Audio Diarization for Biodiversity Monitoring

Conference Paper · September 2015

DOI: 10.1109/AFRCON.2015.7331986

Ciira wa Maina
Dedan Kimathi University of Technology

Audio Diarization for Biodiversity Monitoring

Ciira wa Maina
Department of Electrical and Electronic Engineering
Dedan Kimathi University of Technology
Nyeri, Kenya
Email: ciira.maina@dkut.ac.ke

Abstract—**Biodiversity monitoring is important in assessing the state of an ecosystem and determining if conservation actions are required. This is particularly important when conservation resources are scarce. However, traditional methods of biodiversity monitoring are labour intensive and cannot be applied in every ecosystem where there is need. In order to expand the application of biodiversity monitoring, there is need to automate this important task. In this work we present an application of audio diarization methods for biodiversity monitoring and show how these methods can be used to measure the abundance of indicator taxa in areas of interest. The use of audio recordings has the potential to reduce the time and effort spent in biodiversity monitoring. The experiments are performed on a freely available dataset of bird song recordings with the birds serving as indicator taxa in the ecosystem of interest. We are able to estimate the number of bird species in the recordings and this information can be used to estimate the species richness in an ecosystem.**

I. Introduction

The world's rich biodiversity faces a number of threats including climate change, poaching and human encroachment into wildlife habitats. It is important to continuously monitor the state of different species in our ecosystem in order to prevent species loss and also to determine conservation priorities [1]. Biodiversity monitoring involves the measurement of the abundance of different species in the ecosystem with the aim of determining the state of the ecosystem [2].

Traditionally, biodiversity monitoring is carried out by conducting surveys that determine the abundance of various species in the ecosystem of interest. This task requires experts who are able to identify different animals and this makes traditional approaches to biodiverity monitoring very labour intensive and impractical to apply on a large scale. However, with increasing pressure on natural resources, it is important to increase biodiversity monitoring efforts. To this end modern technology can be brought to bear on this problem to automate as many tasks involved in biodiversity monitoring as possible. In addition to deploying appropriate technology, techniques for rapid biodiversity assessement (RBA) have been developed [3]. These techniques involve the selection of indicator taxa whose abundance is representative of the species abundance of other species present in the ecosystem.

A number of researchers have explored the use of bioacoustic techniques for biodiversity monitoring. These methodologies involve using audio recordings obtained in the wild to determine species abundance. In [4], Sueur *et al.* presented a method for acoustic biodiversity monitoring based on acoustic entropy. The idea is to compute the temporal and spectral entropy of audio recordings obtained in the wild and to associate the entropy with the number of species present in a recording. A number of studies have shown that the entropy of a recording is correlated to the number of species present in that recording [5].

In addition to measuring the entropy of recordings for biodiversity monitoring, acoustic recordings can be used to survey indicator taxa when these indicator taxa produce vocalizations. It has been shown that birds can serve as indicator taxa and since most birds produce vocalizations, audio recordings of bird species can be used to determine the biodiversity where these recordings are obtained. By integrating technology used to identify bird species from their recordings and methods for counting the number of individuals in a recording, systems that automate biodiversity monitoring can be developed. In [6], Adi *et al.* proposed a system capable of determining the abundance of a single bird species, the Norwegian ortolan bunting, from audio recordings. The method proposed extended work used in speaker diarization to perform automatic censusing of the birds from their recordings.

Speaker diarization aims at assigning various segments of an audio recording to the speakers that produced them [7]. It has numerous applications such as transcription of meetings and broadcast news. The methods employed in speaker diarisation can be used to perform biodiversity assessment by segmenting an audio recording obtained in the wild and assigning various segments to the species that produced them. In this work we aim to demonstrate the application of speaker diarisation methods to biodiversity monitoring. We extend the work of Adi *et al.* to recordings of multiple species and we show that the number of individuals identified in a recording using speaker diarisation techniques can be used to determine the number of species in the recording.

II. Speaker Diarization

Given an audio recording with multiple speakers, speaker diarization solves the problem of 'who spoke when' [7]. In this work we assume that the entire audio stream has been parameterised as a sequence of feature vectors or frames. In speech applications, Mel frequency cepstral coefficients (MFCCs) have been very successful and are widely used [7]. The speaker diarization system assigns the individual feature vectors to the speakers who spoke them.

Speaker diarization systems consist of several subsystems [7]. First, most diarization systems have a speech detection system that discards non-speech frames. Simple energy based voice detectors have been applied successfully to this task.

