Automatic Detection of Bird Species from Audio Field Recordings using HMM-based Modelling of Frequency Tracks

Peter Jančovič and Münevver Köküer
Department of Electronic, Electrical & Systems Engineering, University of Birmingham, UK
{p.jancovic, m.kokuer}@bham.ac.uk

***Abstract*—This paper presents an automatic system for detection of bird species in field recordings. A sinusoidal detection algorithm is employed to segment the acoustic scene into isolated spectro-temporal segments. Each segment is represented as a temporal sequence of frequencies of the detected sinusoid, referred to as frequency track. Each bird species is represented by a set of hidden Markov models (HMMs), each HMM modelling an individual type of bird vocalisation element. These HMMs are obtained in an unsupervised manner. The detection is based on a likelihood ratio of the test utterance against the target bird species and non-target background model. We explore on selection of cohort for modelling the background model, z-norm and t-norm score normalisation techniques and score compensation to deal with outlier data. Experiments are performed using over 40 hours of audio field recordings from 48 bird species plus an additional 16 hours of field recordings as impostor trials. Evaluations are performed using detection error trade-off plots. The equal error rate of 5% is achieved when impostor trials are non-target bird species vocalisations and 1.2% when using field recordings which do not contain bird vocalisations.**

Index Terms: bird species detection, field recording, hidden Markov model, HMM, score normalisation, cohort, outlier, vocalisation, element, unsupervised training, sinusoid detection, sinusoidal modelling, frequency track

I. Introduction

Monitoring biodiversity can provide important information on environmental health, migration routes and population status of species for conservation planning and management. In ornithology, this is traditionally conducted by point call counts with field observers. However, point counts are prone to many problems, for instance, the assessment is very limited, the presence of observer may affect vocal activity of birds, as well as it is expensive, tedious and time consuming. An attractive alternative to the use of field observers is to automatically detect bird species from recordings made in the field.

There have been a number of studies on automatic bird species recognition and detection. Typically, the first stage of an automatic system is to parse the acoustic signal into isolated spectro-temporal segments. This is often performed using an energy-based thresholding that requires an estimate of noise level, e.g., [1], or by decomposition into sinusoidal components [1], [2], [3], [4]. A variety of approaches to feature representation of the spectro-temporal segments and their modelling were explored. The use of features extracted from entire frequency range, such as, conventional Mel-frequency cepstral coefficients which were used in a number of studies, e.g., [1], is problematic in the presence of other concurrent vocalisations or noise. The use of a set of statistical descriptors to characterise detected segment, as employed in [1], [2], [5], 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 refer to as frequency track. This representation was employed in a few earlier studies [1], [6] and also in our recent works [3], [4], [7], [8], [9], [10]. Among the acoustic modelling approaches, the most commonly used are Gaussian mixture models (GMM) [1], [3], hidden Markov models (HMMs) [1], [4], [6], [11], and decision trees [12]. Several studies focused on detection of specific bird species [13], [14], [15]. Bardeli et al [13] used an energy-based detection in pre-defined frequency region and autocorrelation for detecting repetition patterns of vocalisations of two endangered bird species. Digby et al [14] presented an assessment of an automatic detection to manual field surveys of a single species. They used autocorrelation to detect repeating calls, then extracted a set of statistical descriptors for each call and used these in a decision tree classifier. The authors in [15] used cepstral coefficients and GMM-based detector of a single bird species.

This paper presents an extension of our previous studies, which focused on acoustic modelling and were performed in the context of a closed-set bird species identification, to detection of bird species from field recordings. We use the sinusoidal detection algorithm introduced in [16] to parse the acoustic signal into isolated time-frequency segments and represent each segment using a temporal sequence of the frequency track features [4]. We employ HMM-based modelling of individual vocalisation elements, which are trained in an unsupervised manner. This model was shown to reduce species identification error rate by over 70% in comparison to the use of a single HMM per bird species [10]. The detection is performed using likelihood ratio on the target bird species model and a background model. We explore ways of cohort selection for the background model, weighting the likelihoods and score normalisation techniques which have been successfully employed in speaker verification research, e.g., [17],

ISBN 978-0-9928626-7-1 © EURASIP 2017

[18], [19]. We also introduce a modified score calculation to deal with the problem of data not seen during the training. Experimental evaluations are performed using over 40 hours of field recordings from 48 bird species from [20] plus 16 hours of non-bird audio from 'freefield1010' dataset [21]. Results are examined for case of impostor trials being vocalisations of different bird species and non-bird audio.

