BIRD SPECIES RECOGNITION COMBINING ACOUSTIC AND SEQUENCE MODELING

Martin Graciarena, Michelle Delplanche, Elizabeth Shriberg, Andreas Stolcke

SRI International, Menlo Park, CA, USA
www.speech.sri.com

ABSTRACT

The goal of this work was to explore modeling techniques to improve bird species classification from audio samples. We first developed an unsupervised approach to obtain approximate note models from acoustic features. From these note models we created a bird species recognition system by leveraging a phone n-gram statistical model developed for speaker recognition applications. We found competitive performance from the note n-gram system compared to a Gaussian mixture model baseline using the same acoustic features. We found an important gain by doing score-level combination relative to the best individual system results. We verified that on most of the bird species under study there was a gain from system combination.

Index Terms— Bird species recognition, phone n-gram modeling, Gaussian mixture model.

1. INTRODUCTION

In this work we aim at bird species recognition based only on bird song audio samples. First some bird song definitions: song syllables are the units that help discriminate bird species. Syllables are composed of notes or elements [1]. A note is a unit of acoustic realization similar to a phone in acoustic modeling of speech. Authors like Härmä [2] have approached the problem of bird species identification by using a specific model of bird song syllables. In [3], Härmä and Somervuo extended this work by using harmonic structure. In [4], Somervuo and Härmä employed a song-level modeling approach using syllable pair histograms. Recently, Somervuo *et al.* [5] compared three feature representations of bird sounds for automatic bird species recognition. In [6], Kwan *et al.* explored Gaussian mixture models (GMMs) and hidden Markov models (HMMs) of acoustic features with the goal of bird species classification to reduce bird strikes to airplanes in airports.

Some of the publications above use hand labeled syllable and note segments. However to the best of our knowledge there is no public database with note or syllable markings that covers a broad range of bird species. We addressed this problem by using unsupervised techniques to obtain note models from a big set of bird species.

The goal of this work is to do bird species recognition based on modeling note sequences. The sequence modeling is based on the phone n-gram statistical model [7] developed for speaker recognition. Instead of defining syllables based on note sequences we directly created bird species models from note sequence statistics. We additionally explore score level system combination with an acoustic feature GMM bird species model.

2. BIRD SONG DATA

In our experiments we used the following publicly available collections of bird song data:

- Macaulay Library of Natural Sounds, *Bird Songs of California*, Cornell Laboratory of Ornithology, Geoffrey A. Keller, 3-CD, 2003
- Peterson Field Guides: Bird Songs: *Western North America, A Field Guide to Western Bird Songs*, Second Ed., Cornell Laboratory of Ornithology Interactive Audio, 1992
- Peterson Field Guides: Bird Songs: *Eastern and Central North America, A Field Guide to Bird Songs*, Third Ed., Cornell Laboratory of Ornithology Interactive Audio, 1990
- *Common Bird Songs* (Audio CD), by Donald J. Borror, Dover Publications, 2003
- *Common Birds and Their Songs* (Book and Audio CD), by Lang Elliott and Marie Read, Houghton Mifflin, 1998
- Stokes Field Guide to Bird Songs: *Western Region* (Audio CD), by Kevin Colver *et al.*, Hachette Audio, 1999

These CD collections contain bird song vocalizations from multiple bird species and were captured using different types of recording equipment. We extracted the bird vocalization segments in the waveforms from background signals using a simple voice activity detection system, with acoustic models trained with bird vocalization data. We additionally discarded very short calls.

Additionally for experiments we used a database [8] of the Borror Laboratory of Bioacoustics (Borror Lab) at the Ohio State University. This database contains multiple high quality recordings bird song samples from nine different bird species (shown in Table 4). The bird song waveforms were hand-segmented into distinct songs and labeled for the song type.

3. UNSUPERVISED NOTE MODELING

Our initial goal was to produce a set of acoustic note models from bird song waveforms. We approached the problem using unsupervised clustering techniques, since there is no formal definition of note boundaries nor is there a public database with note segmentations for different bird species.

These are the steps to train the note models:

1) Do vector quantization of the acoustic features
2) Write cluster index alignments
3) Train note models from step 2) alignments
4) Write note model index alignments
5) Train note models from step 4) alignments

The purpose of doing a two-pass approach is that the models trained in step 3) are better that the clusters trained in 1). The alignments produced should also be better.

Note models are three-state HMMs, similar to speech phone models. For training we used data from the CD collection described in Section 2 which contains data from 93 bird species.

