Automatic Identification of Bird Species: a Comparison Between kNN and SOM Classifiers

Dorota Kamińska
Technical University of Lodz
Institute of Mechatronics and Information Systems
Poland, 90-924 Lodz
Email: kaminska.dorota@o2.pl

Artur Gmerek
Technical University of Lodz
Institute of Automatic Control
Poland, 90-924 Lodz
Email: artur.gmerek@p.lodz.pl

***Abstract* — This paper presents a system for automatic bird identification, which uses audio input. The experiments have been conducted on three groups of birds, which were created basing finishing on classification, the system is fully automated. The main problem in automatic bird recognition (ABR) is the choice of proper features and classifiers. Identification has been made using two classifiers – kNN (k Nearest Neighbor) and SOM (Self Organizing Maps). System has been tested using data extracted from natural environment.**

Index Terms* — *birds, kNN, HMM, recognition, identification, self organizing maps, SOM

I. INTRODUCTION

The main goal of this paper was to develop an automatic system for bird recognition using audio input. This kind of system could be valuable for biological research and environmental monitoring. It is possible to recognize bird species using audio recording, basing on the fact that birdsongs have a grammatical structure and are composed of notes, syllables, phrases and calls (including alarm calls, distress calls, territorial calls and others). A set of one or more syllables and phrases arranged in a regular pattern is referred to as a song.

Classification of bird species by their sound is not a challenging task when they belong to different families. However, practical systems should be able to distinguish birds belonging to the same family but different species. Thus, experiments have been made on three different groups. The similarity between bird sounds in each group differ respectively: small, medium and significant difference. Groups have been created, based on correlation between the most descriptive features.

In order to chose proper methods of classification, literature study on whole spectrum of algorithms has been made. SOM and kNN classifiers, which gave satisfactory results, have been chosen and compared in this paper.

The majority of scientists who conduct research in this field use manual syllables division. Thus, currently existing systems are not fully automated. This paper presents a fully automated algorithm. Moreover, there is no problem with adjusting the system for new birds recognition using training module.

II. RELATED WORK

Analysis of bird sounds can be divided into three main parts: segmentation of bird sounds (e.g. to syllables), features extraction and classification. Scientists usually use manual or semi-automatic syllable segmentation. The last two stages of identification often differ greatly depending on individual approach.

Most of researchers use simple statistical features i.e. mean value, frequency bandwidth, duration of syllable, signal amplitude. Sometimes more sophisticated features are used e.g. Linear Predictive Coding (LPC), LPC Cepstral Coefficients (LPCCs), Mel-Frequency Cepstral Coefficients (MFCCs) [1] or wavelet coefficients.

Statistical classifiers like k nearest neighbors, bayesian classifiers and decision trees can be used for the purpose of bird recognition. Some methods, which are common for human voice identification like Dynamic Time Warping (DTW), Hidden Markov Models (HMMs) [2]–[4], Gaussian Mixture Models (GMM) and Vector Quantization (VQ) have been also used for birds species identification.

Lakshminarayanan et al. have introduced probabilistic models based on birdsong syllables [5]. Their Independent Frame Independent Syllable (IFIS) and Markov Chain Frame Independent Syllable (MCFIS) models achieved better results than Support Vector Machine (SVM) classifier.

Aki Harma has performed identification using sinusoidal modeling [6] basing on the fact that syllables can be approximated as varying amplitude and frequency brief sinusoidal pulses.

In recent years neural networks like Multilayer Perceptron (MLP) [7], Time Delay Neural Networks (TDNN) [8], Autoregressive Time-Delay Neural Networks (AR-TDNN) [9] and Self Organizing Maps [10] have been used by many scientists.

The most valuable are those publications, in which different methods are compared [11]. For example Briggs and others have presented a different statistical manifold approach [12].

McIlraith and Card have compared between backpropagation learning in two-layer perceptrons and discriminant analysis [13]. They have used simple statistical features (duration, mean, standard deviations, power spectral densities) and more complicated e.g. LPC. They achieved performance range from 82% to 93%, but experiments have been made merely on six different species.

III. METHODS

Representation of the signal in time or frequency domain is a complex projection. Therefore features are sought to determine signal properties. In this study following features have been used: duration, bandwidth, fundamental frequency, power spectral density of a syllable and formant and antiformant frequencies. Also other features have been used, which are described in following subsections.

