Acoustic recognition of multiple bird species based on penalised maximum likelihood

Jancovic, Peter; Kokuer, M

DOI:
10.1109/LSP.2015.2409173

Acoustic recognition of multiple bird species based on penalised maximum likelihood

Peter Jančovič[1]* and Münevver Köküer[2,1]

***Abstract*—Automatic system for recognition of multiple bird species in audio recordings is presented. Time-frequency segmentation of the acoustic scene is obtained by employing a sinusoidal detection algorithm, which does not require any estimate of noise and is able to handle multiple simultaneous bird vocalisations. Each segment is characterised as a sequence of frequencies over time, referred to as a frequency track. Each bird species is represented by a hidden Markov model that models the temporal evolution of frequency tracks. The decision on the number and identity of bird species in a given recording is obtained based on maximising the overall likelihood of the set of detected segments, with a penalisation applied for increasing the number of bird models used. Experimental evaluations used audio field recordings containing 30 bird species. The presence of multiple bird species is simulated by joining the set of detected segments from several bird species. Results show that the proposed method can achieve recognition performance for multiple bird species not far from that obtained for single bird species, and considerably outperforms majority voting methods.**

***Index Terms*—bird species recognition, multiple bird species, maximum likelihood, penalisation, partition, BIC, hidden Markov models, frequency track, sinusoid detection**

I. Introduction

Automatic recognition of bird species from their vocalisations typically starts with segmenting the audio signal into isolated segments. In many works, this is performed based on estimating noise level and using an energy-based threshold decision, e.g., [1], [2]. An approach based on decomposing the acoustic scene into sinusoidal components was used in [3], [1], [4], [5], [6], [7], [8]. In [3], [4], this was performed using a threshold-based assessment of the continuity in frequency and amplitude of all the peaks in the short-time spectrum and included a manual or automated energy-based pruning of the obtained segments. We introduced in [9] a probabilistic method, which does not require any noise estimate, to detect only those spectral peaks that correspond to sinusoidal components. This was employed in [6], [7], [8] to obtain time-frequency segmentation and also representation of each segment as a temporal sequence of frequencies, which we refer to as frequency track. This feature representation, unlike spectral or cepstral coefficients extracted from a wide frequency bandwidth which has been used in some previous works, is suitable for representing birds vocalising concurrently. Several types of modelling approaches of bird vocalisations have been explored. The use of dynamic time warping [10], [11], [6] and hidden Markov models (HMMs) [1], [2], [7] is compelling as these allow to model the temporal evolution of sequences.

Recordings made in the field often contain vocalisations of multiple bird species. This issue has been addressed only in few recent works. The authors in [12] dealt with the problem of having the training data associated with multiple class labels by employing a multi-instance multi-label (MIML) approach. This required that each segment was represented as a single feature vector and as such did not allow for temporal modelling of segments. On a similar task and data, there have recently been two bird classification challenges. A summary paper presented in [13] provided only a brief description of the methods used by all contributors to the first challenge. The contributions to the second challenge are described in [14]. In both challenges, most of the contributions were based on using MIML approach or a variety of pattern recognition techniques that did not model the temporal evolution of segments.

This paper extends our recent work on HMM-based recognition of single bird species [7] to recognition of multiple species. An HMM modelling frequency track features is used to represent each bird species. Processing a given recording provides a set of variable length segments. The probability of each segment on each bird species HMM is calculated using the Viterbi algorithm. A method for finding the maximum likelihood of a set of segments for a given number of bird species models in an efficient way is proposed. This also allows to incorporate constraints on the minimum length of signal assigned to each species model. The decision on the number and identity of bird species is based on maximum likelihood subjected to a penalisation for increasing the number of models used. Experimental evaluations are performed on field recordings from [15]. Over 33 hours of field recordings from 30 bird species is processed. Experimental data with multiple bird species are created by artificially mixing detected segments of several bird species. Results indicate that the proposed method can achieve performance not far from that obtained when vocalisations of only single bird species are present and outperforms considerably majority voting methods.

