NEW PARAMETRIC REPRESENTATIONS OF BIRD SOUNDS FOR AUTOMATIC CLASSIFICATION

Seppo Fagerlund, Unto K. Laine

Aalto University, Department of Signal Processing and Acoustics

ABSTRACT

Identification of bird species based on their vocalization is studied in this paper. The main focus is introducing a new parametric representation of bird sounds for automatic identification of their species. The method is based on the statistics of local temporal patterns in bird vocalization. Two different sets of bird species are used in the classification tests. The first set contains six species that often produce inharmonic sounds. For the second set, four species that produce very different types of sounds were added. Recognition results using a k-NN-classifier shows improved recognition accuracy over the results obtained by MFCC-features.

Index Terms— bird classification, permutation transformation, feature extraction

1. INTRODUCTION

Interest towards monitoring our surrounding environment using technology has increased during recent years. Automatic identification of bird species based on their sounds has especially gained much attention. However, due to the large number of bird species as well as a multitude of different bird sounds and different noise conditions in the recordings, there are still many challenges to be solved. Bird species classification by their sounds is a typical audio classification problem that starts with bird sound extraction from continuous recordings. Next, the segmented sounds are represented using a set of features that are used to classify segments into different classes of bird species.

Several different parametric representations of bird sounds have been proposed for classification of bird species or even individuals. A majority of the features operate in the frequency domain and include Mel-frequency cepstral coefficients (MFCC) [1, 2, 3, 4], linear predictive coefficients (LPC) [5], and wavelets [6]. Detection of sinusoidal components of bird sounds have efficiently been used to represent tonal and harmonic bird sounds [7, 8, 9, 10]. Descriptive acoustical signal parameters, such as the spectral centroid and signal bandwidth, provide effective representations for different types of bird sounds [11]. These features are often connected with temporal descriptive features. Lee et al. [12] used a totally different representation of bird sounds that exploits the spectro-temporal characteristics of spectrogram images.

A new parametric representation for bird sounds is proposed in this work. The method is based on the statistics of short-term temporal patterns in the signals. The main motivation is to develop a parametric representation that can discriminate between many different types of sound events that are common within birds. Another desirable feature is the robustness to different noise conditions as well as minor differences in the segmentation of bird sounds. We assume that discriminative information of sound events can be found in the distribution of local temporal variations within the signal structure. Also, time domain methods do not suffer from the time-frequency resolution limitation that are exist with spectral methods.

Overall, very few studies have looked at the temporal structure of bird sounds. The work by E. D. Chesmore [13] used 28 pre-defined duration dependent shape patterns to transform bird sounds into a series of codes. The histogram of pairs of these codes represented single sound events and were used to classify different species. Chesmore achieved encouraging results in preliminary bird species recognition tests. Recognition was tested with sounds of 10 species that were recorded in a similar environmental and background noise conditions.

Earlier studies concerning ordering of events have focused on the correlation of ordered observations made by two or more observers [14]. Recently, more attention has been focused on signal analysis and pattern recognition in time series based on temporal structures [15, 16, 17]. These studies have resulted in a wide number of applications, including the detection of abnormalities in aircraft engines [18] as well as the measurement of the complexity in time series [19]. Additionally, ordinal analysis of medical signals have been studied during the past few years. It has been used to detect certain patterns in EEG [20], EMG [21, 22] and neuronal activity signals [23]. Temporal pattern features have also been used for classification of speech consonants [24]. It is considered a difficult task when using traditional tools and methods from speech recognition.

2. BIRD SOUNDS AND SEGMENTATION

Bird sounds can be divided by their function into call-sounds and songs [25]. Calls are short isolated sounds that are associated with some function, e.g., the warning of a nearby predator. Songs are more spontaneous and consist of complex vocalizations that are mostly produced by male birds and are often associated with territorial defense and breeding. Therefore, many bird species sing only during the breeding season. Birds species are often recognized from their songs but can also be recognized from calls and hence no distinction between songs and calls is made in this work.

The hierarchical levels in bird songs are phrases, syllables, and elements [26]. A phrase is a series of syllables that occur in some particular order within the song. Syllables in the phrase are typically similar but can also be different. Syllables are constructed from elements that can overlap in time and frequency. The separation of elements can be both difficult and ambiguous and therefore the syllable is considered as the basic building block of bird sounds. For these reasons recognition of species is performed based on the classification of individual syllables.