A. LPC coefficients

Linear predictive coding is a method used in audio signal processing for representation of spectral envelope of a digital signal in compressed form. Linear prediction, based on the assumption that a signal sample $u(n)$, can be approximated by linear combination of P previous samples for $n > 0$. The predicted signal value is expressed by the formula:

$$\tilde{u}(n) = -\sum_{p=1}^{P} a_p u(n-p)$$

$u(n-p)$ - previous observed values,
a_p - predictor coefficients,

LPC coefficients are determined by autocorrelation criterion. In this method the expected value of the squared error, which is defined as following equation, is minimized:

$$\sigma = E[err^2(n)] = \frac{1}{N-p}\sum_{n=p}^{N-1}\left[u(n) + \sum_{p=1}^{P} a_p u(n-p)\right]^2$$

where N is the number of samples.

To determine the optimal coefficients a_k, $1 \leq k \geq p$, a partial derivative of σ with respect to the variable a_p should be calculated and equated to zero. Afterwards p equations containing p variables are obtained with following solution:

$$Ra = -r$$

where R is a symmetric, autocorrelation matrix called Toeplitz matrix.

Experiments show, that the most optimal number of LPC coefficents is 12, therefore this amount has been used in this study.

B. Mel Frequency Cepstral Coefficients

Currently MFCCs are a standard in speech recognition [14]. The MFCC algorithm is multistage. At first, the signal is multiplied by the Hamming window, presented by the following equation:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\Pi n}{N-1}\right) & , 0 \leq n \geq N-1 \\ 0 & , \text{otherwise} \end{cases}$$

where N specifies window size.
Subsequently FFT is computed. Then the estimation of power spectral density function is calculated and averaged using overlaping triangular weight functions. Design of the triangular functions includes mel scale. Following equations present the same frequency, m is frequency value in mel, f in $hertz$ using natural logarithm:

$$m = 1127,01048\ln(1 + f/700), \quad f = 700(e^{\frac{m}{1127,01048}} - 1),$$

and using decadic logarithm:

$$m = 2595log(1 + f/700), \quad f = 700(10^{\frac{m}{2595}} - 1).$$

The last step is calculation of a Discrete Cosine Transform (DCT) of the logarithmed estimation, using the following formulas:

$$c_k = \sqrt{\frac{2}{L}}\sum_{l=0}^{L-1} ln\tilde{S}(l)cos\Big(\frac{\Pi k}{L}(l+1/2)\Big), k = 0, 1, ...q-1$$

where L is the number of weight functions and q is the number of Mel Coefficients.

IV. CLASSIFIERS

Classification is an algorithm, which assigns objects to groups (called classes) based on object features. Features are usually presented in a vector:

$$x_j = [x^1, x^2, x^3, ...x^d]$$

where d is the number of features, x^k is feature value.

All features values in a task are called the training set CU. It can be said, that the goal of classification is to assign a class $i \in M$ for an individual object x_j.

A. Nearest Neighbor Classifier (kNN)

In kNN algorithm the recognition process involves calculating distances in parameters space X between the unknown x_j object and all objects from the training set $x_k \in CU$ for $k = 1, 2, ..., I$, where I is the number of training examples.

In presented project euclidean distance has been used:

$$d(x_j, x_k) = \sqrt{\sum_{i=1}^{n}(x_{ij} - x_{ki})^2}$$

Obtained distances are sorted in an ascending order. Object x_j is assigned to this class, which is the most common among k nearest objects.

B. Self Organizing Map

Self Organizing Map is type of Artificial Neural Network (ANN). This type of ANN learns without a teacher, using only the observation of the input data (unsupervised learning). Network map, which creates a static grid cell, has a fixed size. It usually has a rectangular or hexagonal structure. Weights of input neurons can be initiated with random values. SOM has two basic methods of changing the neurons weights. The first one - Winner Takes All (WTA): the neuron, whose weights are closest to the input vector components is modified in such a way that its weights are as close as possible to theinput vector. The second one, Winner Takes Most (WTM):neuron with weight most similar to the input value is called the winner.

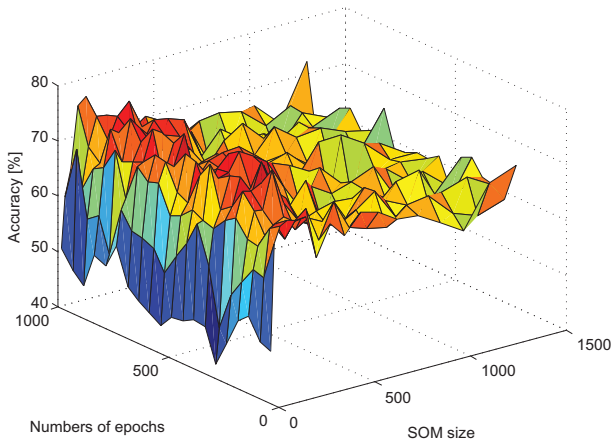

Fig. 1. The classification accuracy of birds sound with respect to the SOM size and number of epochs. The optimal number of neurons for given task was 450.