II. Bird Species Recognition System based on HMM Modelling of Frequency Tracks

This section briefly summarises the developed bird species recognition system, with further details provided in [7].

A. Segmentation and Estimation of Frequency Tracks

The segmentation of the audio signal and estimation of frequency tracks is performed based on detecting sinusoidal

P. Jančovič is with the School of Electronic, Electrical and Systems Engineering, University of Birmingham, UK, E-mail: p.jancovic@bham.ac.uk.

M. Köküer is with the School of Digital Media Technology, Birmingham City University, UK, E-mail: munevver.kokuer@bcu.ac.uk.

components in the signal using a modified version of the method we introduced in [9] and this is summarised below.

The detection of sinusoidal components in a given signal frame is considered as a pattern recognition problem. Each peak in the magnitude short-time spectrum $|S(k)|$ of the signal is considered as a potential sinusoidal component. A peak at the frequency bin k_p is characterised by a feature vector $\mathbf{y}$, formed using M points of the short-time magnitude and phase spectrum around k_p. Specifically, $\mathbf{y}{=}(\mathbf{y}^1, \mathbf{y}^2)$, with $\mathbf{y}^1{=}(|S(k_p - M|/|S(k_p)|, \ldots, |S(k_p + M|/|S(k_p)|)$ and $\mathbf{y}^2{=}(\Delta\phi(k_p - M), \ldots, \Delta\phi(k_p + M))$, where $\Delta\phi(k)$ is the phase difference between the current and the previous signal frame. The distribution of the feature vector $\mathbf{y}$ is modelled using a multi-component Gaussian mixture. A model is obtained for spectral peaks corresponding to sinusoidal signals at various SNRs, denoted by λ_s, and noise, denoted by λ_n. A spectral peak is detected as a sinusoid if $p(\mathbf{y}|\lambda_s) > p(\mathbf{y}|\lambda_n)$.

The following parameter setup is used. The signal, sampled at 48 kHz, is divided into frames of 256 samples with a shift of 48 samples between the adjacent frames. Rectangular analysis window is used. The DFT size is set to 512 points. The parameter M is set to 6 frequency bins. Models consist of 32 Gaussian mixture components.

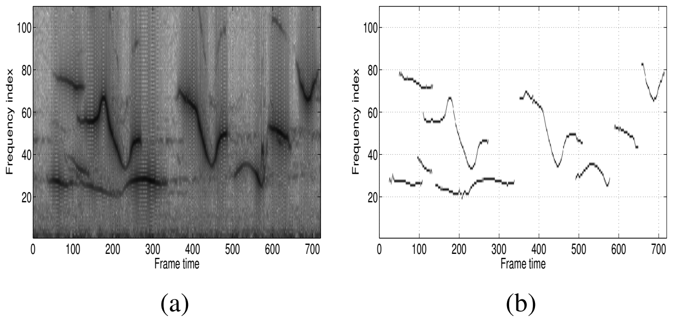
The above provides a set of detected sinusoidal components at each signal frame. In order to determine individual isolated segments, we assess the continuity of the detected components over time. Finally, we discard all segments whose length is too short (set here as 15 ms), whose median frequency is below 2 kHz, or whose average energy is 15 dB lower than the highest average segment energy in each recording. Each detected segment is represented as a sequence of the sinusoidal frequency values, which is referred to as frequency track.

An example of a spectrogram of an audio field recording containing concurrent vocalisations of two bird species and the obtained frequency tracks are depicted in Figure 1. It can be seen that detected frequency tracks correspond well to bird vocalisations.

B. HMM-based Modelling of Frequency Tracks

A model for each bird species is obtained by modelling the temporal evolution of frequency tracks. A single left-to-right (no skip allowed) hidden Markov model (HMM) is built for each bird species by training the model using the entire collection of detected segments from all training recordings of that species. To account for the variety of syllable patterns and the variations of individual instances of vocalisations, the probability density function at each HMM state is a multi-component Gaussian mixture. Gaussian distributions with a diagonal covariance matrix are used due to computational reasons, as is typically done in audio pattern processing.