978-1-4799-7492-4/15/$31.00 ©2015 IEEE

Second, a change point detection system determines the locations in the audio stream where a change in speaker has occurred. This system produces segments of feature vectors that are produced by the same speaker. Third, a clustering step is used to group together disjoint segments that are produced by the same speaker.

A. Change Point Detection

In this work we employ the change point detection algorithm described in [8] where change point detection is treated as a maximum likelihood model selection problem with the change points determined using the Bayesian information criterion (BIC).

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be the set of MFCCs derived from a recording. To determine if there is a change point at frame $\mathbf{x}_i$, we want to chose between the following hypotheses

$$\begin{aligned} H_0 &: \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \sim \mathcal{N}(\mu, \Sigma) \\ H_1 &: \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i \sim \mathcal{N}(\mu_1, \Sigma_1) \\ &\quad \mathbf{x}_{i+1}, \ldots, \mathbf{x}_N \sim \mathcal{N}(\mu_2, \Sigma_2) \end{aligned}$$

Under H_0, all the frames are independent and identically distributed according to $\mathcal{N}(\mu, \Sigma)$. Under H_1, there is a change point at frame $\mathbf{x}_i$ and the frames $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i$ are assumed to be distributed according to $\mathcal{N}(\mu_1, \Sigma_1)$ while the frames $\mathbf{x}_{i+1}, \ldots, \mathbf{x}_N$ are distributed according to $\mathcal{N}(\mu_2, \Sigma_2)$. We treat change point detection as a model order selection problem. In H_0 we model the data using a single Gaussian and in H_1 we model the data using two Gaussians. The difference in BIC values for the two models is given by

$$\Delta BIC(i) = N \log |\Sigma| - N_1 \log |\Sigma_1| - N_2 \log |\Sigma_2| - \lambda P \quad (1)$$

where

$$P = \frac{d(d+3)}{4} \log N$$

is a penalty term used to account for the different number of parameters in the two models. d is the dimension of the feature vector $\mathbf{x}_i$. λ is a penalty weight usually set to 1.

When $\Delta BIC(i)$ is positive, the model with two Gaussians is preferred pointing to the possibility of a change point at frame i. To determine the most likely location of a change point, $\hat{i}$, in the N frames, we compute

$$\hat{i} = \mathsf{argmax}_i \Delta BIC(i) \quad (2)$$

provided the value $\Delta BIC(\hat{i})$ is greater than zero.

B. Clustering

Once change points have been detected, the audio segments must be clustered so that segments produced by a single speaker are clustered together. In this work we follow [7] and use an agglomerative clustering approach. Here the segments are successively merged based on the BIC criterion which determines whether two segments are best modelled as two independent Gaussian processes or a single Gaussian.

Consider two segments 1 and 2 with N_1 and N_2 frames respectively. Initially the segments will be in separate clusters. We represent each cluster by a single full covariance Gaussian and merge the two segments if they are better modelled using a single Gaussian. In this case the change in the BIC value ΔBIC is given by

$$\begin{aligned} \Delta BIC = \frac{1}{2}\Big\{ &(N_1 + N_2) \log |\Sigma_{12}| \\ &- N_1 \log |\Sigma_1| - N_2 \log |\Sigma_2| \Big\} - \lambda P \end{aligned} \quad (3)$$

where

$$P = \frac{d(d+3)}{4} \log(N_1 + N_2).$$

Σ_{12} is the covariance matrix of the single Gaussian distribution modelling the merged clusters. The ΔBIC value is computed for all pairs of clusters and the two clusters for which ΔBIC is lowest are merged. This is repeated until the lowest ΔBIC is greater than zero [7].

III. Experimental Validation

To demonstrate the use of speaker diarization methodology for biodiversity monitoring, we tested the ability of a speaker diarization system to segment and cluster an audio recording obtained in the wild containing several bird species. The data set we used was recorded in the H. J. Andrews Long-Term Experimental Research Forest in Oregon and contained recordings of 19 bird species (HJA dataset)[9]. The data set consists of 645 10 second recordings each with between one and six bird species. The recordings are in wav format and sampled at 16kHz. The data set was used in the 2013 Machine Learning for Signal Processing (MLSP) competition and is freely available[1].

We implemented the change point algorithm of [8] and the agglomerative clustering approach described in [7] in python. Our implementation is freely available on github as a package called BirdPy which can be downloaded here https://github.com/ciiram/BirdPy. Our implementation makes use of the open toolkit Bob to compute MFCCs [10].

