# HMM-BASED MODELLING OF INDIVIDUAL SYLLABLES FOR BIRD SPECIES RECOGNITION FROM AUDIO FIELD RECORDINGS

*Peter Jančovič*[1] **, Masoud Zakeri*[1]*, Münevver Köküer*[2,1] *and Martin Russell*[1]

[1] School of Electronic, Electrical & Systems Engineering, University of Birmingham, UK
E-mail: {p.jancovic, mxz848, m.kokuer, m.j.russell}@bham.ac.uk
[2] Faculty of Computing, Engineering & the Built Environment, Birmingham City University, UK
E-mail: munevver.kokuer@bcu.ac.uk

## ABSTRACT

This paper presents an automatic system for recognition of bird species from audio field recordings. The acoustic signal is first segmented into isolated time-frequency segments, each corresponding to an individual detected sinusoidal component. Each segment is represented by a temporal sequence of the frequency values of the detected sinusoid, referred to as frequency track. Hidden Markov models (HMMs) are employed to model the temporal evolution of frequency track features. Individual syllables of bird vocalisations are discovered using an unsupervised method based on dynamic time warping and agglomerative hierarchical clustering. The outcome of this is then employed to create individual HMMs for syllables of each species. Experiments are performed on over 33 hours of field recordings, containing 30 bird species. Evaluations demonstrate that the use of individual syllable HMMs provides over 40% error rate reduction over the use of single HMM for each bird species of the same complexity. The syllable HMM-based system recognises bird species with accuracy over 95% using 3 seconds of detected signal.

***Index Terms***— bird species recognition, hidden Markov model, syllable, unsupervised clustering, DTW, segmentation, frequency track, sinusoid detection

## 1. INTRODUCTION

Automatic processing of bird acoustic signals usually starts with segmentation of the audio signal into isolated segments. An automated segmentation has been performed using an energy-based threshold decision, with threshold set based on estimating noise level, e.g., [1] or by decomposing the acoustic scene into sinusoidal components [1, 2, 3, 4, 5, 6]. The works in [1, 2, 3] employed the sinusoidal decomposition method proposed in [7]. We proposed in [8] a probabilistic method for the detection of sinusoids and employed this in our recent studies in bird pattern processing [4, 5, 6, 9] and also here.

Several types of feature representations and modelling approaches of bird acoustic signals have been explored. Many previous studies, inspired by features used in the field of speech processing, employed Mel-frequency cepstral coefficients (MFCC), e.g., [10, 11, 1, 12]. Since the conventional MFCCs capture the entire frequency band, they are prone to background noise and presence of other birds/animals concurrently vocalising in other frequency regions. A set of statistical descriptors to characterise the detected spectro-temporal segments were used in [1, 2, 3, 13]. Although this provides a single feature vector, usually of a low dimensionality, it may not be able to describe well more complex types of syllables and may be susceptable to any variations in segmentation. In few other studies, including our recent works, [1, 14, 4, 5, 6, 9], the segments were obtained based on sinusoidal detection and then represented as a temporal sequence of frequencies, which we here refer to as frequency track. The frequency track features, if extracted well, have a good potential, especially, in processing field recordings of bird vocalisations which usually contain various background noise and often also other birds/animals vocalising concurrently. We demonstrated in [4] that frequency track features obtained considerable performance improvements for recognition of bird sounds in noisy background conditions than the use of MFCCs. The most commonly used modelling approaches include dynamic time warping [15, 10], Gaussian mixture modeling [1, 4], and hidden Markov models (HMMs) [1, 14, 16, 6].

In this paper, we extend our study of automatic bird species recognition using field recordings by incorporating modelling of individual bird syllables. Audio signal is first segmented using an improved version of the method introduced in [8] and each segment is represented using frequency track features. The temporal evolution of these features is modelled using hidden Markov models. Unlike our previous work in [6], in which all syllables were modelled using a single HMM, in this paper we investigate the modelling of each individual syllable. Since there is no syllable-level label information available with the data, we first employ an unsupervised clustering approach as presented in [5] to discover a set of syllables for each species. Recognition is performed using the Viterbi algorithm to calculate probability of each detected segment on each bird species model and aggregating the probabilities from all segments within a given duration of the signal. Experimental evaluations are performed on field recordings provided by Borror Laboratory of Bioacoustics [17]. The syllable-based system achieved over 95% bird species recognition accuracy, which is over 40% error rate reduction in comparison to the single model system of the same complexity.

## 2. SEGMENTATION AND ESTIMATION OF FREQUENCY TRACKS

The segmentation of the audio signal and estimation of frequency tracks is performed based on detecting sinusoidal components in the signal. This is performed using the method we introduced in [8], with further modifications, and this is summarised below.

