INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Recognition of Multiple Bird Species based on Penalised Maximum Likelihood and HMM-based Modelling of Individual Vocalisation Elements

Peter Jančovič, Münevver Köküer

Deptartment of Electronic, Electrical & Systems Engineering, University of Birmingham, UK
{p.jancovic, m.kokuer }@bham.ac.uk

Abstract

This paper presents an extension of our recent work on recognition of multiple bird species from their vocalisations by incorporating an improved acoustic modelling. The acoustic scene is segmented into spectro-temporal isolated segments by employing a sinusoidal detection algorithm, which is able to handle multiple simultaneous bird vocalisations. Each segment is represented as a temporal sequence of frequencies of the detected sinusoid. Each bird species is represented by a set of hidden Markov models (HMMs), each HMM modelling a particular vocalisation element. A set of elements is discovered in an unsupervised manner using a partial dynamic time warping algorithm and agglomerative hierarchical clustering. Recognition of multiple bird species is performed based on maximising the likelihood of the set of detected segments on a subset of bird species models, with a penalisation applied for increasing the number of bird species. Experimental evaluations used audio field recordings containing 30 bird species. Detected segments from several bird species are joined to simulate the presence of multiple bird species. It is demonstrated that the use of improved acoustic modelling in conjunction with the maximum likelihood score combination method provides considerable improvements over previous results and the use of majority voting.
Index Terms: multiple bird species recognition, HMM, vocalisation, element, unsupervised training, sinusoid detection

1. Introduction

Automatic recognition of bird species from their vocalisations usually starts with a segmentation of the acoustic signal into isolated segments. This has been performed using an energy-based threshold decision, which requires an estimate of noise level, e.g., [1], or by decomposing the acoustic scene into sinusoidal components [1, 2, 3, 4, 5]. A variety of feature representations of and modelling approaches to bird vocalisations have been explored. The use of conventional Mel-frequency cepstral coefficients, employed by a number of studies, e.g., [1], is problematic in presence of concurrent vocalisations of other birds/animals. Characterising a detected spectro-temporal segment using a set of statistical descriptors, as employed in [1, 2, 3, 6], may not capture well a more complex types of vocalisation elements and may be susceptable to inaccuracies in segmentation. In a case of tonal bird vocalisations, the use of a sinusoidal detection for segmentation also offers a natural way of representing the segment as a temporal sequence of the frequencies of the detected sinusoid, which we here refer to as frequency track. This representation was employed in few earlier studies [1, 7] and also in our recent works [4, 8, 5, 9]. The most commonly used modelling approaches include dynamic time warping [10, 11], Gaussian mixture modeling [1, 4], and hidden Markov models (HMMs) [1, 7, 12, 5].

Audio field recordings often contain vocalisations of multiple bird species. This issue has been addressed only in few recent works. To tackle the problem of having multiple bird species in the training data, the authors in [6] employed a multi-instance multi-label (MIML) approach. This approach requires that each segment is represented as a single feature vector, which prevents the use of temporal modelling of segments. On a similar task and data, there have recently been two bird classification challenges. The methods used by all contributors to the first challenge are only briefly outlined in a summary paper [13]. The contributions to the second challenge are described in [14]. The contributions to these challenges were based on using the MIML approach and a variety of pattern recognition techniques that did not model the temporal evolution of segments.

This paper presents an extension of our recent study on recognition of multiple bird species by incorporating an improved acoustic modelling. While our paper [15] presented a method for recognition of multiple bird species, the acoustic models of each species consisted of a single HMM, with the variety and variability of bird vocalisations being accounted for by using several Gaussian mixture components per state. In this paper, we represent each bird species by a set of HMMs, each HMM modelling a particular type of vocalisation element. It was shown in our previous work [9, 16] that the use of element modelling can reduce the bird identification error rate by over 40%. The effect of such improved acoustic modelling on the recognition of multiple bird species is unknown and this paper aims to explore on this. As no element-level label information is available, we first employ an unsupervised clustering approach as presented in [8] to discover a set of vocalisation elements for each species and then train each individual element HMM. For a given audio recording used during the recognition, the sinusoidal-based segmentation provides a set of variable-length segments. The probability of each segment on each bird species HMM calculated. The identity of the multiple bird species is obtained based on maximising the likelihood of the set of segments on a subset of bird species models using the method introduced in [15]. Experimental evaluations are performed using field recordings from 30 bird species [17]. Experimental data with multiple bird species are created by artificially mixing detected segments of several bird species. Results are compared to those obtained by representing each bird species using a single HMM and to majority voting score combination methods.