Figures 1 through 3 present examples of the note model index alignments. Each figure comprises two graphs. The upper graph contains the bird song waveform for a given bird species, and the lower graph contains the note alignment for a note model with 60 classes.

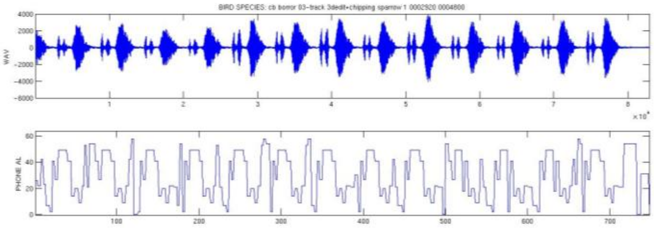

Figure 1 – Bird Song Waveform (upper graph) and Note Model Indexes (bottom graph) for *Chipping Sparrow* Bird Species

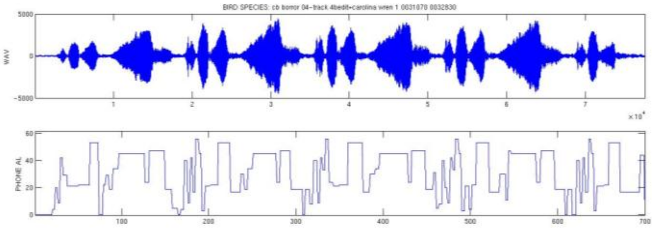

Figure 2 – Bird Song Waveform (upper graph) and Note Model Indexes (bottom graph) for *Carolina Wren* Bird Species

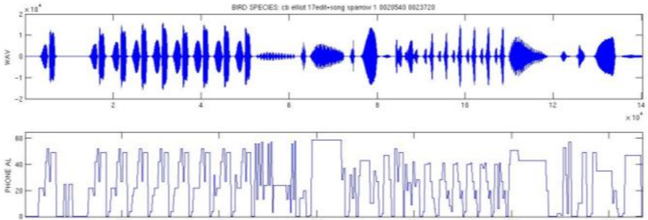

Figure 3 – Bird Song Waveform (upper graph) and Note Model Indexes (bottom graph) for *Song Sparrow* Bird Species

Figure 1 illustrates the waveform and note model indexes or alignments for the *Chipping Sparrow* bird species. As shown, the bird song comprises a repetitive sequence of syllables. The note model alignment somewhat follows the syllable patterns with similar note model index sequences. In Figure 2 for the *Carolina Wren* bird species, the note model index sequence again follows the syllable patterns, just like in Figure 1. However the note model indices from Figures 1 and 2 are different. Figure 3 is the *Song Sparrow* example. This bird species is known for having a very rich repertoire of different syllables. This example illustrates that the complex song comprises some repetitive sub-songs, and those sub-songs have repetitive note index sequences as well.

Syllable units can be derived from note sequences. However, in this work we did not formally define syllable units. Rather, we computed statistics from note sequences, such as unigrams and bigrams. In some sense these statistics capture information similar to syllables, since if a note sequence is repeated often it will have a high probability in our model.

4. NOTE N-GRAM SYSTEM

Once the note models were obtained we used a probabilistic framework to create bird species models. We used the phone n-gram system developed for speaker recognition and called it the note n-gram. The phone n-gram can model phone sequences in a discriminative framework using support vector machines (SVMs). The analogy is that notes in bird songs are the similar to phones in speech.

The note n-gram model has several advantages over the GMM. First, it can model longer time dependencies than the GMM, which operates only at the frame level with some context information given by the deltas. In addition, the SVM provides a discriminative framework where each bird species model is trained to maximize the margin between the correct bird species samples and the impostors. Another advantage is the multiscale modeling provided by using unigrams, bigrams, and trigrams, which basically enables modeling at different lengths. Finally, it provides a framework to model short syllables without explicitly defining them. To train and test with the note n-gram model, we first compute the features for the train and test samples:

1) Compute note-loop lattices from bird song waveforms
2) Extract expected note n-gram statistics from lattices
3) Normalize features (we used rank normalization)

To build bird species-dependent models, we trained a bird species-specific model by training the SVM with correct bird species features and with negative features (from background model data).

Finally, during testing we score the bird species SVM with features computed from the test sample.