The static frequency track features are appended by their temporal derivatives, referred to as delta and acceleration features, which capture local temporal dynamics. The included delta and accelleration features were calculated as in [16] with window set to 3 and 2, respectively. This resulted in a sequence of 3 dimensional feature vectors. Each HMM consists of 13 states, with each state output probability density function having 80 Gaussian mixture components.



(a) (b)

Fig. 1. An example of a spectrogram of an audio field recording (a) and the corresponding estimated frequency tracks (b).

III. Recognition of Multiple Bird Species

We consider the identification of bird species from a finite set of N bird species models based on a given 'utterance' of signal recording. For a given utterance, the segmentation and frequency track feature extraction step, as described in Section II-A, provides a set of R detected segments $O{=}\{O_s\}_{s=1}^{R}$, with each segment being represented by a sequence of features $O_s{=}(\mathbf{o}_s^1, \ldots, \mathbf{o}_s^{T_s})$, where T_s is the number of frames in the segment s. For each segment s, the likelihood $p(O_s|\lambda_{b_i})$ on each bird species HMM λ_{b_i} is calculated. We used the Viterbi algorithm to provide an approximation of this likelihood. Considering that a given utterance may contain vocalisations of one or more bird species, we are facing the problem of how to combine the scores obtained for each individual segment by each bird species model in order to obtain the decision on the number and the identity of the recognised bird species.

To indicate the difficulty of the problem we are dealing with, we analysed recognition results obtained for each individual segment when vocalisations of only single bird species are present. Figure 2 shows the histogram of the rank of the correct bird species model, where the statistics were collected over all the segments of all bird species. Results indicate that the correct model was ranked as the one achieving the highest probability for only around 28.5% of the segments.

A possible approach to deal with this score combination problem could be based on counting the number of segments classified to each bird species model. The identity of the recognised bird species would be obtained as the first top most used models, with some threshold-based decision needed to estimate the number of species present. We refer to this method as 'majority' combination. As this method disregards the differences in lengths of the segments, all segments would have an equal contribution to the final decision, which may not be desirable. A modified version of this majority combination method could count the accummulated length of all the segments classified to each bird species model. Although the majority approach may work well in some situations, the fact that it uses for each segment only the information about the single best model may prove problematic in more realistic scenarios when there may be a larger ambiguity in recognising individual segments. This is also the case here as demonstrated by the statistics presented in Figure 2. Such ambiguity typically increases with increasing the number of classes and with noise presence in input features.

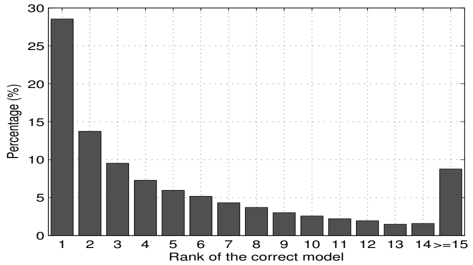


Fig. 2. Histogram of rank statistics of the correct bird species models, collected over all segments of all bird species.

The authors in [12] used the maximum posterior probability over all segments in the utterance as the utterance-level score for each bird species b_i, i.e., $\max_s P(b_i|O_s)$, and then used a threshold-based decision on this posterior probability to decide the number and identity of the bird species present. This approach is problematic because the decision is based on a single segment, which may accidentally be recognised incorrectly with a high posterior probability.

