BIRD SONG RECOGNITION BASED ON SYLLABLE PAIR HISTOGRAMS

Panu Somervuo[*]

Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT
Finland
email: panu.somervuo@hut.fi

Aki Härmä[†]

Laboratory of Acoustics and
Audio Signal Processing
Helsinki University of Technology
P.O.Box 3000, FIN-02015 HUT
Finland
email: aki.harma@hut.fi

ABSTRACT

Bird song can be divided into a sequence of syllabic elements. In this paper we investigate the possibility of bird species recognition based on the syllable pair histogram of the song. This representation compresses the variable-length syllable sequence into a fixed-dimensional feature vector. The histogram is computed by means of Gaussian syllable prototypes which are automatically found given the song data and the dissimilarity measure of syllables. Our representation captures the use of the syllable alphabet and also some temporal structure of the song. We demonstrate the method in bird species recognition with song patterns obtained from fifty individuals belonging to four common passerine bird species.

1. INTRODUCTION

The work reported in this paper is related to the development of technology for automatic recognition of bird songs. Technology for sound-based identification of birds would be a significant addition to the research methodology in ornithology, and biology in general. There is also significant commercial potential for such systems because bird watching is a popular hobby in many countries. Extensive international programs such as the Global Biodiversity Information Facility (www.gbif.org) which are building biological multimedia databases facilitating automatic classification and identification of species are also boosting the activity in the area of bioacoustic signal processing and pattern recognition. Nevertheless, relatively little has been done previously in the field. In a few studies the feasibility of automatic recognition of bird species [1, 5, 6] using sound has been demonstrated. This article is a follow-up work to [3] which presented promising results in automatic recognition of fourteen Finnish song bird species. That work was based on the use of separate syllables as the recognition unit. A subset of the data was also investigated by means of the Self-Organizing Map in [9]. The species used in the present study are listed in Table 1.

Bird song is typically divided into four hierarchical levels: notes, syllables, phrases, and song [2]. Syllables can be seen as elementary building blocks of bird vocalization [1]. Fig. 1 shows an example of three songs. In many species there is high individual and regional variability in phrases and song patterns. This can be seen both as a drawback and advantage, depending on the application. It may facilitate the identification of bird individuals, but at the same time it makes the bird species recognition more challenging.

A histogram collected from single syllables would be invariant to the temporal structure of the song. But the histogram based on consecutive syllables, like in our case the syllable pairs, is able to reveal some temporal structure of the song. The main contributions of our present work are the representation of the variable-durational song by means of the Gaussian syllable prototypes, how to construct the prototype bases, how to avoid sparseness of the histogram representation, and finally, how to compare the histograms.

2. METHODS

The representation of the song in this study is based on the syllable histograms. In order to form the histograms, the syllable space must first be divided into bins. This is done by finding a set of syllable prototypes. First the dissimilarity measure between the syllables is defined and the prototypes are then automatically found based on this measure.

[*] Dr. Somervuo was supported by the Academy of Finland, project no. 44886 “New information processing principles” (Finnish Centre of Excellence Programme 2000-2005)

[†] Dr. Härmä was supported by the Academy of Finland and the graduate school GETA

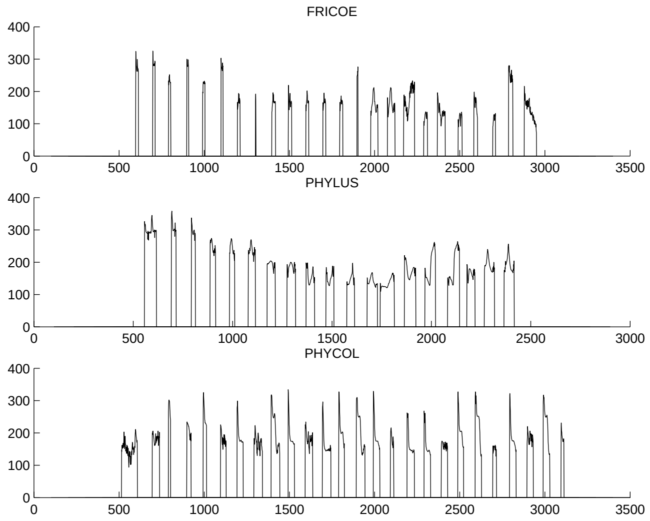

Fig. 1. Examples of bird songs after sinusoidal modeling [3]. Frequency trajectories of consecutive syllables are from three birds. The two songs from FRICOE and PHYLUS (on the top) are melodic where the structure of song is spread over several consecutive syllables whereas the song of PHYCOL (on the bottom) is more "binary"; lower- and higher-frequency syllables alternate in the sequence.

