Analyzing bird song syllables on the Self-Organizing Map

Panu Somervuo
Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, Finland
tel. +358 9 451 3284, fax +358 9 451 3277
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
tel. +358 9 451 6027, fax +358 9 460 224
email: aki.harma@hut.fi

Keywords: bird song, dynamic time warping, sequence prototype, sinusoidal modeling, self-organizing map, visualization

***Abstract*—** In this work, we present two methods for analyzing the syllables of the bird song on the Self-Organizing Map (SOM). Dynamic time warping is used for computing the distances between the data sequences. In the first method, the pairwise distances are first computed between the data sequences and each row of the distance matrix is then considered as a feature vector. The conventional SOM with fixed-dimensional model vectors can then be used. The second method is based on online learning of variable-length sequence prototypes. In both cases the SOM is used for constructing a low-dimensional visualization space for the data. We give results analyzing the syllables from five bird species belonging to the Phylloscopus family of passerine birds.

1 Introduction

The work reported in this paper is related to the development of technology for automatic recognition of bird species and even individuals by sounds they produce [4]. 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, 7, 10] or even individual males of a given species [3, 5] using sound has been demonstrated. This article is a follow-up work to [4], which presented promising results in automatic recognition of fourteen Finnish song bird species.

In this work we analyze the data using the Self-Organizing Map [8, 9]. The emphasis is not in recognition, but the organization and visualization of the data. We believe that this will increase our understanding of the data and form the basis for improving the methods also for the recognition purposes.

In earlier work where SOM has been used in this field, only fixed-dimensional feature vectors taken from the spectrograms of the bird songs have been used [6]. In this work we present two additional methods for the construction of the SOM. The first method utilizes the pairwise distances of the syllables computed by dynamic time warping (DTW). Fixed-dimensional feature vectors can then be obtained from the rows of the distance matrix. The second method is based on learning sequence prototypes on the SOM [13]. This forms variable-length sequence prototypes by means of an unsupervised online learning process.

2 Syllables in bird song

Automatic recognition of bird sounds is a typical pattern recognition problem resembling speech and speaker recognition in some sense. However, there are also many important differencies. Sound production mechanism in bird's vocal organ, *syrinx*, is very different from speech production in humans. In addition, we do not have sufficient knowledge of the *language* and structural properties of bird vocalization, which is available for speech recognition. This makes the problem an ideal case for unsupervised learning methods.

Bird song is typically divided into four hierarchical levels: notes, syllables, phrases, and song [2]. In many species there is high individual and regional variability in phrases and song patterns. Syllables can be seen as more elementary building blocks of bird vocalization [1] and may therefore be more suitable for automatic identification of bird species than song patterns. 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.

The basic methodology in [4] was to decompose a bird song recording to a set of brief frequency and amplitude modulated sinusoidal pulses. Each pulse represents one individual isolated syllable and syllables are not overlapping in time or frequency, see Fig. 1. This is a highly simplified model for bird song but it is a reasonable as a baseline feature especially for songbirds as many of their sounds are clearly sinusoidal. The results in recognition were promising and indicate that this is indeed a very useful representation for recognition of species. We may assume that results and their generality could be significantly improved by using a more sophisticated features incorportating information about harmonic components, modulation, and transient sounds. It would be also beneficial to use context information, e.g., differentiate cases where a syllable is a part of a song or it is an individual call or warning sound. However, in this study we use the trajectories of single sinusoids as features and single syllable is the basic unit being analyzed.

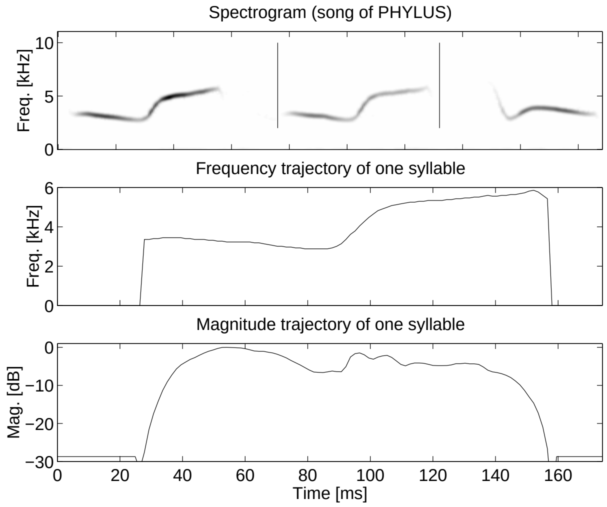


Figure 1: *Frequency trajectory and amplitude trajectory of one syllable of Phylloscopus trochilus after sinusoidal modeling. The spectrogram of the original syllable is shown between two vertical bars on the top image.*

In [4], the comparison of syllables was based on computing the Euclidean distances between feature vector trajectories. It is expected that better results can be obtained by using more appropriate distance measure. In this work we use DTW, which tolerates durational differences between sequences being compared.