The detection of sinusoidal components is tackled as a pattern recognition problem. It is performed on a signal frame basis. Each peak in the magnitude spectrum of signal frame is considered as a potential sinusoidal component. A given spectral peak $k_p$ is characterised by a feature vector $\mathbf{y}$, which is formed using $M$ points of the short-time magnitude and phase spectrum around the peak. Specifi-

cally, $\mathbf{y}=(\mathbf{y}^1, \mathbf{y}^2)$, where $\mathbf{y}^1=(|S(k_p - M|/|S(k_p)|, \ldots, |S(k_p + M|/|S(k_p)|)$ and $\mathbf{y}^2=(\Delta\phi(k_p - M), \ldots, \Delta\phi(k_p + M))$. The $\Delta\phi(k)$ is the phase difference between the current and the previous signal frame, with the shift between signal frames being accounted for. The distribution of the multivariate feature vector $\mathbf{y}$ is modelled using a multi-component Gaussian mixture. A model is obtained for spectral peaks corresponding to noise, denoted by $\lambda_n$, and to sinusoidal signals, denoted by $\lambda_s$, at various SNRs. The decision whether a spectral peak corresponds to a sinusoidal signal or not is based on the maximum likelihood criterion, i.e., the peak is detected as a sinusoid if $p(\mathbf{y}|\lambda_s) > p(\mathbf{y}|\lambda_n)$.

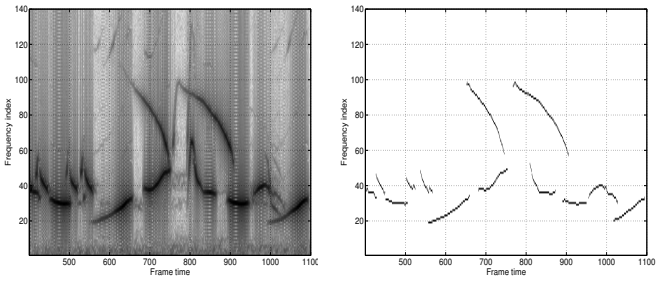
The following parameter setup is used. The signal, sampled at 48 kHz, is divided into frames of 256 samples with a shift of 48 samples between the adjacent frames. Rectangular analysis window is used and the DFT size is set to 512 points, i.e., the signal is appended by 256 zeros in order to provide a finer sampled DFT spectrum. The parameter $M$ is set to 6 frequency bins. The training of the models of sinusoidal signals was performed using simulated sinusoids, with a range of linear frequency modulation. The models consist of 32 Gaussian mixture components.

The above provides a set of detected sinusoidal components at each signal frame. This can be considered as an initial segmentation of the acoustic scene. The following steps are performed to further refine this segmentation result. We first discard all segments of a very short length, specifically those of less than 4 frames, considering that these were detected accidentally by error. Then, interpolation between the beginning and the end point of two detected segments is performed for all segments which are separated by up to two frames and two frequency bins from each other. This was performed in order to avoid accidental split of a segment due to a missed detection of few frequency bins. After this, we discard all segments whose length is less than 14 frames, as it is unlikely to have bird vocalisations of such short lengths. Since we are using field recordings, there are co-vocalisations of other birds and animals present in the background. However, there is no label information available that would indicate the vocalisations of the bird of interest. In order to avoid these background co-vocalisations, we consider that vocalisations of bird species being recorded are of a higher energy than any other present co-vocalisations. Thus, we discard all segments whose average energy is 15 dB below the highest average segment energy in each recording. Finally, we discard all segments whose median frequency is below 2 kHz. This low frequency region does not correspond to bird vocalisations in our data and this is performed to avoid detection of segments corresponding to human speech which is also present in the recordings.

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

## 3. HMM-BASED BIRD SPECIES RECOGNITION SYSTEM

The segmentation and frequency track feature extraction step, as described in Section 2, provides for a given audio recording a set of detected segments. A model of each bird species is obtained based on modelling the temporal evolution of frequency tracks of detected segments using a left-to-right, no skip allowed, HMM. The HMM state output probability density functions are using Gaussian distribution(s) with a diagonal covariance matrix. The following sections describe the baseline model, which employs only a single HMM for each bird species, and the proposed model, which employs individual HMMs for syllables for each bird species.




(a) (b)

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

### 3.1. A single model for each bird species

We use the model presented in our previous work in [6] as the baseline model in this paper. This consists of a single HMM built for each bird species by training the model using the entire collection of the detected segments from all training recordings of that species. To account for the variety of syllable patterns and variations of individual instances of vocalisations, the probability density function at each HMM state is modelled with a mixture of Gaussians.

