RAZPRAVE IV. RAZREDA SAZU | XLVII-3 | 185–204 | LJUBLJANA 2006

BIRD SOUND CLASSIFICATION AND RECOGNITION USING WAVELETS

KLASIFIKACIJA IN PREPOZNAVANJE PTIČJIH GLASOV S POMOČJO VALČKOV

ARJA SELIN, JARI TURUNEN & JUHA T. TANTTU

ABSTRACT

Bird sound classification and recognition using wavelets

In this paper a new method for automatic classification and recognition of bird sounds is presented. Our main idea is to study, how inharmonic and transient bird sounds can be recognized efficiently. The data consisted of sounds of eight bird species. Five species, the Mallard (*Anas platyrhynchos*), the Graylag Goose (*Anser anser*), the Corncrake (*Crex crex*), the River Warbler (*Locustella fluviatilis*), and the Magpie (*Pica pica*) have inharmonic sounds, whereas the remaining three reference species, the Quail (*Coturnix coturnix*), the Spotted Crake (*Porzana porzana*), and the Pygmy Owl (*Glaucidium passerinum*) have harmonic sounds. The wavelet analysis was selected due to its ability to preserve both frequency and temporal information, and its ability to analyse signals which contain discontinuities and sharp spikes. The feature vectors calculated with the proposed algorithm from the wavelet coefficients were used as the inputs of two neural networks, the self-organizing map (SOM) and the multilayer perceptron (MLP). The results were encouraging, for the unsupervised SOM network recognized 78% and the supervised MLP network 96% of the test sounds correctly.

Keywords: Transient sounds, wavelet transform, automated categorization, neural networks.

Klasifikacija in prepoznavanje ptičjih glasov s pomočjo valčkov

Predstavljamo novo metodo samodejne klasifikacije in prepoznavanja ptičjih zvokov. Zastavili smo si problem, kako prepoznati neharmonične in neponavljajoče se vzorce ptičjih zvokov. Uporabili smo podatke osmih ptičjih vrst. Pri petih vrstah, mlakarica (*Anas platyrhynchos*), siva gos (*Anser anser*), kosec (*Crex crex*), rečni cvrčalec (*Locustella fluviatilis*) in sraka (*Pica pica*), je zvok neharmoničen, medtem ko je pri drugih treh, prepelica (*Coturnix coturnix*), grahasta tukalica (*Porzana porzana*) in mali skovik (*Glaucidium passerinum*), zvok harmoničen. Analiza valčkov je bila izbrana zaradi njene značilnosti, da ohrani frekvenčne in časovne informacije in zaradi sposobnosti analizirati signale, ki vsebujejo prekinitve in ostre vrhove. Vektorje zunanje oblike, ki smo jih izračunali na osnovi predlaganega algoritma iz koeficientov valčkov, smo uporabili kot vhodne podatke v dve nevronski mreži, samoorganizirano mrežo (SOM) in mnogoslojni perceptron (MPL). Rezultati so bili zelo vzpodbudni, saj smo z nenadzorovano SOM nevronsko mrežo pravilno prepoznali 78 % in z nadzorovano MLP nevronsko mrežo 96 % testiranih vzorcev.

Ključne besede: neponavljajoči se vzorci zvokov, transforamcija valčkov, samodejna klasifikacija, nevronske mreže.

Addresses – Naslovi

Arja SELIN
Tampere University of Technology
Pori, Pohjoisranta 11
P.O.Box 300
FIN-28101 Pori
Finland
E-mail: arja.selin@tut.fi

Jari TURUNEN
Tampere University of Technology
Pori, Pohjoisranta 11
P.O.Box 300
FIN-28101 Pori
Finland
E-mail: jari.j.turunen@tut.fi

Juha T. TANTTU
Tampere University of Technology
Pori, Pohjoisranta 11
P.O.Box 300
FIN-28101 Pori
Finland
E-mail: juha.tanttu@tut.fi

INTRODUCTION

Birds have many different types of sounds varying from short and simple call notes to long and complex songs (CATCHPOLE & SLATER 1995). Human ear and brain constitute an effective voice recognition system; therefore for the human ear it is relatively easy to notice even subtle differences in bioacoustic sounds. In bird sound research, the typical methods for classification was listening and visual assessment of spectrograms. However, the human decision is always subjective and the automatization of this classification process would be an important new tool for bioacoustic research (DEECKE & JANIK 2006). During past two decades several methods of general signal processing and speech recognition have been applied to animal vocalizations.

However, only a few studies deal with the automatic identification of bird species using sounds. For example HÄRMÄ (2003) has studied automatic identification of bird species using sinusoidal modelling of syllables. In their further study (HÄRMÄ & SOMERVUO 2004) the syllables were modelled using a parametric line spectrum estimation method, the sounds being divided into four classes according to their harmonic structure. MESGARANI and SHAMMA (2003) have shown that with a multiresolution spectrotemporal auditory model it is possible to classify birdcalls. TANTTU et al. (2003, 2006) have proposed a method for automatic classification of flight calls of Crossbill species (*Loxia* spp.) based on the tracking of the first harmonic components of the spectrogram. SOMERVUO and HÄRMÄ (2004) have shown the possibility of bird species recognition based on the syllable pair histogram of the song, where the nearest neighbour classifier was used. BAKER and LOGUE (2003) have used three bioacoustical analysis methods for comparing the calls of the Black-capped Chickadee (*Poecile atricapillus*) among different populations. FAGERLUND (2004) has studied how inharmonic bird sounds can be classified using 19 low level parameters of syllables and k-Nearest-Neighbour nonlinear classifier.

