Automatic Birdsong Recognition Based on Autoregressive Time-Delay Neural Networks

S-.A. Selouani
Université de Moncton,
Shippagan, Canada
selouani@umcs.ca

M. Kardouchi
Université de Moncton,
Moncton, Canada
kardoum@umoncton.ca

E. Hervet
Université de Moncton,
Moncton, Canada
hervete@umoncton.ca

D. Roy
Université de Moncton,
Moncton, Canada
royd@umoncton.ca

***Abstract*-A template-based technique for automatic recognition of birdsong syllables is presented. This technique combines time delay neural networks (TDNNs) with an autoregressive (AR) version of the backpropagation algorithm in order to improve the accuracy of bird species identification. The proposed neural network structure (AR-TDNN) has the advantage of dealing with a pattern classification of syllable alphabet and also of capturing the temporal structure of birdsong. We choose to carry out trials on song patterns obtained from sixteen species living in New Brunswick province of Canada. The results show that the proposed AR-TDNN system achieves a highly recognition rate compared to the baseline backpropagation-based system.**

I. Introduction

The analysis of bioacoustic signals that aims to recognize bird species has been used for many years in order to identify and locate migrating and resident individuals. However, carrying out manual surveys requires expertise that could be time consuming and most of the time needs a long and continuous monitoring in inaccessible regions. It is now possible to develop a technology for automatic identification of birdsongs to achieve the same task with numerous advantages including continuous long term unattended operation. It should be noted that in addition to the added value provided to the research in ornithology, and biology in general, there is a significant commercial potential for such systems because bird watching becomes popular in many countries. For instance, hand-held used systems can be considered in order to perform a rapid biodiversity assessment especially in acoustically rich regions.

The production of birdsong involves the syrinx, which is a unique organ involved in birdsong. The large diversity of the syrinx structures and functions observed within the bird species leads to a wide range of birdsong spectrum, which brings challenges to the design of automatic birdsong classification systems. A popular terminology for vocalizations divides birdsongs into hierarchical levels composed of notes, syllables, phrases and calls [1], as illustrated in the example given in Figure 1. Syllables are made of one or more notes, and are considered as suitable units for recognition of bird species because they can be more accurately detected from continuous recordings than notes. Phrases and calls contain more regional and individual variability than syllables.

Automatic recognition of bird species (ARBS) from their sounds is a typical pattern recognition problem. The patterns are represented with a few sound acoustical features. Recognition is done by trying to match features issued from the parametric representation with the models of sounds produced by species. Features should be selected in such a way that they maximize the ability to distinguish sounds that are produced by different individuals.

The main principle of ARBS consists of building sound models based on large birdsong corpora which attempt to include in their construction common sources of variability that may occur in practice. Nevertheless, not all variabilities can reasonably be covered. In order to achieve a better accuracy, different approaches have been studied. The state–of–the–art methods can be summarized in two major approaches. The first approach consists of improving the front-end techniques prior to the pattern matching in an attempt to maximize the classification ability. For this purpose, spectrograms are considered as the most efficient parameterization of bird sounds. However, comparison of spectrograms is computationally demanding and often includes environmental information that is not relevant to recognition of bird species [2]. Nelson in [3] uses canonical discriminant analysis to determine and select features that maximize the recognition rate of Field Sparrow (*Spizella pusilla*) and Chipping Sparrow (*Spizella passerina*) among 11 other bird species. In [4] syllables are parameterized with one simple sinusoidal model and by time-varying frequency and amplitude trajectories. This front-end was found adapted to species that produce regularly tonal and harmonic sounds. The second approach attempts to establish an accurate pattern matching that takes into account the variability that may occur. Methods in this approach include dynamic time warping, hidden Markov models [5], and neural networks [6]. Training a neural network is computationally demanding, but classification with the network is relatively fast, what could facilitate a commercial implementation.