3 Methods

3.1 Comparison of two sequences

The comparison of two feature vector sequences can be most naturally done using dynamic time warping (DTW) [11, 12]. It allows durational differences between the sequences. 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. Dynamic programming is utilized for finding the best warping function for each sequence pair. The procedure can be illustrated in the two-dimensional trellis where the elements (feature vectors) of two sequences are located along the two axes of the trellis, see Fig. 2. The warping function is a path from the origo to the point of the trellis which corresponds to the end points of the sequences. The cumulative distance can be divided by the length of the warping path or the sum of the lengths of the sequences.

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, respectively.*

Lat. Abbr.	Latin name	Common name
PHYBOR	Phylloscopus borealis	Arctic Warbler
PHYCOL	Phylloscopus collybita	Comm. Chiffchaff
PHYDES	Phylloscopus trochiloides	Greenish Warbler
PHYLUS	Phylloscopus trochilus	Willow Warbler
PHYSIB	Phylloscopus sibilatrix	Wood Warbler

In the classification task it may be advantageous to restrict the warping path from having too rapid changes. Various slope constraints can be applied to the warping function [11]. Besides improving the recognition performance, they also reduce the amount of computation since the search space for finding the warping is restricted to be inside the area of the slope constraints in the trellis.

3.2 Average of two sequences

The average of two sequences can be computed by sampling the warping path at desired time instants [12, page 159]. This is illustrated in Fig. 2. If we compute the weighted average of the sequences A and B with the weights q and $1-q$, the kth element of the warping path (in Fig. 2) corresponds to the time instance $qt_i + (1-q)t_j$, t_i and t_j being the time instances of elements $\mathbf{a}_i$ and $\mathbf{b}_j$. The averaged feature vector for time instance t_k is $q\mathbf{a}_i + (1-q)\mathbf{b}_j$. In order to get constant time intervals for the elements of the sequence average, we can interpolate the desired time instance between two points in the warping path. The corresponding feature vector is then the result of the interpolation between two weighted feature vector averages at these time instances. Linear interpolation was used in the current work.

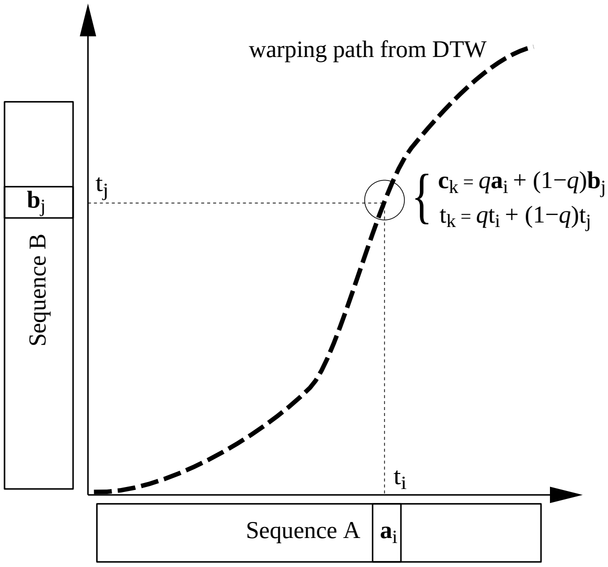


Figure 2: *Computing the weighted average between two variable-length feature vector sequences. See Sec. 3.2 for details.*

3.3 The Self-Organizing Map

The Self-Organizing Map (SOM) [8, 9] is a clustering and visualization tool which enables the organization of the database in an unsupervised manner. The SOM consists of the set of models which are located at the nodes of the low-dimensional regular grid. In case of two-dimensional map it enables easy visualization of the data.

The construction of the SOM is based on competitive learning and the use of neighborhood when adapting the models. In this work we use the SOM both for fixed-dimensional input vectors and also variable-length feature vector sequences.