Most of the above-mentioned studies of bird sound recognition focused on tonal and harmonic sounds. These methods are not necessarily suitable for transient and inharmonic bird sounds. Thus, the aim of our research was first to study how inharmonic and transient bird sounds can be classified and recognized efficiently and second to verify if the wavelet analysis can be used for this purpose. Wavelet analysis is a general mathematical tool, which can track time and frequency information in a signal (BOGGESS & NARCOWICH 2001). The wavelet analysis has an advantage over traditional Fourier analysis for signals showing spikes and discontinuities (BERRY 1999). As the Fourier analysis breaks down a signal into constituent sinusoids of different frequencies, the wavelet analysis uses the shifted and scaled versions of the original wavelet (DAUBECHIES 1992). Wavelets have gained a great deal of attention in the field of digital signal processing (RIOUL & VETTERLI 1991). In the wavelet packet transform the original signal is converted into wavelet coefficients. Because the parameter assessment using all wavelet coefficients will often turn out to be tedious or leads to inaccurate results, the extraction of the most important features is essential (PITTNER & KAMARTHI 1999). LEARNED (1992) has applied the wavelet packet transform to transient signals, and studied the classification of the whale clicks and the snapping shrimp.

Artificial neural networks (ANN) have been applied to an increasing number of real-world problems of considerable complexity. They are good pattern recognition engines and robust classifiers, with the ability to generalize in making decisions about imprecise input data. ANNs have successfully been applied to the automated classification of bioacoustic signals. For example, using artificial neural network, PHELPS and RYAN (1998) have studied the call of the Tungara Frog (*Physalaemus pustulosus*), DEECKE et al. (1999) have measured the similarities of discrete calls of Killer Whales (*Ornicus orca*), and PLACER and SLOBODCHIKOFF (2000) have illustrated that Gunnison's Prairie Dogs (*Cynomys Gunnisoni*) have different alarm calls for different species of predators. In (MCILRAITH & CARD 1997) the bird song recognition was made using backpropagation learning in two-layer perceptrons and using several methods from multivariate statistics. TERRY and MCGREGOR (2002) have studied the call of Corncrake (*Crex crex*) with the backpropagation and the probalistic network and the self-organizing map (SOM). Thorn (2003) has used the SOM to classify the vocal repertoire of the Barbary Macaque (*Macaca sylvanus L.*), and SOMERVUO and HÄRMÄ (2003) have used the SOM in song syllable analysis.

The main idea of this paper is to develop efficient classification methods for the inharmonic and transient bird sounds. The automatic recognition of inharmonic sounds is a difficult task, because in the spectral domain there are no visible trajectories which the a computer can track and identify. Thus, the wavelet packet transform is applied for feature extraction. Two commonly known neural networks, unsupervised self-organizing map (SOM) and supervised multilayer perceptron (MLP) are then used as classifiers, and their relative performance compared. The proposed method is tested with eight bird species, five producing inharmonic sounds, and the remaining three harmonic sounds.

SOUND DATA AND METHODS

Sound data

The sound data were analyzed in the MATLAB environment (MathWorks 2006), and the Wavelet Toolbox was utilized. The data was CD quality (44 100 Hz sampling frequency F_s, 16-bit accuracy).

Sounds were recorded in Finland by Pertti Kalinainen, Ilkka Heiskanen and Jan-Erik Bruun. They represent a total of 3132 sounds, which were divided into training data (2278 sounds), and testing data (854 sounds) training and testing data being from different tracks. For the SOM network, the training data were reduced to 113 samples per species, because the SOM network yields better results when the same number of training data of each group is used.

Table 1 illustrates the selected set of bird sounds used in this paper.

The idea was to choose bird species whose sounds are inharmonic. The sounds of the Mallard, the Graylag Goose, the Corncrake, the River Warbler and the Magpie are inharmonic and transient. Territorial song of male Pygmy Owl was chosen in order to test the performance of the method with harmonic sound.

Typical spectrograms and corresponding wavelet coefficient figures of the eight species used in this paper are presented in Fig. 1. In the spectrograms, the darker colours present the higher energies of the sound. Correspondingly, the larger absolute values of the coefficients are presented with the darker colours in the wavelet coefficient figures.

Three-dimensional bar charts of the sounds of the Corncrake and the Quail are presented in Fig. 2. The higher energies of the sound and correspondingly, larger squared values of wavelet coefficients are illustrated now with higher bars. So, the values of the most important wavelet coefficients are emphasized. As it can be seen from Fig. 2, the wavelet transform graphic representation compresses the energy of the coefficients more than traditional Fourier transform in spectrograms. Only the essential information is preserved after the wavelet transform.

METHODS

In Fig. 3 the whole classification process is presented. First the soundtracks were segmented into smaller pieces, called sounds in sequel. During the preprocessing the noise was reduced. All the sounds were then decomposed into the wavelet coefficients using the wavelet packet decomposition (WPD). The features were calculated from these wavelet coefficients and the feature vectors were built. The feature vectors of the training data were introduced to the multilayer perceptron (MLP) and to the self-organizing map (SOM) network during the training phase. Finally, both the networks were tested with the testing data and the recognition results were examined.

Segmentation and preprocessing

In the segmentation phase the zero mean soundtracks were normalized in the range [-1, 1], and the low frequency wind noise was reduced. The noise threshold level of each soundtrack was calculated adaptively from long-term mean energy value. All soundtracks were segmented into smaller pieces, so that the onset of the sound exceeded the adaptive threshold value and the end of the sound dropped under that value. The results of this automatic segmentation all sounds were checked manually, and sounds recorded in a very noisy environment and overlapping sounds were rejected. In the preprocessing phase the broadband noise was reduced from the sounds using eight-band filter bank.