II. Automatic Bird Species Detection System

This section provides the description of individual components of the presented bird species detection system. It starts with a brief review of the approach we employed for segmentation of the audio signal and extraction of frequency track features and then follows with unsupervised HMM-based modelling of individual vocalisation elements of each bird species. These two components of the system were introduced in our recent publications [4], [8], [10] where we refer the reader to for further details. We then describe the methods we employed for detection of bird species, including score normalisation techniques.

A. Segmentation and estimation of frequency tracks

Audio signal is automatically parsed into isolated time-frequency segments, each segment corresponding to a temporal evolution of a sinusoidal component in the signal. The detection of sinusoidal components is performed in the short-time spectral domain using the method we introduced in [16]. This method considers each peak in the magnitude spectrum of a signal frame as a potential sinusoidal component and characterises it using a set of magnitude and phase spectral features extracted around the peak. The detection is performed based on the maximum likelihood criterion using trained models of sinusoidal signals and noise. The initial segmentation obtained from the sinusoidal detection is further refined by discarding very short segments and segments of a low energy, which are considered to be accidental detection errors or other background sinusoidal components.

B. HMM-based modelling of bird vocalisation elements

The modelling of vocalisations of each bird species is performed using left-to-right hidden Markov models (HMMs). A single HMM could be used to represent the entire set of vocalisations of a given bird species. However, we have demonstrated in our previous research [8], [10] that a better approach is to obtain an individual model of each type of vocalisation element. This is not straightforward if the element-level label information is not available and the set of element vocalisations produced by each bird species is unknown. As such, we first employed an unsupervised procedure to find a set of vocalisation elements. This performed an agglomerative hierarchical clustering of the detected segments based on a similarity score between each pair of segments obtained using a modified dynamic time warping (DTW) algorithm [7]. The modified DTW allowed for a partial match between segments. The similarity score for a given pair of segments was based on the cumulative DTW distance, length of the matching path and the ratio of the length of the matching path to the total length of the path. The number of clusters, i.e., number of vocalisation elements, could be estimated for each bird species, for instance, based on assessing the change in a cluster similarity score or cluster occupancy but in this paper it was set to a fixed value for all bird species. An additional cluster referred to as 'remainder' was used to cover segments which were not assigned to any of the element vocalisation clusters. The above provided an initial element-level label information for each detected segment which was used to estimate initial parameters of each element HMM using conventional Baum-Welch algorithm. This was followed with an iterative training procedure which consisted of an update of the element-level label information based on the current trained models and then Baum-Welch reestimation of the models. Further details of this procedure are provided in [8], [10]. 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. In addition to individual element HMMs, we also have a single 'remainder' HMM whose state pdf consists of several Gaussian mixture components as this model is to cover a variety of remaining segments. An example of the state output pdf of nine trained individual element HMMs of a given bird species is depicted in Figure 1. It can be seen that each model provides a distinctive pattern.

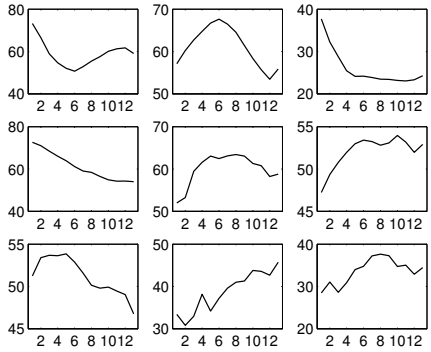

Fig. 1. An example of the mean values of the state output Gaussian pdf, modelling frequency track features, for nine trained element HMMs of bird species *Northern Cardinal*. The x- and y-axis denotes the HMM state and frequency index, respectively.

C. Detection of bird species

The objective in bird species detection is to determine whether a particular bird species of interest b is present in a given utterance of recording.

The training stage provides the model λ_b for the target bird species. For a given test 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. Each detected segment s is considered as an isolated vocalisation element. An approximation of the probability $p(O_s|\lambda_b)$ is obtained using

ISBN 978-0-9928626-7-1 © EURASIP 2017

the Viterbi algorithm on each individual element model, with the highest one being used.