Table 1. Birds in the current study. Columns give an abbreviation derived from the Latin name (a widely used convention), the Latin name, and a common English name. The numbers of FRICOE, PHYLUS, PHYCOL, and PARMAJ individuals in the current data set are 12, 14, 13, and 11, respectively.

Lat. Abbr.	Latin name	Common name
FRICOE	Fringilla coelebs	Common Chaffinch
PHYLUS	Phylloscopus trochilus	Willow Warbler
PHYCOL	Phylloscopus collybita	Common Chiffchaff
PARMAJ	Parus major	Great Tit

2.1. Dissimilarity of syllables

A typical duration of a syllable is in the range of a few to a few hundred milliseconds and it may feature rapid changes in the spectrum. In some cases there may be dozens of different syllables per second in bird song.

In [3], the comparison of syllables was based on computing the Euclidean distances between the trajectories of sinusoid parameters. Variable-length sequences were first zero padded to equal lengths and aligned so that the frames corresponding to the maximum values of the amplitude envelope in two sequences corresponded to each other [3]. This facilitated the use of Euclidean distance without time axis warping. However, in the present work the similarity between two syllables is defined using dynamic time warping (DTW) [8]. The cumulative distance between the feature vectors of two sequences is computed along the warping function which changes the time axis of the sequences nonlinearly so that the maximum fitting between the sequences is attained. DTW handles well the durational differences between sequences. In [9] it was found that DTW outperformed the maximum amplitude alignment based Euclidean distance between syllables when the task was single-syllable based bird species recognition.

Durational differences of the syllables are ignored since the cumulative distances obtained by DTW are normalized by the lengths of the syllables. But although the timing information was not used in thie present study, the durations of syllables and especially the durations of silence regions between the syllables could be used as an auxiliary information in future studies.

2.2. Gaussian syllable prototypes

Based on the DTW-distances between syllables, k-means type clustering algorithm can be used [7]. The following algorithm for finding k syllable prototypes is used:

Step 1. Select randomly k syllables in the data set and use them as initial prototypes.

Step 2. For each data syllable in the data set: compute its dissimilarity against all prototypes and add the data syllable to the list of its 'closest' prototype (having the smallest DTW-distance).

Step 3. For each prototype: replace the old prototype by the centermost data syllable in its list. The centermost data syllable is defined having the smallest sum of DTW-distances to other syllables in the list.

Step 4. Repeat Steps 2 and 3 until convergence.

In our experiments, the algorithm typically converged in five iterations. Several random initializations can be used and the syllable prototypes which give the smallest quantization error is then selected. Quantization error is computed by summing the smallest DTW-distances between the data syllables and their closest prototypes.

This kind of k-means type clustering gives a codebook which already can be used as a basis of song representation. But in our work we used the result as a basis for building Gaussian prototypes. Based on the quantization errors of syllable data we can add a variance parameter to each prototype. This enables us to compute the posterior probabilities of prototypes $i = 1 \ldots k$ for any data syllable x in the following way:

$$p(i|x) = \frac{1/(\sqrt{2\pi}\sigma_i)\exp(-d_{ix}^2/2\sigma_i^2)}{\sum_{j=1}^{k} 1/(\sqrt{2\pi}\sigma_j)\exp(-d_{jx}^2/2\sigma_j^2)} \;, \qquad (1)$$

where d_{ix} is the DTW-distance between data syllable x and prototype i, and σ_i^2 is the variance parameter. The variance parameter of prototype i can be set by computing the average of the squared DTW-distances between the prototype i and the data syllables to whom the prototype i is the closest prototype.

Fig. 2 illustrates the posterior probabilities of prototypes for data syllables using Eq. (1). Already from this picture it can be noticed that the song structure of FRICOE and PHYLUS are mutually more similar than the two other species. In fact, typical melodic line in territorial singing of PHYLUS and FRICOE are strikingly similar and the species are easy to confuse. However, temporal and spectrum structure of a typical syllable is different. For example, in [4] it was found that 74 % of syllables from PHYLUS are almost pure sinusoidal chirps. In FRICOE, syllables have a strong harmonic structure and the second harmonic of the fundamental frequency is dominant in almost 33 % of the syllables. In the current work, however, the syllable representation was based on only single time-varying sinusoid as in [3].

