NOISE ROBUST BIRD SONG DETECTION USING SYLLABLE PATTERN-BASED HIDDEN MARKOV MODELS

Wei Chu

Department of Electrical Engineering
University of California, Los Angeles
Los Angeles, 90024
weichu@ee.ucla.edu

Daniel T. Blumstein

Department of Ecology and Evolutionary Biology
University of California, Los Angeles
Los Angeles, 90024
marmots@ucla.edu

1. ABSTRACT

In this paper, temporal, spectral, and structural characteristics of Robin songs and syllables are studied. Syllables in Robin songs are clustered by comparing a distance measure defined as the average of aligned LPC-based frame level differences. The syllable patterns inferred from the clustering results are used for improving the acoustic modelling of a hidden Markov model (HMM)-based Robin song detector. Experiments conducted on a noisy Rocky Mountain Biological Laboratory Robin (RMBL-Robin) song corpus with more than 75 minutes of recordings show that the syllable pattern-based detector has a higher hit rate while maintaining a lower false alarm rate, compared to the detector with a general model trained from all the syllables.

2. INTRODUCTION

Bird songs play a vital role in the communication between individuals and species. A bird may listen to other birds and classify them as conspecific or heterospecific, neighbour or stranger, mate or non-mate, kin or non-kin [1]. It may also sing to other birds for mate attraction, or territory defense [2]. Ecological and behavioral studies can benefit from automatically detecting and identifying species or individuals from audio recordings.

The motivation of this study is to automatically detect the existence of the Robin songs from continuous recordings collected from Colorado.

Machine learning methods, such as back propagation and multivariate statistics [3], artificial neural networks [4], evolving neural networks [5], dynamic time warping and hidden Markov models [6] [7] [8], are effective for classifying bird and other animal sounds given pre-segmented acoustic recordings; however, for continuous audio stream in which no boundary information is available, it is important to have a recognizer that can both detect the songs and classify the species.

An HMM-based detector with a general model trained from all the syllables is designed as a baseline system. In an improved system, syllable patterns are first inferred from similar syllables observed in the recordings; HMMs of the inferred syllable patterns are then trained to allow finer acoustic modelling of the syllables. According to our experimental results, the proposed syllable pattern-based detector is promising in terms of the hit rate and false alarm rate.

The paper is organized as follows. Section 3 discusses the characteristics of the Robin song and syllable. Section 4 introduces the recorded Robin database. Section 5 shows how the syllable patterns are inferred. Section 6 describes the acoustic modelling and network of the HMM-based detection system. Section 7 discusses the experimental results.

Supported in part by NSF grants DBI-0754247 and DBI-0821369.

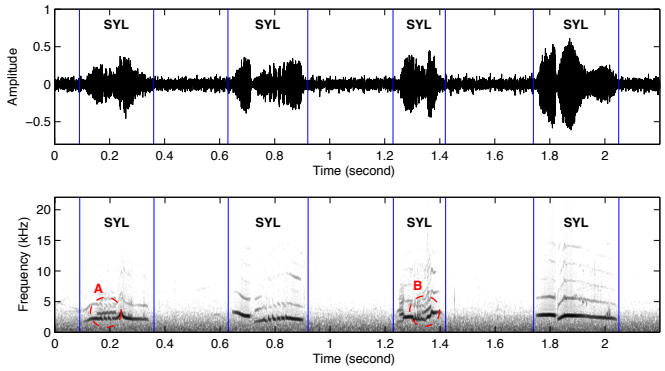

Fig. 1. Time waveform and spectrogram of a typical Robin song. **SYL** refers to the syllable units.

3. ROBIN SYLLABLE AND SONG

A typical Robin song is shown in Fig. 1. It can be seen that the song is composed of several different syllables. Note that these units are sometimes referred as phrases or song types. Although these syllables have similar harmonic structures as the voiced speech of human beings, there are three main differences. The first is that the pitch of the Robin is higher than that of human with fundamental frequencies ranging between 1500 and 4500 Hz. The second is that Robins can only intermittently vocalize syllables, but not continuously as can humans. The third is that Robins may produce two pitches simultaneously during vocalization as shown in the circled regions A and B in Fig. 1. The phenomena can be attributed to how birds produce songs [9]. During Robin vocalizations, air flows from two different syrinxs are controlled by the lateral labium and the medial tympaniform membranes. These membranes are located on the medial walls of the bronchus and these morphological structures enable Robins to have two voicing sources. When the controllers of the two sources are vibrating at different speeds, two different pitch frequencies are produced simultaneously.

