Parametric Representations of Bird Sounds for Automatic Species Recognition

Panu Somervuo, Aki Härmä, and Seppo Fagerlund

***Abstract*—This paper is related to the development of signal processing techniques for automatic recognition of bird species. Three different parametric representations are compared. The first representation is based on sinusoidal modeling which has been earlier found useful for highly tonal bird sounds. Mel-cepstrum parameters are used since they have been found very useful in the parallel problem of speech recognition. Finally, a vector of various descriptive features is tested because such models are popular in audio classification applications, and bird song is almost like music. We briefly introduce the methods and evaluate their performance in the classification and recognition of both individual syllables and song fragments of 14 common North-European Passerine bird species.**

***Index Terms*—Bird song, dynamic time warping (DTW), feature extraction, Gaussian mixture model (GMM), hidden Markov model (HMM), sinusoidal modeling.**

I. Introduction

In SPEECH recognition, one may assume a source model and expect that the signal obeys the laws of a specific spoken language with a vocabulary and a grammar. In other than speech signals, such as music or environmental sounds, this is not always clear. However, bird vocalization is a good example of a class of natural sounds where we can expect to find a vocabulary and other structural elements. Bird song can be often seen as an organized sequence of brief sounds from a species-specific vocabulary. Those brief sounds are usually called elements or syllables [1].

Technical analysis of bird sounds has a long history. This field was revolutionized by the introduction of spectrogram, or sonogram, in the 1950s. Thorpe's book [2] on bird sounds can be considered as a parallel to the classic book on spectrographic speech analysis by Potter *et al.* [3]. A majority of studies on the analysis of bird vocalization has been, and is still, based on visual inspection of spectrograms. The analysis and stylization of spectrograms has developed to a sophisticated art but it is time-consuming and suitable only for relatively small sets of data. It has been predicted that wide-scale application of automatic pattern recognition techniques to bird vocalization research could have similar effects as the introduction of the sonogram earlier [1].

Pattern recognition techniques have been used in most studies to find specific predefined sounds from recordings. Various types of sound-specific parameter sets have been used by many researchers to study geographical variation [4], social ranking in a population of one species of birds [5], imitation [6], or other specific issues [7], [8]. Different parametric representations for bird sounds have been proposed, e.g., in [9] and [10].

Kogan *et al.* [11] introduced a system for the recognition of sounds of two species. The method was based on template matching of spectrograms. A set of spectrogram templates was defined by hand [12], and the recognition was then based on matching the templates with the spectrogram frames computed from a continuous recording. In many practical cases, there are limited possibilities to select *templates* or prototype sounds by hand. Therefore, it is highly preferable to use methods which allow the use of large amount of training data and provide automatic ways to estimate the parameters of the classifier. Kogan *et al.* compared the manually selected template method to a traditional speech recognition system based on mel-frequency cepstrum coefficients (MFCC) and hidden Markov models (HMMs). In their study, HMM was found more robust to changes in background noise. The HMM with mel-cepstrum features is one of the techniques studied also in the current article. It has also been used recently in [13] to develop a bird recognition system for preventing collisions of birds to planes at airport.

McIlraith and Card [14] were among the first to apply automatic classification to a larger number of bird species. They studied the recognition of *songs* from six song bird species. Two methods were proposed which were both based on parameterization of the short-time spectrum of the signal and a feed-forward neural network with back-propagation training. In the first method, the spectral coefficients were computed over a song with a regular 46-ms framing. In the second method, the signal was decomposed into *elements* and silent segments. There, each vector of spectrum parameters represented the spectrum of an element. Temporal parameters characterized the average durations of silences, elements, and the song, but not the temporal order of the elements.

In [15], we studied the automatic recognition of fourteen bird species common in Northern Europe. The working hypothesis was that it would be possible to recognize bird species directly from *syllables* (or elements), which are building blocks of bird song [1]. Typically, the duration of a syllable ranges from few

Manuscript received February 2, 2005; revised November 1, 2005. The work of P. Somervuo was supported by the Academy of Finland Project 44886 (Finnish Centre of Excellence Programme 2000–2005). The work of S. Fagerlund was supported by the Academy of Finland (The AveSound Project). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Schuller.

P. Somervuo was with the Neural Networks Research Centre, Helsinki University of Technology, 02150 Espoo, Finland. He is now with the University of Helsinki, 00790 Heksinki, Finland.

A. Härmä is with the Philips Research, 5656 AA Eindhoven, The Netherlands.

S. Fagerlund is with the Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, 02150 Espoo, Finland.

Digital Object Identifier 10.1109/TASL.2006.872624

to few hundred milliseconds. If this were possible, the recognition of species could then be performed even in a noisy environment using brief clean periods. The alternative approach of recognizing song melodies is more challenging. The main problem is that in a typical habitat there may be several birds singing simultaneously. Also, the high regional variability in the songs of many species and imitation of the songs of other species makes it difficult to define characteristic song patterns for each species.

In [15] the parameterization of bird sounds was based on sinusoidal modeling of syllables. Recognition results were encouraging even though the signal model was clearly oversimplified: each syllable was represented by the frequency and amplitude trajectory of a single time-varying sinusoid. It was also shown that a time-varying frequency model is significantly better in recognition of syllables than just a center frequency. In [16], four additional parameters were introduced which characterize the harmonic structure of the syllable. This model is revised in the current article. The work on pairs of stylized syllables in [17] demonstrated a method to use information about consecutive syllables to improve recognition accuracy. In the current paper, the main focus is in the comparison of different parametric representations.

This paper is organized as follows. The problem of decomposing the bird song into recognizable units is discussed in Section II. A method for performing the segmentation of a continuous environmental recording is also introduced. In Section III, three parametric models are introduced and compared, and in Section IV, we review the properties of the classification methods used in this study. In Section V, we provide the main results of the recognition experiments. In addition to the recognition of individual syllables of bird vocalization, we also construct models for sequences of syllables. Finally, after the discussion in Section VI, the conclusions from the current work are summarized in Section VII.

II. Segmentation of Bird Song

Bird vocalization is usually considered to be composed of calls and songs. Calls are most commonly brief isolated sounds which are usually associated with a specific communicative function, e.g., they may represent a warning for an approaching predator. Songs are more complicated patterns of vocalization which are most commonly associated with territorial singing of male birds and mating. In this study, no distinction is made between calls and songs. Bird vocalizations are often divided into hierarchical levels of phrases, syllables, and elements [1]. For example, the levels of a song of the Common Chaffinch (*Fringilla coelebs*) are illustrated in Fig. 1. A phrase is a series of syllables that occur in a particular pattern. Usually, syllables in a phrase are similar to each other, but sometimes they can be also different as in the last frame of the song presented in Fig. 1. Syllables are constructed of elements. In simple cases, syllables are equal to elements, but complex syllables may be constructed from several elements. Separation of elements is often difficult and can be ambiguous. Call sounds are usually composed of only one syllable and the phrase level cannot be detected. The phrase level is also commonly missing in songs of certain species.

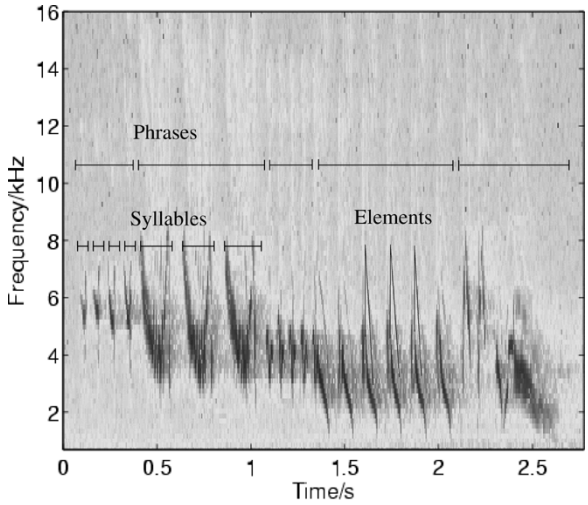

Fig. 1. Hierarchical levels of song of the Common Chaffinch.