Wavelet packet decomposition

The wavelet packet analysis was used for the signal decomposition. The following discussion of the wavelet decomposition is based on (BOGGESS & NARCOWICH 2001, DAUBECHIES 1992). In the wavelet decomposition a signal is broken down with the scaling function φ and the wavelet function ψ. The discrete scaling family function $\varphi_{j,k}(t)$ is defined as

$$\varphi_{j,k}(t) = 2^{j/2}\varphi\left(2^{j}t - k\right) \qquad j,k \in Z \tag{1}$$

and correspondingly, the discrete wavelet family function $\psi_{j,k}(t)$ is defined as

$$\psi_{j,k}(t) = 2^{j/2}\psi\left(2^j t - k\right) \qquad j,k \in Z \tag{2}$$

where t indicates time, j the scaling level and k the amount of shifting. Using the two discrete filters H_0 (low-pass) and H_1 (high-pass), and downsampling by two ($\downarrow 2$) after the filtering the decomposition can be done effectively (cf. Fig. 4).

The impulse responses of H_0 and H_1 are defined as

$$h_0 = \frac{1}{2}\overline{p_{-k}}\ , \quad \text{and} \qquad h_1 = \frac{1}{2}(-1)^k p_{k+1}\ , \tag{3}$$

where $\overline{p}$ is a complex conjugate of the p, and the coefficients p_k come from

$$p_k = 2\int_{-\infty}^{\infty} \phi(t)\overline{\phi(2t-k)}\,dt\ . \tag{4}$$

The whole wavelet decomposition can be done with approximation and wavelet coefficients, the a's and d's. The iteration of the decomposition algorithm can now be formulated in the convolution form as

$$a^{j-1} = (\downarrow 2)\left(h_0 * a^j\right), \tag{5}$$

$$d^{j-1} = (\downarrow 2)\left(h_1 * a^j\right), \tag{6}$$

where the sequences a^j and the d^j are defined as

$$a^j = \left(\ldots, a_{-1}^j, a_0^j, a_1^j, \ldots\right), \tag{7}$$

$$d^j = \left(\ldots, d_{-1}^j, d_0^j, d_1^j, \ldots\right). \tag{8}$$

It is important to notice that the discrete filters and downsampling operator ($\downarrow 2$) do not depend on the level j. The accuracy of the approximation increases with increasing j.

Hence, it is important to chose *j* large enough so that all of the coefficients are accurately calculated.

In the WPD the signal s is split into approximation (A) and detail (D) parts. The approximation coefficients *a*′s are in the part A, and respectively, the wavelet coefficients *d*'s are in the part D. Due to the downsampling, aliasing occurs in the WPD tree. This aliasing exchange the frequency ordering of some branches of the tree (AKANSU & HADDAD 1992). The symmetric wavelet decomposition tree is illustrated in Fig. 5, where the WPD tree is put in an increasing frequency order from the left to the right.

In our case, after preliminary tests it turned out that the best decomposition level (N) was six. Thus, the signal *s* was split into $2^6 = 64$ parts, called bins in sequel. The bin number 1 contained so low frequencies that they proved to be irrelevant for the recognition. The bins 33-64 also proved to be irrelevant; so over 11.025 kHz ($F_s/4$) frequencies were not needed in the recognition. Thus, the wavelet coefficients were calculated from bins 2-32, which are marked grey in Fig. 5.

There are several wavelet families that have proven to be particularly useful. The Daubechies db10 was selected for the wavelet function, because in preliminary tests this wavelet function compromised the best decomposition results of the tested alternatives with the selected set of bird sounds.

Feature calculation

The number of the WPD coefficients of each bin is denoted as n_c. The main disadvantage of the wavelet transform is its time dependence. In the feature calculation phase, the coefficient data of each sound were further reduced to four shift invariant features, *maximum energy*, *position*, *spread*, and *width*. They are illustrated in Fig. 6.

The bin energy $E_B(r)$ of the wavelet coefficients *c* of bin *r* was calculated as

$$E_B(r) = \sum_{n=1}^{n_c} c^2(n,r) \qquad r = 1,\ 2,\ldots,\ 31 \tag{9}$$

and then the average energy of each bin *r* was calculated as

$$\tilde{E}_B(r) = \frac{E_B(r)}{n_c}, \tag{10}$$

The largest average energy value

$$E_m = \max_r(\tilde{E}_B(r)) \tag{11}$$

was searched as, and it is called the *maximum ener*gy E_m of the sound. The *position* P represents the number of the bin r, in which the maximum energy was located.

After preliminary test with the data the threshold value $T_{h1}(r)$ was calculated as

$$T_{h1}(r) = \frac{\tilde{E}_B(r)}{6} \tag{12}$$

from the average energy of bin r. The *spread* S was calculated as

$$S = \frac{1}{\#J} \sum_{(q,r) \in J} c^2(q,r) , \tag{13}$$

where q is the number of the sample and r the number of the bin. J is a set of index pairs (q,r) for which $c^2(q,r) > T_{h1}(r)$. In (14) $\#J$ is the number of elements (cardinality) of the set J. So, the spread S is a sum of the average energies of those coefficients, whose energy exceeded the threshold value T_{h1}. The other threshold value T_{h2} was selected as 1.3 after preliminary tests with the data. The number of bins, which satisfy the inequality

$$T_{h2} = 1.3 < E_B(r) \tag{14}$$

was computed. It is called the *width* W of the sound.