The general approach used in detection is to base the decision on the likelihood ratio of the test utterance O against the target bird species model λ_b and the non-target model $\lambda_{\bar{b}}$, i.e., $p(O|\lambda_b)/p(O|\lambda_{\bar{b}})$. The bird species b is then detected if the ratio, which we refer to as score, is above a given threshold θ, and not detected otherwise. The decision threshold θ is set to adjust the trade-off between rejecting the true target bird species utterances, i.e., false rejection errors, and accepting non-target bird species utterances, i.e., false acceptance errors.

In order the decision threshold to be independent of the utterance length, the likelihood terms $p(O|\lambda)$ need to be normalised for the number of frames. As our utterance consists of a number of segments O_s, this could be performed in two ways. We have observed that better results are obtained by accumulating the likelihood of segments in the utterance and then normalising the total likelihood by the total number of frames, i.e., $[\prod_s p(O_s|\lambda)]^{\frac{1}{\sum_s T_s}}$, as opposed to using the mean likelihood of each segment.

While the calculation of the likelihood $p(O|\lambda_b)$ is clearly defined, as the model λ_b is available from the training stage, it is less so for the likelihood $p(O|\lambda_{\bar{b}})$. The model $\lambda_{\bar{b}}$ is usually referred to as 'world' or 'background' model.

1) 'Background' modelling and dealing with outliers: Ideally, a large amount of data covering well all the possible vocalisations of non-target bird species and all other sounds in the real world should be used. A single 'background' model can be built using all the non-target bird species sounds. An alternative approach, also adopted in this paper, is to use a collection or cohort of background models. The score for the utterance O and target bird species b, denoted as $\Lambda(O;\lambda_b)$, is then calculated in the log domain as

$$\Lambda(O;\lambda_b) = \log p(O|\lambda_b)^\gamma - \log\big(\frac{1}{N_{coh}} \sum_{c\in coh} p(O|\lambda_c)^\gamma\big) \quad (1)$$

where the likelihood terms $p(O|\lambda)$ are length-normalised as mentioned above, and N_{coh} is the number of models used in the cohort. We have also explored the use of a scaling factor γ for the likelihoods, which was employed in [19] for speaker verification.

However well we attempt to model background sounds, there may become situations when an impostor sound is not well covered by the background model. Such utterance of recording could be seen as an outlier. The likelihood of such utterance on each of the models would then be a small random value. This could result in the score becoming a large random number and causing false acceptance error. Various ways could be used to tackle this problem, e.g., [22]. Here, we propose to modify the score calculation to

$$\Lambda(O;\lambda_b) = \Lambda(O;\lambda_b) - Kf(p(O|\lambda_b)) \quad (2)$$

where $f(\cdot)$ is a sigmoid function of the form $f(x) = 1/(1+\exp(-\beta(\log p(O|\lambda_b) - \alpha)))$. The parameter α and β defines the shift and slope of the function, respectively. Suitable values could be set based on examining the distribution of values of $\log p(O|\lambda_b)$ on training data. The value K represents the penalty to be attributed to the score for an outlier data.

2) Score normalisation: The aim of score normalisation is to normalise the distribution of the scores. We employed zero-normalisation, z-norm, and test-normalisation, t-norm, score normalisation techniques. These have been extensively employed in the area of automatic speaker detection/verification [18]. The z-norm and t-norm both use the same score normalisation formula

$$\Lambda_{norm}(O;\lambda_b) = \frac{\Lambda(O;\lambda_b) - \mu_{norm}}{\sigma_{norm}} \quad (3)$$

where μ_{norm} and σ_{norm} are the mean and standard deviation normalisation parameters, respectively. However, they differ in how the normalisation parameters are computed. In the z-norm, the μ_{norm} and σ_{norm} parameters are estimated during the training stage based on a set of scores obtained when using the target bird species model against a set of impostor utterances. The t-norm is performed during the testing stage – the test utterance O is scored against a cohort of non-target (impostor) models to obtain a set of impostor scores, which are then used to estimate the μ_{norm} and σ_{norm} as the sample mean and standard deviation of the log-likelihood ratio scores.

