BIRD SPECIES RECOGNITION USING HMM-BASED UNSUPERVISED MODELLING OF INDIVIDUAL SYLLABLES WITH INCORPORATED DURATION MODELLING

Peter Jančovič, Münevver Köküer, Masoud Zakeri and Martin Russell*

School of Electronic, Electrical & Systems Engineering, University of Birmingham, UK
E-mail: {p.jancovic, m.kokuer, mxz848, m.j.russell}@bham.ac.uk

ABSTRACT

This paper presents an HMM-based automatic system for recognition of bird species from audio field recordings. It includes an improved unsupervised modelling of individual bird syllables and duration modelling. The acoustic signal is decomposed into isolated segments, each segment containing a temporal evolution of a detected sinusodal component. Modelling of bird syllables is performed using Hidden Markov models (HMMs). A set of syllables of bird vocalisations is discovered in an unsupervised manner by employing dynamic time warping and agglomerative hierarchical clustering. A novel iterative maximum likelihood procedure is used to train individual HMMs for syllables of each species. Modelling of the state duration is employed in a post-recognition stage by combining the likelihood of the acoustic and duration modelling. Experiments are performed on over 33 hours of field recordings, containing 30 bird species. Evaluations demonstrate that the use of the proposed unsupervised iterative HMM training procedure and the duration modelling provides in average 45% error rate reduction. The presented system recognises bird species with accuracy of 97.8% using 3 seconds of the detected signal.

Index Terms— bird species recognition, hidden Markov model, HMM, syllable, unsupervised, duration modelling, DTW, segmentation, frequency track, sinusoid detection, audio, field recordings

1. INTRODUCTION

Bird vocalisations can be considered to be composed of stereotyped acoustic units, which we refer here to as syllables. Each syllable has a distinct time-frequency structure.

The first step in automatic processing of bird vocalisations is usually to segment the audio signal into isolated segments. There have been two main approaches to automatic segmentation. One is based on using an energy-based threshold decision, applied to time-domain signal or time-frequency representation of the acoustic signal, which is followed by some filtering, e.g., [1, 2] The other approach aims at decomposing the acoustic scene into sinusoidal components [1, 3, 4, 5, 6, 7]. The works in [1, 3, 4] used the sinusoidal decomposition method presented in [8]. In our recent bird pattern procesing studies [5, 6, 7] as well as in this paper, we employed a probabilistic method, presented in [9], for the detection of sinusoids.

A variety of feature representations and modelling approaches of bird acoustic signals have been explored. Some studies employed Mel-frequency cepstral coefficients (MFCCs), e.g., [10, 11, 1, 12]. As MFCCs normally capture the entire frequency band, they are prone to background noise and presence of other birds/animals concurrently vocalising in other frequency regions. The use of various statistical descriptors to characterise the detected time-frequency segments was employed in [1, 3, 4, 2]. Although the use of a single feature vector may seem attractive, it may not be inadequate to capture well the time-frequency structure of syllables as well as it may be susceptible to even minor inaccuracies in segmentation. In studies based on the sinusoidal decomposition approach, including our recent works, [1, 13, 5, 6, 7, 14], segments are represented as a temporal sequence of frequency values. This representation, which we refer to as frequency track, has a good potential, especially, in processing field recordings which typically contain various background noise and other birds/animals vocalising concurrently. The frequency track features were shown to obtain considerable performance improvements over MFCCs in recognition of bird sounds in noisy conditions [5]. The most commonly used modelling/classifier approaches include dynamic time warping (DTW) [15, 10], Gaussian mixture modelling [1, 5], hidden Markov models (HMMs) [1, 13, 16, 7, 17], support vector machines [18, 19], and decision trees [20]. The use of HMMs is compelling as they allow to model the temporal evolution of vocalisations.