### 3.2. Modelling of individual syllables for each bird species

Instead of using a single model for each species, here we propose to obtain an individual HMM to model each type of bird syllables. This would be straightforward if the label information for syllables was available or if the set of syllable patterns produced by each bird species was known. However, none of these is available for our data and is unlikely to be publicly available in general. As such, we are facing the problem of how to train the individual syllable HMMs in an unsupervised manner. The approach we have taken to deal with this problem is described in the following subsections.

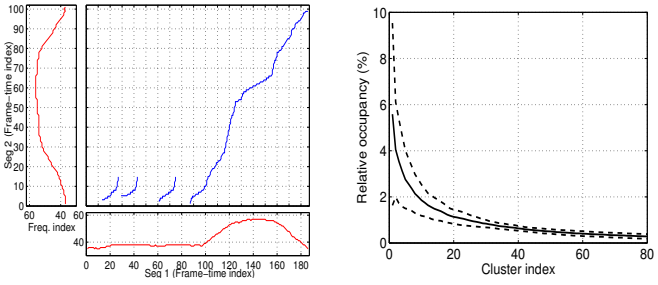
#### *3.2.1. Unsupervised discovery of syllable patterns*

There could be several ways to approach the problem of unsupervised discovery of bird vocalisation patterns. In this paper, we employed an extension of the method introduced in [5], where it was evaluated on a single bird species only. This consists of first searching for matches, possibly partial, between each pair of detected segments by employing a modified dynamic time warping (DTW) algorithm, and then using the obtained similarity values in hierarchical clustering.

Unlike conventional DTW, which calculates similarity of whole sequences, the modified DTW allows to search for partial and multiple matches within segments. This is useful for suppressing the effect of possible detection errors at the beginning and end of segments and also for dealing with situations when the detected segment actually contains several syllables. The modified DTW is implemented by calculating several DTW searches in parallel, each considering a different starting point on one of the sequence and allowing the start anywhere on the other sequence [5]. For a given pair of segments, the outcome of the DTW search is a set of partially matching paths. Out of these, only the match with the highest similarity score is used in further stages. The similarity score is calculated based on a combination of the cummulative distance of the DTW path match, length of the matching path and the ratio of the length of the matching path

to the total length of the segment. After processing of all the detected segments, we have the similarity score for all the segment pairs. This is then used in agglomerative hierarchical clustering approach to arrive at a set of syllable clusters. Initially, each segment is assumed to be a distinct cluster. At each clustering level, two clusters with the highest similarity score are merged into a new joint cluster. The similarity score is caclulated as the average similarity score over all the segments from each of the clusters. Only the clusters for which all the segments assigned to them so far have an overlap in path with each other are being merged.

An example of result obtained by the partial DTW search on a pair of detected segments from field recordings is given in Figure 2 (a). It can be seen that 4 partial matches in the given two sequences were found, with the match in bold to be used in further stage. Figure 2 (b) depicts statistics of the clustering procedure outcome. It shows the relative occupancy of each cluster (in decreasing order), averaged over all the bird species, with the standard deviation above and below. It is observed that the average relative occupancy above 0.5% was obtained for the first 50 clusters.




(a) (b)

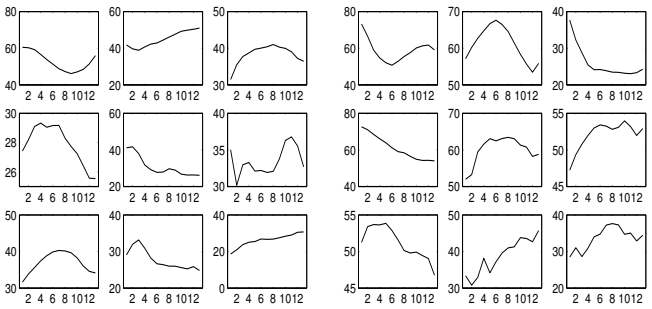
**Fig. 2**. An example of the output of the partial DTW search (a). Mean of the relative occupancy of the first 80 clusters (ordered from highest to lowest) calculated over all bird species as the output of the clustering procedure (b).

