Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 38637, 8 pages
doi:10.1155/2007/38637

Research Article

Bird Species Recognition Using Support Vector Machines

Seppo Fagerlund

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, P.O. Box 3000, 02015 TKK , Finland

Received 13 November 2006; Revised 20 February 2007; Accepted 31 March 2007

Recommended by Satya Dharanipragada

Automatic identification of bird species by their vocalization is studied in this paper. Bird sounds are represented with two different parametric representations: (i) the mel-cepstrum parameters and (ii) a set of low-level signal parameters, both of which have been found useful for bird species recognition. Recognition is performed in a decision tree with support vector machine (SVM) classifiers at each node that perform classification between two species. Recognition is tested with two sets of bird species whose recognition has been previously tested with alternative methods. Recognition results with the proposed method suggest better or equal performance when compared to existing reference methods.

1. INTRODUCTION

Interest towards automatic recognition of bird species based on their vocalization has increased and many recent studies have been published [1–5]. Bird species identification is a typical pattern recognition problem and most studies include signal preprocessing feature extraction and classification sections. Bird vocalization segmentation into smaller recognition units is performed by hand or automatically. The number of species has ranged between 2 and 16 in previous studies.

The works of Anderson et al. [6] and Kogan and Margoliash [7] were among the first attempts to recognize bird species automatically by their sounds. They applied dynamic time warping and hidden Markov models for automatic song recognition of Zebra Finche (*Taeniopygia guttata*) and Indigo Punting (*Passerina cyanea*). In these studies, syllables were represented by spectrograms and classification was performed by matching the spectrograms to predefined prototypes. Comparison of spectrograms is computationally demanding, and in the case of field recordings, they often also include environmental information that is not relevant to recognition of bird species.

Neural network classifiers were used in [1, 8]. Mcllraith and Card [8] tested recognition of songs of six species common to Manitoba, Canada. In this work, songs were represented by spectral and temporal parameters. The dimensionality of the feature space was reduced by selecting features for classification by means of their discriminative ability. Selouani et al. [1] improved the neural network approach by adding a feedback loop to the multilayer perceptron (MLP) network. They tested classification of sixteen Canadian bird species, whose manually extracted syllables were represented by linear prediction coefficients. Similar to SVM classifiers, the training of artificial neural networks is computationally demanding, but the classification phase is relatively fast for both methods.

Kwan et al. [2] used Gaussian mixture models (GMM) to classify 11 bird species. Bird sounds were represented with mel-frequency cepstral coefficients (MFCC). Kwan et al. also introduced a system for automatic monitoring of birds in field conditions. Tyagi et al. [4] introduced a new representation for bird syllables which was based on the average spectrum over time and classification was based on template matching. Tyagi et al. introduced four reference recognition systems that were based on dynamic time warping and GMM with three different feature representations. Different approaches to bird species recognition were introduced in the work of Vilches et al. [3]. They used data mining techniques for classification and analyses were performed on a pulse-by-pulse basis in contrast to traditional syllable-based systems.

This work was performed within the AveSound project [9]. The objective of this research is to develop a fully automatic system for bird species recognition from their sounds made in field conditions. The system is based on the recognition of syllables that are the building blocks of bird songs and calls [10]. In [11] bird vocalization was modeled using

only one sinusoid while in [12] the harmonic structure was incorporated into the model. In [13] recognition was based on the comparison of syllable histograms. Previous works have studied only birds whose vocalization is mostly tonal or harmonic. However, many birds produce also inharmonic or noise-like sounds [14]. In [15] recognition of species that produce regularly inharmonic sounds were studied. Selin et al. [16] studied species that produce tonal, harmonic, and inharmonic sounds. Different parametric representations of bird syllables were studied in [17]. The main emphasis and focus of this article is in applying support vector machine classifiers to the recognition of bird species and to compare its performance to alternative pattern recognition tools already tested within the AveSound project. Fundamental parts of the recognition system are also revised in this article. Recognition was tested using two different datasets previously used in the AveSound project.

This article is organized as follows. Categories of bird vocalization are introduced in Section 2. Also, a method for segmentation of bird sounds into basic elements of the recognition system is introduced. Section 3 describes parametric representations of bird vocalization while Section 4 introduces the support vector machine classification method and system used for classification in this work. Recognition results with bird data are presented and compared to previous work in Section 5. Finally, Section 6 concludes the work.