1-4244-0020-1/05/$20.00 â2005 IEEE

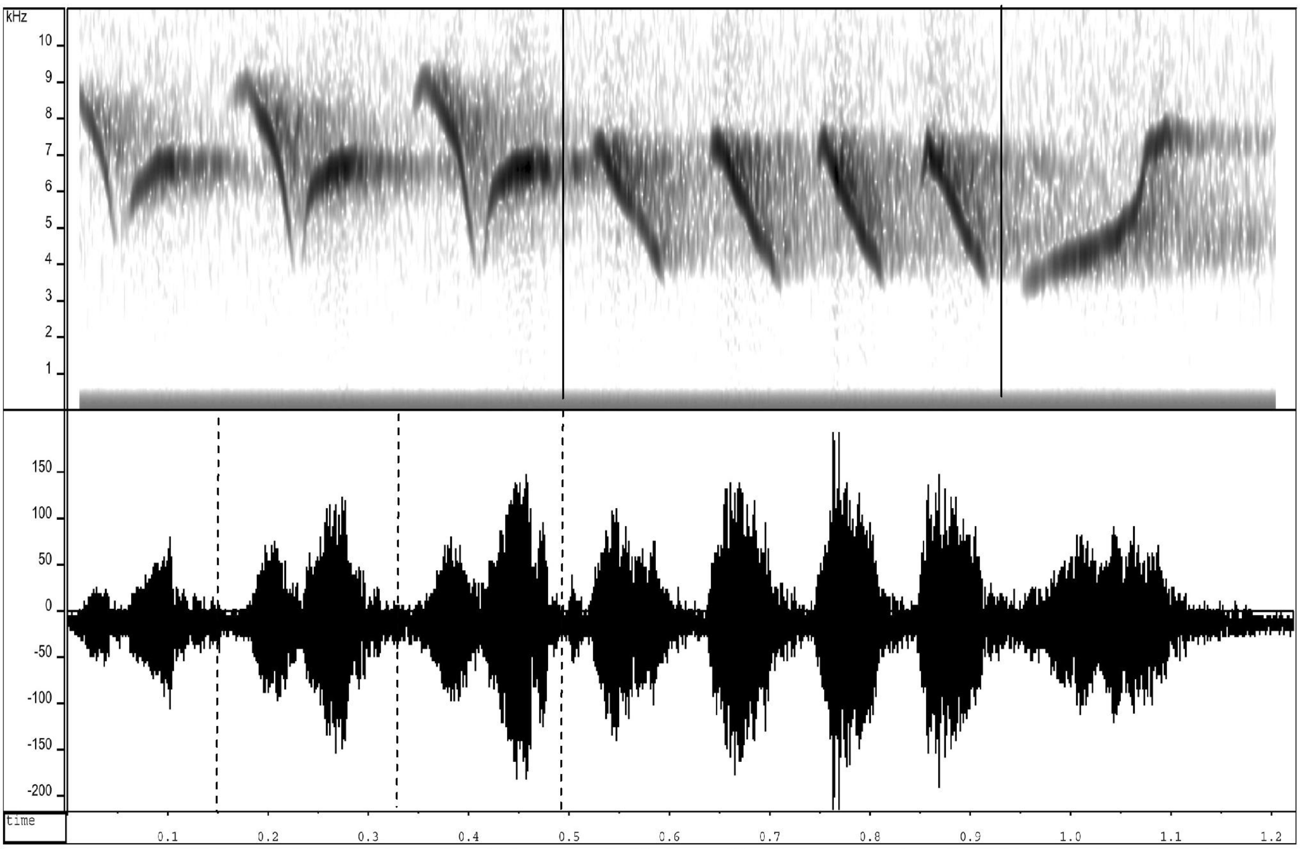

Fig.1. a song of Yellow Warbler (*Dendroica Petechia*) composed of 3 phrases. The first phrase contains 3 syllables separated by dotted lines.

Nevertheless, despite the efforts to address various aspects of accurate birdsong identification, adaptation to the time variability of bird sound events remains one of the most challenging problems. It often requires some innovative computation techniques that can cope with pattern recognition and memory, both on a short time scale as well as on longer time scales. Such a problem occurs especially in auditory tasks such as bird song and speech comprehension.

This paper presents a new bioacoustic signal recognition based on a neural networks configuration which introduces an autoregressive version of backpropagation algorithm and simultaneously incorporates a time delay component in the input layer. This combination gives the neural network context memorization ability and increases its capacity to discern the ‘alphabet’ patterns even in a high time varying context.