In this paper, we call the smallest unit a syllable. A syllable is basically a sound that a bird produces with a single blow of air from the lungs. This is also somewhat inaccurate definition as many birds are capable of complicated circular breathing cycles during singing [18]. The rate of events in bird vocalization may also be so high that the separation of individual syllables is difficult to perform in a natural environment due to reverberation.

The segmentation of a recording to individual syllables is performed using an iterative time-domain algorithm [19]. First, a smooth energy envelope of the signal is computed and the global minimum energy is selected as the initial background noise level estimate N_{dB}. Initial threshold T_{dB} is set to the half of the initial noise level, which is set to the lowest signal envelope energy. Noise level and threshold are updated using the following algorithm until convergence so that the noise level is sufficiently stable.

Algorithm 1

1) Find syllable candidates, i.e., regions that are above syllable threshold T_{dB}
2) Update N_{dB} from gaps between syllable candidates.
3) Update the threshold, e.g., $T_{\mathrm{dB}} = N_{\mathrm{dB}}/2$ and return to step 1.

Once the algorithm has converged, syllable candidates that are very close to each other are grouped together in order to prevent border effect [20]. In this paper, syllable candidates which are less than 15 ms apart from each other are connected.

Segmentation efficiency was evaluated by selecting randomly 50 songs from each of the species used in this study. These songs were manually segmented, and the results were compared to the automatic segmentation. The manual segmentation was performed by visual inspection of the waveform and the automatic segmentation was done using the proposed algorithm. Automatic segmentation found a total of 959 syllables, and manual segmentation found 1068 syllables. One-hundred fifty-seven syllables were not found by automatic segmentation

and 26 syllables which contained temporally distinct pulses were detected as multiple syllables. Otherwise, wrongly detected syllables were 40, thus 93% of syllables detected by automatic segmentation were correctly segmented. Hit rate for syllable detection was 90%. There is, however, high variability in the segmentation accuracy between different species. For example, syllables of songs of the Great Tit (*Parus major*) are almost all correctly segmented, whereas the segmentation of Blackcap (*Sylvia atricapilla*) contains lots of errors.

III. Parametric Modeling of Elements

The segmented regions of bird song are parameterized using the three models introduced below. The sinusoidal model is specially designed for highly tonal time-varying sounds. One of the benefits of the sinusoidal model is that it allows the synthesis of the original signal easily. However, it is clear that a large number of sounds are not purely tonal and, therefore, other representations may be also worth consideration. In this paper, the comparison is made with the mel-cepstrum model which is commonly used in speech recognition and a parametric signal representation which is based on a set of descriptive features.

A. Sinusoidal Model

In sinusoidal modeling, a signal is represented by parameters corresponding to a set of time-varying sinusoidal components. The algorithm that is used in the current article was introduced in detail in [16]. Segmented sounds are modeled using a parametric line spectrum estimation method which is often called analysis-by-synthesis/overlap-add (ABS/OLA) when referring to an efficient frequency-domain algorithm proposed by George and Smith [21]. The analysis is performed in the frequency domain in frames. In this paper, we use a 256-sample Hanning window with 50% overlap between consecutive frames. The fast Fourier transform (FFT) size of 1024 samples is obtained using zero-padding. The sampling rate of the data was 44.1 kHz.

The ABS/OLA algorithm runs so that in the frame n, the frequency of the maximum, ω_n is found by picking the highest peak of the FFT spectrum. In the current case, only one largest frequency component per frame is selected. Then, the frequency domain algorithm proposed in [21] is used to find the phase ϕ_n and magnitude m_n term corresponding to a sinusoidal pulse which is optimal in the minimum mean square error sense. The details of this procedure are given in [21]. The operation applied to each signal frame can be expressed as a function call

$$[m_n, \phi_n, \hat{s}_n] = \text{absola}(\hat{x}_n, \omega_n) \tag{1}$$

where $\hat{x}_n$ is a windowed signal segment corresponding to the nth frame of the original signal x, and $\hat{s}_n$ is a sinusoidal signal which can be used to synthesize a sinusoidal representation of the signal, denoted s_I, in the overlap-add sense. The modeling error in a frame can be computed directly as $e_I = x - s_I$.

The estimation of the parameters of the dominant sinusoid is followed by the analysis of the harmonic structure of the signal. The goal is to divide the sounds into four idealized classes by the properties of their harmonic structure. Class I representation is the original single-sinusoid model. For example, a syllable from the Willow Warbler (*Phylloscopus trochilus*) illustrated in Fig. 2, is a good example of a pure sinusoidal syllable. In Class II representation, the single sinusoid is a fundamental of a harmonic series. For example, the top left spectrogram of Fig. 3 representing a syllable from Blackbird (*Turdus merula*) has the first and the second harmonic of the estimated sinusoidal component clearly visible.

Fig. 4 represents a typical Class III syllable from the Icterine Warbler (*Hippolais icterina*). In this class, the fundamental component is weak, and the sinusoidal component with the highest amplitude is the first harmonic of the series. In Fig. 5, from the March Warbler (*Acrocephalus palustris*), the sinusoid with the highest amplitude is the second harmonic of the series which is characteristic for Class IV syllables.

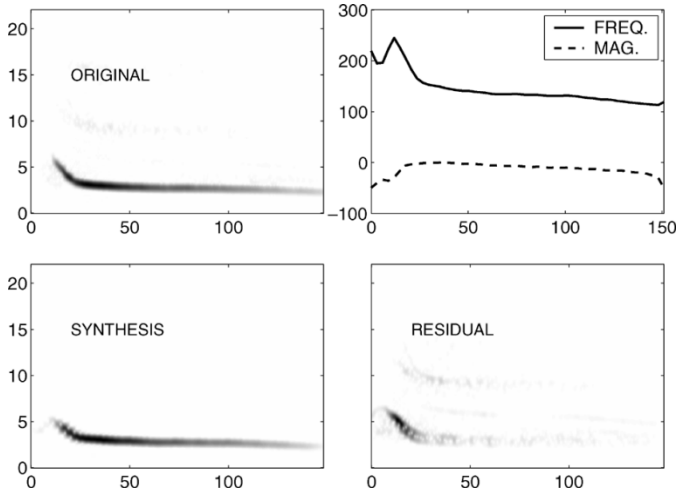

Fig. 2. Top left panel shows a spectrogram of a typical syllable from the Willow Warbler. Top right panel shows frequency and amplitude trajectories of the one-sinusoid model (in FFT-bins and decibels). The two lower panels show spectrograms of a synthesized signal and the residual after subtracting the sinusoid from the original signal. The y-axis represents frequency in kilohertz and the x-axis is time in milliseconds.

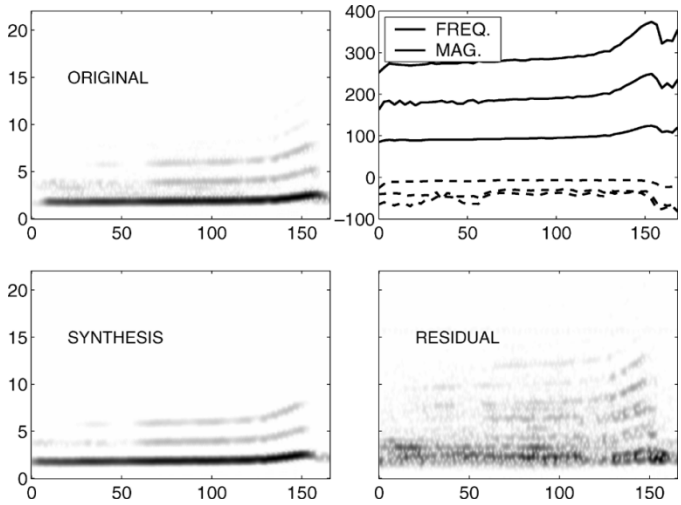

Fig. 3. Representations of a syllable from Blackbird with two harmonic components of the fundamental sinusoidal. Panels as in Fig. 2.