2. SEGMENTATION OF BIRD SOUNDS

Bird sounds are typically divided into categories of songs and calls depending upon their function. Generally, songs are longer and more complex than calls and occur more spontaneously. The main function of songs is related to breeding and territorial defense. Many bird species sing only during the breeding season and is generally further limited to males only. Call sounds are typically short vocalizations that carry a function, for example, an alarm, flight, or feeding. Distinguishing between songs and calls can sometimes be ambiguous and hence the separation of bird sounds into these categories is not studied in this work.

Bird sounds can also be divided into hierarchical levels of phrases, syllables, and elements [10]. For example, the levels of a typical song from the Common Chaffinch (*Fringilla coelebs*) are illustrated in Figure 1. A phrase is a series of syllables that occurs in a particular pattern. Usually syllables in a phrase are similar to each other, but sometimes they can also be different as in the last frame of the song presented in Figure 1. Syllables are constructed from elements but in simple cases syllables and elements are one and the same. However, complex syllables may be constructed from several elements. Separation of elements in complex syllables is often difficult and can be ambiguous. Call sounds are usually comprised of one syllable or a series of similar syllables and the phrase level cannot be detected. The phrase level is commonly also missing in the songs of certain species. In this work the syllable is regarded as the smallest unit of bird vocalization.

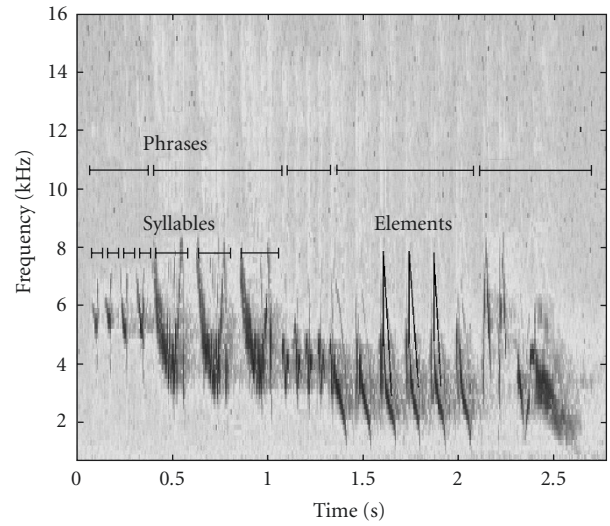

Figure 1: Hierarchical levels of song for the common chaffinch.

(1) Find syllable candidates, that is, regions that are above syllable threshold T_{dB}.

(2) Update N_{dB} from gaps between syllable candidates.

(3) Update the threshold, for example, $T_{\mathrm{dB}} = N_{\mathrm{dB}}/2$ and return to step 1.

Algorithm 1

The segmentation of a recording into individual syllables is performed using an iterative time-domain algorithm [14]. First, a smooth energy envelope of the signal is computed on the decibel scale and the maximum value is set to 0 dB. The global minimum energy is chosen as the initial background noise level estimate N_{dB}. The initial threshold T_{dB} is set to half of the initial noise level, which is itself set to the lowest signal envelope energy level. The noise and threshold levels are updated using Algorithm 1 until convergence is obtained indicating that the noise level is sufficiently stable.

Once the algorithm has converged, syllable candidates that are very close to each other are grouped together in order to prevent a border effect [18]. Also, temporally distinct syllable elements that are detected separately are grouped together. In this work syllable, candidates that are less than 15 milliseconds apart of each other are joined together to become one syllable.

3. FEATURE EXTRACTION OF SYLLABLES

The segmented syllable candidates are represented using two different parametrization methods. The mel-cepstrum model is a common parametrization method used frequently in speech recognition. A second parametrization method employs a set of descriptive signal parameters and is used in many audio classification problems. Descriptive signal parameters include both temporal and spectral features. Both

parametrization methods are presented in the following section in more detail.

3.1. Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients (MFCC) [19] have been a popular signal representation method used in many audio classification tasks, especially in automatic speech recognition (ASR). The basis for the MFCC mel-frequency scale is derived from the human perceptual system. Perceptual systems of birds are not the same as in humans, but exhibit similar characteristics. The calculation of MFCC parameters is efficient and straightforward since they do not involve any tuning parameters.