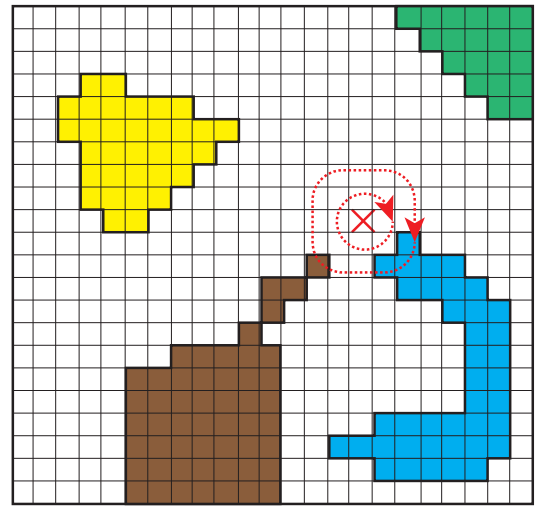

Fig. 2. An example of SOM processing. Colored areas represent different clusters (different bird species). Neighborhood urrounding results of classification (marked here as x) is measured based on euclidean distance. In a presented example result does not overlap with any areas, there are no clusters in the nearest distance of one checkered pattern. Because of that, algorithm checked the next checked patterns. There are 3 of them, which belongs to blue cluster and one, which represent to brown cluster. Consequently results will be classified as blue.

Its weights and neighboring neurons weight are modified. Frequently, this modification is dependent on the distance from the winner.

First step of the learning algorithm is to find the nearest maps element to the $c(x)$ vector.

$$c(x) = argmin||x - m_i||$$

where x is a sample from the training set CU in the step t

Then the winner and its neighbors are modified according to the formula:

$$m_i(t+1) = m_i(t) + h_{c(x),i}(t)[x(t)m_i(t)]$$

where $h_{c(x),i}$ is a neighborhood function given by:

$$h_{c(x),i} = \alpha(t)exp\big(-\frac{||r_i - r_{c(x)}||^2}{2\sigma^2(t)}\big)$$

where $\alpha(t)$ is related to velocity of learning process.

Optimal number of learning epochs and SOM size has been calculated for this problem (Fig. 1). The network does not produce definite results of classification. It rather illustrates links between patterns by projecting them onto n-dimensional plane. After projection, data has to be decoded again in order to achieve accuracy of classification. This process is done by checking the distances between the nearest clusters of data around the point being the result of classification (Fig. 2).

V. EXPERIMENTS

Studies have been conducted on 10 birds species. All files have been downloaded from different Internet sources. The format of these files was PCM WAVE with 44100Hz sampling rate. 70% of files were used as training, 30% as testing set. Both sets were disjoint.

There are 3 groups of birds presented in the Table I. Species from these groups are correlated on a different level. For example the first group consists of birds, whose sounds are in high correlation (only from Passeriformes order). Thus, classification of birds from the first group could be more problematic in relation to other groups.

TABLE I
COMMON AND LATIN NAMES OF BIRD SPECIES OF THE BIRDSONG DATABASE AND THEIR CORRESPONDING AMOUNT OF SYLLABLES

Common name	Latin name	Training Syll.	Test Syll.
Great Tit	Parus major	562	137
Blackbird	Turdus merule	522	143
Eurasian Nuthatch	Sitta europea	530	104
Robin	Erithacus rubecula	543	140
Thrush Nightingale	Luscinia luscinia	370	80
Great Tit	Parus major	562	137
Blackbird	Turdus merule	522	143
Eurasian Nuthatch	Sitta europea	530	104
Grey Partridge	Perdix perdix	246	62
Tengmalm's Owl	Aegolius funereus	277	86
Common Swift	Apus apus	555	83
Wild Duck	Anas platyrhynchos	403	65
Common Cuckoo	Cuculus canorus	330	82
Grey Partridge	Perdix perdix	246	62
Tengmalm's Owl	Aegolius funereus	277	86

In order to prepare data for classification, signals were processed according to the algorithm presented in (Fig. 3). The first step - preprocessing, prepared the signal for features extraction. After that different features were calculated. The process of classification was divided into two stages: learning (teaching SOM classifier and creating a code book for kNN) and testing itself.