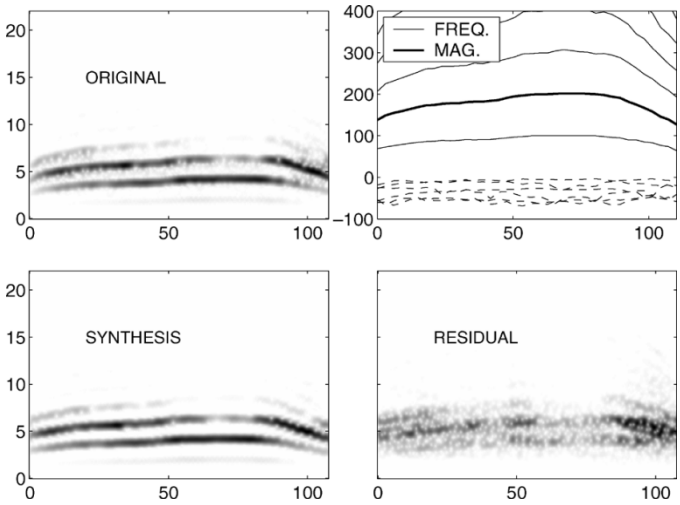

Fig. 4. Representations of a syllable from the Icterine Warbler where the strongest sinusoidal is actually the first harmonic. Panels as in Fig. 2.

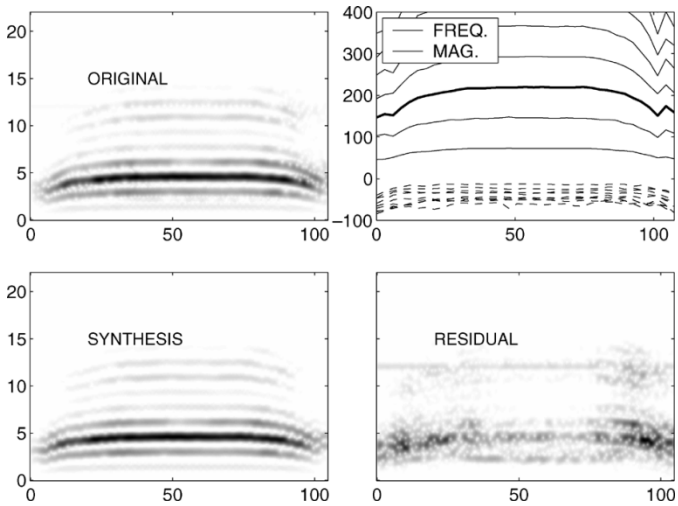

Fig. 5. Representations of a syllable from the Marsh Warbler where the strongest sinusoidal component is the second harmonic of a fundamental. Panels as in Fig. 2.

The frequency estimates corresponding to the dominant sinusoid ω_n are used directly to estimate the level of its kth harmonic. Using the notation from (1), estimation of parameters for a harmonically related component of ω_n is performed by

$$[m_{nk}, \phi_{nk}, \hat{s}_{nk}] = \text{absola}(\hat{e}_n, k\omega_n). \tag{2}$$

For example, frequency curves in Fig. 3 corresponding to Class II were computed using values $k = k_{\text{II}} = 1{,}2{,}3$. In Class III, $k = k'_{\text{III}} = 1/2, 1, 3/2, \ldots$, and finally, Class IV is represented by a harmonic series formed by $k = k'_{\text{IV}} = 1/3, 2/3, 1, 4/3, \ldots$. Synthetic signals s_C whose spectrograms are illustrated in bottom left panels of Figs. 2–5 were created using the overlap-add synthesis of the sum of corresponding synthesized components $\hat{s}_{nC}$, where C denotes a class. Moreover, the residual signals in bottom right panels were given by $e_C = x - s_C$.

TABLE I
HARMONIC PARAMETERS CORRESPONDING TO SYLLABLES IN FIGS. 2–5

Sample	H_{I}	H_{II}	H_{III}	H_{IV}	R
Fig. 2	**0.933**	0.140	0.324	0.43	0.53
Fig. 3	0.226	**0.472**	0.042	0.27	7.02
Fig. 4	0.000	0.005	**0.736**	0.0240	22.66
Fig. 5	0.002	0.001	0.006	**0.951**	18.10

Next, we note that k'_{III} and k'_{IV} intersect with k_{II}. Therefore, in the following, we define Class III and IV harmonic series k_{III} and k_{IV} so that multipliers 2, 3, and 4 have been removed from the sets. In addition, a new group A of harmonics is defined which is obtained as an union of all other classes. That is, $k_{\text{A}} = k_{\text{II}} \cup k_{\text{III}} \cup k_{\text{IV}}$.

The modeling gain corresponding to a class C can be computed in the following way:

$$G_{\text{C}} = 20 \log_{10} \left(\frac{E[x^2]}{E\left[e_C^2\right]} \right) \tag{3}$$

where $E[.]$ denotes expectation. For ABS/OLA, it holds that for any signal $G_{\text{I}} < G_{\text{A}}$, and all other class estimates will fall in $[G_{\text{I}}, G_{\text{A}}]$. Therefore, we define a range measure $R = G_{\text{A}} - G_{\text{I}}$ which gives the difference in modeling gain between the cases where only one sinusoid has been modeled (Class I) and where the dominant sinusoid and all its harmonics and subharmonics have been modeled. Finally, a test can be defined which gives a likelihood H_{C} that a certain syllable is a member of Class $\text{C} = \{\text{II}, \text{III}, \text{IV}\}$

$$H_{\text{C}} = \frac{(G_{\text{C}} - G_{\text{I}})}{R}. \tag{4}$$

This is a value between [0,1]. Likelihood values for the sounds in Figs. 2–5 are given in Table I. These four selected sounds fall well into the four classes. However, the harmonic classification is clearly idealization which does not always work that well. For example, the difference between Classes III and IV is often unclear as can be seen from Fig. 4. The existence of pure sinusoidal sounds of the class I is also debatable as most bird sounds have a harmonic structure. The classification of all data into four harmonic classes is shown in Table I.

The sinusoidal modeling algorithm gives a sequence of $[\omega_n, m_n, \phi_n]$-triplets and the four harmonic class likelihood values. This results in a large number of parameters, and we have considered some approaches to reduce the amount of data. First, the phase function ϕ_n was removed as the same information is present in the temporal evolution of the frequency parameter. Then, the frequency ω_n and amplitude trajectories m_n were approximated by a small number of cosine functions. In practice, this was performed applying the discrete cosine transform (DCT) to the parameter sequence. When computing the transform, the lengths of the DCT base vectors were set equal to the length of the given syllable parameter sequence. The comparison between different syllables was then done using the Euclidean distance between the fixed-dimensional DCT projection vectors. The stretch of the DCT base vectors

TABLE II
BIRD SPECIES USED IN THE STUDY AND PERCENTAGE OF THE SYLLABLES BELONGING TO FOUR HARMONICITY CLASSES

Lat. Abbr.	Common name	Latin name	individuals	songs	syllables	I	II	III	IV
ACRSCH	Sedge Warbler	*Acrocephalus schoenobaenus*	6	20	510	18	1	4	78
FICHYP	Pied Flycatcher	*Ficedula hypoleuca*	8	38	364	45	1	0	54
FRICOE	Common Chaffinch	*Fringilla coelebs*	13	51	567	58	0	2	40
PARATE	Coal Tit	*Parus ater*	9	42	580	64	10	3	23
PARMAJ	Great Tit	*Parus major*	12	75	1089	70	5	1	24
PHOPHO	Common Redstart	*Phoenicurus phoenicurus*	8	49	562	68	3	3	27
PHYBOR	Arctic Warbler	*Phylloscopus borealis*	6	85	762	52	0	0	48
PHYCOL	Common Chiffchaff	*Phylloscopus collybita*	14	67	1104	79	3	1	17
PHYDES	Greenish Warbler	*Phylloscopus trochiloides*	6	31	468	95	0	0	5
PHYLUS	Willow Warbler	*Phylloscopus trochilus*	15	81	1207	83	7	1	9
PHYSIB	Wood Warbler	*Phylloscopus sibilatrix*	9	56	829	62	0	0	38
SYLATR	Blackcap	*Sylvia atricapilla*	10	68	1252	59	4	4	33
SYLBOR	Garden Warbler	*Sylvia borin*	12	85	2083	61	0	2	36
TURMER	Blackbird	*Turdus merula*	9	81	639	29	3	10	58

according to the lengths of the syllables has the effect of linear warping between the different time axes of the syllables.