Of the 645 recording in the HJA dataset, data on the number of species present in the recording is available for 179 recordings. For each of these recordings we first obtain 12 dimensional MFCCs with no energy and determine change points in the audio stream. We then cluster the segments to obtain an estimate of the number of species present in this recording. Due to the environmental noise present, we do not employ an initial bird/non-bird sound classification step. Instead we assume that one cluster will represent all the non-bird sounds.

A. Preliminary Experiments on Individual Recordings

Figure 1 shows the spectrogram of a recording which contains a single species (the Stellar's Jay) while Figure 2 shows the spectrogram of a recording which contains two species (the Stellar's Jay and the Warbling Vireo) [9]. From the figures, we see that the location of the vocalizations can be clearly seen from the spectrograms. In the single species recording, a single spectral signature can be seen. In the two species recording, the two spectral signatures can also be identified visually.

The results of applying the change point detection algorithm to the recordings are shown in Figures 3 and 4. We see

[1] https://www.kaggle.com/c/mlsp-2013-birds/data

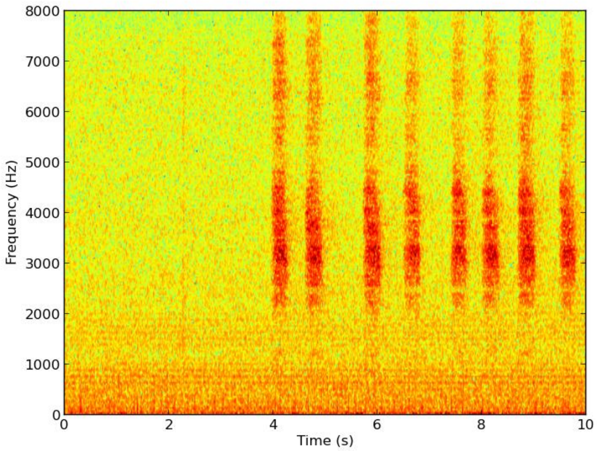

Fig. 1. Spectrogram of a recording which contains a single species namely the Stellar's Jay

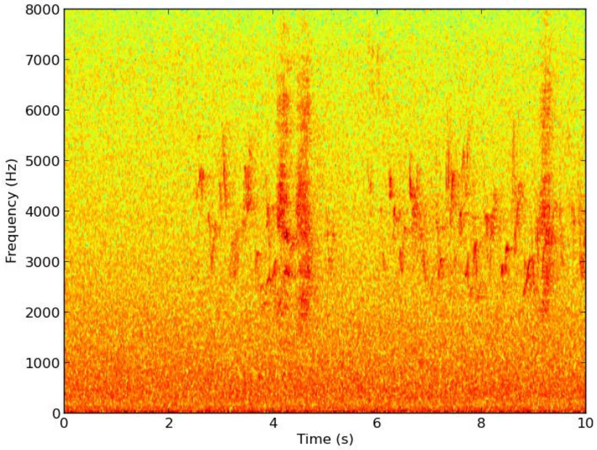

Fig. 2. Spectrogram of a recording which contains two species namely the Stellar's Jay and the Warbling Vireo

that the change points located by the algorithm correspond to the change points that can be visually identified from the spectrograms. However, some spurious change points are also included. These include false detection of change points in regions without vocalization and segmentation of homogeneous regions containing the recording of one species into multiple regions.

Once the change points in the audio stream are determined, the segments are clustered to form homogeneous regions which contain the vocalization of a single species. The result of clustering the segments shown in Figures 3 and 4 are shown in Figures 5 and 6 respectively. From this result we see that homogeneous regions which had been divided into several segments are merged in the agglomerative clustering stage. In addition, different segments produced by the same species are assigned to the same cluster. The segments from the recording which contains one species are assigned to two clusters designated 0 and 1 in Figure 5. One of these clusters corresponds to non-bird sounds and the other corresponds to

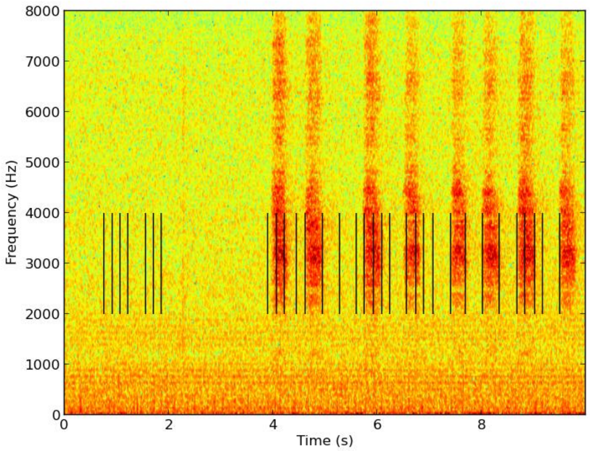

Fig. 3. Result of performing change point detection on the recording of Figure 1. Change points are indicated by vertical black lines.