The calculation of MFCC parameters begins with the segmentation of a signal into overlapping frames. The power spectrum of each frame is transformed into the logarithmic mel-frequency spectrum using a filterbank of 32 triangular filters. The ith MFC-coefficient of each frame is calculated by

$$\mathrm{MFCC}_i = \sum_{k=1}^{K} X_k \cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right], \tag{1}$$

where X_k is the logarithmic energy of the kth mel-spectrum and K is the total number of bands. The discrete cosine transform (DCT) in (1) decreases the dimensionality of the feature vector and decorrelates features as well. In this work a 256 (6 ms) sample frame size was used and adjacent frames overlapped by 50%. Syllables were parameterized using the first 12 MFC-coefficients and the energy term. Also, delta and delta-delta coefficients were calculated to measure temporal change in parameters and delta parameters.

3.2. Descriptive parameters

In many applications in the field of audio signal processing, the specific signal model is unknown and the spectral charasteristics may be quite varied. This is typical especially within the field of animal and natural sounds. In these applications it is common to use many descriptive measures to parametrize sounds, that are derived from both the temporal and spectral domains. In this paper syllables are represented with 11 low-level signal parameters. Seven features are calculated on a frame-to-frame basis providing a short time description of syllables. First, syllables are divided into overlapping frames of 256 samples with 50% overlap. Features are then calculated for each frame and the mean and variance values of the feature trajectories are used as the actual features of the recognition system. Therefore, we have 14 features calculated on a frame basis. Five more features are calculated from the entire syllable duration thus increasing the total number of descriptive parameters to 19. These parameters are listed in Table 1. A detailed description of these features is provided in [14].

Table 1: Descriptive parameters used in this work. An asterisk (∗) in the last column indicates that the feature is calculated on a frame-to-frame basis.

Feature	Abbreviation	Frame feature
Spectral features		
Spectral centroid	mSC, vSC	∗
Signal bandwidth	mBW, vBW	∗
Spectral roll-off frequency	mSRF, vSRF	∗
Spectral flux	mSF, vSF	∗
Spectral flatness	mSFM, vSFM	∗
Frequency range	range1, range2	
Temporal features		
Zero crossing rate	mZCR, vZCR	∗
Short time energy	mEN, vEN	∗
Syllable temporal duration	T	
Modulation spectrum	MSm, MSf	

4. SUPPORT VECTOR MACHINE (SVM) CLASSIFICATION

Support vector machines and other kernel-based methods have become a popular tool in many kinds of machine learning tasks. In audio processing, SVMs have been used, for example, in phonetic segmentation [20], speech recognition [21], and general audio classification [22]. One advantage of SVMs is their accuracy and superior generalization properties they offer when compared to many other types of classifiers. SVMs are based on statistical learning theory and structural risk minimization [23]. In the following section a brief introduction to SVM classification operation is presented when applied to binary and multiclass cases as is done in this work. For a more detailed tutorial covering support vector machines, refer to [24].

4.1. Binary classification

Let $\mathbf{x}_i \in \Re^m$ be a feature vector or a set of input variables and let $y_i \in \{+1, -1\}$ be a corresponding class label, where m is the dimension of the feature vector. In linearly separable cases a separating hyperplane satisfies

$$y_i(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b) \geq 1, \quad i = 1,\ldots,n, \tag{2}$$

where the hyperplane is denoted by a vector of weights $\mathbf{w}$ and a bias term b. The optimal separating hyperplane, when classes have equal loss-functions, maximizes the margin between the hyperplane and the closest samples of classes. The margin is given by

$$\mathrm{d}(\mathbf{w},b) = \min_{\{\mathbf{x}_i, y_i=1\}} \frac{|\langle \mathbf{w}\cdot x_i\rangle + b|}{\|\mathbf{w}\|} + \min_{\{\mathbf{x}_j, y_j=-1\}} \frac{|\langle \mathbf{w}\cdot x_j\rangle + b|}{\|\mathbf{w}\|} \tag{3}$$

$$= \frac{2}{\|\mathbf{w}\|}. \tag{4}$$

The optimal separating hyperplane can now be solved by maximizing (4) subject to (2). The solution can be found

using the method of Lagrange multipliers. The objective is now to minimize the Lagrangian

$$L_p(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) + \sum_{i=1}^{l} \alpha_i, \quad (5)$$

and requires that the partial derivatives of $\mathbf{w}$ and b be zero. In (5), α_i are nonnegative Lagrange multipliers. Partial derivatives propagate to constraints $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$ and $\sum_i \alpha_i y_i = 0$. Substituting $\mathbf{w}$ into (5) gives the dual form

$$L_d(\mathbf{w}, b, \alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle, \quad (6)$$

which is not anymore an explicit function of $\mathbf{w}$ or b. The optimal hyperplane can be found by maximizing (6) subject to $\sum_i \alpha_i y_i = 0$ and all Lagrange multipliers are nonnegative.