4. RMBL-ROBIN DATABASE

The RMBL-Robin database used in this study was collected by using a close-field song meter (www.wildlifeacoustics.com) at the Rocky

Table 1. The details of RMBL-Robin database

	Length (minutes)	Syllable #	Song #
Training Set	45.5	1644	457
Test Set	32.8	970	277

Mountain Biological Laboratory near Crested Butte, Colorado. The sampling rate is 44.1 kHz. The recorded Robin songs are naturally corrupted by different kinds of background noises, such as wind, water and other vocal bird species. Non-target songs may overlap with target songs. Each song usually consists of 2-10 syllables. The dataset is 78.3 minutes long and divided into two sets for training and testing purposes. The details of the database is shown in Table 1. Note that all the analysis is conducted on the training set.

5. INFERENCE OF SYLLABLE PATTERNS

Objectively inferring syllable patterns is not only important in studying the singing behaviour of Robins, but also necessary to improve Robin song detection in the audio stream.

5.1. Distance Measure Between Syllables

A distance measure which was originally used for isolated word recognition is adopted. The distance between two syllables is defined as the minimum accumulative frame-level difference obtained in a dynamic time warping scheme [10]. The difference between two frames is based on the log likelihood ratio of the minimum prediction error [11].

The details of the distance measure is described in the following.

The log likelihood ratio of the prediction error from frame y to frame x, $D(y||x)$, is defined as

$$D(y||x) = \log \frac{E_{yx}}{E_{xx}} = \log \frac{\mathbf{a}_y^T \mathbf{R}_x \mathbf{a}_y}{\mathbf{a}_x^T \mathbf{R}_x \mathbf{a}_x} \tag{1}$$

where E_{yx} denotes the error obtained by feeding frame y into the inverse LPC filter inferred from frame x, E_{xx} is the minimum prediction error for the LPC system inferred from frame x; $\mathbf{a}_x$ and $\mathbf{R}_x$ denote the LPC coefficients and autocorrelation coefficient matrix of frame x, $\mathbf{R}_x$. In this paper, we use a symmetric difference measure, $D_f(x, y)$, defined as

$$D_f(x, y) = \frac{1}{2}\big[(D(x||y) + D(y||x)\big] \tag{2}$$

$D_f(x, y)$ does not satisfy the triangular inequality. Because of its nonzero and symmetric properties, it can still be used as a difference measure between two different analysis frames.

In this study, a fixed frame rate LPC analysis is firstly conducted on the training set to acquire the distribution of the difference $D_f(x, y)$ between two adjacent frames. There are some frames between which the distances change slowly. Downsampling of the LPC analysis over these frames is essential to remove redundant information. When the distances are changing rapidly between other frames, an upsampling of the LPC analysis is also necessary to capture the rapidly changing pitch information. In essence, a variable frame rate (VFR) [12] LPC analysis is then applied on each syllable.

Then, a symmetric distance measure between the two syllable $\mathbf{X}$ and $\mathbf{Y}$ denoted by $D_s(\mathbf{X}, \mathbf{Y})$ is defined as

$$D_s(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}\big[D_s(\mathbf{Y}||\mathbf{X}) + D_s(\mathbf{X}||\mathbf{Y}))\big] \tag{3}$$

where $D_s(\mathbf{Y}||\mathbf{X})$ denotes the distance from syllable $\mathbf{Y}$ to syllable $\mathbf{X}$. It is obtained through dynamic time warping (DTW) [13], i.e. minimizing the accumulative aligned frame-level differences defined in Eq. 2.

Although the defined distance $D_s(\mathbf{X}, \mathbf{Y})$ does not satisfy the triangular inequality, it was used as a distance measure for isolated word recognition [10], and can be used as the distance measure for the Robin syllable clustering in the following section.

5.2. Hierarchical Clustering Analysis

The objective of clustering analysis in this section is to search common patterns which allow fine acoustic modelling of the Robin syllables compared to only using one single general pattern for all the syllables. Training different models or templates for different keywords has been proved to be effective for keyword spotting [14] in which phoneme level transcription is available. However, for the training set of Robin songs, only boundary information of the syllables is annotated. Thus, it is necessary to infer the numbers of the common syllable patterns from the training set, and then train acoustic models for those patterns.