5. BIRD SPECIES RECOGNITION EXPERIMENTS

We used a speaker verification paradigm in our experiments. For a given bird species model we used two types of testing data – one from other samples of the same bird species (called *true trials*) and the other using samples from other bird species (called *impostor trials*). In this paradigm the task is to make a decision on whether to accept or reject the trial sample as being from the same bird species as the sample in the training model. If an impostor trial is accepted, it is called a *false acceptance error.* If a true trial is rejected, it is called a *false rejection error*. The equal error rate (EER) is the point at which the percent of false acceptance errors and percent of false rejection errors are equal. Typically, the number of impostor trials is one or two orders of magnitude larger than the number of true trials.

For training and testing experiments we used a database [8] of the Borror Laboratory of Bioacoustics (Borror Lab) at the Ohio State University. This database contains bird song samples from nine different bird species (see Table 4 below for the bird species). The bird song waveforms were hand-segmented into distinct songs and labeled for the song type.

For note model training and background GMM training data we used data from the CD collection described in Section 2. We used data from 93 bird species and four bird song waveform samples per bird species. These recordings had varying sound quality. Some were clean, but some were noisier than the Borror Lab data.

We report results using all songs (which involves testing and training with songs regardless of whether they are the same as or different from the songs used in training) for all nine bird species. They are called All Songs results. We called the case of training and testing with the same song type of a given bird species the Same Song case. Training and testing with a different song was called Different Song case. Only five bird species had different song types, the rest of the bird species produced a single song type.

5.1. GMM Results

A GMM system [9] was used to model mel frequency cepstral coefficient (MFCC) features computed on bird vocalization waveforms. The system is based on the GMM-UBM model paradigm, in which a bird species model is adapted from a universal background model (UBM). Maximum a posteriori (MAP) adaptation was used to derive a bird species model from the UBM. The GMM has either 1024 or 2048 Gaussian components. The front end uses utterance-level mean and variance normalization [7]. The parameters of the front end were the optimal parameters described in our earlier paper on this subject [10].

Table 1 shows the GMM system EER results for All, Same and Different song conditions for two different number of Gaussians.

Table 1 – GMM Results

System - # Gaussians	EER %		
	All Songs	Same Song	Diff Song
GMM – 1024 Gaussians	17.5	14.1	19.7
GMM – 2048 Gaussians	16.7	13.4	19.3

We conclude from Table 1 that in the All Songs case there is a big EER reduction from using a bigger GMM. We also see that the EER of the Same Song case is lower than the EER of the Different Song case. Clearly, it is much more difficult to identify a bird species when the bird species model was trained and tested with Different Songs from the same bird species. However, comparing the EERs from the Same Song and from the Different Songs condition, it is clear that most of the gain comes from the Same Song case.

5.2. Note n-Gram Results

We trained the note models on the birdsong CD data described earlier. Several note models of different number of notes were trained. The acoustic features used to train the note models were the same as those used in the GMM modeling.

We used a full n-gram vocabulary; that is, all possible note index combinations were used.

Table 2 shows the EER results of the note n-gram system. We first explored the number of different notes. We present results comparing the use of unigrams, then unigram and bigrams, and finally unigrams, bigrams, and trigrams. Different statistics are combined by simply concatenating the normalized frequency vectors for each analysis level. Finally, we show results for all songs, same song, and different song condition. The best result is indicated in bold in each column.

Table 2 – Note n-gram Results

System - # Notes Models	n-gram Orders	EER %		
		All Songs	Same Song	Diff Song
Note n-gram - 10 Note Models	1	21.9	17.2	24.3
	1+2	21.2	14.2	24.2
	1+2+3	19.9	13.0	23.3
Note n-gram - 30 Note Models	1	20.3	14.6	23.0
	1+2	17.5	**12.8**	20.0
	1+2+3	16.5	13.5	18.5
Note n-gram - 60 Note Models	1	18.7	14.0	20.8
	1+2	17.3	13.7	19.0
	1+2+3	**16.5**	13.5	18.3
Note n-gram - 90 Note Models	1	18.6	13.9	21.0
	1+2	17.0	14.7	18.5
	1+2+3	17.0	14.8	**18.1**

We see in Table 2 that for the All Songs condition the 60 note model performs the best, and its performance compares well to the GMM system in Table 1. In addition, there is an advantage in all cases of combining unigrams, bigrams, and trigrams over either using unigrams or combining unigrams and bigrams. Comparison of the Same and Different Song conditions showed an opposite trend. For the Same Song condition, using a model of fewer than 60 notes seems optimal. However, for the Different Song condition, a model of more than 60 notes is better. This means that when there are fewer note classes, it is better to model the same song per bird species. However, having more classes improves generalization across different songs of a given bird species. The optimal 60 note classes seem to be a balance between these two trends.