A. Preprocessing

The goal of preprocessing is adaptation and simplification of the signal for further analysis. It is divided into three steps:

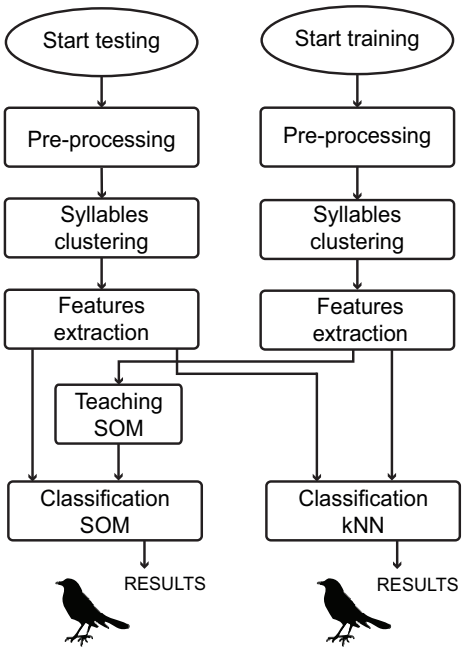

Fig. 3. Algorithm for processing audio files. Experiments have been conducted on 3 different groups of birds species.

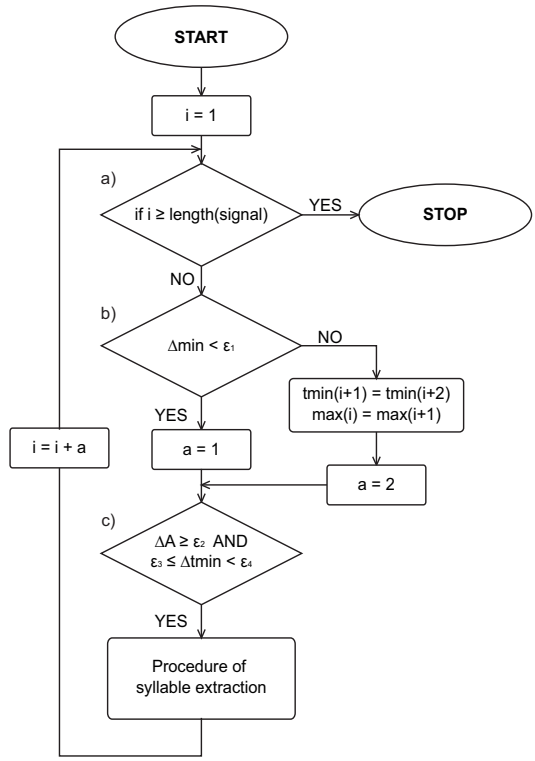

Fig. 4. Division into syllables algorithm

filtration, normalization and wavelet decomposition. The aim of filtration, done by the use of band-pass filter, was to remove higher frequencies.

After filtration data were normalized. The goal of normalization was to eliminate the influence of the amplitude from the further analysis. Different amplitudes may be the result of various conditions during signal registration. In this study signal was normalized to fit $[-1, 1]$ value interval. Unfortunately normalization also decreased distances between classes. However, this was a necessary step, before proceeding to the next stages.

After normalization wavelet analysis was used for signal de-noising. Noise usually comes from recording apparatus, as well as from the environment. The first step of the method is decomposition. After selecting its level L and the type of wavelet functions, signal is divided into L decomposition levels according to the equation:

$$s(t) = \sum_{j=1}^{L} \sum_{k} d_j(k)\psi_j(t) + \sum_{k} c_L(k)\varphi_L(t)$$

where:
$\varphi_L(t)$ is a scaling function of the L-level,
$\psi_j(t)$ for $j = 1, 2...L$ are wavelet functions for the L levels.

B. Division into syllables

The definition of syllable is a problem in phonetics and phonology of human speech. This problem becomes even greater when it comes to birds. Therefore one of the biggest challenge of this study was syllables extraction. Physically, the syllable is defined as a segment which has higher intensity than its neighborhood. In this paper, following considerations are based on the signal time domain.

Division into syllables was divided into three parts. The first part was approximation, which reduced the noise and dimensionality of signal samples. After that local maxima and minima were designated, based on the gradient of signals polynomial approximation.