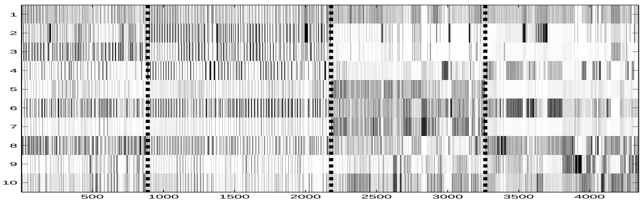

Fig. 2. Prototype posteriors for song data. Vertical axis represents the posterior value of 10 Gaussian prototypes for each of the 4344 syllables from 257 songs in the horizontal axis. Vertical dashed lines separate four species, from left to right: FRICOE, PHYLUS, PHYCOL, and PARMAJ.

2.3. Histogram of consecutive syllables

N-gram is a sequence of N consecutive symbols. It is simple to extract the N-grams from a symbol string. For syllable sequence, we can use the indices of the best-matching prototypes in order to build the N-grams. But representing the syllable using only one prototype in time gives quite crude representation of the original syllable. The more accurate representation is gained if the number of prototypes is increased, but this results in sparse N-grams and the comparison of sequences is not robust. For some songs in our data set, the number of syllables is relatively small and therefore we need proper smoothing for the N-grams. This is our main motivation for using Gaussian prototypes instead of simple vector-quantization type histogram bins. By means of Gaussians we can represent the syllables smoothly using all prototypes (not just the closest prototype) and thus avoid the sparseness of the N-grams. The number of consecutive syllables N can be arbitrary, but in the current work we have used syllable pairs (bigrams, N=2).

Let x_{t-1} and x_t denote two consecutive syllables of the song and $\mathbf{p}_{t-1}$ and $\mathbf{p}_t$ the corresponding posterior probability vectors of Gaussians, respectively. The value of bigram for prototype pair i, j and syllables x_{t-1}, x_t is:

$$h_{i,j}(t) = \frac{p_{t-1,i}p_{t,j}}{\sum_{i',j'} p_{t-1,i'}p_{t,j'}} \ . \qquad (2)$$

The smooth bigram $h_{i,j}(t)$ for all i, j can be conveniently expressed by the product of $\mathbf{p}_{t-1}$ and $\mathbf{p}_t^T$ ($\mathbf{p}$ is a column vector and T denotes the transpose so $\mathbf{p}_{t-1}\mathbf{p}_t^T$ is a k-by-k matrix) . The histogram representation for the entire song is then obtained by summing the instantaneous bigram values over the song duration L:

$$\mathbf{H} = \sum_{t=2}^{L} \mathbf{p}_{t-1}\mathbf{p}_t^T / |\mathbf{p}_{t-1}\mathbf{p}_t^T| \ , \qquad (3)$$

where the start index $t = 2$ is for bigrams (denoting the time index of the first syllable by $t = 1$). The numerator in Eq. (3) is the normalization term for instantaneous bigram values (the sum of the elements in matrix $\mathbf{p}_{t-1}\mathbf{p}_t^T$) .

An example of histograms is shown in Fig. 3. It can be seen qualitatively that FRICOE and PHYLUS are more similar than PHYCOL.

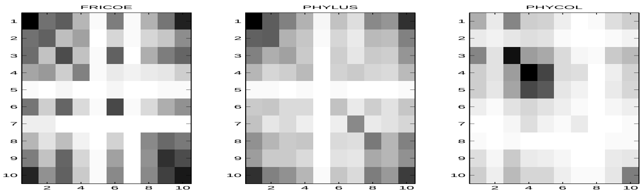

Fig. 3. Syllable pair histograms computed over the entire data set for FRICOE, PHYLUS, and PHYCOL. Ten Gaussians (not the same ones as used in Fig. 2) were used as syllable prototypes. Dark shade of gray represents high value.

2.4. Comparison of histograms

Histograms can be compared quantitatively by computing their mutual correlations. Let $\mathbf{h}$ be a column vector representing a histogram (the column vectors of histogram matrix $\mathbf{H}$ being concatenated). In case of k prototypes and if two consecutive syllables are used as a building block of the histogram, the dimension of $\mathbf{h}$ is k^2. The correlation coefficient c between two histograms $\mathbf{h}_1$ and $\mathbf{h}_2$ is:

$$c(\mathbf{h}_1, \mathbf{h}_2) = \frac{\mathbf{h}_1^T\mathbf{h}_2}{\sqrt{\mathbf{h}_1^T\mathbf{h}_1}\sqrt{\mathbf{h}_2^T\mathbf{h}_2}} \ . \qquad (4)$$

The correlation coefficient is 0 is two histograms are totally disjoint, i.e. there is no common histogram bin where both histograms contain data. The value is closer to 1 the more the two histograms have data in the common histogram bins. Correlations computed between song-wise histograms are shown in Fig. 4.