This paper is organized as follows. Section II gives a background of birdsong classification using feedforward neural networks. Section III describes the Autoregressive Time-Delay Networks (AR-TDNN) architecture. Section IV reports trials using AR-TDNN-based and standard systems for the classification of sixteen Canadian bird species. Finally, we summarize our major findings and discuss perspectives of this work in section V.

II. Birdsong Identification Using Neural Networks

An Artificial Neural Network (ANN) classifier provides a powerful technique for bioacoustic signal analysis and recognition. The ANN standard is the multilayer perceptron (MLP) which uses a backpropagation algorithm for its training. The MLP takes a set of selected features as input and has a different output for each species or individual to be recognized. Identification operates in two steps, training step and test step. In the training step, utterances of sounds to be identified are used to train the MLP so that the correct MLP output is activated. Training occurs by repeatedly presenting known sounds to the network and iteratively modifying the weights within the network in such a way as to reduce the overall error between given and desired outputs. Training continues until the overall error is below a given threshold. Once trained, the system is ready to be used with unknown sounds to be recognized. Each of the outputs will give a value between 0.0 (zero match) and 1.0 (perfect match); the unknown sound being classified as the output with the highest value.

In both human speech and bird vocalizations, the communication is encoded into patterns that must be recognized both for their short timescale features as well as for longer ones. Unfortunately, solving multiple time scale problems has generally proven to be very difficult for an ANN. This is mainly due to a potential lack of sufficient neural network architectures as well as to a weakness in the basic backpropagation algorithm. In order to take advantage of the well-proven discrimination ability of ANN, one strategy consists of incorporating a time delayed component, either in the sense of a time delayed edge, which forces an afferent unit to use older activations of its efferent units, or in the sense of a decay factor on an unit activation function, which allows to hold information for a longer period of time. The delays allow information separated by relative amounts of time to be combined for simultaneous processing. Thus a network using time delays and recurrence for solving an identification task of bird species through their syllables would be expected to have small time delays closer to the inputs to deal with quick information changes, and larger time scales in the network to deal with slower changes in the input pattern. These aspects must be integrated by both the network structure and the learning/identification algorithm as it is the case in the proposed AR-TDNNs.

III. AUTOREGRESSIVE TIME-DELAY NEURAL NETWORKS

Because birdsong is a temporarily unstable phenomenon, we consider Recurrent Networks (RNs) to be more adequate than feedforward networks in the case of any classification task dealing with bird vocalizations. RNs are generally trickier to work with, but they are theoretically more powerful, having the ability to represent temporal sequences of unbounded length. Another consideration related to the syllable context leads us to use a particular RN: the one proposed by Russel and Bartley [7], which uses an autoregressive version of the backpropagation algorithm. In speech processing, however, even if RNs using AR perform very well in the context dependent labeling, this power turns out to be source of disappointment in the case of event's time shifting. The approach we are investigating proposes to integrate in addition to the AR component, a delay component similar to the one used by Waibel's time delay neural networks (TDNN) [8]. Through this combination, we expect that the ability of the system to discern patterns even in a strong time misalignment will be increased.

A. AR-TDNN Unit

The model described by Russel and Bartley includes an autoregressive memory which constitutes a form of self-feedback where the output depends on the current output plus a weighted sum of previous outputs. Then, the AR node equation is:

$$y_i(t) = f\left(bias_i + \sum_{j=1}^{P} w_{i,j} x_j(t)\right) + \sum_{n=1}^{M} a_{i,n} y_i(t-n), \qquad (1)$$

where $y_i(t)$ is the output of node i at time t, $f(x)$ is the $tanh(x)$ bipolar activation function, P is the number of input units. M is the order of autoregressive prediction. Weights $w_{i,j}$, biases and coefficients $a_{i,n}$ are adaptive and optimized in order to minimize the output error. Our proposition consists in incorporating a time delay component on the input nodes of each layer, then equation 1 becomes:

$$y_i(t) = f\left(bias_i + \sum_{l=0}^{L} \sum_{j=1}^{P} w_{i,j,l} x_j(t-l)\right) + \sum_{n=1}^{M} a_{i,n} y_i(t-n), \qquad (2)$$