5.3. System Combination Results

In evaluating the score-level combination of the GMM and the note n-gram system, we used a simple equal weight at the score level.

Table 3 shows the system combination EER results using the best GMM system from Table 1 and the best note n-gram from Table 2. We present results for all songs, same song, and different songs condition.

Table 3 – System Combination Results

Systems	EER %		
	All Songs	Same Song	Diff Song
GMM – 2048 Gauss	16.7	13.4	19.3
Note n-gram – 60 Note Mods, 1+2+3 n-grams	16.5	13.5	18.3
Score Combination	**14.1**	**11.7**	**15.5**

Table 3 shows an important EER reduction in all songs from system combination. We found similar relative EER reduction of 14.5% in All Songs, 12.7% in Same Song and 15.3% in Different Song conditions over the best EER in each condition. The important EER reductions indicate that these systems are complementary.

5.4. Results by Bird Species

Table 4 shows the EER results for each bird species, for the GMM, the note n-gram, and the combined systems.

Table 4 – GMM and Note n-gram Results by Bird Species

Bird Species	EER %		
	GMM 2048	Note n-gram 60 /1+2+3	Comb.
Acadian Flycatcher	1.8	3.6	**0.2**
Brown-headed Cowbird	**13.7**	18.3	14.0
Carolina Wren	14.2	14.5	**11.8**
Kentucky Warbler	12.9	12.3	**11.6**
Northern Cardinal	17.0	20.4	**14.6**
Red-winged Blackbird	5.3	1.6	**1.5**
Song Sparrow	25.7	20.1	**19.5**
Swamp Sparrow	24.3	**17.9**	19.2
White-crowned Sparrow	3.4	14.0	**2.2**

From Table 4 we can conclude that for most of the bird species there is an important gain from combining both of the proposed systems compared to the best individual system. Thus having complementary systems is an advantage for most of the bird species.

6. CONCLUSIONS

We explored two modeling techniques and score-level system combination for bird song recognition. We first developed an unsupervised approach to obtain note acoustic models. From these note models we created a bird species recognition system by leveraging the phone n-gram model developed for speaker recognition applications. We found competitive performance from the note n-gram system compared to a GMM baseline using the same acoustic features. We found important gains from using score-level combination relative to the individual systems. When analyzing the bird species-specific results, we found that the combination leads to gains for most of the bird species used in the experiments.

7. ACKNOWLEDGMENTS

We thank Michelle Hewlett for helping with segmentation and song labeling of the Borror Laboratory bird song data. The views are those of the authors and do not represent the views of the funding agency. This work was supported by the Sandia National Laboratories.

8. REFERENCES

[1] C. K. Catchpole and P. J. B. Slater, Bird Song: Biological Themes and Variations, Cambridge University Press, Cambridge, MA, 1995.

[2] A. Härmä, “Automatic Identification of Bird Species Based on Sinusoidal Modeling of Syllables,” in *Proc. ICASSP*, Hong Kong, 2003.

[3] A. Härmä and P. Somervuo, “Classification of the Harmonic Structure in Bird Vocalization,” in *Proc. ICASSP*, Montreal, Canada, 2004.

[4] P. Somervuo and A. Härmä, “Bird Song Recognition Based on Syllable Pair Histograms,” in *Proc. ICASSP*, Montreal, Canada, 2004.

[5] P. Somervuo, A. Härmä, and S. Fagerlund, “Parametric Representations of Bird Sounds for Automatic Species Recognition,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2252–2263, Nov. 2006.

[6] C. Kwan, G. Mei, X. Zhao, Z. Ren, R. Xu, V. Stanford, C. Rochet, J. Aube, and K.C. Ho, “Bird Classification Algorithms: Theory and Experimental Results,” in *Proc. ICASSP*, Montreal, Canada, 2004

[7] S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, “The SRI NIST 2008 Speaker Recognition Evaluation System,” in *Proc. ICASSP*, Taipei, Taiwan, 2009.

[8] Borror Laboratory of Bioacoustics, Ohio State University. http://blb.biosci.ohio-state.edu

[9] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker Verification Using Adapted Mixture Models,” *Digital Signal Processing*, vol. 10, pp.181-202, 2000.

[10] M. Graciarena, M. Delplanche, E. Shriberg, A. Stolcke, and L. Ferrer, “Acoustic Front-end Optimization for Bird Species Recognition,” in *Proc. ICASSP*, Dallas, TX 2010.