Finally all four features were normalized, in order to be comparable with each other. The normalization levels were defined after preliminary tests with the data. The maximum energy E_m was normalized as

$$\widetilde{E}_m = \frac{E_m}{n_B} \tag{15}$$

where n_B is the number of those coefficients of the bin, where the coefficients exceeded the T_{h1}. The position P was normalized as

$$\tilde{P} = \frac{P}{2^N / 4} = \frac{P}{16} . \tag{16}$$

The spread S was normalized

$$\tilde{S} = \frac{S}{100}, \tag{17}$$

and the width W as

$$\tilde{W} = \frac{W}{20}. \tag{18}$$

These four normalized features formed the final feature vector for sound recognition. The main reason for the normalization was the SOM, which yields better recognition results, if the inputs are in the same scale. The training time of the SOM network is also shorter with normalized inputs.

Network training and testing

Two well-known neural networks, the self-organizing map (SOM) and the multilayer perceptron (MLP), were used for the recognition. Our purpose was to compare the relative performance of these two different networks. The SOM is based on unsupervised learning and the MLP on supervised learning. In unsupervised learning there is no a priori knowledge of the categories into which the feature vectors are to be classified. In supervised learning the network has to undergo the training session first, which means that a set of input patterns along with the known class is repeatedly presented to the network (HAYKIN 1994).

The SOM is a clustering and visualization tool that enables the organization of the database in an unsupervised manner (KOHONEN 2001). In the training phase the SOM network was trained using the feature vectors of 904 sounds as inputs (cf. Table 1). These feature vectors were introduced to 10 x 10 -size SOM. The other sizes, for example 6 x 6, 8 x 8, and 12 x 12, of the network were also tested, but the chosen size yielded best recognition results. When initialising the SOM, the random layer topology function (randtop) was used. That specified the topology for the original neuron locations. The training data was introduced to the SOM network 3000 times. Other number of epochs was also tested, but the results did not improve.

The second network, which was used in the experiments, was the MLP. There were 2278 sounds in the training phase of the MLP network (cf. Table 1). The chosen MLP architecture was 4-15-40-3, which means that the MLP network contained the four features (maximum energy, position, spread, width) at the input layer. The first hidden layer had 15 and the second hidden layer had 40 nodes. Between the input and the first hidden layer, and the first and the second hidden layer hyperbolic tangent sigmoid function (tansig) was used as the activation function. The range of the tansig activation function was [-1, 1]. Finally, in the output layer three outputs were used, and log sigmoid function

(logsig) was used as activation function between the second hidden layer and the output layer. The range of the logsig function is [0, 1]. After that each output was rounded to 0 or 1. Thus, it was possible to present decimal numbers 0-7 with three output bits, and that was enough for classes of eight bird sounds. Finally, the binary output was converted into numbers 1-8. The MLP network was trained for up to 65 epochs[1] and the mean square error (MSE) goal was 0.0001. After the training, it was examined that all the nodes, and the weighting and bias parameters of the MLP network were needed, which means that none of the outputs of the nodes was too close to zero.

In the testing phase both the networks were tested with the testing data. The testing data were from different tracks than the training data, and contained totally 854 sounds (cf. Table 1).

RESULTS

The clustering result of the SOM network after the training phase is illustrated in Fig. 6. The areas marked with letters present how sounds of each bird species were situated in the 10 x 10 SOM network after the over-lapping nodes have been analysed. The SOM network was examined node by node and the outliers were labelled. The species which had most sounds in a particular node won and the possible other sounds were classified as outliers. If two or more different species had the same number of sounds in the particular node, they all were classified as outliers. If no species won or any sound did not situate in the particular node, it was classified as unspecified node. Unspecified nodes are marked with black colour in Fig. 7. The SOM clustered 87% of training sounds into their own nodes.

The confusion matrix of Table 2 illustrates the recognition results of the SOM network after the testing phase. The rows of the confusion matrix show how each species is recognized. All the test sounds of the River Warbler (LOCFLU) were recognized correctly, as it can be seen from the examination of the diagonal of the matrix. Altogether, 7% of the test sounds were unspecified and 15% were recognized wrongly. Altogether, 92 sounds among the 854 tested sounds were recognized wrongly. Altogether, 78% of the test sounds were recognized correctly with the SOM network.

Table 3 contains the recognition result of the MLP network. All the test sounds of the Quail (COTCOT) and the Spotted Crake (PORPOR) were recognized correctly. Twenty-four sounds of all the test sounds were recognized wrongly. Altogether, 96% of the test sounds of the eight bird species were recognized correctly with the MLP network.

DISCUSSION

Our aim was to study how inharmonic and transient bird sounds can be recognized and classified efficiently. The wavelet analysis was selected for its ability to analyse signals showing discontinuities and sharp spikes, such as those found in inharmonic bird

sounds. By wavelet analysis, it is possible to present the essential information of the sounds with only few features. In this paper, the two well-known neural networks, the self-organizing map (SOM) and the multilayer perceptron (MLP) were used for classification. The test data consisted of sounds of eight bird species: the Mallard, the Graylag Goose, the Corncrake, the River Warbler, the Magpie, the Quail, the Spotted Crake, and the Pygmy Owl.

In automatic sound based bird monitoring the soundscape may contain hundreds of other natural sounds as well as noise. Segmentation methods based on energy or spectral content of the sound data operate poorly in low signal to noise ratio situations or with overlapping sounds. This is often the case with bioacoustic data. Accurate segmentation is very important, because incorrectly segmented sounds will probably be classified wrongly. Segmentation is the most time consuming part of the whole recognition process. All segments of sounds must be checked before calculating the features. In many cases a manual checking is the most reliable way but it is time consuming. Thus, the development of new automatic segmentation methods for sound data analyses needs more investigation. Noise reduction goes hand in hand with successful segmentation: the segmentation is more difficult if the sound tracks are very noisy. Thus, efficient but subtle noise reduction constitutes a crucial part of the pre-processing, as the original sound information of the target species should remain intact.