The diversity of different sounds that birds can produce is large. Tonal and voiced bird sounds are characterized by fundamental frequency components and possibly their harmonics. The strongest component of a harmonic sound can also be some harmonic component [7]. Tonal and voiced sounds can also be modulated in both fre-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Lat. Abbr.	Common name	Individuals	songs	Syllables
CORRAX	Common Raven	7	33	128
CORNIX	Hooded Crow	8	61	391
PICPIC	Magpie	7	61	398
GARGLA	Eurasian Jay	9	80	260
ACRSCH	Sedge Warbler	6	20	514
ACRRIS	Marsh Warbler	4	20	819
FICHYP	Pied Flycatcher	8	38	365
PHYBOR	Arctic Warbler	6	85	764
PARATE	Coal Tit	9	42	580
PHYCOL	Common Chiffchaff	14	67	1104
		78	507	5323

Table 1. Birds considered in the current work. Columns: i) widely used abbreviation derived from the Latin name, ii) common English name, iii) the number of individuals from different species, iv) the total number of songs, and v) syllables.

quency and amplitude [27]. In addition to these factors, bird sounds can also be noisy, broadband or chaotic in structure [28].

Two groups of bird species are used in this work. The first group includes the six topmost species in table 1 and are the same species that were used in [2]. Many sounds of these birds are noisy and inharmonic. Four bird species are added to the second group of birds and the sounds of these additional species are largely different from the first group of birds since they are often tonal or harmonic albeit frequency modulated. All species with their number of recordings and total number of syllables are presented in table 1.

2.1. Segmentation

The bird sounds in this work have been collected from several different sources, which include different CD-collections and recordings from Finnish nature recordists. The original audio files also include other undesired environmental sounds. Together with varying background noise levels from one recording to another, segmentation turns out to be a challenging task. In this work segmentation of bird sounds is made using an iterative time domain algorithm that reduces the effect of different noise conditions. First, a decibel scale energy envelope of the signal is calculated using a 3 ms frame size and 50% overlap between frames. IThe initial noise estimate is the global energy minimum while the threshold for syllables is set to half of the noise level. During the following iterations the noise estimate is calculated from non-syllable frames and new syllables are searched for according to the new threshold. In figure 1 an example of the segmentation of a song by an Arctic Warbler (PHYBOR) is shown. As can be seen in the figure, the segmentation algorithm detects all eight syllables in the song, but detection of syllable boundaries is not always accurate. However, we assume that minor segmentation inaccuracies have only a small effect on the recognition accuracy since the proposed method collects structural information from very short time windows as will be seen in section 3.1.

3. METHODS

3.1. Permutation transformation coefficients

The parametric representation of bird syllables is based on the distribution of short temporal patterns in the signal. Patterns are represented using the permutation from ranking the signal values in short windows (window size n). The permutation of a set corresponds to an ordered arrangement of its elements, which are in this case the ranking of the signal amplitude values. The rankings of n different

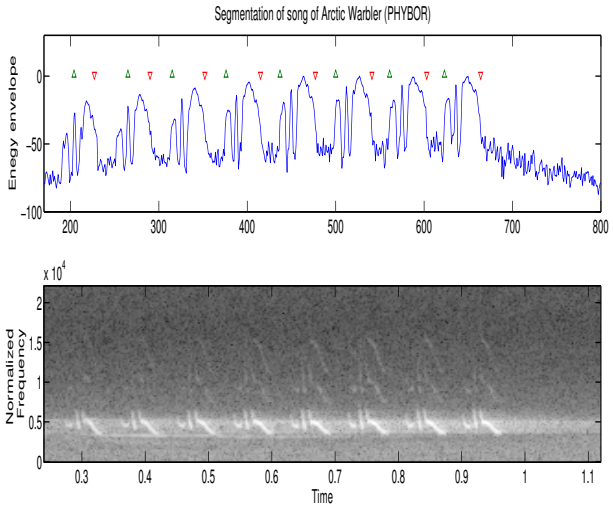

Fig. 1. Segmentation of a song from an Arctic Warbler. Curve in upper panel is the energy envelope while the lower panel is the spectrogram of the song. Triangles pointing up and down in the upper panel are segment boundaries.

signal values can be ordered in $n!$ different ways and thus $n!$ different possible permutations are possible. The ranking of sample values in each window is represented by a permutation code index. Indices from subsequent windows forms a symbolic sequence of permutation codes. Permutation codes can be seen as pointers to the corresponding ranking (permutation).