However, in most real world situations classes are not linearly separable and it is not possible to find a linear hyperplane that would satisfy (2) for all $i = 1, \ldots, n$. In these cases a classification problem can be made linearly separable by using a nonlinear mapping into the feature space where classes are linearly separable. The condition for perfect classification can now be written as

$$y_i (\langle \mathbf{w} \cdot \Phi(\mathbf{x}_i) \rangle + b) \geq 1, \quad i = 1, \ldots, n, \quad (7)$$

where Φ is the mapping into the feature space. Note that the feature mapping may change the dimension of the feature vector. The problem now is how to find a suitable mapping Φ to the space where classes are linearly separable. It turns out that it is not required to know the mapping explicitly as can be seen by writing (7) in the dual form

$$y_i \left(\sum_{j=1}^{l} \alpha_j y_j \langle \Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i) \rangle \right) + b \geq 1, \quad i = 1, \ldots, n, \quad (8)$$

and replacing the inner product in (8) with a suitable kernel-function $K(\mathbf{x}_j, \mathbf{x}_i) = \langle \Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i) \rangle$. This form arises from the same procedure as was done in the linearly separable case, that is, writing the Lagrangian of (7), solving partial derivatives, and substituting them back into the Lagrangian. Using a kernel trick, we can remove the explicit calculation of the mapping Φ and need to only solve the Lagrangian (6) in dual form, where the inner product $\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$ has been transposed with the kernel function in nonlinearly separable cases. In the solution of the Lagrangian, all data points with nonzero (and nonnegative) Lagrange multipliers are called support vectors (SV).

Often the hyperplane that separates the training data perfectly would be very complex and would not generalize well to external data since data generally includes some noise and outliers. Therefore, we should allow some violation in (2) and (7). This is done with the nonnegative slack variable ζ_i:

$$y_i (\langle \mathbf{w} \cdot \Phi(\mathbf{x}_i) \rangle + b) \geq 1 - \zeta_i, \quad i = 1, \ldots, n. \quad (9)$$

The slack variable is adjusted by the regularization constant C, which determines the tradeoff between complexity and the generalization properties of the classifier. This limits the Lagrange multipliers in the dual objective function (6) to the range $0 \leq \alpha_i \leq C$.

Any function that is derived from mappings to the feature space satisfies the conditions for the kernel function. However, this approach requires the design of a suitable feature map and it also restricts the number of possible kernel functions. A more common approach is to find functions that fulfill the characterization of a kernel function. A symmetric function in the input space is a kernel function if a kernel matrix $\mathbf{K} = [K(\mathbf{x}_j, \mathbf{x}_i)]_{i,j=1}^{n}$ is positive semidefinite, that is, its eigenvalues are nonnegative. Probably the most commonly used kernel function is the Gaussian

$$K(\mathbf{x}_j, \mathbf{x}_i) = \exp\left(- \frac{||\mathbf{x}_j - \mathbf{x}_i||^2}{2\sigma^2} \right). \quad (10)$$

The Gaussian kernel function is translation invariant and it generalizes well for different shape classes in the feature space. Also, the Gaussian kernel has only one tuning parameter σ which adjusts the kernel's width.

4.2. Multiclass classification

The above discussion only covers the binary classification case, which is insufficient for our situation. There are several ways to construct SVM classifiers for more than two classes. Methods can be divided into submethods that use only one decision function, or into methods that solve many binary problems, the latter being more common. Furthermore, methods comprising multiple binary classifiers can be constructed in many ways. In [25] a good review of different methods is presented.

In this work, we use a binary decision tree that consists of binary SVM classifiers at each node [26]. Each classifier performs classification between two classes ignoring all other classes. At each layer of the decision tree one class is rejected. Finally, at the bottom, the last remaining class is considered as the winning class. Figure 2 indicates the topology of the SVM decision tree classifier for the species listed in table 2.