where L is the delay order of the input. Feedforward and feedback weights are initialized from a uniform distribution in the range [-0.8, 0.8]. A neuron of the AR-TDNN configuration is shown in Figure 2. AR backpropagation learning algorithm performs the optimization of feedback coefficients in order to minimize the mean squared error noted $E(t)$ defined as:

$$E(t) = \frac{1}{2}\sum_{i}\left(d_i(t) - y_i(t)\right)^2, \qquad (3)$$

where d_i is the desired value of the i^{th} output node. The weight and feedback coefficient changes, noted respectively $\Delta w_{j,i,l}$ and $\Delta a_{i,n}$ are accumulated within an update interval $[T_0, T_1]$. In the proposed AR-TDNN version, the update interval $[T_0, T_1]$ is fixed such as it corresponds to the time delay of the inputs. The updated feedback coefficients are written as follows:

$$a_{i,n}^{new} = a_{i,n}^{old} + \frac{1}{T_1 - T_0}\sum_{t=T_0}^{T_1} \Delta a_{i,n}(t), \qquad (4)$$

and if T is the frame duration, the weights are computed as follows:

$$w_{i,j,l}^{new} = w_{i,j,l}^{old} + \frac{1}{L.T}\sum_{t=T_0}^{T_1} \Delta w_{i,j,l}(t). \qquad (5)$$

The $\Delta w_{j,i,l}$ variations are accumulated during the update interval after accumulating Time-Delay frames at the input. This type of networks which combines both input delays and feedback of outputs can perform context sensitive decisions. An AR-TDNN system is trained by using the Nguyen-Widrow initialization conditions [9]. The TDNN part of the system consists of three layers. Each neuron in the first hidden layer receives input from the coefficients in the three-frame window of the input layer. The input is centered on the hand-labeled syllables.

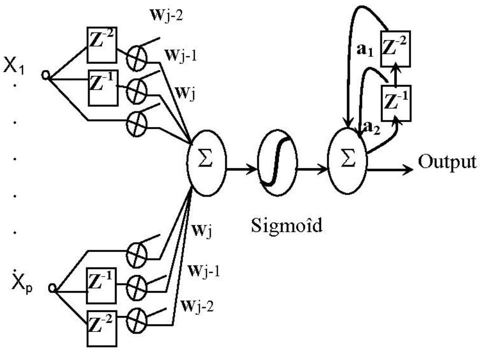

Fig. 2. A neuron from a hidden layer of an autoregressive time delay neural network

B. Principle of Learning and Test by AR-TDNNs

The learning phase is performed as follows: if a sequence of the targeted syllable appears in the vocalization continuum, the network activation arises gradually in the output corresponding to the targeted species. For instance, in the case of American Kestrel (*Falco Sparverius*) detection/classification, as illustrated in Figure 3, the task is to learn to recognize this sequence: LCA-AK-RCA: LCA is the left acoustic context of the syllable (noted AKS) and RCA is its right acoustic context. The AR-TDNN receives three input tokens at a time *t* and it must detect a syllable of a targeted specie sequence from any other sequence combination. Ideally, the learning consists of setting the desired output at the high level (+1) when the sequence LCA-AKS-RCA is encountered, otherwise the low level (-1) is set. However, due to time variability, each state of the sequence can be represented by a various number of frames. The length of a frame and the delay order of AR-TDNN play an important role in the determination of the possible sequence configurations. Table I gives desired values of the output corresponding to the American Krestel (AK) identification, depending on the sequence configuration that may occur in our application. The other outputs are set at -1.0.

IV. Experiments and Results

A. Data Collection and Preprocessing

We chose a set of templates that include many examples of syllables manually extracted from recordings of 16 bird species selected among the 395 resident species of New Brunswick. Vocalizations were sampled at 22 kHz with 16-bit resolution. Each resulting waveform resides in a file that represents several seconds of singing. These files may include numerous vocalizations, as well as noises caused by movement, beak wiping, etc., separated by periods of silence.