Like in the traditional vector based SOM, the online training of the learning sequence prototype SOM consists of two steps which are iterated for samples taken from training data:

(1) Find the best-matching unit (BMU) for the current input

(2) Update all models using the neighborhood function of the best-matching unit

For data consisting of feature vector sequences, the distance between the model and data is computed using DTW. Let $X(t)$ denote the input feature sequence at training cycle t, $M_i(t)$ the model of the feature sequence at the SOM node i, and $h(c(t), i)$ the neighborhood function where $c(t)$ denotes the index of the BMU. The updated model of the ith node can then be expressed as:

$$M_i(t+1) = (1 - h(c(t), i))M_i(t) + h(c(t), i)X(t). \quad (1)$$

This formula has exactly the same form as the original SOM algorithm for single feature vectors. The weighted average between the model sequence and the current input feature vector sequence is computed according to Sec. 3.2 the weight q being $h(c(t), i)$.

4 Data analysis on the SOM

A subset of data from [4] was chosen for the current study. This is a set of syllables from five species in the genus of *Phylloscopus*. Most of these birds look very similar and they can hybridize. In fact, vocalization is actually the largest discriminative feature between many of these species. For each of the five species we picked up randomly 50 syllables from four different individuals, i.e., total of 1000 syllables. For each syllable we estimated frequency and amplitude trajectories of a single time-varying sinusoidal component using an algorithm described in [4]. Isolation of syllables from continuous singing is an organic part of the estimation of these features.

In order to select proper distance measure, we performed nearest neighbor classification for the data, see Table 2. For computing the Euclidean distances, the syllables were zero-padded so that the sequence lengths became equal. Before that the syllables were aligned so that the frames corresponding to the maximum value of the amplitude envelope sequence were in the same position of the vector [4].

The results show that dynamic time warping is clearly better than Euclidean distance, and the DTW with slope constraints is better than the basic DTW. The differences between the two columns, Species and Species2 in Table 2, can be partly explained by the small number of individuals per class.

Table 2: *Nearest-neighbor classification of 1000 syllables from 20 individuals and five species. Correct classification per cent. In the first two columns, each test syllable was compared to the remaining 999 syllables. In the column Species2, all reference syllables belonging to the test individual were removed. DTW2 denotes dynamic time warping with slope constraints.*

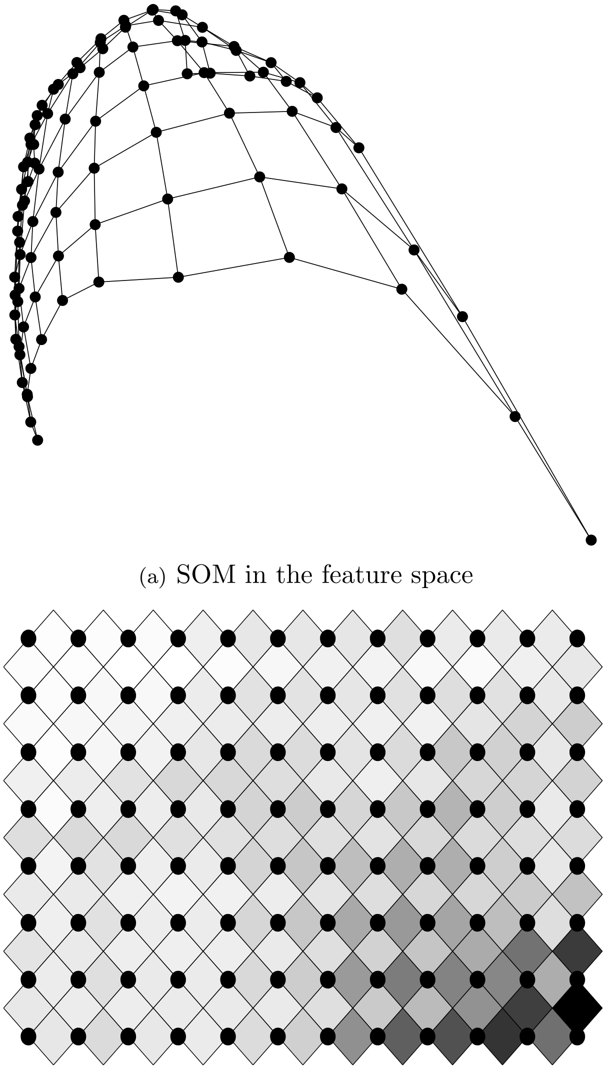
Distance	Target class		
measure	Individual	Species	Species2
Euclidean	74.5	86.2	52.6
DTW	89.1	93.8	56.0
DTW2	90.4	95.9	67.2

4.1 Feature vectors from distance matrix

In order to find a fixed-dimensional feature representation for the data, we computed the distance matrix, i.e., all pairwise distances between the bird syllables, using DTW. The ith row vector of the distance matrix became then the feature vector for the ith data syllable. Since there were 1000 syllables, the dimension of the feature vector was 1000. But eigenvalue decomposition revealed that 99.1 % of the variance could be represented by only seven components. Each 1000-dimensional feature vector was then projected to the seven eigenvectors of the covariance matrix of the 1000-dimensional data. These eigenvectors corresponded to the seven largest eigenvalues.

