BIRD SPECIES RECOGNITION FROM FIELD RECORDINGS USING HMM-BASED MODELLING OF FREQUENCY TRACKS

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

[1] School of Electronic, Electrical & Computer Engineering, University of Birmingham, UK
E-mail: {p.jancovic, m.kokuer, m.j.russell}@bham.ac.uk
[2] Faculty of Technology, Engineering & 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 field audio recordings. The proposed system employs a novel method for detection of sinusoidal components in the acoustic scene. This provides a segmentation of the signal and also feature representation of each segment in terms of frequencies over time, referred to as frequency track. We employ hidden Markov models (HMMs) to model the temporal evolution of frequency tracks. We demonstrate the effect of including local temporal dynamics of frequency tracks and HMM modelling parameters. Experiments are performed on over 33 hours of field recordings, containing 30 bird species. Evaluations demonstrate that the HMM-based temporal modelling provides considerable performance improvement over a system based on Gaussian mixture modelling. The proposed HMM-based system is capable of recognising bird species with accuracy over 85% from only 3 seconds of detected signal.

Index Terms— bird species recognition, hidden Markov models, segmentation, frequency track, sinusoidal detection

1. INTRODUCTION

Identification of bird species is an important issue for biological research and environmental protection. Birds are sensitive to environmental changes and their presence can indicate the biodiversity and the health of the environment. Bird species identification currently relies on expert ornithologists who identify birds by sight and, more often, by their songs and calls. In recent years there has been an increased interest in automatic recognition of bird species using the acoustic signal.

Automatic processing of bird vocalisations usually starts with segmenting the continuous audio signal into syllables and then extracting some features to represent the syllable. Some studies involved manual segmentation [1, 2]. Many works that employed automated segmentation used an energy-based threshold decision in time or time-frequency domain, with the threshold set based on estimated noise level, e.g., [3, 4, 5]. The approach of decomposing the acoustic scene into sinusoidal components was employed in [2, 3, 6, 7]. The works in [2, 6], employing the method proposed in [8], considered all the spectral peaks at each frame time and used a threshold-based assessment of frequency and amplitude continuity of peaks over adjacent frames. The obtained peak tracks underwent further automated energy-based pruning. In [6], the obtained tracks were further hand-pruned for training data. Similar sinusoidal-based segmentation was also used in [3] where the pruning was effectively performed by using only the dominant sinusoid. We proposed in [9, 10] a probabilistic method that enables to detect only those peaks corresponding to sinusoidal components from all the spectral peaks at a given frame time, without requiring any noise estimate, and employed this for segmentation of bird vocalisations in [7, 11].

Several types of feature representations and modelling approaches of bird acoustic signals have been explored. The modelling approaches used include dynamic time warping [1, 12], Gaussian mixture models [3, 7], hidden Markov models (HMMs) [3, 13], neural networks [14], and support vector machines [4]. The use of HMMs is compelling as they can model the temporal evolution of sequences. For feature representation, many previous studies were inspired by features used in the field of speech processing, for instance, filter-bank energies were used in [12], and Mel-frequency cepstral coefficients (MFCC) in [12, 15, 3, 4, 16]. 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. The works in [2, 3, 6, 17, 4] used a set of statistical descriptors to characterise the detected spectro-temporal regions of syllables. Such descriptors may not be able to describe well a more complex types of syllables and may be susceptable to any variations in segmentation. Few other studies aimed at representing the bird detected segments as a temporal sequence of frequencies, which we here refer to as frequency track [3, 13, 7, 11]. Note that although the works in [2, 6] obtained frequency tracks, they did not use the sequence directly but characterised the tracks by a set of statistical descriptors. In [3], the frequency track features were shown to perform worse than MFCC features. However, our recent study in [7] has demonstrated that the frequency track features can provide significantly better performance than MFCCs in noisy background conditions. 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.

In this paper, we study automatic bird species recognition using real-world field recordings. This study extends our previous works by investigating the use of the frequency track features, obtained using the sinusoidal detection algorithm from [10], on large-scale field recordings and studying the effect of temporal modelling of the frequency track features for bird species recognition. We model the temporal evolution of frequency tracks for each bird species using hidden Markov models. The recognition is performed based on recognising individual detected segments on each bird species model and aggregating the probabilities from all segments to reach a decision. Experimental evaluations are performed on field recordings provided by the Borror Laboratory of Bioacoustics [18]. We demon-

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

strate the effect of including information on local temporal dynamics of the frequency tracks and the effect of modelling with various number of HMM states and mixture components at each state. Evaluations are also presented for different lengths of detected signal. Experimental results show that over 85% recognition accuracy can be achieved with using only 3 seconds of detected signal.