The wavelet analysis has many advantages, particularly for its ability to accurately track both frequency and temporal information and to analyse signals, which contain discontinuities and sharp spikes. These properties are appropriate for inharmonic and transient bird sounds. However, the wavelet transform is time variant, which is its main disadvantage in bird sound recognition. To avoid this problem, shift invariant features were used in this paper. The selection of the wavelet function and the decomposition level are the most important phases of the wavelet packet decomposition (WPD). In this study the db10 was selected experimentally for the wavelet function.

The data of the eight bird species was divided so that there were about 70% training data and 30% testing data. This kind of division is common when neural networks are used. Although the neural networks have many benefits, like their ability to learn and therefore generalize, there is a long way to go before the recognition system beats the human ear. The disadvantage of the neural networks is the fixed number of output classes. If more species must to be classified, the network has to be retrained all over again before it can be tested with the new set of birds.

The training data contained probably sounds of seven Mallard, nine Graylag Goose, eight Corncrake, two River Warbler, six Magpie, three Quail, three Spotted Crake, and five Pygmy Owl individuals. Respectively, the testing data contained sounds of two Mallard, four Graylag Goose, two Corncrake, one River Warbler, one Magpie, two Quail, one Spotted Crake and two Pygmy Owl individuals. More data will be needed in order to generalize the results.

After the MLP network testing session, all wrongly classified sounds were manually examined and labelled. It turned out that 24 sounds were classified wrongly. The shape of the coefficient pattern of the highest, 6^{th} level of the wavelet packet tree proved to be very

significant. It turned out that the coefficient pattern of the outlier Greylag Goose resembles the Mallard more than the Greylag Goose. Similar coefficient pattern error might be the reason for the misrecognition in 18 other cases. The wrong recognition was presumably caused by the bit pattern error in six cases. One bit might have flipped due to false segmentation or poor sound quality. Hence, that can be the reason, why for example three sounds of Greylag Goose (bit pattern 001) were recognized as River Warblers (bit pattern 101).

Also the SOM network was checked after the testing session. It turned out that 32 sounds were classified as outliers. Most of the SOM network misrecognitions might result from the variation of the coefficient pattern shape of the highest level (level 6) in the wavelet packet tree. The value of the feature vectors varied too much and it might have caused the outliers. Another presumable reason was the segmentation error, which might cause the misrecognition. It should be noticed that the SOM classified all the ten outliers of the Mallard as the Greylag Goose and 12 of 14 outliers of the Greylag Goose as the Mallard. That was very interesting, because both of those species belong to the *Anseriformes* and so they belong to the same order and are relatives.

The SOM classified 78% and the MLP 96% of the test sounds correctly. The difference between the results might be a consequence of the MLP being a supervised and the SOM an unsupervised network. The supervised network may learn and generalize the data better and faster than unsupervised network. On the other hand, the data might have not been entirely suitable for the SOM network. Although the data had been checked, there were a few sounds, which had false segmentation or poor sound quality. That might have disturbed the unsupervised SOM more than supervised MLP.

In conclusion, the presented results are very encouraging. It turned out that it is possible to recognize bird sounds using neural networks with only four features calculated from the wavelet packet coefficients. The automatic classification in general presents a new method for identifying and differentiating bird species by their sounds, and may offer new important tools for ecological and evolutional bird researches.

ACKNOWLEDGEMENTS

The authors thank Pertti Kalinainen, Ilkka Heiskanen and Jan-Erik Bruun for their recordings. The authors also wish to thank nature-recording expert, Pertti Kalinainen for his specialized comments and Docent Mikko Ojanen for helpful discussions. This research was funded by the Academy of Finland under research grant 206652.

REFERENCES

AKANSU, A.N. & HADDAD, R.A., 1992: Multiresolution signal decomposition: transforms, subbands, and wavelets.- Boston: Academic Press.

BAKER, M.C. & LOGUE, D.M. 2003: Population Differentiation in a Complex Bird Sound: A Com-parison of Three Bioacoustical Analysis Procedures.- *Ethology*, 109 (3), 223-242.

BERRY, S., 1999: Practical wavelet signal processing for automated testing.- In: Proceedings of the IEEE Systems Readiness Technology Conference (AUTOTESTCON '99).- San Antonio, 30 August - 2 September 1999, pp. 653-659.

BOGGESS, A. & NARCOWICH, F.J, 2001: A first course in wavelets with Fourier analysis.- New Jersey: Prentice-Hall, Inc.

CATCHPOLE, C.K. & SLATER, P.J.B., 1995: Bird song: biological themes and variations.- Cambridge: Cambridge University Press.

DAUBECHIES, I., 1992: Ten lectures on wavelets.- Philadelphia: Society for Industrial and Applied Mathematics.

DEECKE, V.B., FORD, J.K.B. & SPONG, P., 1999: Quantifying complex patterns of bioacoustic variation: use of a neural network to compare Killer Whale (*Ornicus orca*) dialects.- *Journal of Acoustical Society of America*, 105 (4), 2499-2507.

DEECKE, V.B. & JANIK, V.M., 2006: Automated categorization of bioacoustic signals: avoiding perceptual pitfalls.- *Journal of the Acoustical Society of America*, 119 (1), 645-653.

FAGERLUND, S., 2004: Automatic recognition of bird species by their sounds.- Master's Thesis, Helsinki University of Technology.

HAYKIN, S., 1994: Neural networks: a comprehensive foundation.- New York: Macmillan College Publishing Company, cop.

HÄRMÄ, A., 2003: Automatic identification of bird species based on sinusoidal modelling of syllables.- In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03).- Hong Kong, 6-10 April 2003, pp. V-545-548.