The syllables were clustered between two neighboring minima and usually had one maximum. However if a time period of a syllable was to small or differences between extrema were to low (what means, that this observation is a part of the same syllable), it was added to the previous syllable (Fig. 4).

Values of factors in this algorithm have great influence on classifiers performance. At the beginning the values of factors were established basing on the observation of the system, and after that, factors were optimized basing on the highest results of accuracy.

All the research and analysis was carried out on isolated syllables. Timing and syllable spectrogram are presented in the Fig. V-B.

After automatic division the features were extracted and clusters have been classified. During classification the methods described in the previous section have been used.

C. Results

The classification accuracy for different features shows that spectral features are the best for ABR task (Table II).

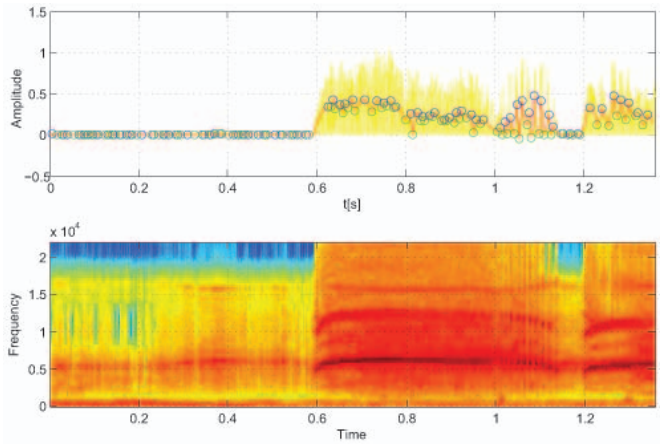

Fig. 5. (Upper Figure)Graph presents approximated signal (red line), as well as selection of minimal and maximal values (in circles). (Bottom Figure) Corresponding spectrogram of process signal.

TABLE II
CLASSIFICATION ACCURACY (CA) FOR SELECTED FEATURES

Feature name	CA(%)
PLP (Perceptual Linear Prediction)	79.89
LPC (Linear Predictive Coding)	74.6
MFCC (Mel Frequency Cepstral Coefficients)	80.42
Histogram	53.43
Formant	46.82
Antyformant	36.77
Bandwidth	26.19
Duration	23.28
FF (Fundamental Frequency)	27.78
PSD (Power Spectrum Density)	23.81
Sum of features	89.95

Table III presents the results of accuracy of classifiers for different birds species. One can observe that some birds species provide high classification accuracy, no matter to which group they were assigned to (Euroasian Nuthach, Tengmalms Owl). This means that the sound of those birds differ significantly from others in a particular group. One can also observe correlation between results accuracy and the ratio of training syllables to test syllables. At this point it is worth to note that a 30% of files were designated to test collection and not 30% of syllables. That is why the amount of bird syllables is different. Of course, if the number of training syllables was greater then test syllables, the accuracy could be higher. This regularity can be seen while comparing Eurasian Nuthatch (530 training syll., 104 test syll.) and Blackbird (522 training syll., 143 test syll).

VI. DISCUSSION

Achieved results are relatively satisfactory. It is difficult to exactly compare different works, because accuracy of classification depends greatly on the type of audio files and compared bird species. It is usually not a problem to identify birds, which sounds differ greatly, the problem is with similar

TABLE III
CLASSIFICATION ACCURACY OF BIRDS SPECIES FOR 2 DIFFERENT CLASSIFIERS

Common name	CA_{kNN} (%)	CA_{SOM} (%)
Great Tit	59.85	43.79
Blackbird	43.36	30.07
Eurasian Nuthatch	70.19	61.54
Robin	31.43	15.00
Thrush Nightingale	57.5	21.25
Great Tit	59.12	37.96
Blackbird	44.06	36.36
Eurasian Nuthatch	73.08	59.62
Grey Partridge	66.12	40.32
Tengmalm's Owl	96.51	89.53
Common Swift	90.36	85.54
Wild Duck	93.85	73.85
Common Cuckoo	85.37	54.88
Grey Partridge	80.65	58.06
Tengmalm's Owl	97.67	86.05

bird sounds (e.g. from the 1st group). Results show that highest accuracy was achieved by the 3rd group, in which sounds of birds species differ emphatically.

There are unfortunately a few disadvantages of the system. First one is connected with the automatic syllables segmentation algorithm - the system has low immunity for various interferences. There may be a problem, when identifying bird sings on the same time with others. This problem can be considerable because birdsong is almost always connected with others (birds answer to each other).