2. Multiple bird species recognition employing modelling of individual elements

This section provides description of individual components of the recognition system, specifically, the approach we employed

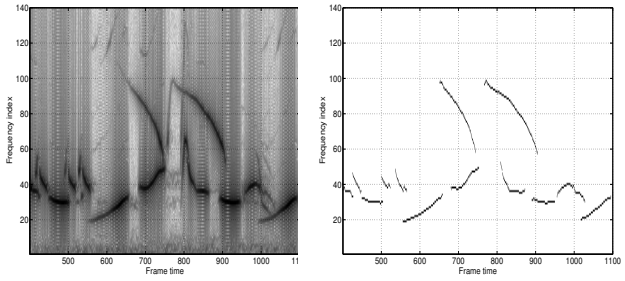
http://dx.doi.org/10.21437/Interspeech.2016-669

for segmentation of the audio signal and extraction of frequency track features, the acoustic modelling based on using a single HMM and multiple individual element HMMs for each bird species, and the method for recognition of multiple bird species. Each of these components was introduced separately in our recent publications [5, 9, 15] where we refer the reader to for further details. Throughout this paper, we consider 'element' as the smallest structurally distinct unit of bird vocalisations, visible as a continuous line on a spectrogram [18].

2.1. Segmentation and estimation of frequency tracks

The segmentation of the audio signal and estimation of frequency tracks is performed based on detection of sinusoidal components in signal using the method we introduced in [19]. In brief summary, each peak in the magnitude spectrum of a signal frame is considered as a potential sinusoidal component. A peak is characterised using a set of magnitude and phase spectral features extracted around the peak. A model for sinusoidal signals and noise is built and the detection is performed based on maximum likelihood criterion. This segmentation is further refined by discarding very short segments and segments of a low energy – all these are considered to be detected by an accidental error or to correspond to other vocalisations in the background.

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 the detected frequency tracks correspond well to vocalisations of birds and that this method can detect well vocalisations which are concurrent in time but in different frequency regions.



(a) (b)

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

2.2. HMM-based modelling of bird vocalisations

The temporal evolution of frequency tracks of detected segments is modelled using a left-to-right HMMs, which are used to represent the bird species.

As the baseline model, which was also used in our initial work on multi bird species recognition in [15], we use a single HMM to represent each bird species. This model is trained using the entire collection of the detected segments from all training recordings of that species. To account for the variety of element patterns and variations of individual instances of vocalisations, the probability density function at each HMM state is in this case modelled with a mixture of Gaussians.

To represent each bird species using a set of individual element HMMs was not straighforward as the element-level label information was not available and the set of element patterns produced by each bird species was unknown. As such, we first employed an unsupervised procedure to find a set of vocalisation elements and provide label information for the data and then trained the individual element models using the conventional Baum-Welch algorithm. The unsupervised procedure was based on a modified dynamic time warping (DTW) algorithm and an agglomerative hierarchical clustering. The modified DTW allows to search for parial and multiple matches between two given segments. For each found partial match, an overall score is calculated based on the cummulative distance obtained from the DTW, length of the matching path and the ratio of the length of the matching path to the total length of the path. Only the match with the highest overall score is used if matches overlap. The overall similarity score is then used in an agglomerative hierarchical clustering to arrive at a set of clusters of vocalisation elements. We refer the reader for further details to [8, 9]. As the obtained clusters of vocalisation patterns are expected to be homogenous, the state output probability density function (pdf) of each individual element 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 element 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 element HMMs of a bird species is depicted in Figure 2. It can be seen that each model provides a distinctive pattern.

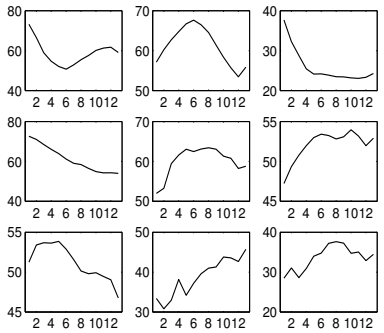


Figure 2: An example of the mean values of the state output Gaussian pdf, modelling frequency track features, for nine trained element HMMs of bird species *Northern Cardinal*. The x- and y-axis denotes the HMM state and frequency index, respectively.

2.3. Recognition of multiple bird species

We consider the identification of multiple bird species from a finite set of species based on an utterance of test signal of a given length. For a given utterance, the segmentation and frequency track feature extraction step provides a set of R detected segments $O=\{O_s\}_{s=1}^{R}$. Each segment s is 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. We consider each detected segment as an isolated vocalisation element. An approximation of the probability of each segment s on each bird species model λ_b, i.e., $p(O_s|\lambda_b)$, is obtained using the Viterbi algorithm. In the case of using the system based on individual element models, the probability is calculated on each element model and the highest

one is used.

The set of segments may consist of vocalisations of a single or multiple bird species. As such, we are facing a 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.