In this paper, we extend our studies of automatic recognition of bird species by introducing an improved unsupervised modelling of individual bird syllable HMMs and incorporating duration modelling. The audio signal is first segmented in time-frequency plain into isolated sinusoidal segments and each segment is represented using frequency track features. The temporal evolution of these features is modelled using hidden Markov models. We employ an unsupervised procedure, based on DTW and agglomerative hierarchical clustering, to discover a set of bird vocalisation patterns. This provides labels for initial training of individual syllable HMMs, as presented in [17]. This paper introduces a novel iterative-label maximum likelihood training procedure to improve the quality of the trained individual syllable HMMs. We also introduce an incorporation of the state duration modelling, which is performed in a post-recognition stage by combining the likelihood from the acoustic model and the duration model. 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 audio field recordings provided by Borror Laboratory of Bioacoustics [21]. The proposed improved acoustic modelling and incorporation of duration modelling in a syllable HMM-based system achieved 97.8% bird species recognition accuracy, which is over 48% error rate reduction in comparison to the previous syllable-based system.

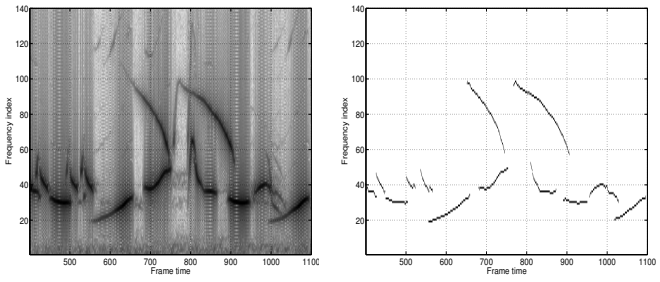
2. HMM-BASED BIRD SPECIES RECOGNITION SYSTEM

The proposed HMM-based bird species recognition system consists of the following parts: i) decomposition of the acoustic scene into

individual segments and extraction of frequency track features for each segment; ii) unsupervised modelling of individual bird syllables; iii) incorporating duration modelling. These are described in following subsections.

2.1. Segmentation and estimation of frequency tracks

Segmentation of the audio signal and estimation of frequency tracks is performed based on decomposing the entire acoustic scene into individual sinusoidal components. 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 a signal frame is considered as a potential sinusoidal component. Assessment whether the peak is a sinusoid or noise is performed based on short-time local magnitude and phase spectral features. The probability of a peak belonging to sinusoid and noise is calculated based on trained statistical models. More details of the sinusoidal detection method we employed are presented in [9]. Isolated segments are obtained based on the temporal evolution of detected sinusoidal components, with filtering applied to deal with accidental errors – the procedure is the same as used in our recent papers [7, 17] where further details of the procedure as well as the parameter setup are available. An example of a spectrogram of an audio field recording, containing concurrent bird vocalisations, and the final estimated individual segments are depicted in Figure 1. It can be seen that frequency tracks detected correspond well to bird vocalisations.



(a) (b)

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

2.2. Acoustic modelling

We use the model introduced in [17] as the baseline model in this paper – this presents considerable improvement over our previous research presented in [7]. For each bird species, we obtain a set of hidden Markov models (HMMs). This consists of a number of models for individual bird syllables plus one additional general model. To obtain the individual syllable models is not straighforward as there is no bird syllable label information available for our data (and is unlikely to be available for large datasets in general) as well as the set of syllable patterns produced by each bird species is not known. We introduced an approach to deal with the problem of obtaining the individual syllable HMMs in an unsupervised manner in [6, 17] and this is summarised in Section 2.2.1. The novel parts introduced in this paper consist of an improved iterative procedure for training syllable HMMs and incorporation of duration modelling and these are presented in Sections 2.2.2 and 2.3, respectively.

2.2.1. Unsupervised modelling of individual bird syllables

The first step towards obtaining individual syllable models in an unsupervised manner is to find a set of bird vocalisation patterns. There could be several ways to deal with this problem. The approach we employed in this paper is based on a modified dynamic time warping (DTW) algorithm to search for partial and multiple matches between each pair of detected segments and then use the obtained similarity values in an agglomerative hierarchical clustering [6].