Using the standard method, the classifiers in the nodes of the decision tree have identical model parameters. However, this may lead to a nonoptimal binary classifier for some nodes, especially when the classes are not equally spaced in the feature space, as is the case with this problem. In this paper, customized classifiers for each node of the decision tree are used. Each node contains a binary SVM classifier with a Gaussian kernel function where the regularization constant and width of the Gaussian kernel are different for each classifier.

4.3. Training SVMs

Construction of SVM classifiers includes two phases. The first phase requires finding optimal model parameters, that is, the regularization constant C and the width of the Gaussian kernel σ. Actual training of the classifier is performed

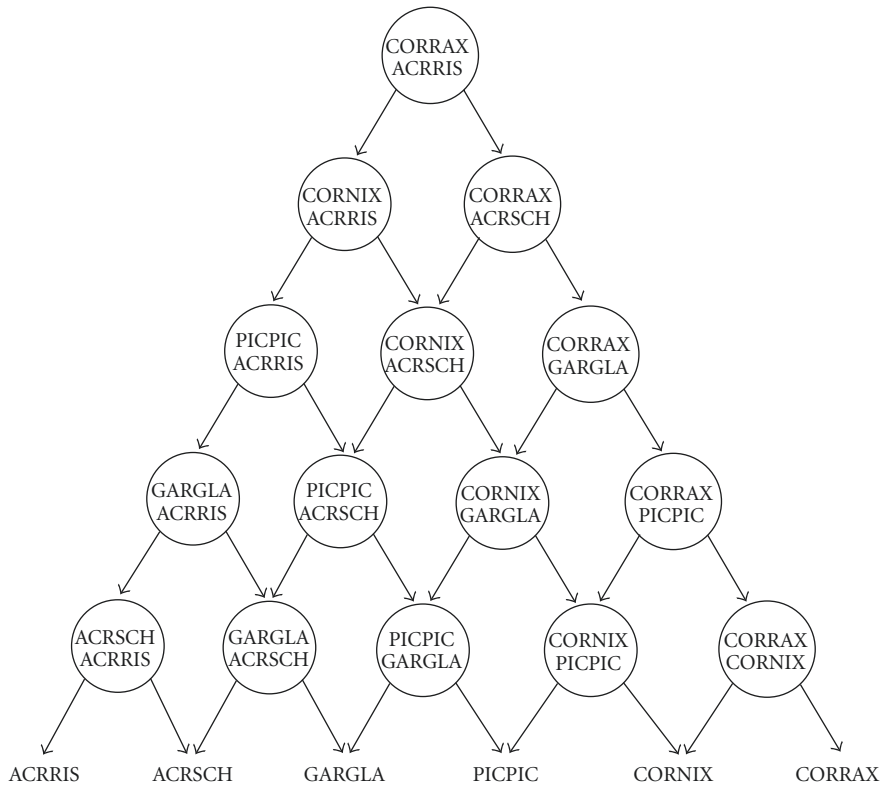

Figure 2: Topology of the decision tree classifier.

during the second phase. These two phases are repeated separately for each pair of classes in the decision tree.

N-fold cross validation is used to find the optimal values for the model parameters. In this work, N depends on the number of individuals within species for dataset 1 (Table 2). For all pairs of classes in the decision tree, the data points are divided into the training and test subsets such that the test subset contains all data vectors from one individual. The training subset is used to construct an SVM classifier and its performance is evaluated with a test subset. The classification error is the average of the test errors of the subsets. For dataset 2 (Table 3) a 10-fold cross validation in training data was used to select optimal model parameters. The validation procedure is repeated for a grid of parameter values C and σ. Parameters that produce the lowest classification error are selected as the final model parameters. Limits for the parameter values are chosen such that they contain extreme values at all ends of the scale and the resolution of values is suitable.

Actual training of SVM classifiers is performed using the sequential minimal optimization (SMO) algorithm [27]. The MATLAB support vector machine toolbox [28] implementation of the SMO algorithm was used to train individual SVM classifiers. The SMO algorithm decomposes the original large-scale optimization problem into several smaller problems that can be solved analytically. The SMO algorithm solves the Lagrangian for two vectors at each iteration. The vectors are selected from the set of vectors that violates the optimality condition.

Table 2: 1st set of bird species used for recognition in this work. The last column indicates the total number of syllables.

Lat. Abbr.	Common name	Individuals	Syllables
CORRAX	Common Raven	7	91
CORNIX	Hooded Crow	8	160
PICPIC	Magpie	7	312
GARGLA	Eurasian Jay	9	99
ACRSCH	Sedge Warbler	6	331
ACRRIS	Marsh Warbler	8	277

Table 3: 2nd set of bird species studied in this work. The last two columns indicate the number of syllables in training and testing datasets, respectively.