### *3.2.2. Modelling individual bird syllables*

The outcome of the partial DTW search and hierarchical clustering is a set of clusters of vocalisation patterns for each bird species. Consequently, this also provides the label information for each detected segment of the data. Using this label information, we can train the individual syllable HMMs of each species. As the obtained clusters of vocalisation patterns are expected to be homogenous, the state output probability density function (pdf) of each individual syllable HMM consists only of a single Gaussian distribution. As we use only a given number of clusters based on their occupancy, there will be remaining clusters whose segments are not assigned to any of the selected clusters. Thus, in addition to the individual syllable HMMs, we also have a single HMM to model all these remaining segments. To cover the variety of these remaining segments, the state pdf of this model consists of several Gaussian mixture components. An example of the state output pdf of nine trained individual syllable HMMs of two bird species is depicted in Figure 3. It can be seen that each model provides a distinctive pattern.

## 3.3. Recognition of bird species

We consider the identification of bird species from a finite set of species based on an utterance of test signal of a given length.



(a) (b)

**Fig. 3**. An example of the mean values of the state output Gaussian pdf, modelling frequency track features, for nine trained syllable HMMs of bird species *House Finch* (a) and *Northern Cardinal* (b). The x- and y-axis denotes the HMM state and frequency index, respectively.

For a given utterance of audio recording, the segmentation and frequency track feature extraction step, as described in Section 2, provide 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 segment $s$. We treat each detected segment individually. The Viterbi algorithm is used to calculate an approximation of the probability of each segment $s$ on each bird species model $\lambda_b$, i.e., $p(O_s|\lambda_b)$. In the case of recognition based on individual syllable models, the probability is calculated on each syllable model and the maximum is taken. Considering that vocalisations of only a single bird species are present in the signal, we can calculate the probability of the utterance being produced by each bird species $b$ as the product of the individual segment probabilities, i.e., $p(O|\lambda_b)=\prod_{s=1}^{R} p(O_s|\lambda_b)$, and obtain the recognised bird species as $b^*=\arg\max_b p(O|\lambda_b)$. To account for a possible presence of other birds/animals which do not exist in our bird species vocabulary, i.e., outliers, we explored the calculation of the overall probability $p(O|\lambda_b)$ by ommitting from the product those segments whose average frame probability was below a given value on all models, which is similar to approach presented in [18], but no improvements were observed.

# 4. EXPERIMENTAL EVALUATIONS

## 4.1. Data description

Experimental evaluations were performed using field recordings from [17]. These are recordings in real world natural habitats of birds, collected over several decades, mostly in the western United States. The recordings are encoded as mono 16-bit wav files, with sampling rate of 48 kHz. There are several files for each bird species, and each file is typically between one to ten minutes long. As these are field recordings, the audio contains also background environmental noise, vocalisations of other birds/animals and human speech. For each recording, there is a label indicating the single bird species vocalising but there is no label information that would indicate the start and end times of each bird vocalisation.

From the available data, we chose randomly a set of 30 bird species. In total, we used over 33 hours of audio recordings, with between 28 to 95 minutes per bird species. The total length of detected and used frequency track segments was 2.2 hours. For experimental

evaluation, each recording is 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.

### 4.2. Experimental setup

The frequency track features extracted as presented in Section 2 provide the frequency value at each frame time but do not include any information about how the frequency track evolves over time. In order to include local dynamic information, we calculated temporal derivatives of the frequency track features, referred to as delta and acceleration features. These were obtained as in [19] with the window set to 3 and 2, respectively, and added to the frequency track features, resulting in 3 dimensional feature vectors. In all experiments, the number of states in HMMs was set to 13, which reflects the minimum allowed length of the detected segment and results observed in our previous experiments.

### 4.3. Experimental results

First, we present evaluations of the baseline HMM-based bird species recognition system, i.e., system which uses only a single model for each bird species. Results achieved as a function of the number of Gaussian mixture components at each HMM state are presented in Table 1. These results are obtained using utterances of length of only 1 second. It can be seen that the recognition accuracy is quite high for 10 mixture components. It keeps increasing little but steadily as the number of mixture components increases up to 80 and then flattens.

**Table 1**. *Bird species recognition accuracy (RA) obtained by the baseline HMM-based system using a single model for each bird species with a given number of mixture components per state. Utterances of 1 second length used.*

| | Number of mixture components per state | | | | | | |
|---|---|---|---|---|---|---|---|
| | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
| RA (%) | 75.1 | 78.4 | 80.9 | 82.2 | 82.4 | 83.3 | 83.3 |

Now, we present results obtained with the proposed recognition system, having a set of individual syllable models for each bird species. In order to have comparable conditions of the new models to baseline model presented above, we first set the number of individual models to be the same for all bird species. Experiments are performed with several different values for the number of individual models. As stated in Section 3.2.2, the state output pdf of each individual syllable HMM consisted only of a single Gaussian distribution. The state output pdf of the additional model, used to cover the segments not assigned to any of the individual models, used mixture of Gaussians with the number of components set such that the total number of parameters is the same across all the proposed and baseline models. The baseline model with 80 Gaussian mixture components per state is used. Results are presented in Table 2. It can be seen that the proposed model achieves considerable recognition accuracy improvements over the use of single HMM, while the complexity of the models is the same. As the number of used individual syllable models increases, the recognition accuracy increases.