The permutation transformation starts by dividing a real-valued signal into permutation windows; a rank number replaces the signal values in each window. The permutation of a real valued signal $x(t)$ at location t is denoted by π_n^τ where τ is time delay between the signal samples (this delay is not necessarily equal to one; the samples picked up from the signal can be undersampled as well) and n is the size of the permutation window. Formally, the permutation of a time series is defined as

$$\pi_n^\tau = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ r_1 & r_2 & r_3 & \cdots & r_n \end{pmatrix} \quad (1)$$

satisfying

$$x_{t-r_1\tau} \geq x_{t-r_2\tau} \geq \cdots \geq x_{t-r_{n-1}\tau} \geq x_{t-r_n\tau} \quad (2)$$

where r is the ranking of the sample within the permutation window. Note that when, e.g., $\tau = 2$, the permutation window consists of every second sample in the original time domain signal. Equal amplitude values of the original signal within the permutation window are assumed to be very rare but if they occur we define $r_t > r_{t-1}$.

Each permutation pattern is coded with its symbolic reference and these placed in series form a permutation code sequence. Each value in this index sequence points to a permutation corresponding to the ranking or ordering of the present signal sample values within the permutation window. In this work, the permutation code corresponds to relative signal values of five values, thus there are $5! = 120$ possible permutations and also 120 possible values that the permutation code may obtain.

3.2. Permutation code pairs

Next, a permutation pair frequency (PPF) matrix is constructed from the permutation code sequence. It is a parametric representation for audio events that describes the histogram of the frequency of permutation transitions (permutation pairs) in the sequence. The PPF matrix can also be created using different time lags, which describe the time delay between two permutations (corresponding to the window hop size) in the permutation code sequence and delay (in samples) between permutation windows in the original signal. PPF matrices are then used as statistical models of the signals in the classification tests. The PPF matrix is formally defined as

$$A(\pi_i, \pi_j) = \frac{n(\pi_i, \pi_j)}{N} \tag{3}$$

where $n(\pi_i, \pi_j)$ is the total number of permutation code pairs at the chosen time lag of the sequence, and $N = \sum n(\pi_i, \pi_j)$ is a normalization coefficient.

In order to increase the robustness of the method, PPF matrixes are smoothed using a spatial filter. The PPF matrix cannot be smoothed directly because each element in the matrix refers to a pair of permutation codes and neighbouring elements in the PPF matrix are typically not neighbouring permutations. The spatial filter is applied to the PPF matrices through Kendall's distance d_k that can be solved from the corresponding Kendall's tau (correlation of two permutations) [14] defined as

$$\tau_k(\pi_i, \pi_j) = 1 - \frac{2d_k(\pi_i, \pi_j)}{d_{k_{max}}} \tag{4}$$

where d_k is Kendall's distance (also known as *Kendall's metric*). It is defined as the minimum number of local, elementary permutations needed to reorganize a permutation π_b to form the permutation π_a. The elementary permutation interchanges neighboring elements, e.g., $\{1234\} \rightarrow \{2134\}$ has $d_k = 1$. The term $d_{k_{max}}$ in Eq. (4) is the minimum number of elementary permutations needed to organize a permutation into its reverse permutation, e.g., $\{2134\} \rightarrow \{4312\}$ and it is given by

$$d_{k_{max}} = \frac{n(n-1)}{2} \tag{5}$$

The PPF matrix is smoothed by a weighted spatial filter by

$$G(\pi_i, \pi_j) = \sum_{d_k}^{d_{k_{max}}} w(d_k) A(\pi_i, \pi_j) \tag{6}$$

where $w(a)$ is a weighting function and d_k is Kendall's distance. The applied weighting function is:

$$w(d_k) = \begin{cases} 1 & d_k = 0 \\ \frac{1}{2d_k} & d_k = 1, 2, 3, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where k is the order of the spatial filter.

3.3. Classification

In this work the classification was done using a k-Nearest Neighbours (k-NN) method. The nearest neighbour of a test feature vector is a vector in the training data set that has the minimum distance to the test features. In the k-NN method, the test vector is assigned to the class which is most often represented in the k-nearest neighbour. Euclidean distance measures were used to calculate the distances between feature vectors.

The n-fold cross validation was used in splitting the segmented syllables from the database into training and testing data sets. With this method it is possible to use all available data for training and testing and still maintain the individual independence between training and testing data sets. Syllables left out from the training data set were selected so that those were never compared with syllables from the same recording (individual). Syllables from the same individual are most likely correlated and by including those in the training and testing data sets would have resulted in an overly optimistic error probability. Thus the number of folds varied between different species according to the number of recordings.