Lat. Abbr.	Common name	Syllables train	Syllables test
ANAPLA	Mallard	138	60
ANSANS	Greylag Goose	135	59
COTCOT	Quail	190	83
CRECRE	Corncrake	443	110
GLAPAS	Pygmy Owl	113	48
LOCFLU	River Warbler	890	328
PICPIC	Magpie	203	97
PORPOR	Spotted Crake	166	69

Table 4: Recognition results for datasets 1 and 2 (upper and lower panel, resp.). Values indicate the percentage of correctly classified syllables for each species using different parametric representations.

species	comp	MFCC	MFCC Δ	MFCC Δ Δ	mixture	reference
CORRAX	89	95	89	92	95	92
CORNIX	76	87	84	88	89	66
PICPIC	85	82	84	87	91	63
GARGLA	89	83	84	81	92	80
ACRSCH	64	73	85	82	86	57
ACRRIS	75	88	92	90	92	86
overall	79	85	88	87	91	74

species	comp	MFCC	MFCC Δ	MFCC Δ Δ	mixture	reference
ANAPLA	93	98	98	98	100	98
ANSANS	76	75	90	90	85	83
COTCOT	100	96	96	96	100	100
CRECRE	100	100	100	100	99	96
GLAPAS	75	100	100	100	90	96
LOCFLU	100	100	100	100	100	100
PICPIC	98	87	87	87	96	94
PORPOR	100	100	100	100	100	100
overall	96	96	97	97	98	96

5. RESULTS

Recognition performance was tested with datasets used in [15, 16]. Species in dataset 1 are listed in Table 2. Recognition was tested separately for each individual by arranging the test so that syllables in the testing dataset were not used during the training phase. The recognition results indicate the percentage of correctly classified syllables. Information regarding dataset 2 is described in Table 3. In this dataset, manually segmented syllables were distributed into training and testing subsets. Syllables from single individuals were part of either datasets but not both, thus recognition was also individually independent for the second dataset.

Recognition results for dataset 1 (Table 2) are shown in the upper panel in Table 4. Columns indicate recognition results with a different parametric representation. A mixture model includes all MFC-coefficients (including delta and delta-delta coefficients) as well as descriptive parameters. The reference produces the best recognition performance as obtained in [15], where MFCC parameters were used for syllable representation and nearest-neighbor classification with the Mahalanobis distance measure used for recognition. The best recognition results were obtained using a mixture model, but the feature vector dimension was also the highest with this representation.

Results for dataset 2 are shown in the lower panel of Table 4. The reference results are from [16] where syllables were represented with four parameters derived from a wavelet decomposed signal representation and where neural networks were used for classification. Results show only a slight difference in performance between different parametric representations. Compared to the reference method, the SVM classifier performs equally well when compared to other parametric representations. Also, in this dataset the best overall recognition result was obtained with a mixture model.

6. CONCLUSIONS

In this paper, support vector machine classification methods were applied to automatic recognition of bird species. Recognition was tested with two datasets previously used in this project in order to obtain references for the new methods. Results suggest that equal or better performance, compared to the reference methods, was achieved. However, recognition results for two datasets cannot be directly compared since dataset 2 includes more species with a larger spectrum of different sounds than dataset 1. Species in the dataset 1 are also more closely related when compared to the species in dataset 2.

In the proposed method the decision tree topology is invariant to the ordering of the species (classes) and the same result would have been arrived at by changing the ordering of the species in the tree. This topology is efficient and straightforward to construct and it does not require any additional information regarding the relations between different species. However, a hierarchical topology that utilizes the relationships of the sound between different species could lead to a more robust and computationally efficient classifier.

In the proposed method all syllables are represented with the same parameters. However, the decision tree topology in the classifier enables the use of weighting of features in each subproblem separately. For example, when weighting is not used, in dataset 2 the recognition results for the Pygmy Owl (GLAPAS) (lower panel in Table 4, row 5) using the descriptive parameter model is 75% while using MFCC-models

100% accuracy is achieved. The method thus produces a lower recognition result (90%) in the mixture model when compared to the MFCC-models. Future work will investigate the use of feature weighting, for example, its use would have produced 100% accuracy in the case of the mixture model.