The SOM was trained using the seven-dimensional feature vectors. Model vectors on the 12-by-8 unit SOM were initialized by random entries and batch-SOM training [9] was applied with Gaussian neighborhood function. Thirty batch-cycles were used. The effective width of the neighborhood function decreased linearly from 10.0 in the beginning of the training to 1.0 in the end of the training. The resulted SOM is illustrated in Fig. 3. Gray-scale image of adjacent model vector distances reveals that there are no clearly separate clusters, the data is like a continuum.

The seven-dimensional data vectors were then projected on the SOM. Fig. 4 shows the histograms of the best-matching units computed for each species and individual separately. Five species divide the SOM display roughly into five areas. There are some individual birds which are located on several parts on the map. From Fig. 3 it can be seen that the map is very smooth. The reason for the syllables of some individuals being located at several areas on the SOM is therefore not explained by the folding of the SOM in the feature space, but because of the true intra-bird variability of the syllable data. We can observe that the scattering of the syllables belonging to PHYDES species is quite large. The scattering may be explained that those birds have larger vocabulary than other species in our data set.



(a) SOM in the feature space

(b) Model vector distances

Figure 3: *Self-Organizing Map for fixed-dimensional feature vectors. Seven-dimensional feature vectors were based on the eigenvalue decomposition of the 1000-by-1000 distance matrix. Figure (a): 8-by-12 unit SOM in the feature space spanned by first two components of the model vector. Figure (b): Distances between adjacent model vectors. Euclidean distances were computed using all seven components. Dark shade of gray represents large distance.*

4.2 Learning sequence prototypes

Using the rows of the distance matrix as a feature vector enables data organization and clustering, but it is not easy to directly interprete the characteristics of the data from the resulting model vectors on the SOM. For this purpose we trained learning sequence prototypes on the SOM. The procedure was explained in Sec. 3.3.

It is possible to initialize the sequence prototypes with random values. However, here the following initialization scheme was used. In the beginning of training each data sequence was represented by the single average feature vector of the entire sequence. The model sequences consisted then of single feature vectors only. After the single-vector model SOM had been trained, we used the full 1000 feature vector sequences instead of their temporal averages. The Gaussian neighborhood function was used in Equation (1);

$$h(c(t), i) = \alpha(t) \exp[0.5 d(c(t), i)^2 / \sigma^2], \qquad (2)$$

where $d(c(t), i)$ is the Euclidean distance between the coordinates of the nodes $c(t)$ and i on the map grid. During 10.000 training cycles $\alpha(t)$ decreased linearly from 0.1 to 0.001 and σ decreased linearly from 10.0 to 1.0.

The learned sequence prototypes are shown in Fig. 5. The characteristics of the data are clearly visible there. Although the model sequences were initialized having