III. Experimental evaluations

A. Experimental setup

Experimental evaluations were performed using over 40 hours of field recordings from the Borror Laboratory of Bioacoustics [20] and nearly 16 hours recordings from 'freefield1010' collection used in the Bird Audio Detection challenge [21]. The Borror collection contains recordings of bird vocalisations made 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. Each recording was split into training and testing part in proportion of two to one, respectively. The 'freefield1010' collection contains audio with 'field-recording' tag selected from the Freesound audio archive. From these data, only recordings marked as not containing bird vocalisations were used as impostor trials. The data used for testing was further split into utterances, where each utterance consisted of signal containing approximately 1 second of detected segments.

From the Borror collection, data from a set of randomly chosen 48 bird species were used. From this set, a sub-set of 18 bird species was used for training the 'background' model, and another sub-set of 6 bird species was used as impostor utterances for calculation of the statistics used in the z-norm score normalisation. The remaining sub-set of 24 bird species was used in a leave-one-out methodology – at a time, one bird species was used as the target bird species and data of the other 23 bird species were used for impostor trials.

ISBN 978-0-9928626-7-1 © EURASIP 2017

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 [23]. The parameter setup used for HMM-based modelling of vocalisations was based on our previous research outcomes [8], [10]. We used a left-to-right HMMs with 13 states and no skip allowed. The overall model of each bird species consisted of a set of 70 individual element HMMs using a single Gaussian per state plus an additional single general HMM having 10 components Gaussian mixture model per state.

Performance is evaluated using detection error trade-off plots, which have been used as the main performance measure for speaker verification tasks in NIST evaluations [24].

B. Experimental results

The first set of presented experimental results is obtained using the Borror dataset only. We analyse the effect of cohort selection and different schemes of score normalisation. Figure 2 presents results achieved by using different set of non-target bird species models in the cohort in Eq. 1 to represent the 'background' model. We can see that the use of all bird species, i.e., no cohort selection, performed worse then using unconstrained cohort selection with only few models achieving the highest likelihood. The performance when using only the best 1 model, which is not shown for clarity of the figure, was slightly worse than the use of the best 3 models. Next, Figure 3 presents results achieved when using different weight γ of the likelihoods in Eq. 1 for the score calculation. These experiments were performed with using all the models in the cohort. It can be seen that a suitable choice of the weight parameter γ can provide similar performance to the use of cohort selection. Finally, Figure 4 presents results obtained using the z-norm and t-norm score normalisation techniques again when all models in the cohort were used. It can be seen that the use of z-norm gives worse performance than the baseline. The t-norm, when all models are used for calculating the normalisation parameters, performs also poorly but with the use of only the best 3 models it provides improvement over the baseline model and achieves performance similar to that obtained using the cohort selection or likelihood weighting schemes. A combination of cohort selection and score weighting and normalisation did not provide further improvement.

The next set of experiments differ from the above only in the set of impostor trials used – instead of using the recordings of non-target bird species from the Borror dataset, now non-bird data from the 'freefield1010' collection are used as impostor trials. The model using likelihood weighting with γ=6 was used. We explore here also the effect of the outlier score compensation as in Eq. 2. Results are presented in Figure 5. It can be seen that while the use of bird vocalisations of other species as impostor trials, as in the first set of experimental evaluations above, gave the equal error rate (EER) just under 5% (red dotted line), the same system but with non-bird impostor trials (black dotted line) performs considerably worse, with EER increased to 12.5%. The other lines in the figure show the

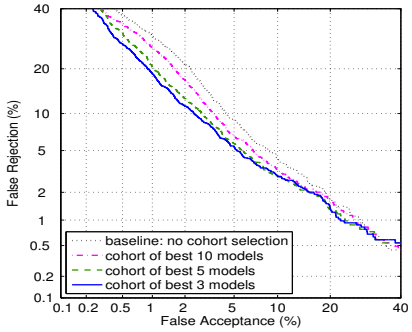

Fig. 2. Bird species detection results obtained using different unconstrained cohort selection.

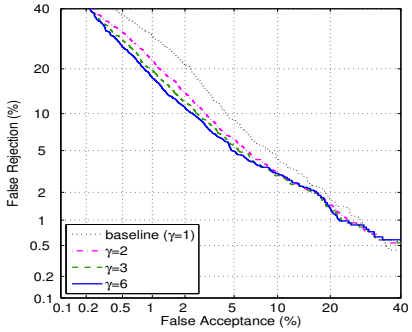

Fig. 3. Bird species detection results obtained using different likelihood weighting parameter γ when no cohort selection was used.

performance when employing the outlier score compensation. It can be seen that the performance improves by an order. The EER is now reduced to only around 1.2%. This demonstrates that the vocalisations of other non-target bird species present a considerably bigger challenge to the detection system than non-bird sounds. Note that the use of this compensation had negligible effect on results when bird vocalisations of other bird species from the Borror dataset were used as impostors.