B. Mel-Cepstrum Model

Cepstrum belongs to the class of homomorphic representations [22], [23], and it has been found useful in various types of recognition tasks. In particular, the mel-cepstrum [24] has been a popular signal representation in the automatic speech recognition (ASR). Although several alternative feature extraction methods have been presented during the long history of ASR, the mel-cepstrum still constitutes the basis of several state-of-the art speech recognizers. It is simple and robust and its computation does not require any particular parameter tuning. In an ASR comparison, mel-cepstrum performed well against a more complex human auditory model, the latter performed slightly better only under noisy conditions when SNR was below 30 dB [25].

Computation of mel-cepstrum involves discrete cosine transform applied to the logarithmic mel-spectrum. The role of DCT is to reduce the dimensionality of the original mel-spectrum vector. It also decorrelates the feature components. The ith mel-cepstral coefficient is computed as [24]

$$\mathrm{MFCC}_i = \sum_{k=1}^{K} X_k \cos\left(\frac{i(k-0.5)\pi}{K}\right) \tag{5}$$

where X_k is the logarithmic energy of the kth mel-spectrum band, and K is the total number of the mel-spectrum bands. In ASR, typical number of MFC coefficients computed from one speech frame is 8–12. This representation can be named as static MFCC feature vector since it is computed from one time frame only. The sequence of static feature vectors computed from the consecutive time frames represents the evolution of the features in time. Local temporal dynamics can be represented in a fixed-dimensional feature vector by computing so called delta features. These are temporal differences or regression coefficients computed from predefined number of consecutive feature vectors. Typically, the first-order delta (Δ) features in ASR are computed from 50-ms time span. It is also almost a standard procedure to use second-order delta ($\Delta\Delta$) features in ASR. These represent the temporal change of the first-order deltas.

The basic idea to use mel-cepstrum in this paper is to represent the entire spectrum of the signal in contrast to the sinusoidal modeling which picks only the most dominant spectral peak of the frame. We do not claim that the mel-scale would give the optimal accuracy for representing avian vocalizations since it compresses the spectral information by approximating the resolution of the human hearing system. However, there are ornithologists who are able to recognize bird species using only the human auditory system. However, the main point here is to compress the spectral information. In case MFCC misses something essential, like e.g., pitch information, this can be compensated by using a set of additional descriptive parameters explained in the following.

C. Descriptive Parameters

The sinusoidal model and the mel-cepstrum model are both based on a specific signal model. The sinusoidal model assumes strongly tonal signal with possibly a few harmonics. The mel-cepstrum, on the other hand, assumes that the spectra are characterized by relatively wide formant regions like in speech signals and most of the relevant detail is at low frequencies.

In many applications, it is not possible to identify a specific signal model fitting all target signals. The problem of automatic audio classification is a typical application where the sound material may be composed of many different types of sounds. In these applications, it is typical to use various types of descriptive measures which relate both to the temporal and spectral characteristics of the signal. In this paper, we represent syllables with 19 low-level signal parameters. Seven features are calculated on the frame basis which provide a short-time description of the syllable. First, the syllable is divided into overlapping frames of 256 samples with 50% overlap. The actual features used in the recognition system are the means and variances of these features computed over the duration of the syllable. This gives 14 features which are concatenated to the five other features. Descriptive parameters used in the present study are listed in Table III. Their detailed description is provided in [19].

IV. CLASSIFICATION METHODS AND MODELS

In this paper, the recognition of individual syllables is based on the nearest neighbor classifier. This method involves no training process. All samples in the training data are used as

TABLE III
DESCRIPTIVE PARAMETERS USED IN THE CURRENT STUDY. ASTERISK (*) INDICATES THAT THE FEATURE IS CALCULATED ON THE FRAME BASIS

Feature	Frame feature
Spectral features	
Spectral centroid	*
Signal bandwidth	*
Spectral roll-off frequency	*
Spectral flux	*
Spectral flatness	*
Minimum frequency	
Maximum frequency	
Temporal features	
Zero crossing rate	*
Short time energy	*
Syllable temporal duration	
Modulation spectrum magnitude	
Modulation spectrum frequency	

such for representing the classes. In the classification phase, the test syllable is compared against all syllables of the training data, and the class label is determined by the training data sample which has the largest similarity/smallest dissimilarity to the test syllable. Here, we experimented two approaches for representing the syllables: using a variable-length trajectory model and a fixed-dimensional feature vector.

When recognizing the species based on song fragments, i.e., sequences of consecutive syllables, both GMMs and HMMs were used. GMM involves no temporal modeling of the syllable sequence unless that information is embedded in the feature vectors. HMMs, instead, allow more explicit modeling of temporal dynamics.

A. Dynamic Time Warping

Syllables have typically different durations. Dynamic time warping (DTW) algorithm [26] can be used for comparing variable-length sequences. Its basic idea is to warp the time axes of two sequences nonlinearly so that the maximum fitting between the sequence elements is attained. The computation can be done in a two-dimensional trellis. Here, the word element refers to the generic element of the feature vector sequence, it should not be confused with the syllable/element decomposition of the song described in Section II.

In the following, two syllables are represented by the trajectory models A and B. The elements of the sequences are frame-based feature vectors and the sequence lengths are denoted by L_A and L_B. The distance between the sequence elements $A(i)$ and $B(j)$ is denoted by $d(i,j)$, and the cumulative distance at trellis coordinate (i,j) is denoted by $g(i,j)$. First, the trellis is initialized

$$g(0,j) = \begin{cases} 0 & j = 0 \\ \infty & j = 1 \dots L_B \end{cases}$$
$$g(i,0) = \begin{cases} 0 & i = 0 \\ \infty & i = 1 \dots L_A \end{cases}. \quad (6)$$

Cumulative distances are then computed using dynamic programming as follows:

$$g(i,j) = \min \begin{cases} g(i,j-1) + d(i,j) \\ g(i-1,j-1) + w_d d(i,j) \\ g(i-1,j) + d(i,j) \end{cases} \quad (7)$$

where index i goes from 1 to L_A and index j from 1 to L_B. Parameter w_d is the weight of the diagonal movement in the trellis. DTW distance is defined to be

$$D(A,B) = \frac{g(L_A, L_B)}{(L_A + L_B)}. \quad (8)$$

Here, the cumulative distance is divided by the sum of the lengths of the sequences, but other choices are also possible [26]. In order to use DTW, the distance measure must first be defined for the sequence elements. In the sinusoidal modeling, there are two parameters per sequence element: amplitude and the frequency. We can now consider these two parameters separately and have a two-dimensional vector or use the amplitude information to weigh the importance of the frequency information. These two approaches were compared.

The parameter w_d (7) is the cost of the diagonal movement in a trellis. If w_d is smaller than two, the warping favors linear warping. We did experiments to see how this parameter affects the single syllable recognition, the results are in Section V-A.

