -35db, with which some tiny noises in the mute area are skipped, thus improving the quality in a sense of signal/noise ratio. After this procedure, the total sample length is shrinked by about $5\% - 15\%$.

4 Feature extraction

The first feature extraction strategy is to take the maximum amplitude $A_{max,t}$ and the corresponding frequency $f_{max,t}$ in each time frame t in the spectrogram. This will thus form a single line-shaped contour of the spectrogram, and roughly describe the shape of the specific pattern in spectrogram. The feature vector sequence then become the observation sequence $O = O_1, O_2, \cdots, O_t$ with a length of t , in which $O_t = (A_{max,t}, f_{max,t})$.

The second strategy adds a series of quantiles to the former one, thus can show the detail of the peak-shape in a more refined way(Figure 2). However this raises the total feature dimension from 2 to 10, resulting in a difficulty to fit the data with HMM. The feature vector generated in this way is

$$O_t = (A_{max,t}, f_{max,t}, f_{0.8ql}, f_{0.6ql}, f_{0.4ql}, f_{0.2ql}, f_{0.8qr}, f_{0.6qr}, f_{0.4qr}, f_{0.2qr})$$

In which q stands for quantile, l stands for left-side, and r stands for right-side.

The third strategy is to take the first n largest peak values and their corresponding

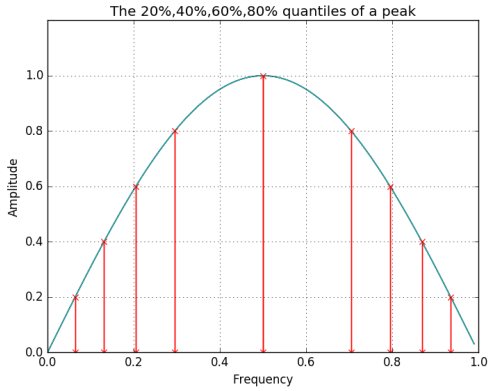

Figure 2: Method 2, the idea of taking quantiles as feature

frequencies and their neighbour values for each time frame, aiming to gain a better description of the overtones as well as the pitch(Figure 3). This is supposed to be able to generate a multi-line shaped contour of the spectrogram and should perform better than the first one. This method is similar to the ideas in [6]. The feature vector generated in this way is

$$O_t = (A_{peak1,max,t}, f_{,peak1,max-1,t}, f_{peak1,max,t}, f_{peak1,max+1,t}, \cdots, f_{peakN,max+1,t})$$

Here N stands for the N-th largest peak in the times frame of spectrogram. In this article the number is set to $N = 5$, thus the dimension of the feature vector is 20.

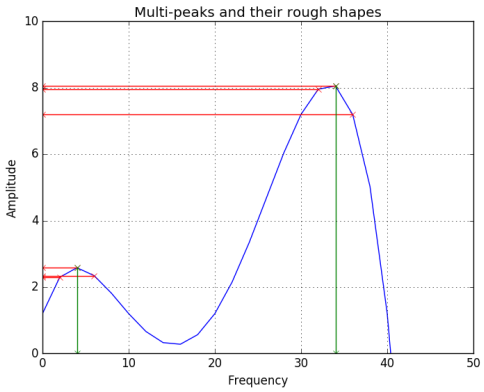

Figure 3: Method 3, extracting features from multiple peaks

5 HMM-based bird species recognition

5.1 Brief introduction to HMM

After segmentation, we get a sequence of vectors $O = O_1, \cdots, O_t$ for each data sample. With analysable data prepared, we are able to choose an appropiate probabilistic model to address the question of bird sound recognition. To our best knowledge, the technique of HMM-based probabilistic modeling can be a good fit for sequential data analyses like the current one[7].