A possible approach to deal with this score combination could be to use for each segment only the information about the best bird species model, i.e., the model achieving $\max_b p(O_s|\lambda_b)$. The number and identity of the bird species recognised could be obtained based on majority counting, with the criteria being, for instance, the number of segments or the accummulated length of segments classified to each bird species models. This approach may not work well in situations when there is a larger ambiguity in recognising individual segments.

We approach this score combination problem as a problem of partitioning of the entire set of segments and assigning 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, where the value of K corresponds to the number of bird species. Let us denote each subset by B_i, i.e., $O{=}\cup\{B_i\}_{i=1}^{K}$ and $B_i{\cap}B_j{=}\emptyset$, and consider that the subset B_i is assigned to a bird species model b_i. The maximum overall likelihood of the set O, denoted by $P^{(K)}$, is

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

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 a set of $R{=}15$ segments into $K{=}3$ sub-sets is over 2.3 million. However, the maximisation in Eq. 1 can be split into two steps. First, for a given subset of models $\{b_1,\ldots,b_K\}$, calculate the likelihood of the best partitioning of O, which we denote by $P^{(K)}_{b_1,\ldots,b_K}$. This likelihood can be calculated simply by assigning each segment s, $s{=}1,\ldots,R$ to a model from the subset $\{b_1,\ldots,b_K\}$ that achieves the highest likelihood. The calculation of $P^{(K)}_{b_1,\ldots,b_K}$ is then repeated for all K model combinations out of the number of bird species. Finally, the likelihood $P^{(K)}$ is obtained as $\max_{b_1,\ldots,b_K} P^{(K)}_{b_1,\ldots,b_K}$.

An incorporation of constraint on the minimum length of signal assigned to each bird species can be useful to reduce some accidental errors. However, the above procedure does not allow for this. The calculation of the probability $P^{(K)}$, subject to constraints, can be performed using binary linear programming. An alternative procedure, which can find a close approximation in a faster way, was introduced in [15].

Finally, the parameter K, i.e., number of bird species in signal, is selected based on principles 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 $\alpha(K)$ was chosen for each K based on experiments on simulated mixture using the training data.

3. Experimental evaluations

3.1. 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. There are several files for each bird species, each file is typically between one to ten minutes long. 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. As these are field recordings, the audio contains also background environmental noise, vocalisations of other birds/animals and human speech.

Data from randomly chosen 30 bird species was used (list available at [20]). 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, vocalisations of multiple bird species were created by randomly mixing set of detected segments from several bird species.

3.2. Experimental setup

Each detected segment was characterised by a sequence of 3 dimensional frequency track features, containing the frequency value of the detected sinusoid and its temporal derivatives obtained as in [21]. A left-to-right HMMs with no skip allowed were used and these were built using the HTK [21]. The number of HMM states was set to 13, which reflects the minimum allowed length of the detected segment. The following setup was based on our results presented in [9]. The baseline model, i.e., single HMM per bird species, used 80 Gaussian mixture components per state. In the case of the individual element HMMs, the number of individual element models was set to 70. Each element model used a single Gaussian per state. In conjunction with the element models we also used a single general HMM, having 10 components Gaussian mixture model per state.

Performance is evaluated in terms of recognition correct, $100{\cdot}N_c/N$, 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.

3.3. Experimental results

First, we analysed recognition results when using only individual detected vocalisation segments in the case of single bird species presence. Figure 3 depict histograms of the rank of the correct bird species model obtained when using the baseline single HMM per bird species model and the individual vocalisation element models. It can be seen that the correct model was ranked as the one achieving the highest probability for only 27.2% of the segments when using the baseline model and this increased to 43% in a case of the individual element models. This shows that there is still a large proportion of the segments for which the correct model is not the best recognised model.

Bird species recognition performance when only single bird species is present for utterances containing three, two and one seconds of the detected signal is 92.0%, 88.8% and 83.3% in the case of using the baseline single HMM and 95.5%, 94.4% and 90.2% in the case of using the element HMMs [9].

Now, we present results when there are multiple bird

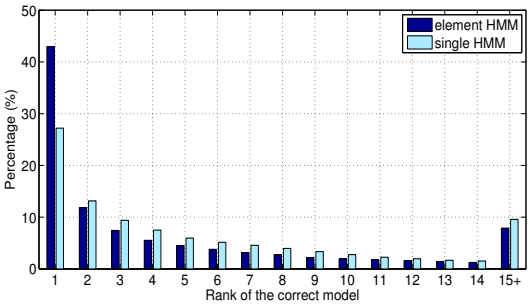


Figure 3: Histogram of rank statistics of the correct bird species models, collected over all segments of all bird species.

species present. First, we consider separately the case with 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. Table 1 presents results obtained by using the conventional majority voting method when employing the baseline single HMM and individual element HMMs acoustic modelling. It can be seen that the use of improved acoustic modelling, i.e., element HMMs, provides significant performance improvements. Using the accummulated length as the criteria function, the performance drops only marginally when the number of bird species increases. Except for the single bird species case, this criteria function provides better performance than the use of counts.