The recognition system based on DTW consists of the templates which are the reference sequences of the classes. Each sequence consists of feature vectors, so each template is a point trajectory in the feature space. In order to expand the point trajectory representation into probability distributions, there are two alternatives. The templates can be divided into segments, and each segment is represented by some probability distribution of the feature vectors. The information about the temporal order of the feature vectors inside each segment will then be lost, but the order of the segments can be maintained by forming chains of segments. This is essentially the concept of HMM [27]. Another alternative to bring the basic DTW algorithm into probability domain is to model the distribution of the DTW distances between the training data and the template. This is equal to adding the probability density function (pdf) to each element of the DTW template and then computing the probability of the data sequence given the chain of pdfs by means of the Viterbi algorithm [28]. The difference between these two alternatives is that in the HMM the temporal resolution is usually relatively small, i.e., the number of the states in the HMM is smaller than the number of the elements in the typical DTW template. The benefit of the HMM approach is that different states can have different pdfs and thus the changing variance of different parts of the sequence can be taken into account in the model. When only the distribution of the cumulative distance is modeled, the same pdf is applied to all parts of the sequence. These alternatices are illustrated in Fig. 6.

The motivation for the use of DTW was the desire to compute the distances between syllables with varying lengths. The

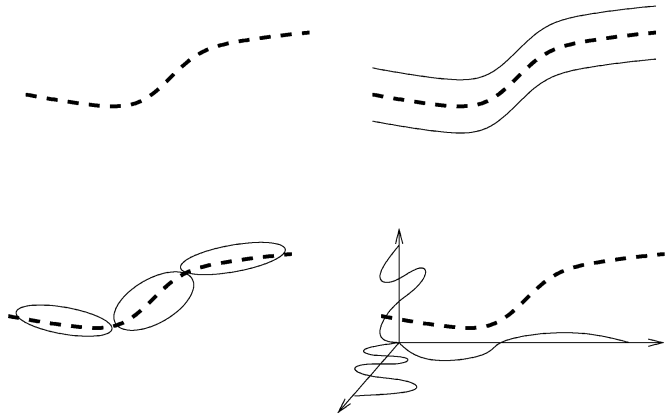

Fig. 6. Examples of trajectory models. Top left corner: template of basic DTW. Each point of the trajectory represents a feature vector. Bottom left corner: three-state hidden Markov model, ellipsoids represent the Gaussian pdfs of the states. Top right corner: modeling the distribution of DTW distance by some pdf corresponds to the trajectory model where each trajectory point is the mean of that pdf. Bottom right corner: fixed-dimensional representation of the trajectory can be obtained using trajectory projection. Three base trajectories are shown along the three axes.

comparison would be easy if the sequences had fixed-dimensional feature vector representations. For this reason, we experimented the method where the data sequences are represented by the means of the base trajectories. The syllable is projected against the base trajectories and each projection gives one component of the feature vector. The comparison between syllables is then done by means of the fixed-dimensional projection vectors. In this paper, the base trajectories had the shape of the cosine base vectors. The lengths of the base trajectories were determined individually to each syllable being the length of the syllable. Effectively, this performs linear warping when comparing two syllables.

B. Gaussian Mixture Model

Mixture models provide a flexible framework for modeling probability density functions in pattern recognition. In this paper, Gaussian mixtures were used. GMMs were trained using the standard expectation maximization (EM) algorithm [29], [30]. One could argue that maximum-likelihood (ML)-based training is not the best approach for training a classifier since in the recognition task the goal is to discriminate different classes. ML-based models give optimal classifiers only when the underlying models are correct. This is almost never the case in practice, since the models are only approximations and abstractions of the reality. The computational resources should, therefore, be put to the modeling of the class borders instead of the general shapes of the class distributions. However, in the practical application of bird song recognition, the number of the bird classes may vary. If the classifiers are trained using the ML approach, each classifier is trained separately from other classes and it is then easy to add and remove new classes to the system without any retraining. If discriminative training is used, this freedom is lost and there must be access to all training data from all classes when adding new classes to the system.

Initialization of the GMM is an important issue in practice. There are several local maxima in the likelihood function which results that different initializations may give different models which are all suboptimal. A common strategy is to initialize the centroids of the Gaussians by the K-means algorithm [31], [32]. However, also the K-means algorithm is sensitive to the initialization. Therefore, in this paper, we used the self-organizing map (SOM) [33] to initialize the code vectors of the K-means. SOM is similar to K-means, but the main difference is that the code vectors are located at the nodes of a structured lattice. In K-means algorithm, one data vector adapts only one code vector (the nearest one) at a time, but in the SOM algorithm, the code vectors in the neighborhood of the nearest code vector are also adapted. The neighborhood which is defined in the lattice space controls the stiffness of the map. It is typically wide in the beginning and then shrinks as the training proceeds. This results in that the SOM fits a flexible grid of code vectors to the feature space covering the most dense areas of the data. An important aspect is that all code vectors will be moved toward the area of input data despite their initializations. Both online learning and batch algorithms can be used for training the SOM. We used in this paper the batch algorithm. In the end of the training, the neighborhood function was shrunk to contain only the closest code vector to the data so that the algorithm reduced essentially to K-means. The code vectors were then initialized with variance parameters and mixture weights so that GMM training could begin. Ten iterations of EM algorithm were applied to each GMM.

C. Hidden Markov Model

The HMMs used in the current study are slightly different from the traditional ones used in speech recognition. Usually, the HMM topology is determined first and the training is performed so that the state label sequence corresponding to the observations, i.e., feature vector sequences, is known. The commonly used Baum–Welch training utilizes the EM algorithm principle so that the state indices are considered as latent variables and the model parameters are trained using all possible state alignments weighted by their probabilities [27]. Alternatively, Viterbi training can be used where the model parameters are adapted only along the single best state alignment [27]. In both cases, only the state label sequence needs to be known, the alignment of the observations into states is embedded in the training algorithm.

HMMs have been used in the bird song recognition earlier e.g., in [11], where HMM approach was compared against DTW. In that work, the bird songs used in the HMM training were manually transcribed into elementary segments so that the standard HMM training with labeled HMM states could be used. In that work, no species recognition was performed since the interest was in the recognition of the song elements. The recognition systems in that study were tested separately for two bird species (Zebra Finches and Indigo Buntings).

In our data set, we have only the knowledge of the bird species as an identifier of the song. The songs have not been further transcribed or manually segmented. Since we can only use the class identifier of the entire song, our HMM training is more unsupervised in nature than the training methodology commonly used in human speech recognition systems where the data have usually been transcribed in word or subword level. Our goal is now to

TABLE IV
SINGLE-SYLLABLE-BASED SPECIES RECOGNITION USING SINUSOID TRAJECTORY MODEL. COLUMNS CORRESPOND TO DIFFERENT DEFINITIONS OF THE DISTANCE BETWEEN INSTANTANEOUS TIME POINTS OF THE SYLLABLE. 2-DIM REPRESENTATIONS TREAT AMPLITUDE AND FREQUENCY INFORMATION SEPARATELY WHEN COMPUTING EUCLIDEAN DISTANCE AND 1-DIM REPRESENTATION USES AMPLITUDE WEIGHTED DISTANCE BETWEEN FREQUENCY COMPONENTS. PARAMETER w_d IS THE WEIGHT FOR DIAGONAL MOVEMENT IN THE DTW TRELLIS

w_d	2-dim	2-dim squared	1-dim	1-dim squared
0.1	43.6	43.5	38.2	39.9
0.5	45.1	44.3	39.8	41.0
1.0	43.9	44.0	39.0	40.4
2.0	38.3	41.8	34.6	38.7

find and extract the appropriate model topology, i.e., the structure of the state connectivity graph, rather than predetermine it. In this paper, we model each state of the HMM by a single Gaussian pdf. Contrasted with the GMM, each Gaussian component acts now as a separate state so the transitions between the Gaussians are added to the model. The transition probabilities as well as the means and variances of the Gaussians are trained by ten iterations of Viterbi training. Baum–Welch training could be used as well, in fact Viterbi training can be considered as its approximation [34]. We did not manually predetermine the topology of the HMM, the state transitions were totally determined by the data. In the end of the training, the states and their transitions modeled the trajectories of the songs as a one bird species specific graph. Nothing prevents using more than one Gaussian per state, but there is evidence that typically in a Gaussian mixture model only one or few mixture components are "active" at the same time, i.e., the nearest Gaussians to the observation vector dominate the value of the entire pdf [35]. When only one Gaussian is used per state and the total number of the Gaussians is fixed, the modeling of the temporal dynamics of the song is emphasized and the resulting state graph should then model the song trajectories most accurately.