Classifications were also performed for entire songs by integrating the classification of the individual syllables in songs. Songs were assigned to the class where most of the syllables in the song were classified.

The PPF matrix is $n! \times n!$ elements in size, which becomes computationally infeasible for large values of n. In general, PPF-matrixes become sparse or its elements have very low values. The dimension of the PPF-matrix can be decreased by selecting only features that have the highest values for classification. Dimension reduction decreases the computational load and it may also increase the classification accuracy by removing noisy elements from the PPF-matrices.

4. RESULTS

Recognition performance was tested for two sets of bird species described in Section 2. The average classification rates for species classification are presented in figures 2 and 3. Results indicate the percentage of correctly classified syllables and songs. Recognition results show an increasing recognition rate when the spatial filter is applied, especially for the classification of songs or a series of call sounds. However, increasing the feature vector size does not always increase recognition accuracy, especially when smoothing is not applied. For comparison, the recognition rate was also calculated using MFCC-features using a 256 sample frame size with a 50% overlap of adjacent frames. Recognition rates for the first group of birds were 65% and 57% for correctly recognized syllables and songs (or series of call-sounds) while for the second set of birds the recognition results were 61% and 59%, respectively.

5. CONCLUSIONS

Birds can produce a myriad of different types of sounds. An optimal feature representation for each type of sound could be reasonable, but in automatic classification this would require the detection of the type of sound first. In this work a method that is able to represent different types of sounds was introduced. The classification accuracy of syllables was almost the same for permutation features and MFCC-features, but for song classification the permutation methods performed significantly better. This seems to indicate that PPF-matrixes are more robust in representing sounds that may have been incorrectly segmented.

PPF-matrix smoothing clearly increases recognition accuracy in both syllable and song classification. Smoothing reduces noise in PPF-matrixes and emphasizes the influence of an often existing temporal pattern pair. However, at the same time smoothing reduces the

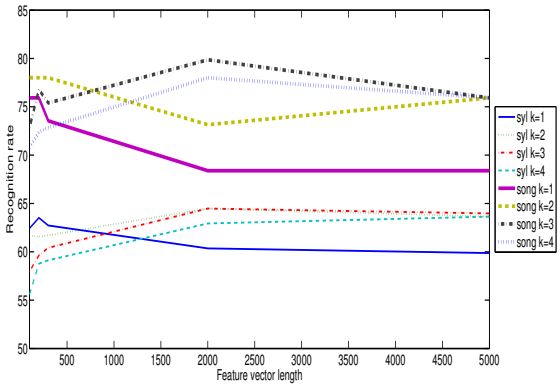

Fig. 2. Average recognition accuracy for the first set of bird species as a function of the size of the feature vector (PPF-matrix). Results indicate classification for syllables (syl) and songs or series of calls (song) varying the order of the spatial filter (k).

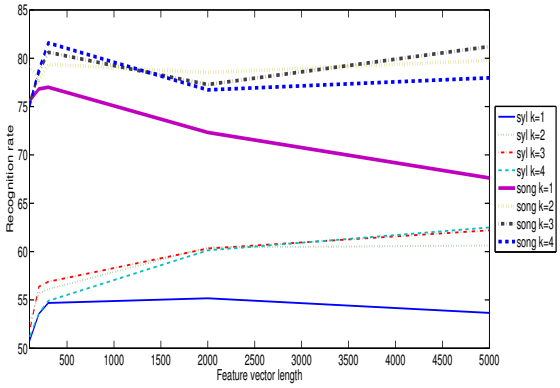

Fig. 3. Average recognition accuracy for the second set of bird species as a function of the size of the feature vector (PPF-matrix). Results indicate classification for syllables (syl) and songs or series of calls (song) varying the order of the spatial filter (k).

sparsity of the PPF-matrixes and increases computational complexity.

Overall classification results showed that essential information for accurate classification of different types of audio events can be found in short time windows. in future work more attention should be paid to finding only those temporal patterns that discriminate between different classes. This would probably increase classification accuracy while at the same time would decrease computational complexity.

6. REFERENCES

[1] 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. Am.*, vol. 103, no. 4, pp. 2185–2196, April 1998.

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

[3] 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, pp. Article ID 96706, 19 pages, 2006.

[4] C Lee, C Han, and C Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1541–1550, 2008.