To deal with this score combination problem we propose a method that partitions the entire set of segments and assigns each partition to a bird species model in a way that the overall likelihood is maximised. Let us consider that the set of segments O is to be partitioned into K subsets, each subset being denoted as B_i, i.e., $O{=}\cup\{B_i\}_{i=1}^K$ and $B_i \cap B_j{=}\emptyset$. Each subset B_i is considered to belong to a different bird species b_i and the value of K corresponds to the number of bird species present. The aim is to calculate the maximum overall likelihood of the set O, denoted by $P^{(K)}$, i.e.,

$$P^{(K)} = \max_{\forall B_i; b_1,\ldots,b_K} \prod_{i=1}^{K} \prod_{O_s \in B_i} p(O_s|\lambda_{b_i}) \qquad (1)$$

where the maximisation is over all the possible partitionings of the set O into K subsets as well as over all the K partial permutations out of the total number bird species models. The direct implementation of Eq. 1 is computationally not feasible. For instance, the number of ways to partition the set O into K subsets is the Stirling number of the second kind and this increases exponentially with increasing both the number of segments R in the set O and the value of K, e.g., it is over 2.3 million when $R{=}15$ and $K{=}3$. However, we can split the maximisation in Eq. 1 into two steps. First, calculate the likelihood of the best partitioning of O for a given subset of models $\{b_1, \ldots, b_K\}$, which we denote by $P^{(K)}_{b_1,\ldots,b_K}$, and then maximise over all the K model combinations.

The likelihood $P^{(K)}_{b_1,\ldots,b_K}$ can be calculated simply by assigning each segment O_s, $s{=}1,\ldots,R$ to a model from the subset $\{b_1,\ldots,b_K\}$ that achieves the highest likelihood. If constraint on the minimum length of signal assigned to each bird species is required, binary linear programming can be employed. This finds the values of binary variables, which express the assignment of each segment to each model. The cost function to optimise is the summation of terms consisting of each of the binary variable multiplied by the corresponding log-likelihood of the segment on the model. The length of each segment is used to formulate the minimum length criteria constraint plus constraints to ensure that each segment is assigned to only one model are used. Alternatively, we observed that the following procedure can find $P^{(K)}_{b_1,\ldots,b_K}$, or its close approximation, in a faster way. This procedure needs to be performed for all model permutations. First, for each segment in O, calculate $\log(p(O_s|\lambda_{b_1})){-}\max_{i=2}^K \log(p(O_s|\lambda_{b_i}))$. Then, a subset from O where the above difference is positive is assigned to B_1, subject to leaving enough segments for the remaining $K{-}1$ bird models. If the minimum signal length constraint is not satisfied, next segments with the above difference being least negative are included in B_1. This procedure is repeated with the remaining subset of segments $O \backslash B_1$ and models until the set O is partitioned into K subsets.

The likelihood $P^{(K)}$ is obtained by maximising over the likelihoods $P^{(K)}_{b_1,\ldots,b_K}$, calculated for all K model combinations (or permutations), using one of the above ways.

The final step is to select parameter K, i.e., number of bird species present in signal. This can be performed based on principles used in model selection research, e.g., Bayesian information criterion (BIC). Increasing the value of K effectively means that we are allowing a more complex model to fit the data. As such, the likelihood $P^{(K)}$ needs to be subjected to a penalisation. The estimated K^* can be obtained as

$$K^* = \arg \max_{K \in <1,\ldots,K_{max}>} \log P^{(K)} - \alpha(K) \qquad (2)$$

and the set of recognised bird species $\{b_1,\ldots,b_K\}^*$ is then obtained as corresponding to $P^{(K^*)}$. The value of the penalisation $\alpha(K)$ increases with increasing the number K of models used. Various ways of setting the penalisation function have been proposed, e.g., [17], [18]. As we may in general have different amount of signal being assigned to each model, we based the penalisation $\alpha(K)$ on segmental BIC [18] and calculated it as $\lambda C^{(K)} \sum_{i=1}^{K} \log T(i)$, where $C^{(K)}$ denotes the number of parameters of using K models, $T(i)$ is the number of signal frames assigned to the i^{th} model, and λ is a tuning factor. We observed that using a different value for the tuning factor λ for each K provided slightly better performance than using a fixed value. Values of λ used during testing are found based on the best performance obtained on simulated mixture using the training data.