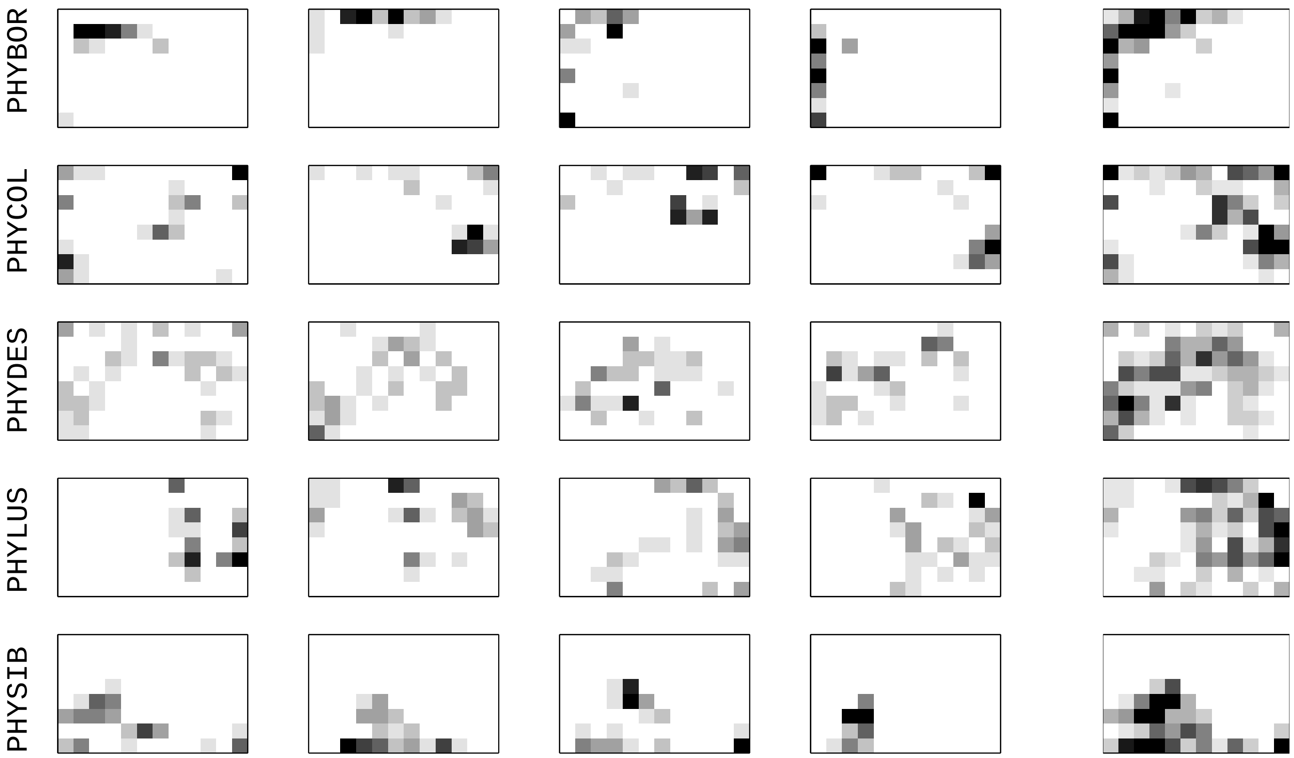


Figure 4: *Distributions of the syllables on the vector-SOM. Fixed-dimensional feature vectors were obtained as explained in Sec. 4.1. Shade of gray represents the number of times the map unit is the best-matching unit for the data, dark shade denotes high value. Each row corresponds to one species. Four columns on the left represent four different individuals and the column on the right is the bmu-histogram of all individuals from one species.*

only one feature vector as their element, the durations of the trained prototype sequences now vary and correspond to the durations found in the data set.

The organization of the syllables can be roughly described so that the falling-tone syllables are represented on the left hand side of the SOM display and the rising-tone syllables on the right hand side. Longer syllables are located at the bottom of the map.

5 Conclusions

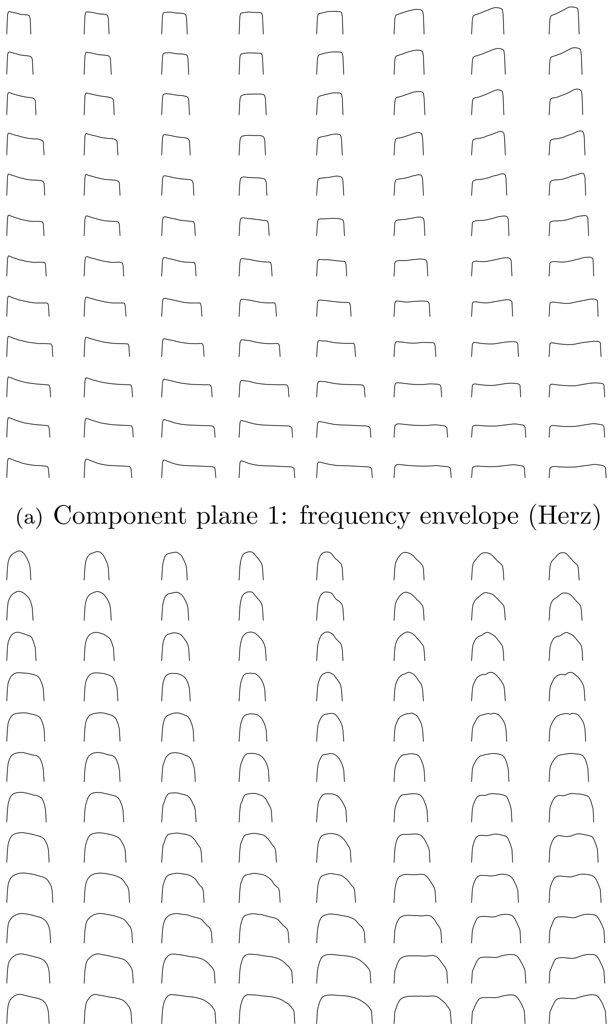
In this work, we presented two methods for analyzing the syllables of the bird song on the SOM. The data samples were from twenty birds belonging to the five species of the Phylloscopus family. Sinusoidal modeling was used in the feature extraction. Each syllable was represented as a sequence of two-dimensional feature vectors, one component representing the instantaneous frequency and the other component representing the amplitude.