[5] C-F Juang and T-M Chen, "Birdsong recognition using prediction-based recurrent neural fuzzy networks," *Neurocomputing*, vol. 71, no. 1-3, pp. 121 – 130, 2007.

[6] A. Selin, J. Turunen, and J. T. Tanttu, "Wavelets in automatic recognition of bird sounds," *EURASIP Journal on Signal Processing Special Issue on Multirate Systems and Applications*, vol. 2007, no. 1, 2007.

[7] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP 2004)*, Montreal, Canada, May 2004.

[8] Z Chen and R C Maher, "Semi-automatic classification of bird vocalizations using spectral peak tracks," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2974–2984, 2006.

[9] J R Heller and J D Pinezich, "Automatic recognition of harmonic bird sounds using a frequency tract extraction algorithm," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1830–1837, 2008.

[10] P Jančovič and M Köküer, "Automatic detection and recognition of tonal bird sounds in noisy environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, 2011.

[11] F Briggs, B Lakshminarayanan, L Neal, X Z Fern Z., R Raich, S J K Hadley, A S Hadley, and M G Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4640–4650, 2012.

[12] C Lee, S Hsu, J Shih, and C Chou, "Continuous birdsong recognition using gaussian mixture modeling of image shape features," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 454 – 464, 2013.

[13] E.D Chesmore, "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals," *Applied Acoustics*, vol. 62, no. 12, pp. 1359 – 1374, 2001.

[14] M Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, Jun 1938.

[15] Roberto Monetti, Wolfram Bunk, Thomas Aschenbrenner, and Ferdinand Jamitzky, "Characterizing synchronization in time series using information measures extracted from symbolic representations," *Phys. Rev. E*, vol. 79, pp. 046207, Apr 2009.

[16] L. Zunino, M. C. Soriano, I. Fischer, O. A. Rosso, and C. R. Mirasso, "Permutation-information-theory approach to unveil delay dynamics from time-series analysis," *Phys. Rev. E*, vol. 82, pp. 046212, Oct 2010.

[17] Erik M. Bollt, Theodore Stanford, Ying-Cheng Lai, and Karol Życzkowski, "Validity of threshold-crossing analysis of symbolic dynamics from chaotic time series," *Phys. Rev. Lett.*, vol. 85, pp. 3524–3527, Oct 2000.

[18] N Eklund and K Goebel, "Using neural networks and the rank permutation transformation to detect abnormal conditions in aircraft engines," *IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications 2005*, Jan 2005.

[19] C Bandt and B Pompe, "Permutation entropy: A natural complexity measure for time series," *Physical Review Letters*, vol. 88, pp. 174102, Jan 2002.

[20] K Keller, H Lauffer, and M Sinn, "Ordinal analysis of eeg time series," in *Advanced Methods of Electrophysiological Signal Analysis and Symbol Grounding*, J. Kurths C. Allefeld, P. beim Graben, Ed., chapter 7, pp. 109–119. Nova Science Publishers, 2008.

[21] G Ouyang, Z Ju, and H Liu, "Mutual information analysis with ordinal pattern for emg based hand motion recognition," *Intelligent Robotics and Applications*, vol. 7506, pp. 499–506, Jan 2012.

[22] Yinhe Cao, Wen-wen Tung, J. B. Gao, V. A. Protopopescu, and L. M. Hively, "Detecting dynamical changes in time series using the permutation entropy," *Phys. Rev. E*, vol. 70, pp. 046217, Oct 2004.

[23] D Arroyo, P Chamorro, J Amigó, and F Rodríguez. . . , "Event detection, multimodality and non-stationarity: Ordinal patterns, a tool to rule them all?," *The European Physical Journal Special Topics*, vol. 222, no. 2, pp. 457 – 472, 2013.

[24] S. Fagerlund and U K Laine, "Stop consonant recognition by temporal fine structure of burst.," in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, Florence, Italy, August 2011.

[25] J. R. Krebs and D. E. Kroodsma, "Repertoires and geographical variation in bird song," *Adv. Study Behav.*, vol. 11, pp. 143–177, 1980.

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

[27] G. J. L. Beckers, R. A. Suthers, and C. ten Cate, "Mechanisms of frequency and amplitude modulation in ring dove song," *The Journal of Experimental Biology*, vol. 206, no. 11, pp. 1833–1843, June 2003.

[28] N. H. Fletcher, "A class of chaotic bird calls," *The Journal of the Acoustical Society of America*, vol. 108, no. 2, pp. 821–826, Aug 2000.