V. RESULTS

There were 137 individual birds from 14 species in our data set. The total number of the syllables in 792 songs was 12 016. In the nearest neighbor classification, all syllables from the test bird individual were removed from the reference syllables before classification. Also in the GMM and HMM tests, the training data never included any recordings from the bird individual being tested. This setting can, therefore, be described as a bird individual independent species recognition. All experiments were repeated so that each of the 137 bird individuals was once in the test set. Each recognition result presented below is an average of these 137 tests.

A. Recognition Based on Trajectory Model of a Syllable

The first recognition experiments were based on modeling the syllables as time varying trajectories. The DTW distance between the sequences consists of the sum of the sequence element distances as shown in Section IV-A. In case of syllables, the sequence elements are parameter vectors computed at constant time points over the duration the syllable. In the first experiment, the parameter vectors contained two components of the time varying sinusoid: the instantaneous frequency in hertz and its amplitude in decibels. Two different methods were experimented for computing the distances $d(i,j)$ in (7). The first one used the plain Euclidean distance between the two-component vectors, and the second method used the absolute difference of two sinusoid frequencies weighted by the sum of their amplitudes. Both of these methods were also tested with and without squaring the sequence element distances. The results of these four tests are in Table IV. We see that the two-component representation of the instantaneous sinusoid performs slightly better compared to the amplitude weighted frequency. When investigating the effect of parameter w_d, we can see that the value 0.5 gives the best results for all four distance computation methods. Since the use of small value of w_d favors linear time axis warping and maintains more durational differences in matching, the results confirm that temporal durations of sequence elements contain useful information for classification.

TABLE V
SINGLE-SYLLABLE-BASED RECOGNITION USING TRAJECTORY MODEL WITH 8-, 16-, AND 24-DIMENSIONAL MFCC VECTORS

w_d	8 MFCC	8 MFCC, Δ	8 MFCC, Δ , $\Delta\Delta$
0.1	45.3	46.0	46.1
0.5	50.6	50.7	50.7
1.0	51.1	51.6	51.7
2.0	46.2	47.3	47.5

TABLE VI
SINGLE-SYLLABLE-BASED RECOGNITION USING FIXED-DIMENSIONAL SYLLABLE REPRESENTATIONS

feature	dimension	correct %
descriptive parameters	19	40.7
DCT of sinusoid trajectory, duration, harmonicity	8	42.7
MFCC	8	38.2
MFCC, Δ	16	41.6
MFCC, Δ , $\Delta\Delta$	24	41.8
PCA (MFCC, Δ , $\Delta\Delta$, descriptive parameters)	30	47.7

In the second experiment, mel-cepstrum vectors were used as the elements of the syllable trajectory model. The Euclidean distance was used between the MFCC vectors as the sequence element distance $d(i,j)$. Contrasted with the sinusoidal modeling, now the entire spectrum between 1 and 10 kHz was represented in each time frame. Eight-dimensional MFCC vectors were computed from fixed 20-ms time windows at 10-ms intervals using HTK software [36]. The number of mel-spectrum bands was 36. Nearest neighbor classification results using DTW are shown in Table V. Delta features did not improve the recognition accuracy here.

B. Recognition Based on Single Feature Vector Representation of a Syllable

After DTW based tests, it was investigated how the syllable could be represented using only a single feature vector. These results are in Table VI.

Single-syllable recognition was first tested using descriptive parameters. Syllables were compared using the nearest neighbor classifier with the Mahalanobis distance measure. Advantages of Mahalanobis distance instead of Euclidean is that it automatically scales the coordinate axes of the feature

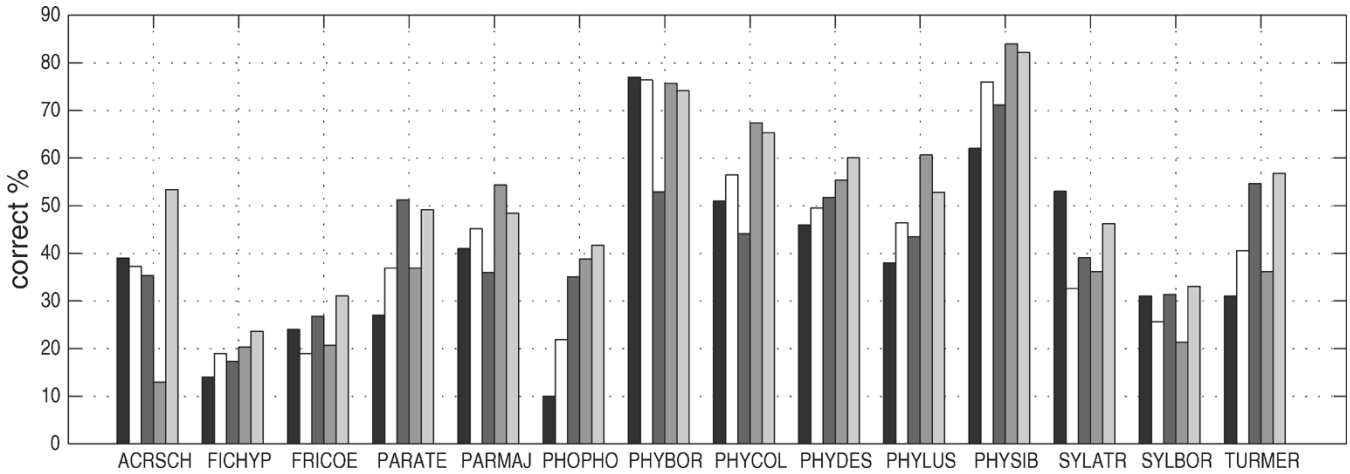

Fig. 7. Single-syllable-based species recognition. Each set of five bars shows the classification accuracy for one species using five methods (from left to right): 1) 19-dim descriptive parameters. 2) 8-dim vector consisting of three DCT components from sinusoid trajectory, two duration components, and three harmonicity classes. 3) 24-dim average MFCC of syllable. 4) Trajectory of 2-dim sinusoid vectors. 5) Trajectory of 24-dim MFCC vectors.

space. It also decorrelates different features. This corresponds to applying principal component analysis (PCA) without dimensionality reduction to the original feature vectors and then computing the Euclidean distances between the projected vectors. The norms of the PCA base vectors must then also be scaled by the square roots of the eigenvalues of the data covariance matrix.

In the DCT-based trajectory projection, the time-varying sinusoid trajectories were projected into fixed-dimensional feature vectors using cosine base trajectories. When computing the first cosine base component, amplitude information was used, otherwise only the frequency information was used. The first projection measures the average frequency content in the syllable. The second projection measures whether the syllable trajectory is falling or rising in frequency, and the third base trajectory models an additional dip in the trajectory. Besides these projection values, the final fixed-dimensional feature vector contained two duration components and three harmonicity class measures. The first duration information was the length of the syllable, and the second one was the time between the current syllable and the previous one. For the first syllable in the song, this missing information was not used when computing the distances between the vectors. Three harmonicity measures were added to the final feature vector as explained in Section III. Harmonicity classes III and IV were merged together and thus the final feature vector contained eight components. All feature components were scaled to have zero mean and unit variance. The comparison between the feature vectors was done using the Euclidean distance. The recognition results are close to those of 1-dim representations in Table IV where the original syllable trajectories were used with DTW, but they are slightly inferior compared to the results of 2-dim representations in Table IV and cepstrum representations in Table V. By comparing MFCC results in Tables VI and V, it can be seen that averaging the syllable content into a single vector degrades the recognition accuracy.