HÄRMÄ, A. & SOMERVUO, P., 2004: Classification of the harmonic structure in bird vocalization.- In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04).- Montreal, Canada, 17-21 May 2004, pp.V-701

KOHONEN, T., 2001: Self-Organizing Maps.- Berlin: Springer.

LEARNED, R., 1992: Wavelet packet based transient signal classification.- Master's Thesis, Massachusetts Institute of Technology.

MATHWORKS, 2006: Matlab software homepage: URL: http://www.mathworks.com/, 2nd February 2006.

MCILRAITH, A.L. & CARD, H.C., 1997: Birdsong recognition using backpropagation and multivariate statistics.- *The Proceedings of the IEEE Transactions on Signal Processing*, 45 (11), 2740-2748.

MESGARANI, N. & Shamma, S., 2003: Bird call classification using multiresolution spectrotemporal Auditory Model.- In: Proceedings of the 1st International

Conference on Acoustic Communication by Animals.- Maryland, USA, 27-30 July 2003, pp. 155-156.

PLACER, J. & SLOBODCHIKOFF, C.N., 2000: A fuzzy-neural system for identification of species-specific alarm calls of Gunnison's Prairie Dogs.- *Behavioural Processes*, 52 (1), 1-9.

PHELPS, S. & RYAN, M.J., 1998: Neural networks predict response biases of female Tungara Frogs.- *Proceedings of the Royal Society of London, Series B, Biological Sciences*, 265 (1393), 279-285.

PITTNER, S. & KAMARTHI, S.V., 1999: Feature extraction from wavelet coefficients for pattern recognition tasks.- *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 21 (1), 83-88.

RIOUL, O. & VETTERLI, M., 1991: Wavelets and signal processing.- *IEEE Signal Processing Magazine*, 8 (4), 14-38.

SOMERVUO, P. & HÄRMÄ, A., 2003: Analyzing bird song syllables on the self-organizing map.- In: Proceedings of the workshop on self-organizing maps (WSOM '03), Hibikino, Japan, 11-14 September 2003. Proceedings on CD-ROM.

SOMERVUO, P. & HÄRMÄ, A., 2004: Bird song recognition based on syllable pair histograms.- In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04), Montreal, Canada, 17-21 May 2004, pp. V-825-828.

TANTTU, J.T., TURUNEN, J. & OJANEN M., 2003: Automatic classification of flight calls of Crossbill species (*Loxia* spp.).- In: Pproceedings of the 1st International Conference on Acoustic Communication by Animals, Maryland, USA, 27-30 July 2003, pp. 239-240.

TANTTU, J.T., TURUNEN, J., SELIN, A. & OJANEN M., 2006: Automatic feature extraction and classification of Crossbill (*Loxia* spp.) flight calls.- *Bioacoustics*, 16 (1).

TERRY, A.M.R. & MCGREGOR, P.K., 2002: Cencus and monitoring based on individually identifiable vocalizations: the role of neural networks.- *Animal Conservation*, 5 (2), 103-111.

Table 1: Selected set of bird sounds used in this paper.

Scientific Abbr.	Scientific name	English name	Sound Type	MLP Trainin	SOM Trainin	Testing
ANAPLA	*Anas platyrhynchos*	Mallard	inharmonic	138	113	60
ANSANS	*Anser anser*	Graylag Goose	inharmonic	135	113	59
CRECRE	*Crex crex*	Corncrake	inharmonic	443	113	110
LOCFLU	*Locustella fluviatilis*	River Warbler	inharmonic	890	113	328
PICPIC	*Pica pica*	Magpie	inharmonic	203	113	97
COTCOT	*Coturnix coturnix*	Quail	tonal	190	113	83
PORPOR	*Porzana porzana*	Spotted Crake	tonal	166	113	69
GLAPAS	*Glaucidium passerinum*	Pygmy Owl	pure harmonic	113	113	48
				2278	904	854

Table 2: The confusion matrix in percentage terms when using the SOM network.

%	ANAPLA	ANSANS	CRECRE	LOCFLU	PICPIC	COTCOT	PORPOR	GLAPAS	Unspecified
ANAPLA	78	20	0	0	0	0	0	0	2
ANSANS	24	51	0	0	0	0	2	0	23
CRECRE	0	0	83	0	1	0	0	0	16
LOCFLU	0	0	0	100	0	0	0	0	0
PICPIC	1	0	1	0	58	2	38	0	0
COTCOT	0	0	0	0	8	87	4	0	1
PORPOR	0	0	0	0	9	0	91	0	0
GLAPAS	0	15	0	0	0	0	0	75	10

Table 3: The confusion matrix in percentage when using the MLP network.

%	ANAPLA	ANSANS	CRECRE	LOCFLU	PICPIC	COTCOT	PORPOR	GLAPAS
ANAPLA	98	2	0	0	0	0	0	0
ANSANS	1,7	83	5,1	5,1	1,7	1,7	0	1,7
CRECRE	1	2	96	0	1	0	0	0
LOCFLU	0	0,3	0	99,7	0	0	0	0
PICPIC	0	0	1	0	94	5	0	0
COTCOT	0	0	0	0	0	100	0	0
PORPOR	0	0	0	0	0	0	100	0
GLAPAS	0	2	0	2	0	0	0	96