In terms of employing the presented detection system for a long-term automatic acoustic monitoring of bird species, the impostor trials in the 'freefield1010' collection consisted of nearly 16 hours of recordings. Out of this, the sinusoidal detection algorithm found around 98 mins of potential vocalisation segments. As such, using the presented detection system with,

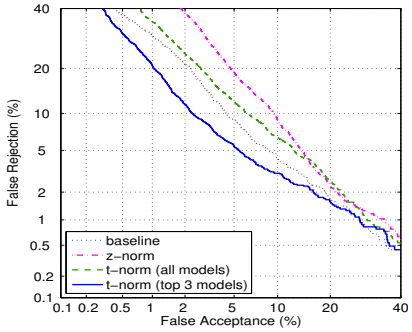

Fig. 4. Bird species detection results obtained using different score normalisation techniques.

ISBN 978-0-9928626-7-1 © EURASIP 2017

for instance, 1% false acceptance error rate setup would mean that less than 1 minute of audio would be incorrectly detected as target bird species in the total of 16 hours of continuous field recordings, while only 1.6% of target bird species vocalisations would be missed.

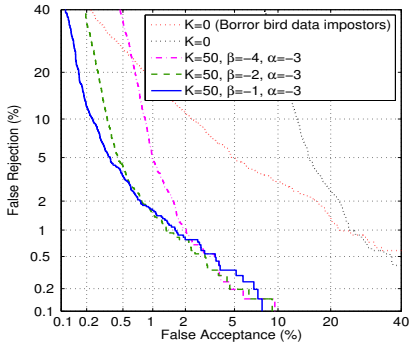

Fig. 5. Bird species detection results obtained using the system with likelihood weighting (γ=6) and employing outlier score compensation, when using the 'freefield1010' data as impostor trials.

IV. Conclusion

This paper presented an automatic bird species detection system. It 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 3-dimensional vectors, consisting of the detected sinusoid frequency and its temporal derivatives. Each bird species was represented using a set of HMMs, each HMM modelling individual type of vocalisation elements. The training of element HMMs was performed in an unsupervised manner. The detection was based on the likelihood ratio of the target model to background model. We explored the effect of cohort selection, likelihood weighting, t-norm and z-norm score normalisation, and outlier compensation on the detection performance. The first set of experiments was performed using over 40 hours of field recordings of bird vocalisations, with impostor trials being non-target bird species vocalisations. Except the z-norm, all the above score calculation techniques showed similar performance when using a suitable parameter setup. The second set of experiments was performed using impostor trials from field recordings not containing bird vocalisations. The equal error rate (EER) of 5% was achieved when impostors are other bird species vocalisations and 1.2% when using field recordings not containing bird vocalisations.

Acknowledgement

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

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] P. Jančovič, M. Zakeri, M. Köküer, and M. Russell, "HMM-based modelling of individual syllables for bird species recognition from audio field recordings," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Brisbane, Australia*, pp. 768–772, Apr. 2015.

[9] 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.

[10] 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.

[11] 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.

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

[13] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1524–1534, 2010.

[14] A. Digby, M. Towsey, B. D. Bell, and P. D. Teal, "A practical comparison of manual and autonomous methods for acoustic monitoring," *Methods in Ecology and Evolution*, vol. 4, no. 7, pp. 675–683, 2013.

[15] O. Jahn, T. D. Ganchev, M. I. Marques, and K.-L. Schuchmann, "Automated sound recognition provides insights into the behavioral ecology of a tropical bird," *PLOS One*, pp. 1–29, Jan. 2017.

[16] 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.

[17] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, pp. 91–108, 1995.

[18] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, Jan.-Jul. 2000.

[19] L. F. Lamel and J. L. Gauvain, "Speaker verification over the telephone," *Speech Communication*, vol. 31, no. 2-3, pp. 141–154, June 2000.

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

[21] D. Stowell and M. D. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," *https://arxiv.org/abs/1309.5275*.

[22] 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.

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

[24] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," *tech. rep., DTIC Document*, 1997.