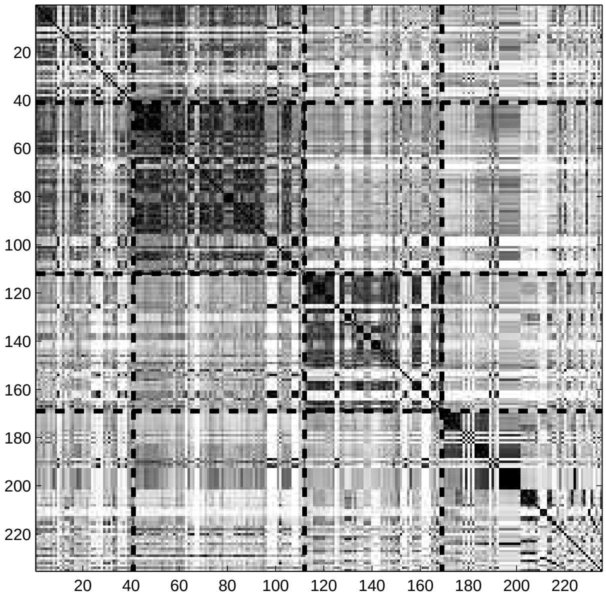

Fig. 4. Correlations between song-wise syllable pair histograms. Dashed lines divide FRICOE, PHYLUS, PHYCOL, and PARMAJ from left to right and top to bottom. Dark shade of gray represents high value.

3. RECOGNITION RESULTS

Songs from 50 bird individuals were used in the recognition experiments. The individuals belonged to four species, c.f. Table 1. There were 257 songs containing altogether 4344 syllables in the data set.

Syllable pair histograms were formed for each song as explained in Sec. 2.3 and the comparison of songs was based on Eq. (4). The "songs" containing only one syllable were removed from the data set and the remaining 235 songs were used in the classification. We used the nearest neighbor classifier in our study. During the classification all songs which belonged to the bird invidual currently being classified were removed from reference songs. The classification was performed for three different histogram representations based on three Gaussian syllable prototype sets containing 10, 30, and 50 Gaussians. The confusion matrices of the classifications are shown in Table 2. The corresponding classification accuracies are 76 %, 79 %, and 80 %. It is interesting that the histogram based on only 10 Gaussians gave comparable results to those of using larger number of Gaussians. Although the optimal number of Gaussians should reflect the variability of the syllables (syllable alphabet), there should be no danger of using even larger number of Gaussians because of the bigram smoothing explained in Sec. 2.3.

Table 2. Confusion matrix for bird species classification. The order of columns and rows from left to right and top to down are FRICOE, PHYLUS, PHYCOL, and PARMAJ. Rows represent the species being recognized and columns represent the target classes. Row-sum of confusion matrix gives the number of bird individuals in each species.

10 Gaussians				30 Gaussians				50 Gaussians			
23	9	4	5	27	3	7	4	26	7	5	3
8	60	2	1	4	65	1	1	1	66	2	2
3	3	45	6	1	4	45	7	1	4	48	4
8	2	5	51	7	2	8	49	5	4	8	49

4. CONCLUSION

In this study we have demonstrated the use of syllable pair histograms as the basis of bird species recognition. Entire songs or parts of them can compared using this representation. The fixed-dimensional representation of variable-length syllable sequences enables also the use of several analysis methods. For example, based on the histogram representations, we can easily cluster the data and find typical songs for different species.

5. REFERENCES

[1] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *J. Acoust. Soc. Am.*, vol. 100, pp. 1209–1219, August 1996.

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

[3] A. Härmä, "Automatic recognition of bird species based on sinusoidal modeling of syllables", *IEEE Int. Conf. Acoust. Speech and Signal Processing (ICASSP'2003)*, Hong Kong, 2003.

[4] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalization", *submitted to this conference*, 2004.

[5] 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," *J. Acoust. Soc. Am.*, vol. 103, pp. 2185–2196, April 1998.

[6] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Trans. Signal Processing*, vol. 45, pp. 2740–2748, November 1997.

[7] L. Rabiner and J. Wilpon, "Considerations in applying clustering techniques to speaker-independent word recognition", *J. Acoust. Soc. Am.*, vol. 66, no. 3, pp. 663-672, 1979.

[8] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43-49, 1978.

[9] P. Somervuo and A. Härmä, "Analyzing bird song syllables on the Self-Organizing Map", *Workshop on Self-Organizing Maps (WSOM'03)*, Kitakyushu, Japan, 2003.