Providing the distance measure between two syllables defined in the previous section, it is possible to conduct a distance measure-based hierarchical clustering analysis. In this study, a modified average-linkage hierarchical clustering is used to reliably cluster syllables into patterns. Before introducing the algorithm, the inter-cluster distance of cluster C, $D_c(C)$, is defined as

$$D_c(C) = \frac{1}{N_C(N_C - 1)} \sum_{i=1}^{N_C} \sum_{j=1}^{N_C} D_s(\mathbf{X}_i, \mathbf{X}_j) \tag{4}$$

where N_C denotes the syllable numbers in the cluster, $\mathbf{X}_i$ denotes the i_{th} syllable in the cluster. The intra-cluster distance between cluster C_a and C_b denoted by $D_c(C_a, C_b)$ is defined as

$$D_c(C_a, C_b) = \frac{1}{N_{C_a} N_{C_b}} \sum_{i=1}^{N_{C_a}} \sum_{j=1}^{N_{C_b}} D_s(\mathbf{X}_i^{C_a}, \mathbf{X}_j^{C_b}) \tag{5}$$

where N_{C_a} denotes the syllable numbers in the cluster C_a, and $\mathbf{X}_i^{C_a}$ denotes the i_{th} syllable in the cluster C_a.

The pseudocode of the modified average-linkage hierarchical clustering algorithm is expressed in the following:

Algorithm 5.1: A MODIFIED HIERARCHICAL CLUSTERING (C)

Set the stopping distance threshold as $D_{\max}^C$
Each syllable is initiated as a cluster.
do {
  Search the closest two clusters, C_{i^*} and C_{j^*}, by comparing $D_c(C_i, C_j)$
  Copy the elements of C_{i^*} and C_{j^*} into a new cluster C^*
  if $D_c(C^*) > D_{\max}^C$
    then Remove C^*, **break** ;
  else
    then Use C^* to replace C_{i^*} and C_{j^*}
while More than one cluster is left
$D_c(C)$: intra cluster distance of cluster C;
$D_c(C_a, C_b)$: inter cluster distance of cluster C_a and C_b;

In this paper, only clusters with numbers of syllables greater than a threshold denoted by N_{th}^C are retained as syllable patterns. The

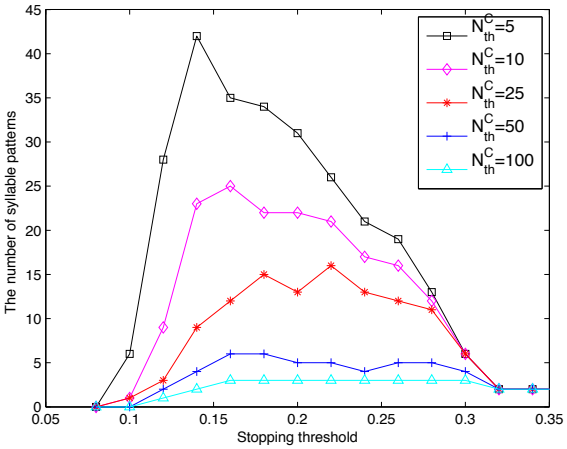

Fig. 2. The relationship between the number of syllable patterns and stopping distance threshold $D_{\max}^C$ given different cluster number threshold N_{th}^C. Only clusters with numbers of syllables great than N_{th}^C are regarded as syllable patterns.

relationship between the number of syllable patterns and stopping distance threshold $D_{\max}^C$ given different N_{th}^C is shown in Figure 2. Under the same clustering stopping threshold, the larger the cluster number threshold is, the fewer syllable patterns there are. Under each cluster number threshold, the number of syllable patterns first increases then decreases when the clustering stopping threshold $D_{\max}^C$ increases. It might be because when $D_{\max}^C$ is small, many small clusters are not regarded as syllable patterns; when $D_{\max}^C$ has a high value, i.e. the allowable maximum intra-cluster distance is high, many syllables are clustered together, which causes the number of patterns to be small.

It is still difficult to infer the actual numbers of the syllable patterns from the clustering results, because biologists are not clear about the repertoire size of the syllable patterns in Robin songs. However, the clustering results are helpful in the sense of training acoustic models from the syllable patterns that are close in a certain feature space, which may improve detection and classification results.

6. ROBIN SONG DETECTION SYSTEM

During training, feature segments required by the template-based approach, i.e. DTW, can be obtained by examining the boundary information contained in the transcriptions. However, the boundary information is no longer available in the test set which implies that the template-based method is not suitable for the detection task, and pre-processing is needed to acquire the boundary information. As an HMM-based system with models of the Robin syllables and background sounds is capable of detecting the boundaries and classifying the sounds, by decoding the continuous feature stream, simultaneously, HMMs are used for acoustic modelling in our detection task. A left-to-right HMM with 3 emitting states is adopted for modelling the syllable patterns; an ergodic HMM with 3 emitting states is used for modelling the background sounds.