Another problem is connected with values of various parameters in automatic syllables division algorithm. They were assign experimentally. One of the solution could be improving the algorithm by automatic adjustment of values of these coefficients basing on the information about expected group of recorded birds.

VII. CONCLUSIONS AND FUTURE WORK

In this article, the results of birds classification, based on their sounds have been presented. An automatic algorithm for division of bird sounds into syllables has been developed. Classification has been made using strictly selected features and 2 different classifiers. Tests have been made on real environment data sets. Mean accuracy of classification was 69,94 % for kNN and 52,92 % for SOM classifier. The highest accuracy has been achieved using MFCC features. The accuracy of classification depends mainly on the type of data sets, but also on used descriptors and classifiers.

Best results were achieved with kNN classifier. The research also shows that results are correlated with the similarity of the birds sounds. The experiments confirm that high accuracy in fully automated systems for ABR is possible, but not easy to achieve.

Future work will focus on adapting this system for handheld devices like cell phones or palmtops. These actions will be connected with optimalization of the algorithm in terms of speed of calculation and memory usage. Also a procedure, which will contrive with some difficulties mentioned in discussion section, should be developed.

New features, which are not connected to spectral construction of syllables, should be also tested. Descriptors from nonlinear dynamics, like fractal dimension or shapes of attractors can serve as an example of such features. Also additional features, extracted from phrases and songs, which show connections between syllables could be used.

ACKNOWLEDGMENT

This work is partly financially supported by the grant awarded to Artur Gmerek by the Ministry of Science and Higher Education of Poland (Grant No. N N514 469339).

REFERENCES

[1] C.-H. Chou, P.-H. Liu, and B. Cai, "On the studies of syllable segmentation and improving mfccs for automatic birdsong recognition," in *Proc. IEEE Asia-Pacific Services Computing Conf. APSCC '08*, 2008, pp. 745–750.

[2] T. S. Brandes, "Feature vector selection and use with hidden markov models to identify frequency-modulated bioacoustic signals amidst noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1173–1180, 2008.

[3] C.-H. Chou, C.-H. Lee, and H.-W. Ni, "Bird species recognition by comparing the hmms of the syllables," in *Proc. Second Int. Conf. Innovative Computing, Information and Control ICICIC '07*, 2007, p. 143.

[4] E. Vilches, I. A. Escobar, E. E. Vallejo, and C. E. Taylor, "Targeting input data for acoustic bird species recognition using data mining and hmms," in *Proc. Seventh IEEE Int. Conf. Data Mining Workshops ICDM Workshops 2007*, 2007, pp. 513–518.

[5] B. Lakshminarayanan, R. Raich, and X. Fern, "A syllable-level probabilistic framework for bird species identification," in *Proc. Int. Conf. Machine Learning and Applications ICMLA '09*, 2009, pp. 53–59.

[6] A. Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 5, 2003.

[7] C. E.D., "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals," *Applied Acoustics*, vol. 62, pp. 1359–1374, 2001.

[8] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, "Sensor network for the monitoring of ecosystem: Bird species recognition," in *Proc. 3rd Int. Conf. Intelligent Sensors, Sensor Networks and Information ISSNIP 2007*, 2007, pp. 293–298.

[9] S.-A. Selouani, M. Kardouchi, E. Hervet, and D. Roy, "Automatic birdsong recognition based on autoregressive time-delay neural networks," in *Proc. ICSC Congress Computational Intelligence Methods and Applications*, 2005.

[10] P. Somervuo and A. Harma, "Analyzing bird song syllables on the self-organizing map," *Proceedings of the Workshop on Self-Organizing Maps (WSOM '03), Kitakyushu, Japan*, 2003.

[11] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895 – 2907, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/B6V15-49207P2-1/2/ef7fc86445321fbc2e001c70ff7e9f65

[12] F. Briggs, R. Raich, and X. Z. Fern, "Audio classification of bird species: A statistical manifold approach," in *Proc. Ninth IEEE Int. Conf. Data Mining ICDM '09*, 2009, pp. 51–60.

[13] A. L. McIlraith and H. C. Card, "A comparison of backpropagation and statistical classifiers for bird identification," in *Proc. Int Neural Networks,1997. Conf*, vol. 1, 1997, pp. 100–104.

[14] D. Niewiadomy and A. Pelikant, "Implementation of mfcc vector generation in classification context," *JOURNAL OF APPLIED COMPUTER SCIENCE*, 2008.