IV. Experimental Evaluations

A. Data Description and Experimental Setup

Experimental evaluations were performed using audio field recordings from [15], collected over several decades, mostly in the western United States. Each bird species contains several audio files, each file being typically several minutes long. There is no annotation of the recordings other than the label indicating the single bird species name. As these are field recordings, there are sometimes vocalisations of other birds and animals. Data from a set of randomly chosen 30 bird species was used. The list of bird species is given at [19]. This contained in total over 33 hours of audio recordings. Each recording was split into training and testing part in proportion of two to one, respectively. The data used for testing was further split into utterances, where each utterance consisted of signal containing approximately a given length of detected segments. In total, there was 2126 utterances. The utterances of one, two, and three seconds of the detected segments contained by average 13, 20, and 40 segments, respectively.

In order to conduct methodological evaluations of the proposed score combination method, vocalisations of multiple bird species were created by randomly mixing set of detected segments from several bird species. This effectively means that the segment detection method is considered to detect the same set of segments it would have detected if recordings contained individual bird species vocalisations.

Performance is evaluated in terms of recognition correctness, $100{\cdot}N_c/N$, and recognition accuracy, $100{\cdot}(N_c{-}N_i)/N$, where N_c, N_i and N is the number of correctly recognised, inserted and total number of bird species in recordings.

B. Experimental Results

First, we report performance for the case when only a single bird species is present. The developed recognition system achieved bird species recognition correctness of 92.0%, 88.8% and 83.3% when using, respectively, utterances containing three, two and one seconds of the detected signal.

Now, we present results with multiple bird species present. First, we consider that there are separately one, two, or three bird species present, each species with 3 seconds of the detected segments and we assume that the number of bird species is known. Evaluations of the proposed maximum-likelihood method without and with constraints on the minimum length of the signal assigned to each bird species model are performed. The latter uses constraint matching the length of the bird signal present, i.e., 3 seconds here, and as such, this represents an idealised best performance the method can achieve. Experiments were also performed using the majority voting method, either based on the number of segments or the cummulated length of segments. Results are presented in Table I. It can be seen that the proposed maximum likelihood method obtains considerably better performance than majority-based methods in all cases of 1, 2 and 3 bird species present. The idealised case of incorporating strong constraints on minimum signal length achieves relatively small performance improvements in comparison to using no constraints.

TABLE I
Bird species recognition correctness (%) as a function of the number of bird species present when each species contains 3 seconds of the detected signal.

Number of bird species present	Score combination method		
	majority		max-likelihood
	count	length	with constraint: yes / no
1 species	63.1	63.7	92.0 / 92.0
2 species	54.9	61.4	84.7 / 81.2
3 species	51.7	61.3	77.6 / 72.5

Next, we assess how the presence of different length of bird vocalisations in the mixture affects the recognition performance. These experiments are performed using two bird species present, with the vocalisation length being 3 seconds for the first species and varying from 1 to 3 seconds for the second species. Results are presented in Table II. It can be seen that the first bird species are recognised in all cases with a similar correctness, in the range from 83.5% and 86.1%. The correctness of recognising the second bird species decreases only little when 2 seconds of vocalisation is present and drops down more when only 1 second of vocalisation is available.

TABLE II
Bird species recognition correctness (%) when two bird species are present, one with 3 seconds and the other with various length of the detected signal.

Length of 2^{nd} bird species (sec)	Rec. Corr. (%)	
	species 1	species 2
3	84.4	84.7
2	83.5	80.1
1	86.1	65.0

Finally, we present experiments demonstrating the performance when also the number of bird species is estimated. For these experiments, the number of bird species in the data was chosen randomly in the range from 1 to 3 and the data contained vocalisations of around 3 seconds of the detected signal as follows: either 3 sec from 1 bird species, 1.5 sec from 2 bird species, or 1 sec from 3 bird species. Constraint on the minimum length of the signal assigned to a bird species model was set to 1 second. Results are presented in Table III. It can be seen that recognition correctness/accuracy of 78.1% is achieved when the number of bird species is known and this drops to 72.5% and 69.4%, respectively, in the case of automatically estimating the number of species.