All single vector representations resulted in similar recognition accuracies on average. However, different representations may capture specific aspects of the syllable and, therefore, perform differently for different species. The variation in the species specific recognition results using five syllable representations is shown in Fig. 7. This suggests that it may be advantageous to combine different features. This was experimented with MFCC and descriptive parameters. 24-dimensional MFCC vector (eight static, eight delta, and eight delta-delta components) and 19-dimensional descriptive parameters vector was concatenated into a 43-dimensional feature vector. PCA was then applied with scaling the components by their standard deviations. When investigating the effect of the number of PCA components, it was found that the improvement in the recognition accuracy saturated at 30 components. The feature combination clearly improved both the MFCC and descriptive parameter based results.

C. Recognition Based on Syllable Sequences

The possibility to use several consecutive syllables in the recognition was investigated next. The data unit being classified is now a song fragment. The number of the syllables in a song fragment was not limited and it varied from song to song. In the first experiment, the fixed-dimensional syllable representation was used. The classifiers were species-specific GMMs and HMMs. In both cases, diagonal covariance matrices were used in the Gaussian kernels. The particular way of training the HMMs was explained in Section IV-C. Cosine trajectory bases with one, two, and three base trajectories were used for converting the variable-length syllables into fixed-dimensional feature vectors. The results are in Table VII. These are considerably better compared to the single-syllable based recognition results using the same features. The single-syllable based recognition accuracy was around 40% and using song-based recognition it increased to almost 60%.

The final experiment was GMM and HMM-based classification with MFCC features. MFCC vectors were computed from overlapping 25-ms time windows at 10-ms intervals. The entire waveform of each song fragment was now used without any syllable/nonsyllable segmentation. The number of mel-frequency

TABLE VII
SPECIES RECOGNITION USING SONG FRAGMENTS WITH EIGHT-DIMENSIONAL SYLLABLE REPRESENTATION. THREE COSINE BASE TRAJECTORY PROJECTIONS WERE USED WITH TWO DURATION INFORMATION AND THREE HARMONICITY CLASS COMPONENTS. COLUMN K IS THE NUMBER OF GAUSSIANS IN THE GMM AND HMM

K	GMM	HMM
10	57.0	58.2
20	58.6	57.2
30	57.7	51.4

TABLE VIII
SPECIES RECOGNITION USING MFCC VECTORS COMPUTED OVER THE ENTIRE SONG FRAGMENT. THE NUMBER OF GAUSSIANS IN GMM AND HMM IS DENOTED BY K

	8 MFCC		8 MFCC, Δ		8 MFCC, Δ , $\Delta\Delta$	
K	GMM	HMM	GMM	HMM	GMM	HMM
10	59.8	61.2	66.3	65.9	69.2	68.9
20	59.9	58.3	68.7	69.6	70.0	70.3
30	58.8	57.1	69.3	69.5	70.9	70.3
40	56.3	56.7	67.5	66.0	71.0	67.7
50	55.5	54.4	65.7	65.6	71.3	69.5

spectral bands was 36 and the number of the cepstral coefficients was eight. The results are in Table VIII. Delta features clearly improved the performance. It can be seen that modeling the local dynamics of the song is thus important. We also investigated the effect of the number of spectral bands and cepstral coefficients. These parameter values seemed not to be very critical. When changing the number of mel-bands between 24 and 48 and the number of MFC coefficients between 8 and 24, all results were close to 70% when delta features were used.

There were only small differences between the results of GMMs and HMMs in general. The main difference between the GM and HM models was that in the HMM, there were additional transitions between the Gaussians which modeled the sequential order of the song elements. However, there seemed to be no benefit from modeling the dynamics of the song in the recognition accuracy. This is partly because of the different numerical ranges of transition probabilities and Gaussian emission probabilities. Since the latter ones have much wider range, they dominate the likelihood computation. However, although the transition probabilities between Gaussians did not improve the species recognition accuracy, HMM state graphs can still be useful for investigating the structure of syllable sequences. They should tell how the syllables follow each other.

VI. DISCUSSION

The recognition accuracy for single syllables is generally low. For some species, the percentage of correct recognition is over 70%, but the average is only around 40%–50% depending on the method. This suggests that the hypothesis [15] that bird species could be recognized from isolated syllables only may be too optimistic. The recognition results improved significantly in song-based recognition. This is because there is more data to support the classification. If a single syllable is confusing or not representative, the preceding or following ones may help in making the correct decision.

The alternative representations of sounds led to very different result in different species. Comparison of the harmonic class probabilities in Table II and recognition results in Fig. 7 suggests that one difference may be related to the tonality of sounds. As expected, the method based on modeling the syllables as time-varying sinusoids gave high recognition accuracies for species with a large number of syllables belonging to the harmonicity class I. These are the Great Tit and four species from the *Phylloscopus* family.

In the following, Methods 1–5 correspond to the methods described in Fig. 7. In Methods 1–3, the MFCC (Method 3) gave the highest accuracy for six out of 14 species. However, the average percentage over all species was slightly higher using Method 2 which was based on the sinusoidal model. Method 1 which was based on descriptive parameters was on the average worse than the two other methods. It gave the highest accuracy for two species, the Common Redstart and Arctic Warbler.

Some of the results suggest that the number of recordings may not be sufficient. For example, the Pied Flycatcher has the lowest overall recognition percentage and the smallest number of syllables. On the other hand, the Garden Warbler has the largest number of syllables and a low recognition accuracy in all methods. The Garden Warbler is one of the most versatile singers in the selection of species and is capable of producing a large variety of different syllables.

Methods 4 and 5 gave, on the average, higher accuracies than Methods 1–3. The trajectory method for MFCC vectors (Method 5) gave the highest overall performance in recognition of single syllables. However, the difference to Method 4 was relatively small considering that the latter was based on a temporal trajectory of two-dimensional vector values while the dimension of the MFCC vector was 24. When comparing the average-MFCC vector (Method 3) to the MFCC trajectory, one can see a systematic trend that Method 5 is 5%–20% better. Comparing the performance of Methods 2 and 4 a similar trend cannot be observed. Method 2 contains the temporal variation over a syllable in the DCT coefficients and, therefore, the difference to the trajectory model based on the same parameters could be expected to be small. However, the differences between Methods 2 and 4 for several species are very different. This may be caused by the use of harmonicity parameters in Method 2 which improve the performance in the case of nontonal sounds.

The large differences in the recognition results between methods and species are not only caused by the fact that a certain method fits better to the sounds of some species. In the nearest neighbor classification, it is possible that if a method does not fit well to the actual physical properties of sounds from one species, it may degrade the recognition accuracy of all species. For example, if Method 4 is applied to an ensemble of typical creaky syllables often heard at the end of the song of Blackbird, it may give very different parametric representations. Some of the feature vectors may randomly fall close to some pure sinusoidal syllable from another species. In the nearest neighbor classification, this will degrade the recognition accuracy of both species. For this reason, it could make sense to use hierarchical classification techniques. For example, if the likelihood of belonging to the harmonicity class I is low, the sinusoidal model should be abandoned and the classification should be made, e.g., using the MFCC parameters.

It was found that the results with song level parameterization are significantly better than the results from the recognition of single syllables. One obvious reason is that there is important species specific information at the song level and that many species produce very similar syllables. However, it is also likely that the recognition of single syllables is difficult because there is large variability in bird vocalization. Some of the sounds are clearly sinusoidal while others have more complex spectrum and temporal envelope. In this light, it is somewhat surprising that the set of descriptive parameters combining spectrum and temporal parameters gave the lowest average-recognition accuracy and the largest variability in performance between different species. Based on the results of the current article, it seems that the MFCC trajectory model is, on average, the best model for a large number of species.