2. SEGMENTATION AND ESTIMATION OF FREQUENCY TRACKS FOR BIRD VOCALISATIONS

The estimation of frequency tracks is performed based on the detection of individual sinusoidal components at each frame time in the entire acoustic scene using the method we introduced in [10], with some modifications as presented here. Based on the detected sinusoidal components, we then characterise the signal in terms of the frequency of each detected sinusoidal component. A continuous sequence of detected sinusoidal components forms what we refer to as a frequency track. As this can provide multiple frequency tracks during the same time periods, it can deal with concurrent vocalisations of multiple birds. The following subsections first give a brief summary of the sinusoidal detection method with a parameter setup and then describe the segmentation procedure.

2.1. Estimation of Frequency Tracks

We tackle the sinusoidal detection problem as a pattern recognition problem. We consider that the signal may consist of an unknown number of sinusoidal components. Each spectral peak is first considered as a potential sinusoidal component. A set of features, extracted from the short-time spectrum, is obtained for each spectral peak. The decision whether the peak is detected as a sinusoid or not is based on calculating the probability of the extracted set of features on a model corresponding to sinusoids and to noise.

2.1.1. Spectral magnitude and phase features

Let us denote the short-time spectrum of the l^{th} frame of the signal by $S_l(k)$. Denote the frequency index of a spectral peak found in the short-time magnitude spectrum by k_p. For each peak, a multivariate feature vector $\mathbf{y}$, capturing the spectral magnitude shape and phase continuity information around the peak, is extracted. The magnitude shape features are obtained by using a normalised spectral magnitude values over the range of frequency bins from $k_p - M$ to $k_p + M$, i.e., $\mathbf{y}^1$=($|\tilde{S}_l(k_p - M|, \ldots, |\tilde{S}_l(k_p - 1)|, |\tilde{S}_l(k_p + 1)|, \ldots |\tilde{S}_l(k_p + M|)$, where $|\tilde{S}_l(k)|$ is the magnitude spectrum $|S_l(k)|$ normalised by the magnitude value at the peak $|S_l(k_p)|$ and M denotes the number of bins considered around the peak. The phase continuity features are obtained by using the spectral phase difference values over the range of frequency bins from $k_p - M$ to $k_p + M$, i.e., $\mathbf{y}^2$=($\Delta\phi_l(k_p - M), \ldots, \Delta\phi_l(k_p + M)$). The phase difference between the current and previous signal frame is defined as $\Delta\phi_l(k) = \phi_l(k) - \phi_{l-1}(k) - 2\pi kL/N$, where $\phi_l(k)$ and $\phi_{l-1}(k)$ denote the phase of the frequency point k at frame-time l and $l-1$, respectively, and L is the frame-shift in samples.

2.1.2. Probabilistic modelling

The distribution of the multivariate feature vector $\mathbf{y} = (\mathbf{y}^1, \mathbf{y}^2)$, representing the spectral magnitude shape and phase continuity, is modelled by using GMM. A large collection of features $\mathbf{y}$ corresponding to spectral peaks of noise and of sinusoidal signals at various SNRs are used as the training data to estimate the parameters of the GMM of noise, denoted by λ_n, and of sinusoidal signal, denoted by λ_s.

A given unknown audio signal is processed as described in the previous section to extract the features for each spectral peak. The decision whether a spectral peak at a given signal frame corresponds to a sinusoidal signal or not is based on the maximum likelihood criterion, i.e., the peak is detected as a sinusoid if $p(\mathbf{y}|\lambda_s) > p(\mathbf{y}|\lambda_n)$.

2.1.3. Parameter setup

The signal, sampled at 48 kHz, was divided into frames of 256 samples with a shift of L=48 samples between the adjacent frames. The frame length and frame shift corresponds to 5.3 ms and 1 ms, respectively. This is considerably shorter than used in most other studies on bird processing. The use of signal frames of similarly short duration were found suitable for processing of bird acoustic signals in our previous research [7] [11]. 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 in Section 2.1.1 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.

2.2. Segmentation

The sinusoidal detection method indicates which spectro-temporal points were detected as sinusoids. This can be considered as an initial segmentation of the acoustic scene. The following steps are performed to further refine this segmentation result.