The outcome of the above step is a set of clusters of vocalisation patterns for each bird species. The resulting assignment of segments into clusters provides the label information for each detected segment. Using this label information, we can use the Baum-Welsh algorithm to 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. In this paper, we use a fixed number of clusters, based on the highest occupancy, for each bird species. As there will be segments which are not assigned to any of the clusters, in addition to the individual syllable HMMs, we also have a single 'general' HMM for each bird species to model all the remaining segments. To cover the variety of these remaining segments, the state pdf of this model consists of several Gaussian mixture components.

2.2.2. Improved modelling of individual bird syllables

The individual syllable models as described in Section 2.2.1 are trained based on the label information obtained as a result of the DTW search and hierarchical clustering. This may, however, contain some errors – there may be some segments incorrectly assigned to clusters or some segments which should be assigned to one of the individual syllable clusters may be assigned by error to the general model. To improve over this, we incorporate an iterative training procedure in which the label assignment is modified based on the maximum likelihood criteria during the HMM training. After training the HMMs with the initial label, for each segment of each bird species we find the model (either individual or general) for that species that achieves the maximum likelihood. This provides a new label for that segment to be used in the next Baum-Welsh training iteration. The iterative process can be repeated several times and the stoping criteria may be based, for instance, on the likelihood change between iterations.

Figure 2 shows examples of the state output pdf of several trained individual syllable HMMs of one bird species. The top and bottom row present models obtained using the initial label information and after 2 iterations of the training procedure, respectively. It can be seen that each model provides a distinctive pattern. The models in the first and second column show modification of their parameters as a result of the iterative training procedure, while the model in the third colum is largely unchanged. The first model mainly decreased its variance at the beginning states, while the second model also modified its means. We have observed that the models changed only little after 2 iterations of this training procedure. Our analysis on the training data also showed that the proportion of segments assigned to the general model was over 40% after the DTW and clustering but decreased to below 7% after two iterations of this training procedure.

2.3. Incorporating duration modelling

While the duration is a key aspect of bird vocalisation pattern structure, the underlying model of duration in standard HMMs is not well

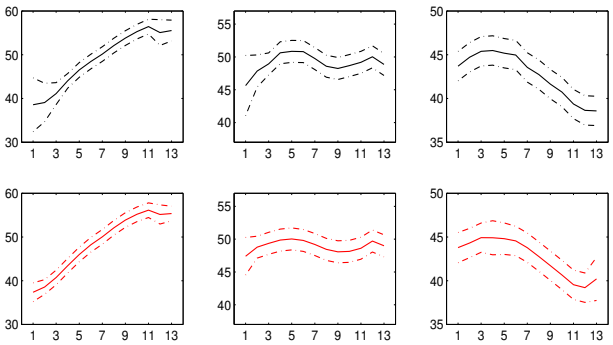


Fig. 2. Examples of the mean values, with variance around the mean indicated using dashed-dotted lines, of the state output Gaussian pdf, modelling frequency track features, for three trained syllable HMMs of bird species *Pine Warbler* after using the initial label information (top row) and after 2 iterations of the training procedure (bottom row). The x- and y-axis denotes the HMM state and frequency index, respectively.

suited. This section presents an approach we employed for modelling the duration of each HMM state and incorporating this in a post-recognition. A similar post-recognition approach, although for different purpose, was also used in [22].

First, for each segment of the training data of each bird species, using the Viterbi algorithm we find the model of that bird species that achieves the highest likelihood. This will be either one of the individual syllable models or the general model. The Viterbi algorithm also provides the alignment of the sequence of frequency track features of the segment on the states of that model, i.e., we obtain the occupancy count for each state, which we denote by a vector $D = (d_1, \ldots, d_S)$, where S is the number of states. Using the whole training set, the state occupancy counts are collected for each individual syllable model and general model of each bird species. These are used to estimate a state-duration probability distribution for each state of each model. A variety of distribution functions could be employed, for instance, in the context of speech processing, Gamma and Poisson distributions have usually been used, e.g., [23]. We have observed that the state occupancies may not follow well a single Poisson distribution. As such, we used a mixture of Poisson distributions, with the parameters being estimated using the EM algorithm.