ACKNOWLEDGMENT

This work is supported by the Academy of Finland under research Grant 206652 (The AveSound project).

REFERENCES

[1] S.-A. Selouani, M. Kardouchi, E. Hervet, and D. Roy, "Automatic birdsong recognition based on autoregressive time-delay neural networks," in *Congress on Computational Intelligence Methods and Applications (CIMA '05)*, pp. 1–6, Istanbul, Turkey, December 2005.

[2] C. Kwan, K. C. Ho, G. Mei, et al., "An automated acoustic system to monitor and classify birds," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 96706, 19 pages, 2006.

[3] E. Vilches, I. A. Escobar, E. E. Vallejo, and C. E. Taylor, "Data mining applied to acoustic bird species recognition," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 3, pp. 400–403, Hong Kong, August 2006.

[4] H. Tyagi, R. M. Hegde, H. A. Murthy, and A. Prabhakar, "Automatic identification of bird calls using spectral ensemble average voiceprints," in *Proceedings of the 13th European Signal Processing Conference (EUSIPCO '06)*, Florence, Italy, September 2006.

[5] E. J. S. Fox, J. D. Roberts, and M. Bennamoun, "Text-independent speaker identification in birds," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '06)*, Pittsburgh, Pa, USA, September 2006.

[6] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *The Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 1209–1219, 1996.

[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," *The Journal of the Acoustical Society of America*, vol. 103, no. 4, pp. 2185–2196, 1998.

[8] A. L. Mcllraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2740–2748, 1997.

[9] S. Fagerlund, "Avesound project web-site," 2006, http://www.acoustics.hut.fi/research/avesound/avesound.html.

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

[11] A. Härmä, "Automatic identification of bird species based on sinusoidal modelling of syllables," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 5, pp. 545–548, Hong Kong, April 2003.

[12] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 5, pp. 701–704, Montreal, Canada, May 2004.

[13] P. Somervuo and A. Härmä, "Bird song recognition based on syllable pair histograms," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 5, pp. 825–828, Montreal, Canada, May 2004.

[14] S. Fagerlund, "Automatic recognition of bird species by their sounds," M.S. thesis, Helsinki University of Technology, Espoo, Finland, 2004.

[15] S. Fagerlund and A. Härmä, "Parametrization of inharmonic bird sounds for automatic recognition," in *Proceedings of the 13th European Signal Processing Conference (EUSIPCO '05)*, Antalya, Turkey, September 2005.

[16] A. Selin, J. Turunen, and J. T. Tanttu, "Wavelets in recognition of bird sounds," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 51806, 9 pages, 2007.

[17] P. Somervuo, A. Härmä, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2252–2263, 2006.

[18] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 533–544, 2001.

[19] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[20] A. Juneja and C. Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '03)*, vol. 1, pp. 675–679, Portland, Ore, USA, July 2003.

[21] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2348–2355, 2004.

[22] C.-C. Lin, S.-H. Chen, T.-K. Truong, and Y. Chang, "Audio classification and categorization based on wavelets and support vector machine," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, part 1, pp. 644–651, 2005.

[23] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[24] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[25] F. Schwenker, "Hierarchical support vector machines for multi-class pattern recognition," in *Proceedings of the 4th International Conference on Knowledge-Based Intellingent Engineering Systems and Allied Technologies (KES '00)*, vol. 2, pp. 561–565, Brighton, UK, August-September 2000.

[26] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin dags for multiclass classification," in *Advances in Neural Information Processing Systems 12*, pp. 547–553, MIT Press, Cambridge, Mass, USA, 2000.

[27] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges, and A. J. Smola, Eds., chapter 12, pp. 185–208, MIT Press, Cambridge, Mass, USA, 1999.

[28] G. C. Cawley, "MATLAB support vector machine toolbox (v0.55β)," School of Information Systems, University of East Anglia, Norwich, Norfolk, UK. NR4 7TJ, 2000, http://theoval.sys.uea.ac.uk/~gcc/svm/toolbox/.

Seppo Fagerlund was born in Pori, Finland, in 1978. He received the M.S. degree in electrical engineering from the Helsinki University of Technology (TKK), Espoo, Finland, in 2004. In 2002, he worked as a Research Assistant in Nokia Research Center. In 2004, he became a Research Assistant and in 2005 a Researcher at the Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology (TKK). His research intrests include signal processing of bioacoustic signals and pattern recognition.