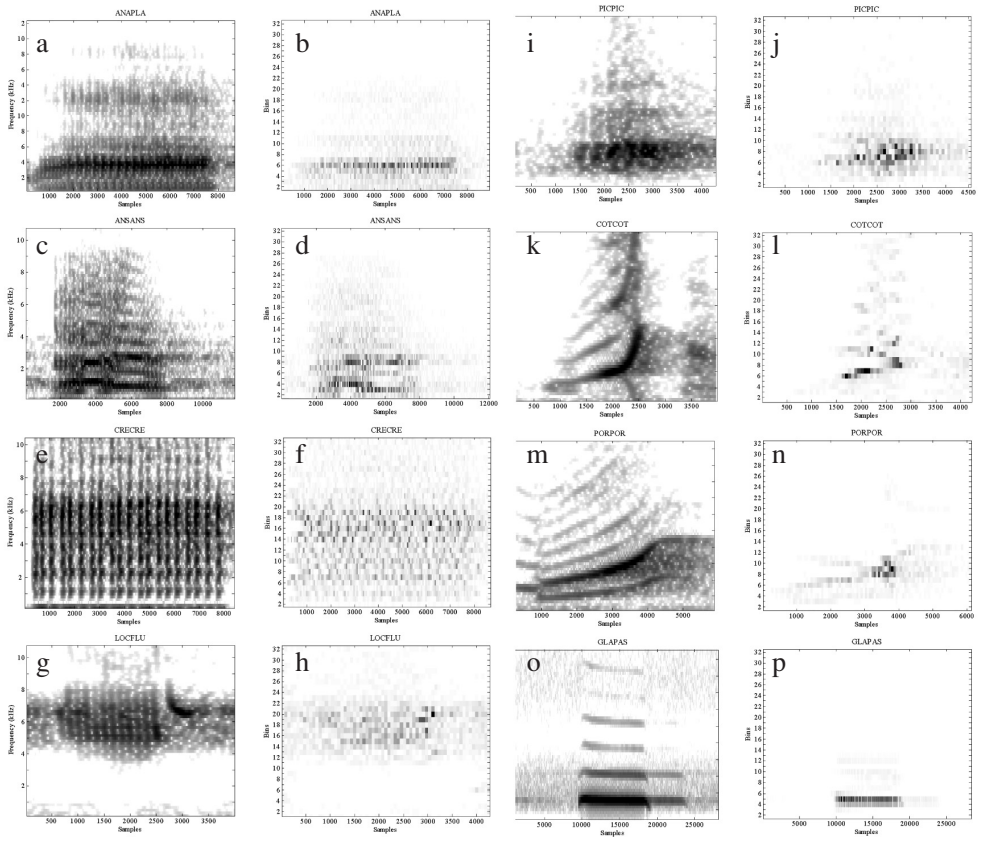

Figure 1: Typical spectrograms (1st and 3rd column) and corresponding wavelet coefficient figures (2nd and 4th column) of the eight species used in this paper.

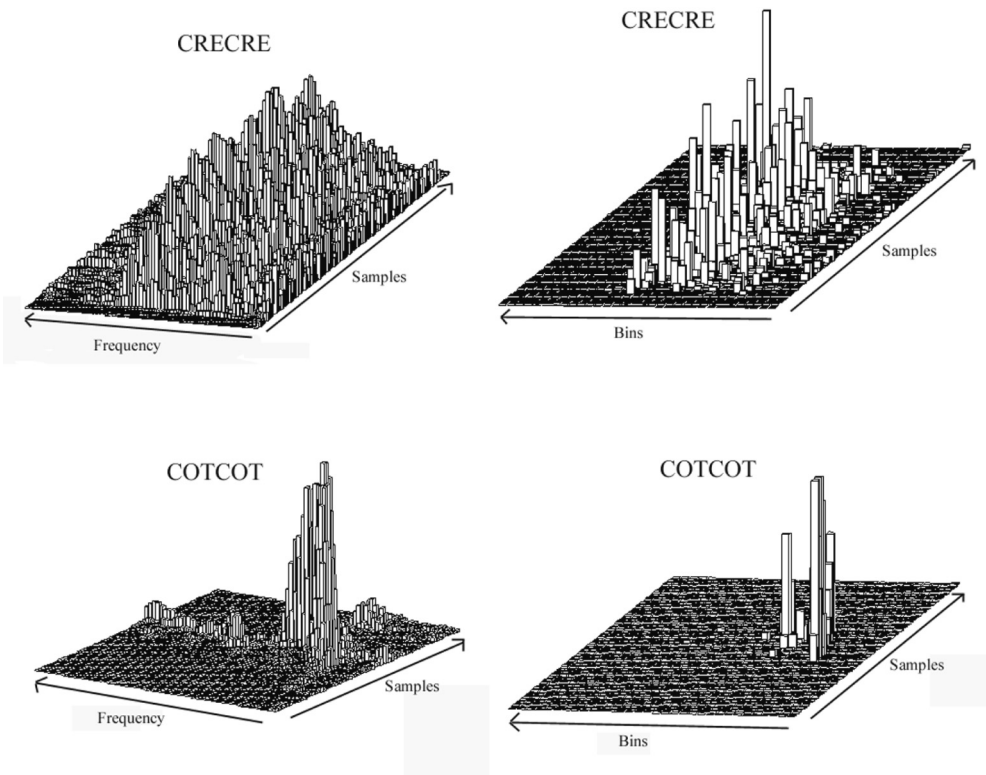

Figure 2: Three-dimensional bar charts of the higher energies of the sounds (left) (cf. left column of Fig. 1) and larger squared values of the wavelet coefficients (right) (cf. right column of Fig. 1) of the sounds of the Corncrake (*Crex crex*) and the Quail (*Coturnix coturnix*).