Under the assumption of HMM, there are several Markov states Z, for which the exact amount is unknown and needs to be assigned by researchers before modeling; each observation O_t in a sequence of observations O happens at a Markov state Z_j at a time point t with probability $B_j(O_t)$, which is emission probability; for a state Z_j at time t, it can be followed by a state Z_k at time $t+1$ with probability $A_{j,k}$, which is the transition probability.

HMMs mainly address three basic problems according to Lawrence R. Rabiner [7]: $Problem1$, given that latent states probabilities π, transition probability matrix A and emission probability matrix B are already known, HMMs are able to calculate the probability that a sequence of observations O can happen, which is achieved with the help of forward-backward algorithm; $Problem2$, given the same conditions as in $Problem1$, HMMs are able to calculate a sequence of states with the maximum likelihood to happen, which can be achieved using viterbi algorithm; $Problem3$, given a sequence of observations, HMMs are able to estimate π, A and B so as to maximize the likelihood for the sequence of observations O to happen, which can be achieved using EM algorithm.

5.2 Model Specification

In practice, fitting an HMM to our dataset is equivalent of finding the answer of $Problem3$. Predicting a testset using an HMM is an attempt to estimate the probability of the given data to happen under a set of given parameters, which is nothing but sovling $Problem1$.

Back to our specific case, we first group training data into subsets by bird type Table 1. Then an HMM is fitted to each subset. As a result, an HMM network is obtained in which each node is a model with respect to a specific bird type. We made three experiments on our model using the three feature extraction methods mentioned above. For the features extracted with the first two methods, we assume their probability distribution follow the multivariate normal distribution, while those from the third method are modeled with a mixture of Gaussian, or say GMM.

In each experiment, we also test our model with numbers of hidden states as 2, 4, 8, and 16, and compare respective results in order to find the optimal number of hidden state. Table 2

Table 2: Model performance summary

Extraction Method	No. of hidden states	Accuracy(%)
Method 1	2	78.6
Method 1	4	71.4
Method 1	8	71.4
Method 1	16	71.4
Method 2	2	64.3
Method 2	4	57.1
Method 2	8	57.1
Method 2	16	50.0
Method 3	2	57.1
Method 3	4	57.1
Method 3	8	57.1
Method 3	16	57.1

With the fitted HMM network, we put each of testing datasets into the network for prediction. The model calculates the log likelihood for a testing dataset to happen under each of HMM nodes, and output the predicted bird type with the maximum log likelihood. The experiment results show that with more hidden states used, the accuracy tends to be decreasing. This may be because the over-fit phenomenon, since the observation sequence has simple patterns. The feature extraction methods with more dimensions tend to perform worse than the first method, the reason is that curse of dimensionality[3] appeared in this case. When more dimensions are added into the model, the feature dataset becomes too sparse for giving a reasonable description of the oringal data, and thus a poor fitting result. However, with our skills limited, we are not able to solve this problem for now.

5.3 Evaluation

Our experiment shows that it is not always needed to use too many dimensions of features as observance sequence in order to obtain a better fitting result. Insteadly we

only need two features, the maximal magnitude and the corresponding frequency. More hidden stats turn out to be useless in improving the results. This is controversial to our previous knowledge of data fitting and a curse of dimensionality is believed to happen. We hope that we can approve this in our further research. However, we believe that the results are already good enough, considering the simple structure and idea of our model.

References

[1] Creative commons. `https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode`. Last accessed: 2016-12-12.

[2] xeno-canto. `http://www.xeno-canto.org/`. Last accessed: 2016-12-12.

[3] R. E. Bellman. *Dynamic Programming*. Courier Corporation, 2003.

[4] R. J. Dooling. Auditory perception in birds. *Acoustic communication in birds*, 1:95–130, 1982.

[5] A. B. Downey. *Think DSP:Digital Signal Processing in Python*. Green Tea Press, 2014.

[6] P. Jančovič, M. Köküer, and M. Russell. Bird species recognition from field recordings using hmm-based modelling of frequency tracks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8252–8256. IEEE, 2014.

[7] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.