TABLE I
Possible Sequence Configurations and their Corresponding Outputs in the Case of American Krestel Identification.

Sequence	Value at AK Output
#-#-RCA	0.1
#-RCA-RCA	0.1
RCA-RCA-RCA	0.1
AK-RCA-RCA	0.5
AK-AK-RCA	1.0
AK-AK-AK	1.0
LCA-AK-AK	1.0
LCA-LCA-AK	0.5
LCA-LCA-LCA	0.1
#-LCA-LCA	-1.0

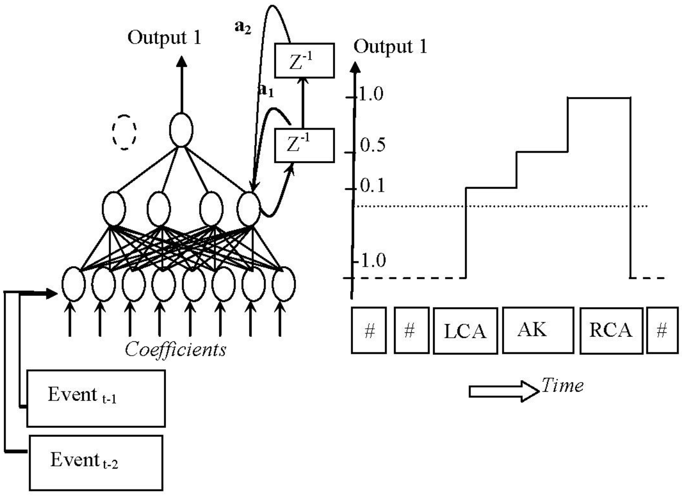

Fig. 3. An AR-TDNN performing an identification of an American Kestrel syllable. The activation of the corresponding output is presented.

The most significant preprocessing step is the creation of templates for the segments to be identified. Selection of templates depends on the sound environment in which recognition will take place. In our experiments the sound environment included song syllables, calls, noises, and periods of silence. Template waveforms of syllables were manually labeled and extracted from several songs using *Wavesurfer* software [10]. In our experiments, we choose to extract the front-end features using a Linear Predictive Coding (LPC) analysis [11]. LPC analysis has been among the most popular methods for extracting spectral information from biological signals. This analysis is efficient for representing the spectral envelope with no harmonic effects. The frame duration is fixed in such a way that it corresponds to the average of expected vocalization lengths of the studied species. There were 80 songs containing altogether 582 syllables in the training data set. A total of 290 syllables, different from those used in the training phase and covering all of the 16 species, were used to evaluate the AR-TDNNs.

It is important that the optimal number of utterances should reflect the variability of the syllables of the studied species (syllable alphabet). The scoring procedure does not consider silence and noise labels. Table II gives the names of the studied species and their corresponding output number in the networks.

B. Recognition Platform

In order to evaluate our proposed approach for ARBS, the *AM7* neural network simulation environment, available freely from the MITRE Corporation, has been used throughout all experiments [12]. The software contains a neural network simulation code generator which generates ANSI C code implementations for various networks structures and learning algorithms.

C. Tests and Results

The performance of the AR-TDNN-based recognizer was compared against a baseline feedforward neural network (ANN) using manually labeled vocalizations. In all our experiments, input features are composed of 20 LPC coefficients. These coefficients were calculated on a 50 msec Hamming window advanced by 20 msec at each frame. The architecture of both AR-TDNN and ANN consists of three layers. The input layer consists of 20 neurons, while the hidden layer and the output layer are composed of 26 and 16 neurons respectively. The values presented as inputs of standard ANN are calculated by averaging the LPC values of all the syllables frames. In the AR-TDNN, the LPC parameters calculated at each frame are sequentially presented as inputs, as described in section III-B. The weights of the networks are calculated during a training phase with a backpropagation algorithm with a learning rate equal to 0.3 and a momentum coefficient equal to 0.07. A second-order autoregressive backpropagation algorithm has been used to train the AR-TDNN. Therefore, the two autoregressive coefficients (cf. section III-A) have been fixed at 0.12 and 0.18 respectively.