First, we consider that any detected segment which is of a very short length was detected accidentally by error. As such, we discard all segments whose length is less than 4 frames. Then, in order to avoid accidental split of a segment due to a missed detection of few frequency bins, 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. 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 real-world recordings obtained from natural environment, there are often co-vocalisations of other birds and animals present in the audio. However, there is no time-stamp label information available that would indicate the location of the vocalisations of the bird of interest in the recordings. As we do not want to have included and model these background co-vocalisations, we can consider that the vocalisations of bird species being recorded are of 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 from the Borror data [18] containing concurrent vocalisations of two bird species and the estimated frequency tracks before and after applying the segmentation procedure are depicted in Figure 1. It can be seen that frequency tracks detected correspond well to vocalisations of birds. We can see that even very weak vocalisations are detected in the originally estimated frequency tracks (i.e., before applying the segmentation), for instance, the high frequency components around frequency index 120 and around frame time index 560, 600 and 1050. Listening confirmed that these were co-vocalisations of other

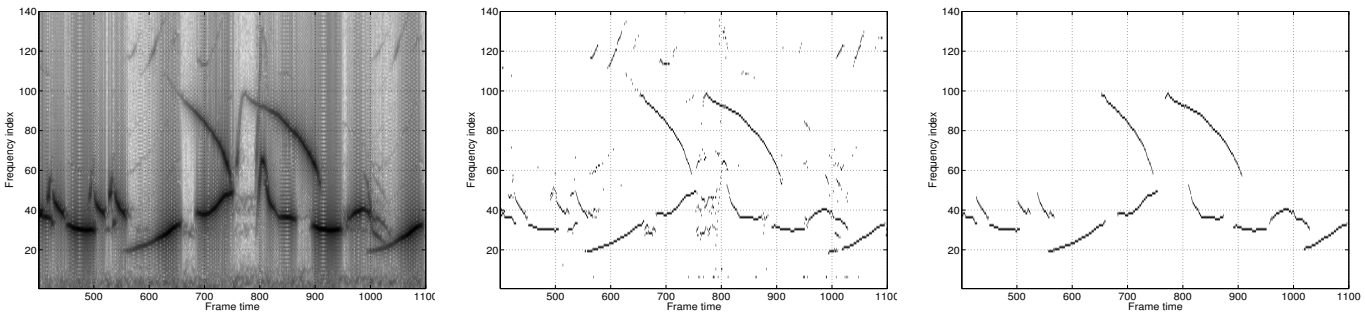

Fig. 1. An example of a spectrogram (a) of audio field recording and the corresponding estimated frequency tracks initial (b) and final (c).

birds in the background. The final frequency tracks capture well the vocalisations of birds of interest.

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 $O = \{O_s\}_{s=1}^{R}$, where R is the number of detected segments, 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. We treat each detected segment individually.

3.1. Modelling the Frequency Tracks

The model of each bird species is obtained by modelling the temporal evolution of frequency tracks of detected segments using a left-to-right (no skip allowed) HMM. In this paper, a single 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. This could be improved in future, and is currently under our investigation, by first discovering the vocabulary set of acoustic patterns produced by each bird species, for instance using the approach we recently presented in [11], and then constructing and training a separate HMM to model each type of vocalisation. 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 modelled with a mixture of Gaussians. Gaussian distributions with a diagonal covariance matrix are used due to computational reasons, as is typically done in speech and audio pattern processing. In experimental section, we demonstrate the effect of using different number of HMM emitting states and mixture components.

3.2. Recognition of Bird Species

We consider the identification of bird species from a finite set of bird species based on a given utterance of testing signal recording.

For recognition, we use an HMM network consisting of a single pass through any bird species HMM model. Using the Viterbi algorithm, we obtain the probability $p(O_s|\lambda_b)$ of each segment s on each bird species model λ_b. Considering that there are vocalisations of only a single bird species 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.,

$$p(O|\lambda_b) = \prod_{s=1}^{R} p(O_s|\lambda_b), \quad (1)$$

and obtain the recognised bird species as $b^* = \arg\max_b p(O|\lambda_b)$.

We explored several variations to calculating the overall probability $p(O|\lambda_b)$ in Eq. 1. First, we used only a sub-set of the detected segments, omitting a given percentage of signal segments that achieved the lowest probability at each bird species models. Then, since among the detected segments there may be detected vocalisations of other birds/animals which do not exist in our bird species vocabulary, i.e., outliers, we also explored approach, similar as presented in [19], in which the segment s is omitted from the product in Eq. 1 if its probability is lower than a given threshold on all bird species models. Both of these modifications led to only negligible improvements in our system.

Note that the outlined approach can also handle recognition of multiple bird species when there are concurrent vocalisations of several bird species present. In such case, a set of segments detected in a given utterance will contain segments from all the birds vocalising in the utterance. As such, this translates in our system to the problem of determining which sub-set of segments from the detected set belong to each bird species, without knowing the identity of the bird species and their number and is subject of our current research.

4. EXPERIMENTAL EVALUATIONS

4.1. Data description

Experimental evaluations were performed using field recordings from [18]. 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 Results

This sections presents experimental evaluations investigating the HMM-based temporal modelling of the frequency track features.