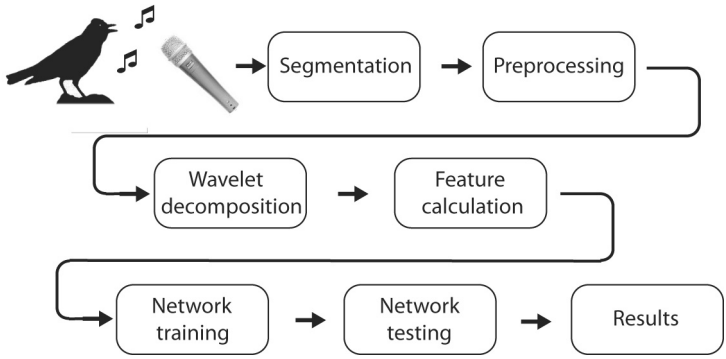

Figure 3: The classification process.

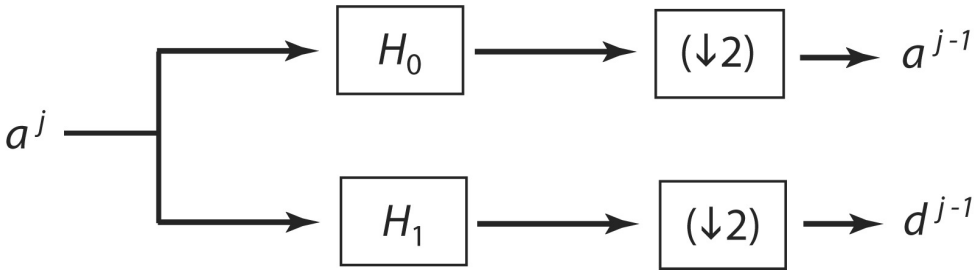

Figure 4: The iteration of the decomposition algorithm using filters H_0 and H_1.

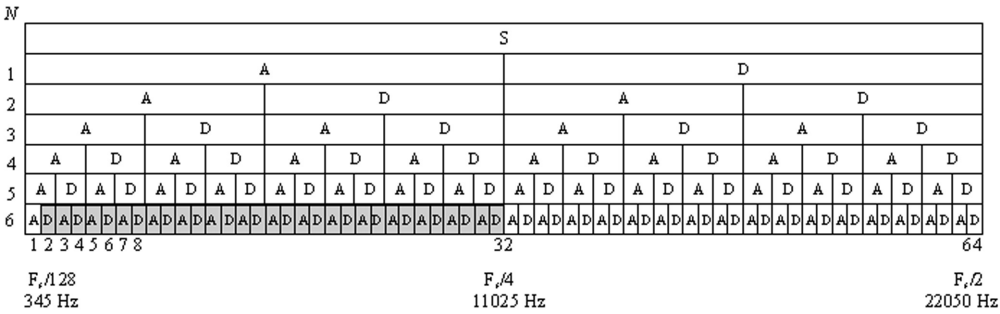

Figure 5: The symmetric wavelet packet decomposition tree. The grey bins are used in the proposed method.

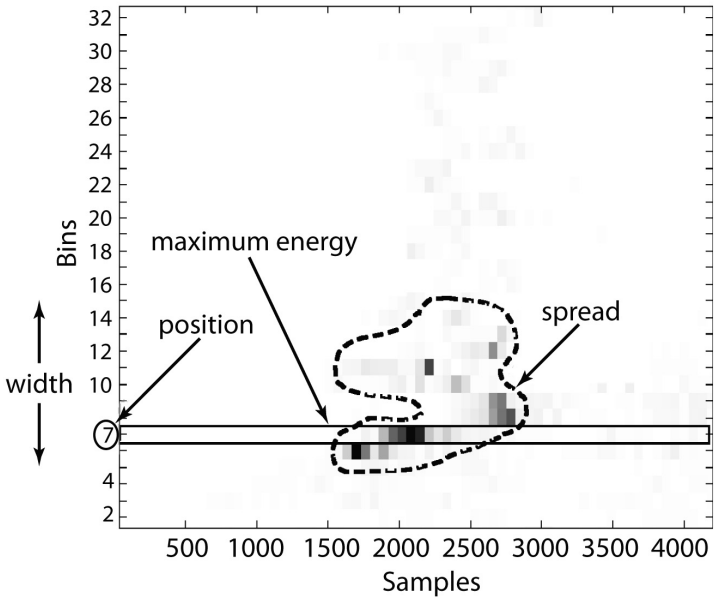

Figure 6: The four feature, maximum energy, position, spread, and width.

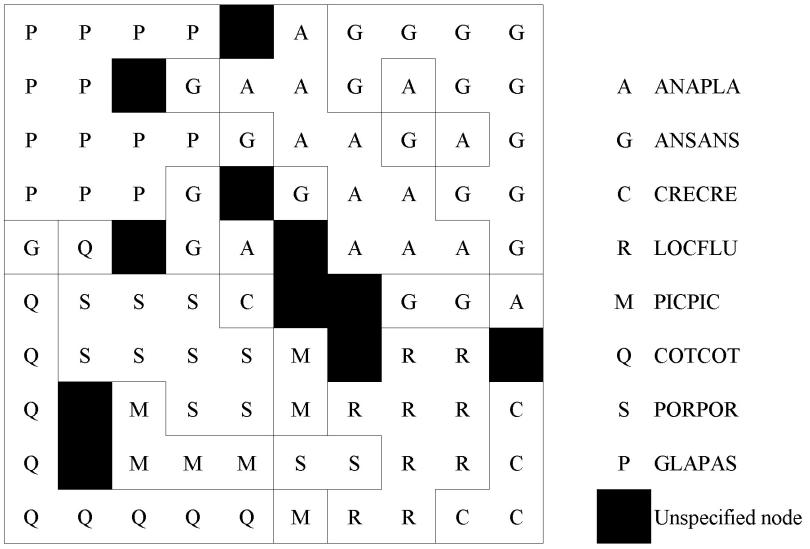

Figure 7: The clustering result of the 10 x 10 -size SOM network after training.