The analysis of the results revealed that the AR-TDNN configuration is more accurate in all cases. Globally, the AR-TDNN system performs with 83% correct rate which represents 16% fewer errors than standard ANN. The confusion matrices of the classifications are given in Table III. We noticed that standard ANN failed dramatically in the classification of long syllables as it is found in the Great Horned Owl and Mourning Dove identifications. The same trend is observed when there are noticeable acoustical differences in syllables. It seems that standard ANNs are accurate when syllables are highly stereotyped, short and contain prominent onset and offset phases. By contrast, AR-TDNNs are more accurate when the events are quietly long and temporally unstable, as noticed in the case of Gray Catbird classification. This confirms the suitability of such networks to capture the time variability information.

TABLE II.
COMMON AND LATIN NAMES OF BIRD SPECIES USED IN THE EXPERIMENTS AND THEIR CORRESPONDING UNIT NUMBER IN THE OUTPUT LAYER OF BOTH AR-TDNN AND STANDARD ANN.

Common name	Latin name	Output Number
Red-winged blackbird	*Agelaius phoeniceus*	1
Great horned owl	*Bubo virgianus*	2
Northern cardinal	*Cardinalis cardinalis*	3
American goldfinch	*Carduelis tristis*	4
Northern flicker	*Colaptes auratus*	5
Blue jay	*Cyanocitta cristata*	6
Yellow warbler	*Dendroica petechia*	7
Gray catbird	*Dumetella carolinensis*	8
American kestrel	*Falco sparverius*	9
Song sparrow	*Melospiza melodia*	10
House sparrow	*Passer domesticus*	11
Black-capped chickadee	*Poecile atricapillus*	12
Tree swallow	*Tachycineta bicolour*	13
American robin	*Turdus migratorius*	14
Mourning dove	*Zenaida macroura*	15
White-throated sparrow	*Zonotrichia albicollis*	16

V. CONCLUSION

We have presented a system which achieves automated discrimination between different bird species by using a new connectionist approach based on Autoregressive Time-Delay Neural Networks. The system has been shown to recognize 16 species of New Brunswick province of Canada with 83% accuracy for good quality sounds. We found through the experiments that using AR-TDNNs leads to an increase in the accuracy of the species identification rate up to 16% comparatively to an ANN-based baseline system. However, since ARBS is an extremely complex task and require large training sets as well as complex networks, only smaller representative problems were used in this study to determine what sort of neural network would be suitable at solving this task. In order to gain more insight into the important issue concerning the capture of the temporal component, we will investigate in a future work the contribution of dynamic parameters and a hybrid approach which combines AR-TDNN and an effective technique of time alignment such as the one based on Hidden Markov Models.

REFERENCES

[1] C. K. Catchpole, and P. J. B. Slater, *"Bird song: biological themes and variations"*, Cambridge, UK: Cambridge University Press, 1995.

[2] A. L. McIlraith and H. C. Card, "Birdsong recognition using back-propagation and multivariate statistics", IEEE Trans. Signal Processing, vol. 45, pp. 2740-2748, 1997.

[3] D. A. Nelson, 'The importance of invariant and distinctive features in species recognition of birdsong', *Condor* 91, 120–130, 1989.

TABLE III.