Two HMM network A and B are constructed for acoustic model training and audio feature stream decoding purposes. Network A shown in Figure 3 models all syllables as a single general HMM, and all background sounds as another general HMM. The difference between networks A and B, shown in the Figure 4, is that different syllable patterns are modeled as different HMMs. As mentioned

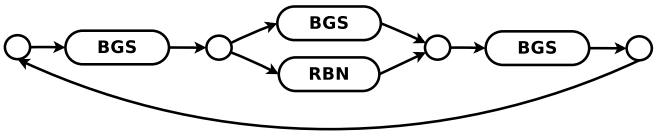

Fig. 3. HMM network A. **RBN**: the general HMM for all Robin syllables. **BGS**: background sound HMM.

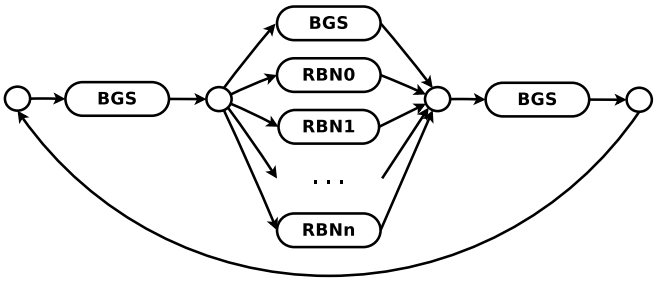

Fig. 4. HMM network B. **RBNn**: the HMM for the n_{th} Robin syllable pattern. **RBN0**: the HMM for the remaining Robin syllables that do not belong to any syllable pattern. **BGS**: background sound HMM.

above, not all syllables can be clustered into a syllable pattern. An extra HMM with the same topology as the syllable pattern HMM is used for modelling unclustered syllables. Syllable patterns are inferred by using the clustering-based method mentioned in the previous section.

Bigram models for both HMM networks are learned from the training set such that each arc in the network is assigned with a transition probability. The integration of the bigram model into the HMM networks implies the occurrence relationship between the syllable and background sounds are taken into consideration.

Unsupervised Maximum Likelihood Linear Regression (MLLR) adaptation [15] is applied to minimize the mismatch between the trained acoustic models and the test cases.

As we are interested in detecting the existence of the Robin songs, the syllable level decoding results are needed to be converted to song level results. According our observation, the duration between the syllables in a Robin song is less than 0.5 seconds most of the time. Therefore, detected syllables that are less than 0.5 seconds in distance are grouped into a single song.

7. EXPERIMENTAL RESULTS

The performance of the Robin song detection is evaluated in terms of the recall rate and precision rate denoted by R and P which can be expressed as

$$R = \frac{N_h}{N_g} \times 100\%, \quad P = \frac{N_h}{N_d} \times 100\%, \tag{6}$$

where N_h is the number of hit songs, N_g is the number of the ground truth songs, and N_d is the number of detected songs. A detected song is regarded as a hit song only if the center of the detected song in time falls into the vicinity (± 0.5 seconds) of the center of a ground truth song.

The objective is to increase the recall rate and precision rate at the same time. Because of the well-known trade-off relationship between the two rates, the F-score, a weighted combination of the two rates denoted by F [16], is defined as:

$$F = \frac{(1+\beta^2)PR}{\beta^2 P + R}, \tag{7}$$

Table 2. the detection results including the Recall Rate (R), Precision Rate (P), and F-score (F) using HMM network A and B. **wo VFR**: only use a fixed frame rate in syllable pattern clustering. **+ adapt**: unsupervised MLLR adaptation.

	R (%)	P (%)	F
Network A	74.2	71.8	0.734
Network B wo VFR	75.5	73.3	0.748
Network B	76.0	73.6	0.753
Network B + adapt	76.0	75.2	0.758

where the β is a weighting factor. Since the recall rate is more important than the precision rate in this study, β is set to be 1.5.

The sampling rate of the recordings is 44.1 kHz. When the microphone is far from the vocalizers during the recording, the high frequency components ($>$ 5000 Hz) of the songs sometimes are lost. As the pitch information of the Robin ranging from 1500 to 4500 Hz are retained most of the time, a band pass filter with cut-off frequencies of 1000 and 5000 Hz is applied to the raw recordings.