First, we demonstrate the effect of temporal modelling of the frequency track features using HMMs. This is compared to the use of Gaussian mixture model (GMM), which models the distribution of the features only, without accounting for any temporal structure. The HMM-based system consists of 13 states; the pdf of each state modelled with 20 mixture components. The GMM-based system consists of 260 mixture components. As such, both systems have the same effective number of mixture components modelling the features. Results are presented in top part of Table 1. It can be seen that the incorporation of temporal modelling using HMMs provides considerable performance improvements over GMMs, from 33.8% to 54.1%.

Table 1. *Bird species recognition accuracy obtained by GMM-based and HMM-based system when using the frequency track features without and with including their temporal derivatives.*

Model	Features	Rec. Acc. (%)
GMM (260 mix)	freqTrack (1d)	33.8
HMM (13 states, 20 mix)	freqTrack (1d)	54.1
GMM (260 mix)	freqTrack+DA (3d)	68.0
HMM (13 states, 20 mix)	freqTrack+DA (3d)	84.3

The frequency track features used in the above experiments provide the frequency value at each frame time but do not include any information about how the features vary over time. However, in speech and audio pattern processing, it is common to append temporal derivatives of features, referred to as delta (D) and acceleration (A) features, which capture local temporal dynamics. We repeated the above experimental evaluations employing the frequency track features and incoporating such temporal derivatives into the representation. The included delta and accelleration features were calculated as in [20] with window set to 3 and 2, respectively, and added to the frequency track features, resulting in 3 dimensional (3d) feature vectors. Results are presented in the bottom part of Table 1. It can be seen that the performance of the GMM-based system improved significantly from 33.8% to 68.0% as a result of including the local temporal dynamics into the representation. A similar performance improvement is also achieved by the HMM-based system.

Now, we evaluate the performance when using different number of states in HMM modelling and different number of mixture components at each state. The number of states varied from 5 to 13, with the upper limit reflecting the minimum length of segment that could be output from the segmentation method. These experiments were performed using the frequency track features with incorporated delta and acceleration features. Results are presented in Table 2. It can be seen that increasing the number of states improves the recognition accuracy, especially when increasing from 5 to 9. This is not surprising as the HMM in effect performs a non-linear quantisation of the frequency track features over time. Results obtained using different number of mixture components show that the recognition accuracy increases as the number of mixture components increases, with the improvement being relatively small when the number of mixtures is above 20. The use of such high number of mixture components is plausible, as it accounts for different types of bird vocalisations and differences within vocalisation type across individual birds. This was also observed by our inspection of examples of data and trained models.

Table 2. *Bird species recognition accuracy obtained by HMM-based modelling of acoustic segments with different number of states and mixture components per state.*

Number of		Rec. Acc. (%)
HMM states	mixture components	
5	15	69.3
9	15	81.1
13	15	82.9
13	5	75.2
13	10	79.8
13	15	82.9
13	20	84.3
13	25	85.2
13	30	85.8

Finally, we evaluated the recognition performance as a function of the length of the detected signal used for testing, which we varied from 1 second to 3 seconds. The results are presented in Table 3. It can be seen that very good performance can be achieved even with such short amount of detected signal of length as 1 second.

Table 3. *Bird species recognition accuracy obtained when using different length of detected signal.*

Length of testing utterance (sec)	Rec. Acc. (%)
3	85.8
2	83.9
1	80.5

5. CONCLUSION

In this paper, we presented an automatic system for recognition of bird species from field audio recordings. The proposed system first employed a novel method for detection of sinusoidal components in the entire acoustic scene. This provided a segmentation of the signal and also feature representation of each segment in terms of the detected sinusoid frequencies over time, referred to as frequency track. Hidden Markov models were employed for modelling the sequences of frequence tracks of each bird species. Experimental evaluations were performed on field recordings provided by the Borror Laboratory of Bioacoustics. Experimental results demonstrated that the use of HMMs for temporal modelling of the frequency track features provided considerable performance improvements over a bag-of-features GMM-based system. We also showed that the inclusion of local dynamic information into the frequency tracks, as temporal derivatives, improved the performance considerably. The evaluations showed that the proposed system is capable of recognising bird species well from only few seconds of detected data.

Acknowledgement

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

6. REFERENCES

[1] 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.

[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] 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.

[4] 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.

[5] A. Selin, J. Turunen, and J.T. Tanttu, "Wavelets in recognition of bird sounds," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. Article ID 51806, Jan. 2007.

[6] 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.

[7] 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.

[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, "Estimation of voicing-character of speech spectra based on spectral shape," *IEEE Signal Processing Letters*, vol. 14, no. 1, pp. 66–69, Jan. 2007.

[10] 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.

[11] 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*, Sep. 2013.

[12] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] "Borror Laboratory of Bioacoustics," *The Ohio State University, Columbus, OH, all rights reserved.*, URL: www.blb.biosci.ohio-state.edu.

[19] 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.

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