CONFUSION MATRIX FOR BIRD SPECIES CLASSIFICATION. THE ORDER OF COLUMNS AND ROWS CORRESPONDS TO THE SPECIES GIVEN IN TABLE III. ROWS REPRESENT THE SPECIES BEING RECOGNIZED AND COLUMNS REPRESENT THE TARGET CLASSES. COMPONENTS (I, J) OF THE MATRIX ELEMENTS CORRESPOND TO THE RESULTS OBTAINED BY ANN (I) AND AR-TDNN (J) RESPECTIVELY.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	(14,16)	(0,0)	(0,0)	(0,0)	(0,0)	(1,0)	(0,0)	(0,0)	(1,0)	(1,1)	(0,0)	(0,0)	(1,2)	(1,0)	(0,0)	(0,0)
2	(0,0)	(9,15)	(0,0)	(1,0)	(2,1)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(1,0)	(0,0)	(1,1)	(0,0)	(3,0)	(1,1)
3	(1,1)	(0,0)	(13,14)	(0,0)	(1,1)	(0,0)	(0,0)	(0,0)	(0,0)	(0,1)	(1,1)	(1,0)	(0,0)	(0,0)	(0,0)	(1,0)
4	(0,0)	(0,0)	(1,0)	(13,15)	(0,0)	(0,0)	(1,0)	(0,0)	(1,0)	(0,1)	(0,0)	(0,0)	(2,2)	(0,0)	(0,0)	(0,0)
5	(0,0)	(1,0)	(1,1)	(0,0)	(13,16)	(1,0)	(0,0)	(0,0)	(1,0)	(1,0)	(0,0)	(0,0)	(0,0)	(0,1)	(0,0)	(0,0)
6	(0,0)	(0,0)	(0,0)	(1,1)	(0,0)	(13,16)	(0,0)	(0,0)	(0,0)	(1,1)	(1,0)	(0,0)	(2,0)	(0,0)	(0,0)	(0,0)
7	(0,0)	(0,0)	(1,1)	(2,2)	(0,0)	(0,0)	(12,14)	(0,0)	(0,0)	(1,0)	(0,0)	(0,0)	(0,0)	(0,1)	(0,0)	(2,0)
8	(1,0)	(0,1)	(0,0)	(1,0)	(1,0)	(0,0)	(0,0)	(10,17)	(1,0)	(0,0)	(1,0)	(0,0)	(0,0)	(0,0)	(0,0)	(3,0)
9	(0,0)	(0,0)	(1,0)	(0,0)	(0,1)	(0,0)	(1,1)	(0,0)	(12,14)	(1,1)	(1,1)	(0,0)	(0,0)	(1,0)	(0,0)	(1,0)
10	(0,0)	(0,0)	(0,1)	(0,1)	(0,0)	(0,0)	(1,0)	(0,0)	(1,1)	(15,16)	(0,0)	(0,0)	(0,0)	(1,0)	(0,0)	(1,0)
11	(0,1)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(1,1)	(0,0)	(13,15)	(2,0)	(1,0)	(0,0)	(0,0)	(0,0)
12	(1,0)	(1,1)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,1)	(1,1)	(1,1)	(11,14)	(1,0)	(1,0)	(0,0)	(1,0)
13	(2,1)	(1,1)	(0,1)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(1,0)	(0,0)	(0,0)	(14,15)	(1,1)	(0,0)	(0,0)
14	(0,0)	(0,0)	(0,0)	(2,1)	(0,0)	(0,0)	(1,1)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(1,0)	(14,16)	(0,0)	(0,0)
15	(0,0)	(5,2)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(1,0)	(1,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(9,15)	(1,0)
16	(2,1)	(0,0)	(1,1)	(0,0)	(0,0)	(1,0)	(0,0)	(3,2)	(0,0)	(1,1)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(11,14)

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

[5] J.A. Kogan and D. Margoliash, "Automated recognition of birdsong elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study", J. Journal Acoust. Soc. Amer. Vol. 103, pp. 2185-2196, 1998.

[6] F. Schwenker, C. Dietrich, H. A. Kestler, K. Riede, and G. Palm, "Radial basis function neural networks and temporal fusion for the classification of bioacoustic time series", Neurocomputing vol. 51, pp.265-275, 2003

[7] R. L. Russel, and C. Bartley, "The Autoregressive Backpropagation Algorithm", Proc. of IJCNN, vol. II, pp.369-377, 1991.

[8] A. Waibel, T. Hanazawa, G. Hinton, and K. Shikano, "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. on Audio Speech and Signal Processing, No 37, pp 328-339, 1989.

[9] D. Nguyen, and B. Widrow, "Improving the Learning Speed of Two Layer Neural Networks by Choosing Initial Values of the Adaptive Weights", Proc. of IJCNN, vol. III, pp. 21-26, 1990.

[10] Wavesurfer tool for sound visualisation and manipulation, http://www. speech.kth.se/wavesurfer/

[11] J. Makhoul, "Linear prediction: A tutorial review", Proc. IEEE Vol. 63, pp. 561-580, 1975.

[12] AM version 7, "Neural Networks toolkit", http://www.elegant-software.com.