Although the best results were obtained using MFCC features computed from song fragments without any syllable segmentation, we still find the idea of syllable segmentation and syllable-specific features tempting. The presence of silence after each syllable is very characteristic to the bird song.

The lengths of the song fragments in our present experiments were determined by the human user. The recordings were manually segmented so that in each recording there was only one bird present. Now when the models have been trained, especially HMMs, they can be used for segmenting recordings where several species are present, provided that syllables of multiple species do not overlap in time. If overlaps occur, additional source separation methods must then be used in the analysis.

VII. Conclusion

In this paper, we have compared three feature representations for avian vocalizations in the context of species recognition. The recognition methods were based on the use of single syllables and song fragments. In the single-syllable based recognition, two syllable representations were compared. One represented the variable-length syllable as a sequence of frame-based feature vectors and the other converted the entire syllable into a single feature vector. The best results were obtained by using MFCC-based syllable trajectory models with DTW matching. The best fixed-dimensional feature vector consisted of the combination of the average MFCC vector and the set of descriptive parameters. Further improvement in the recognition accuracy was gained using song fragments, i.e., sequences of consecutive syllables, as the basis of the classification.

Based on the results of this study, there seem to be certain bird species which are able to be recognized using currently available methods, whereas the variety of the sounds of some species clearly require further analysis. Whether this is because of the lack of data or inadequacy in the current preprocessing and classification methods will be seen in the future. Collecting large databases of avian vocalizations has just started, and we might now be in the similar situation where the automatic speech recognition field was two decades ago. Besides, this new field is interesting per se, the avian vocalization data also provide an interesting test bench for current pattern recognition systems and challenges for developing new machine learning tools.

References

[1] C. K. Catchpole and P. J. B. Slater, *Bird Song: Biological Themes and Variations*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

[2] W. H. Thorpe, *Bird Song*. Cambridge, U.K.: Cambridge Univ. Press, 1961.

[3] R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech*. New York: Van Nostrand, 1947.

[4] B.-S. Shieh, "Song structure and microgeographic variation in a population of the Grey-cheeked Fulvetta (Alcippe morrisonia) at Shoushan nature park, southern Taiwan," *Zool. Stud.*, vol. 43, no. 1, pp. 132–141, 2004.

[5] P. J. Christie, D. J. Mennill, and L. M. Ratcliffe, "Chickadee song structure is individually distinctive over long-broadcast distances," *Behavior*, vol. 141, no. 1, pp. 101–124, 2004.

[6] O. Tchernichovski, F. Nottebohm, C. E. Ho, B. Pesaran, and P. P. Mitra, "A procedure for an automated measurement of song similarity," *Animal Beh.*, vol. 59, pp. 1167–1176, 2000.

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

[8] K. Ito, K. Mori, and S. Iwasaki, "Application of dynamic programming matching to classification of budgerigar contact calls," *J. Acoust. Soc. Amer.*, vol. 100, no. 6, pp. 3947–3956, Dec. 1996.

[9] C. Rogers, "High resolution analysis of bird sounds," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 3011–3014.

[10] A. Härmä and M. Juntunen, "A method for parameterization of time-varying sounds," *IEEE Signal Process. Lett.*, vol. 9, no. 5, pp. 151–153, May 2002.

[11] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study," *J. Acoust. Soc. Amer.*, vol. 103, no. 4, pp. 2185–2196, Apr. 1998.

[12] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *J. Acoust. Soc. Amer.*, vol. 100, no. 2, pp. 1209–1219, Aug. 1996.

[13] C. Kwan, G. Mei, X. Zhao, Z. Ren, R. Xu, V. Stanford, C. Rochet, J. Aube, and K. C. Ho, "Bird classification algorithms: Theory and experimental results," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Montreal, QC, Canada, May 2004, pp. 289–292.

[14] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2740–2748, Nov. 1997.

[15] A. Härmä, "Automatic recognition of bird species based on sinusoidal modeling of syllables," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Hong Kong, China, Apr. 2003, pp. 545–548.

[16] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Montreal, QC, Canada, May 2004, pp. 701–704.

[17] P. Somervuo and A. Härmä, "Bird song recognition based on syllable pair histograms," in *IEEE Int. Conf. Acoust, Speech, Signal Processing (ICASSP)*, Montreal, QC, Canada, May 2004, pp. 825–828.

[18] A. S. King and J. McLelland, Eds., "Larynx and Trachea," in *Form and Function in Birds*. New York: Academic, 1989, vol. 4, pp. 69–103.

[19] S. Fagerlund, "Automatic Recognition of Bird Species by Their Sounds," M.S. thesis, Helsinki Univ. Technol., Espoo, Finland, 2004.

[20] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Lett.*, vol. 22, pp. 533–544, 2001.

[21] B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 389–406, Sep. 1997.

[22] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The frequency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proc. Symp. Time Series Analysis*, Istanbul, Turkey, Jun. 5–9, 1963, pp. 209–243.

[23] A. Oppenheim and R. Schafer, *Digital Signal Processing*. New York: Springer, Prentice-Hall, 1975.

[24] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[25] C. R. Jankowski, Jr., H.-D. H. Vo, and R. P. Lippman, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 286–292, Jul. 1995.

[26] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP–26, no. 1, pp. 43–49, Feb. 1978.
[27] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[28] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, no. 2, pp. 260–269, Apr. 1967.
[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.
[30] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.
[31] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. C-28, no. 1, pp. 84–95, Jan. 1980.
[32] R. Gray, "Vector quantization," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 1, no. 2, pp. 4–29, Feb 1984.
[33] T. Kohonen, *Self-Organizing Maps*. Berlin, Germany: Springer, 1995.
[34] B.-H. Juang and L. R. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 9, pp. 1639–1641, Sep. 1990.
[35] P. Somervuo, "Speech recognition using temporally connected kernels in mixture density hidden Markov models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Istanbul, Turkey, Jun. 5–9, 2000, pp. 3434–3437.
[36] P. Woodland, S. Young, and G. Evermann, HTK, the Hidden Markov Model Toolkit, Version 3.0 [Online]. Available: http://htk.eng.cam.ac.uk/ 2002
[37] A. S. King and J. McLelland, Eds., *Form and Function in Birds*. New York: Academic, 1989, vol. 4.

Panu Somervuo was born in Helsinki, Finland, in 1971. He received the M.Sc. degree in electrical engineering and D.Sc. degree in computer science from the Helsinki University of Technology (HUT), Espoo, Finland, in 1996 and 2000, respectively.

From 1994 to 2002, he was with the Neural Networks Research Centre, HUT. From 2002 to 2003, he was a Visiting Researcher at the International Computer Science Institute, Berkeley, CA. After returning to Finland, he continued his work at the Neural Networks Research Centre. Currently, he is researching bioinformatics at the University of Helsinki, Department of Applied Biology. His general research interests are in machine learning.

Aki Härmä was born in Oulu, Finland, in 1969. He received the M.S. and Ph.D. degrees in electrical engineering from the Helsinki University of Technology (HUT), Espoo, Finland, in 1997 and 2001, respectively.

From 2000 to 2001, he was a Consultant at Lucent Bell Laboratories and later, Agere Systems, Murray Hill, NJ. He started his work on bird sounds in 2001 when he returned to the Laboratory of Acoustics and Audio Signal Processing, HUT, to the position of a Postdoctoral Researcher. In 2004, he joined the Digital Signal Processing Group of Philips Research Laboratories, Eindhoven, The Netherlands. His research interests are mainly in acoustics and audio signal processing.

Seppo Fagerlund was born in Pori, Finland, in 1978. He received the M.Sc. in electrical engineering from the Helsinki University of Technology (HUT), Espoo, Finland, in 2004.

In 2002, he was a Research Assistant with the Nokia Research Center. In 2004, he became a Research Assistant, and in 2005, Researcher at the Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology. His research interests include signal processing of bioacoustic signals and pattern recognition.