The above considers the duration of each state separately. This may not be robust against inaccuracy in the frame-state alignment. This could be improved by considering the duration within several adjacent states, i.e., the duration d_s at state s will be the sum of the durations within the range of states $(s, s+\delta)$. We call this multi-state duration model. Both the single-state and multi-state duration models were explored in our experiments.

An example of the estimated state-duration probability distribution is depicted in Figure 3.

2.4. 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.

For a given utterance of audio recording, the segmentation and frequency track feature extraction step, as described in Section 2.1, provide a set of J detected segments $O=\{O_j\}_{j=1}^{J}$,

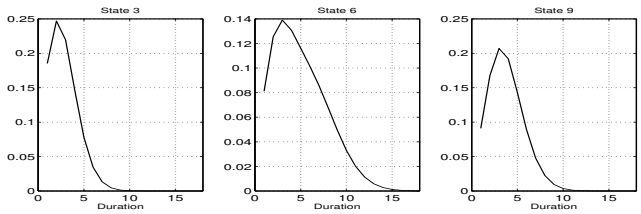


Fig. 3. An example of the state duration model for state 3, 6, and 9 corresponding to the bird syllable model shown in column 1 in Figure 2.

with each segment being represented by a sequence of features $O_j=(\mathbf{o}_j(1), \ldots, \mathbf{o}_j(T_j))$, where T_j is the number of frames in the segment j. Each detected segment is treated individually and we consider that vocalisations of only a single bird species are present in the given utterance. Using the proposed model with incorporated duration modelling, the recognised bird species, denoted by b^*, is obtained as

$$b^* = \arg\max_b \prod_{j=1}^{J} p(O_j|\lambda_b) p(D_j|\gamma_b)^{\alpha} \quad (1)$$

where the term $p(O_j|\lambda_b)$ is the probability of the sequence of frequency track features O_j and $p(D_j|\gamma_b)$ is the duration model probability. The $p(O_j|\lambda_b)$ is obtained for each bird species as the maximum probability over the general and all individual syllable models of that species using the best state sequence s^* found by the Viterbi algorithm as

$$p(O_j|\lambda_b) = \max_i \prod_{n=1}^{T_j} p(O_j(n)|\lambda_{b(i),s^*}) \quad (2)$$

where i is an index going through the set of general and individual syllable models.

The state sequence obtained for segment j from the Viterbi algorithm on the best model of that bird species defines the duration vector D_j. This is used to calculate the duration probability term $p(D_j|\gamma_b)$, where γ_b denotes the duration model. As the duration probability is of a different scale to the acoustic feature probability, we weight the contribution of the duration probability to the overall probability in Eq. 1 using the parameter α. We used the same value of the weighting factor for all segments and models and its value was found based on recognition experiments on the training data.

3. EXPERIMENTAL EVALUATIONS

3.1. Data description

Experimental evaluations were performed using field recordings from [21]. These are recordings of birds in real world natural habitats, 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 few 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.

3.2. Experimental setup

Each detected segment was characterised by a sequence of frequency track features. These contained the frequency value of the detected sinusoid at each frame time, obtained as presented in Section 2.1, and its temporal derivatives, referred to as delta and acceleration features, obtained as in [24] with the window set to 3 and 2, respectively. This resulted in a sequence of 3 dimensional feature vectors. In all experiments, a left-to-right HMMs with no skip allowed were used and these were built using the HTK [24]. The number of HMM states was set to 13, which reflects the minimum allowed length of the detected segment. Based on our results presented in [17], the number of individual syllable models was set to 70. The HMM state output probability density functions are using Gaussian distribution(s) with a diagonal covariance matrix. A single Gaussian is used for individual syllable models while a Gaussian mixture model with 10 components is used for the general model.