Table 1: *Bird species recognition correct (%) achieved by the majority voting method for a given number of bird species present when using single HMM and individual element HMMs. Each species contained 3 seconds of the detected signal.*

Number of bird species present	Majority voting combination method			
	Single HMM		Element HMMs	
	count	length	count	length
1 species	63.1	63.7	91.1	89.6
2 species	54.9	61.4	84.3	89.0
3 species	51.7	61.3	80.0	88.4

Table 2 presents results obtained by using the proposed maximum likelihood score combination method when employing the baseline single HMM and individual element HMMs. Comparing the effect of the acoustic modelling, it can be seen that the use of element models resulted in significant performance improvement, especially when there are 2 and 3 bird species present. In a case of using the constraints, the minimum length of the signal was set to match 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. When using element modelling, the use of constraints resulted in average in 12% relative error rate reduction. This is relatively small performance improvement for the need of using a considerably more computationally demanding algorithm. Comparing the performance of the score combination methods when element modelling is employed (see Table 1 and Table 2), we can see that the maximum likelihood method achieved considerable improvement over the majority voting – over 56% error rate reduction in average over all the number of species.

Finally, experimental results are presented for a given length of the detected signal which may contain a varying number of bird species. The number of bird species was generated randomly in the range from 1 to 3. Then, the set of vocalisation segments of around 3 seconds of the detected signal was considered as follows: either 3 sec from 1 bird species, 1.5 sec from 2 bird species, or 1 sec from 3 bird species. The constraint on the minimum length of the signal assigned to a bird species model was set to 1 second. Results are presented in Table 3. It can be seen that there is only a relatively small drop in recognition correct, from 91.3% to 89.2%, when the number of bird species is estimated as opposed to being known. The recognition accuracy is 85.4% due to insertions.

Table 2: *Bird species recognition correct (%) achieved by the maximum likelihood method for a given number of bird species present when using single HMM and individual element HMMs. Each species contained 3 seconds of the detected signal.*

Number of bird species present	Maximum Likelihood combination method	
	Single HMM (constraints: no / yes)	Element HMMs (constraints: no / yes)
1 species	92.0 / 92.0	95.8 / 95.8
2 species	81.2 / 84.7	94.9 / 95.7
3 species	72.5 / 77.6	93.4 / 94.0

Table 3: *Bird species recognition correct and accuracy (%) achieved by the maximum likelihood method when one, two, or three bird species are present in a given utterance of 3 seconds of the detected signal.*

Number of bird species	Maximum Likelihood combination method with element HMMs	
	Rec. Corr.	Rec. Acc.
known	91.3	91.3
estimated	89.2	85.4

4. Conclusion

This paper presented an extension of our work on recognition of multiple bird species. A method for detection of sinusoidal components was employed to decompose the acoustic scene into isolated time-frequency segments. Each segment was represented as a temporal sequence of 3 dimensional vectors, consisting of the detected sinusoid frequency and its temporal derivatives. Each bird species was represented by a set of HMMs, each HMM modelling individual vocalisation element type. Training of element HMMs was performed in an unsupervised manner. 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 on a given set of detected segments, with a penalisation applied for increasing the number of models used. Experimental results demonstrated that the use of element modelling and maximum likelihood segment score combination provided considerable improvements over previous results and over majority voting methods.

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] 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] J. R. Heller and J. 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, 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.

[6] F. Briggs, B. Lakshminarayanan, L. Neal, X. Fern, R. Raich, S. J. Hadley, A. Hadley, and M. 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. [Online]. Available: http://link.aip.org/link/?JAS/131/4640/1

[7] T. 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.

[8] 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.

[9] 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, Apr. 2015.

[10] S. Anderson, A. 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] W. Chu and D. 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.

[13] F. Briggs, R. Raich, Z. Lei, K. Eftaxias, and Y. Huang, "The Ninth Annual MLSP Competition: Overview," *IEEE Int. Workshop on Machine Learning for Signal Processing*, Sept. 2013. [Online]. Available: http://mlsp2013.conwiz.dk/competition.htm

[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] 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.

[16] P. Jančovič, M. Köküer, M. Zakeri, and M. Russell, "Bird species recognition using HMM-based unsupervised modelling of individual syllables with incorporated duration modelling," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China*, pp. 559–563, March 2016.

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

[18] C. Catchpole and P. Slater, *Bird Song – Biological Themes and Variations*. Cambridge University Press, 2008.

[19] 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.

[20] "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.

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