For Robin syllables, the magnitude of the first harmonic is usually higher than other harmonics, and hence is less susceptible to background noise. As a pair of conjugate poles of the LPC filter is supposed to match one spectral peak, given the fact that there may exist one or two pitch harmonics in the pass-band, i.e. one or two spectral peaks in the spectrum, the order of LPC has to be at least 4 to capture all possible pitches.

In the fixed frame rate LPC analysis, a frame rate of 5 ms is used. In the variable frame rate-based LPC analysis, effective frame rates 5, 10, and 20 ms are used. The low and high thresholds are set as 0.13 and 0.36 respectively, which makes the ratio of the numbers of frames with high, middle, and low frame rates to be 1:1:1. In both analyses, the frame length is 10 ms.

In feature extraction, to be consistent with the LPC-based clustering analysis, a 15-dimension feature composed of the 4th-order LPCs plus logarithm energy and first and second derivatives is computed every frame for model training and testing. The frame step size is fixed to 5 ms. The frame length is 10 ms.

In the variable frame rate-based clustering analysis, the clustering stopping threshold $D^C_{\max}$ ranges from 0.08 to 0.40, the syllable number threshold in a cluster N^C_{th} is set to be 5, 10, 25, 50, or 100. In acoustic modelling, the number of Gaussian mixtures per state is set to be $1, 2, 4, 8, 16$, or 32. For the HMM network B, the highest F-score is achieved when $D^C_{\max} = 0.12$, $N^C_{\text{th}} = 25$, and the number of Gaussian mixtures per state is 8. Changing the number of the states in the HMMs to other than 3 can not improve the F-score. Under this configuration, there are 3 HMMs for syllable patterns and 1 HMM for the background sound. The details of the detection results using HMM networks A and B are shown in Table 2. When replacing the simple HMM network A with the advanced network B, the recall and precision rate are both improved by 1.8% . When the network B is followed by an unsupervised MLLR adaptation module, the precision rate has a gain of 1.6% while the recall rate keeps unchanged. We also found that using a fixed frame rate in the syllable pattern clustering can result in a lower recall and precision rate.

8. CONCLUSIONS

In this paper, syllable patterns of Robin songs can be objectively inferred by performing a hierarchical clustering analysis in which the distance measure is calculated by aligning the LPC-based frame level differences. This HMM-based Robin song detection system with models trained for the syllable patterns has a higher hit rate under the same false alarm rate compared with a system models trained from all syllables.

9. REFERENCES

[1] P. Marler, "A comparative approach to vocal learning: song development in white-crowned sparrows," *J Comp Physiol Psychol*, vol. 71, pp. 1–25, 1970.

[2] C. K. Catchpole and P. J. B. Slater, *Bird Song: Biological Themes and Variations*, Cambridge University Press, New York, 1995.

[3] A.L. McIlraith and H.C. Card, "Bird song recognition using back propagation and multivariate statistics," *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2740–2748, 1997.

[4] A.L. McIlraith and H.C. Card, "Bird song identification using artificial neural networks and statistical analysis," in *Proc. of IEEE Canadian Conference on Electrical and Computer Engineering*, 1997, vol. 1, pp. 25–28.

[5] L. Ranjard and H.A. Ross, "Unsupervised bird song syllable classification using evolving neural networks," *JASA*, vol. 123, no. 6, pp. 4358–4368, 2008.

[6] J.A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," *JASA*, vol. 103, no. 4, pp. 2185–2196, 1998.

[7] V. Trifa, A. Kirschel, and C. E. Taylor, "Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models," *JASA*, vol. 123, no. 4, pp. 2424–2431, 2008.

[8] T.S. Brandes, "Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 16, no. 6, pp. 1173–1180, 2008.

[9] R.A. Suthers, F. Goller, and C. Pytte, "The neuromuscular control of birdsong," *Philos Trans R Soc Lond B Biol Sci*, vol. 354, no. 1385, pp. 927–939, 1999.

[10] L. Rabiner, A. Rosenberg, and S. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 26, no. 6, pp. 575–582, 1978.

[11] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975.

[12] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in *Proc. of ICASSP*, 2000, vol. 3, pp. 1783–1786.

[13] C. Myers, L. Rabiner, and A. Rosenberg, "An investigation of the use of dynamic time warping for word spotting and connected speech recognition," in *Proc. of ICASSP*, 1980, vol. 5, pp. 173–177.

[14] J.R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden Markov modeling for speaker-independent word spotting," in *Proc. of ICASSP*, 1989, pp. 627–630.

[15] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[16] C. J. van Rijsbergen, *Information Retrieval*, Butterworth-Heinemann, 1979.