3.3. Experimental results

We first analyse the effect of the iterative training procedure. Results obtained using utterances of 1 second length with different number of iterations are presented in Table 1. It can be seen that the iterative training gives significant performance improvements over the baseline model. A large improvement is achieved after the first iteration, followed by a further though smaller improvement after the second iteration. The improvement is due to the improved quality of the individual syllable HMMs as well as due to the reduced set of patterns the general HMM can account for.

Table 1. *Bird species recognition accuracy obtained by the HMM-based system using individual syllable models without (baseline) and with the iterative training procedure. Utterances of 1 second length used.*

	Baseline model	Model with iterative-label training	
		Iter 1	Iter 2
Rec. Acc. (%)	89.8	92.5	93.1

Next we analyse the effect of incorporating the HMM state duration modelling. Results, again using utterances of 1 second length, obtained by the baseline model without and with the single- and multi-state duration modelling are presented in Table 2. It can be seen that the use of single-state duration modelling gives good improvement and it is further improved by using multi-state duration, which achieved over 15% relative error rate reduction over the baseline model. Experiments with the multi-state duration were performed using the parameter δ set from 1 to 3 and the presented results are obtained when δ is 2.

Finally, we present results when both of the proposed techniques were employed. Evaluations were performed with different length of the detected signal and results are presented in Table 3. It can be seen that the combination of the two techniques provides further recognition accuracy improvement. The error rate reduction from the baseline models is between 37% to 48%.

Table 2. *Bird species recognition accuracy obtained by the HMM-based system using individual syllable models without (baseline) and with incorporated state duration modelling. Utterances of 1 second length used.*

	Baseline model	Model with state duration modelling	
		Single–state	Multi–state
Rec. Acc. (%)	89.8	90.8	91.2

Table 3. *Bird species recognition accuracy and error rate reduction obtained by the baseline individual syllable HMM-based recognition system and the system with incorporated iterative training procedure and duration modelling when using different length of detected signal.*

Utterance length (sec)	Rec. Acc. (%)		Error Rate Reduction (%)
	Baseline model	Model with iterative training & duration modelling	
1	89.8	93.6	37.2
2	94.3	97.1	48.3
3	95.8	97.8	48.6

4. CONCLUSION

We presented in this paper an automatic system for recognition of bird species from audio field recordings based on improved modelling of individual syllables of species and incorporation of the duration modelling. 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). Unsupervised clustering was employed to discover the set of bird syllable patterns and an individual HMM was obtained for each syllable. A novel iterative procedure, based on the maximum likelihood principle, for training the syllable HMMs was introduced. An HMM state duration modelling was incorporated in a post-recognition approach. Experimental evaluations were performed on field recordings provided by the Borror Laboratory of Bioacoustics. Experimental results demonstrated that the proposed HMM-based system achieved in average 45% error rate reduction in recognising 30 bird species in comparison to our previous system. Using 3 second of the detected signal, the recognition accuracy of 97.8% was obtained.

Acknowledgement

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

5. 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] 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.

[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] 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.

[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] 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] 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.

[14] P. Jančovič and M. Köküer, "Acoustic recognition of multiple bird species based on penalised maximum likelihood," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1585–1589, Oct. 2015.

[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] 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*, pp. 768–772, April 2015.

[18] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. Article ID 38637, Jan. 2007.

[19] K. Kaewtip, L. N. Tan, C. E. Taylor, and A. Alwan, "Bird-phrase segmentation and verification: A noise-robust template-based approach," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Brisbane, Australia*, pp. 758–762, April 2015.

[20] M. Lasseck, "Improved automatic bird identification through decision tree based feature selection and bagging," in *Working notes of CLEF 2015 conference*, 2015.

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

[22] P. Jančovič and J. Ming, "A probabilistic union model with automatic order selection for noisy speech recognition," *The Journal of the Acoustical Society of America*, vol. 110, pp. 1641–1648, 2001.

[23] M.J. Russell and A.E. Cook, "Experimental evaluation of duration modeling techniques for automatic speech recognition," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, Texas, USA*, pp. 2376–2379, 1987.

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