TABLE III
Bird species recognition performance when one, two, or three bird species are present in a given utterance of 3 seconds of the detected signal.

Number of bird species	Rec. Corr. (%)	Rec. Acc. (%)
Known a-priori	78.1	78.1
Estimated	72.5	69.4

V. Conclusion

In this paper, we presented an automatic system for recognition of multiple bird species. The system employed a method for detection of sinusoids to obtain time-frequency segmentation of acoustic signal and extract frequency track features to characterise each segment. Each bird species was represented by a hidden Markov model modelling frequency track features. In a given recording, a set of segments is detected. The recognition decision on the number and identity of bird species was performed based on finding a subset of models that achieved maximum likelihood aggregated over all the segments. An efficient method for finding the maximum likelihood, which also allowed to incorporate constraints in the decision, was proposed. Based on the principles of Bayesian information criterion, the obtained likelihood was penalised according to the number of models used. Experimental results demonstrated that the proposed method performed well and considerably outperformed majority voting approach.

Acknowledgment

Data provided by Borror Laboratory of Bioacoustics, The Ohio State University, Columbus, OH, all rights reserved.

References

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

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

[3] 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.

[4] Jason R. Heller and John D. Pinezich, "Automatic recognition of harmonic bird sounds using a frequency track extraction algorithm," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, 2008.

[5] P. Jančovič and M. Köküer, "Automatic detection and recognition of tonal bird sounds in noisy environments," *EURASIP Journal on Advances in Signal Processing*, pp. 1–10, 2011.

[6] P. Jančovič, M. Köküer, M. Zakeri, and M. Russell, "Unsupervised discovery of acoustic patterns in bird vocalisations employing DTW and clustering," *European Signal Processing Conference (EUSIPCO), Marrakech, Morocco*, Sept. 2013.

[7] P. Jančovič, M. Köküer, and M. Russell, "Bird species recognition from field recordings using HMM-based modelling of frequency tracks," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy*, pp. 8307–8311, May 2014.

[8] P. Jančovič, M. Zakeri, M. Köküer, and M. Russell, "HMM-based modelling of individual syllables for bird species recognition from audio field recordings," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Brisbane, Australia*, p. accepted, April 2015.

[9] P. Jančovič and M. Köküer, "Detection of sinusoidal signals in noise by probabilistic modelling of the spectral magnitude shape and phase continuity," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Prague, Czech Republic*, pp. 517–520, May 2011.

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

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

[12] F. Briggs, B. Lakshminarayanan, L. Neal, X.Z. Fern, 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.

[13] F. Briggs, R. Raich, Z. Lei, K. Eftaxias, and Y. Huang, "The Ninth Annual MLSP Competition: Overview," in *IEEE Int. Workshop on Machine Learning for Signal Processing*, Sept. 2013.

[14] H. Glotin, Y. LeCun, S. Mallat, T. Artieres, O. Tchernichovski, and X. Halkias, "Neural information processing scaled for bioacoustics," *http://sabiod.univ-tln.fr/nips4b/*, 2013.

[15] "Borror Laboratory of Bioacoustics," *The Ohio State University, Columbus, OH*, www.blb.biosci.ohio-state.edu.

[16] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book. V2.2*, 1999.

[17] M. Lavielle, "Using penalized contrasts for the change-point problem," *Signal Processing*, vol. 85, pp. 1501–1510, 2005.

[18] T. Stafylakis, V. Katsouros, and G. Carayannis, "The segmental bayesian information criterion and its applications to speaker diarization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 857–866, Oct. 2010.

[19] "List of bird species used in paper 'Acoustic recognition of multiple bird species based on penalised maximum likelihood' submitted to IEEE Signal Processing Letters," http://www.eee.bham.ac.uk/jancovic/research/Data.htm.