Finally, we performed experiments when the number of individual models used for each bird species is not the same but it is decided based on the relative occupancy of each cluster, with the threshold set in a way that the number of individual models is 60 in average over all bird species. The state output pdfs contained a single Gaussian for each individual HMM and 20 Gaussian mixture components for the additional model. These experiments were performed using different length of the detected signal for testing, specifically, varying from 1 second to 3 seconds. The results achieved by the baseline and proposed models of the same complexity are presented in Table 3. It can be seen that using the varying number of individual models further improved the performance from 89.5% (as presented in Table 2) to 90.2% when using the utterance of 1 second. The recognition accuracy when using 2 seconds is considerably higher than 1 second and smaller improvement is seen for 3 seconds long utterances. In all cases, the use of the proposed individual syllable models showed significant recognition accuracy improvements, with the error rate reduction between 41.3% to 50%.

**Table 2**. *Bird species recognition accuracy (RA) obtained by the HMM-based system employing individual models of bird syllables. Utterances of 1 second length used.*

| | Single HMM | Syllable HMM | | | | | |
|---|---|---|---|---|---|---|---|
| | | Number of individual syllable models | | | | | |
| | (Base) | 20 | 30 | 40 | 50 | 60 | 70 |
| RA (%) | 83.3 | 85.6 | 87.9 | 88.1 | 89.2 | 89.5 | 89.8 |

**Table 3**. *Bird species recognition accuracy and error rate reduction obtained by the baseline single and individual syllable HMM-based recognition system when using different length of detected signal.*

| Utterance length (sec) | Rec. Acc. (%) | | Error Rate Reduction (%) |
|---|---|---|---|
| | Single HMM (Base) | Syllable HMM | |
| 1 | 83.3 | 90.2 | 41.3 |
| 2 | 88.8 | 94.4 | 50.0 |
| 3 | 92.0 | 95.5 | 43.8 |

## 5. CONCLUSION

We presented in this paper an automatic system for recognition of bird species from audio field recordings based on modelling individual syllables of species. The proposed system employed a method for detection of sinusoidal components to decompose the acoustic scene into isolated time-frequency segments. Each segment was represented as a temporal sequence of the detected sinusoid frequency, referred to as frequency track. The temporal evolution of frequency track features was modelled by employing hidden Markov models (HMMs). We developed a baseline system that used only a single HMM for each bird species and the proposed model that used several HMMs to model individual syllables. Unsupervised clustering was employed to discover the set of bird syllable patterns. Experimental evaluations were performed on field recordings provided by the Borror Laboratory of Bioacoustics. Experimental results demonstrated that the proposed individual syllable HMM-based system provided over 40% bird species recognition error rate reduction over the single HMM-based system of the same complexity.

**Acknowledgement**

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

## 6. 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustic, Speech, and Signal Proc.*, vol. 34, pp. 744–754, Aug. 1986.

[8] 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.

[9] P. Jančovič and M. Köküer, "Acoustic recognition of multiple bird species based on penalised maximum likelihood," *IEEE Signal Processing Letters*, under review, Feb. 2015.

[10] 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.

[11] C. Kwan, K.C. Ho, G. Mei, Y. Li, Z. Ren, R. Xu, Y. Zhang, D. Lao, M. Stevenson, V. Stanford, and C. Rochet, "An automated acoustic system to monitor and classify birds," *EURASIP Journal on Applied Signal Processing*, vol. 2006, no. 3, pp. Article ID 96706, 2006.

[12] C.H. Lee, Y.K. Lee, and R.Z. Huang, "Automatic recognition of bird songs using cepstral coefficients," *Journal of Information Technology and Applications*, vol. 1, no. 1, pp. 17–23, May 2006.

[13] 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.

[14] 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.

[15] 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.

[16] W. Chu and D.T. Blumstein, "Noise robust bird song detection using syllable pattern-based hidden markov models," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic*, pp. 345–348, May 2011.

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

[18] P. Jančovič, M. Köküer, and F. Murtagh, "Reliability-based estimation of the number of noisy features: Application to model-order selection in the union models," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Hong-Kong, China*, vol. I, pp. 416–419, 2003.

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