The first analysis method was based on pairwise DTW-based distances between data sequences. Rows of the distance matrix were then considered as feature vectors. Eigenvector decomposition was used for projecting these high-dimensional feature vectors into lower-dimensional space before training the conventional SOM with fixed-dimensional model vectors.

Another approach was to train learning sequence prototypes on the SOM. The model associated with each map node was then the sequence of feature vectors and its length was not fixed during the training. This is relatively new method, and has been demonstrated earlier using only speech data [13]. The current work can be considered as a feasiblity study of that method. The method resulted in the smooth prototype sequence display.

We also made a small comparison between sequence distance methods for the recognition purpose. Dynamic time warping outperformed clearly the computationally simpler method where sequences were first aligned according to their amplitude envelopes and then compared with Euclidean distance. This gives also justification for the use of learning sequence prototypes instead of fixed-dimensional vector representations of the sequences.

In our study the SOM was used for constructing visualization display for the data. Scattering of the projections of data syllables on the SOM revealed the variability of the data. For some bird individuals the projections were concentrated to small, compact regions, whereas for some birds almost the entire SOM display was occupied. Already this information gives us insight to the characteristics of the data and the



(a) Component plane 1: frequency envelope (Herz)

(b) Component plane 2: amplitude envelope (desibels)

Figure 5: *Syllable prototypes on the 12-by-8-unit SOM. Variable-length sequence prototypes were trained using the method explained in Sec. 3.2. Since the bird syllables were modeled by single sinusoids with time-varying amplitudes and frequencies, each element of the sequence prototype is a two-dimensional vector representing instantaneous frequency and amplitude. Horizontal axis in each prototype sequence plot represents time. The frequency envelope component of the sequence is shown in the upper plot and the amplitude envelope in the lower plot. The range of the vertical axis for each model sequence is* $0\ldots 8.6kHz$ *in the upper plot and* $-50\ldots 0dB$ *in the lower plot.*

variations between different species and individuals.

This work served as a feasiblity study of the described methods. In the future we plan to continue our studies with larger database and larger number of species.

Acknowledgements

Panu Somervuo would like to thank Academy of Finland for financial support, project no. 44886 "New information processing principles" (Finnish Centre of Excellence Programme 2000-2005). Aki Härmä's work was supported by the Academy of Finland and the graduate school GETA.

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] P. Galeotti and G. Pavan, "Individual recognition of male Tawny owls (Strix aluco) using spectrograms of their territorial calls," *Ethology, Ecology & Evolution*, vol. 3, no. 2, pp. 113–126, 1991.

[4] 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.

[5] K. Ito, K. Mori, and S. Iwasaki, "Application of dynamic programming matching to classification of budgerigar contact calls," *J. Acoust. Soc. Am.*, vol. 100, pp. 3947–3956, December 1996.

[6] P. Janata, "Quantitative assessment of vocal development in the zebra finch using self-organizing neural networks", *J. Acoust. Soc. Am*, vol. 110, no. 5, pp. 2593–2603, November 2001.

[7] 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.

[8] T. Kohonen,"Self-Organized formation of topologically correct feature maps", *Biological Cybernetics*, vol. 43, pp. 59-69, 1982.

[9] T. Kohonen, *Self-Organizing Maps*, Springer, 1995.

[10] 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.

[11] 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.

[12] D. Sankoff and J. Kruskal, *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, Addison-Wesley, 1983.

[13] P. Somervuo and T. Kohonen,"Self-organizing maps and learning vector quantization for feature sequences ", *Neural Processing Letters*, vol. 10, no. 2, pp. 151-159, 1999.