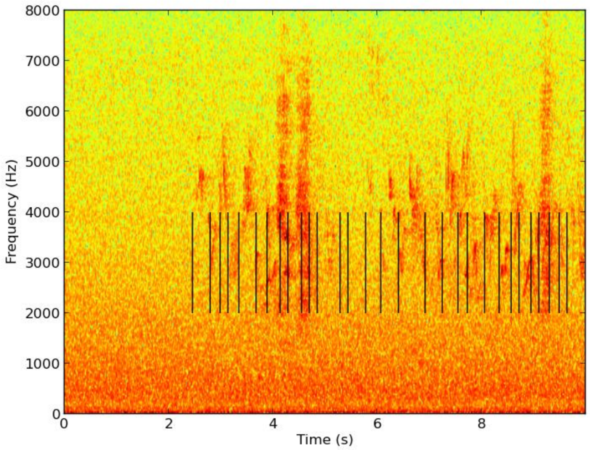

Fig. 4. Result of performing change point detection on the recording of Figure 2. Change points are indicated by vertical black lines.

vocalizations of the species (Stellar's Jay). On the other hand, the segments from the recording with two species are assigned to three clusters designated 0, 1 and 2 in Figure 6.

B. Species Abundance Determination

Based on the results in the previous section, we see that the number of unique clusters determined during the agglomerative clustering step can be used to estimate the number of species present in the recording. Assuming that one cluster contains all the non-bird sound segments, the number of species is obtained by subtracting one from the number of clusters obtained during agglomerative clustering.

Using the entire set of 179 recordings for which data on the number of species present in the recording is available, we segmented the audio stream based on the change points detected and then clustered the segments. The final number of clusters after agglomerative clustering was then used to estimate the number of species present in the recording. Figure

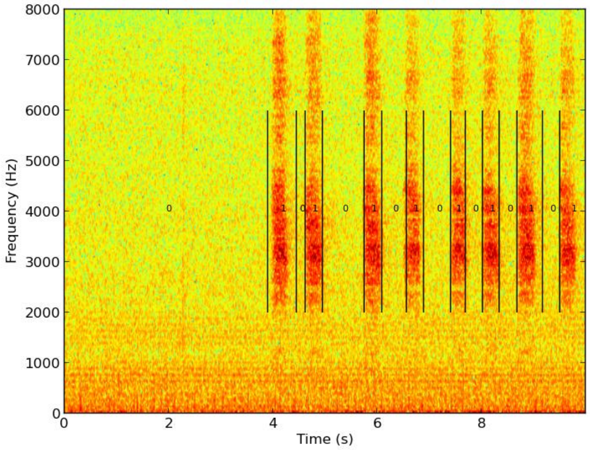

Fig. 5. Result of performing agglomerative clustering on the recording of Figure 1. Merged segment boundaries are indicated by vertical black lines. Cluster numbers are indicated within the segment. Here there are only two unique clusters namely 0 and 1 resulting in a the correct estimate of 1 species present.

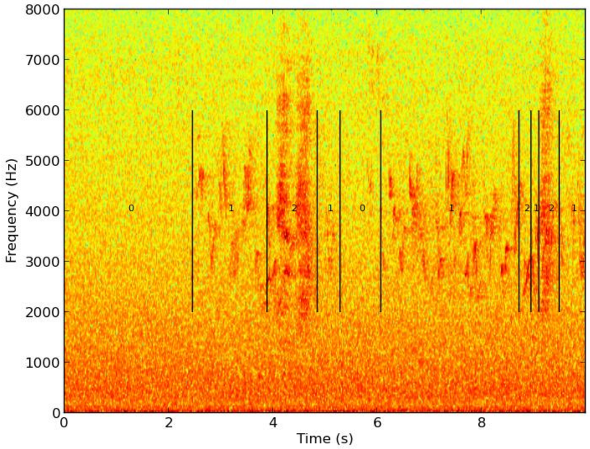

Fig. 6. Result of performing agglomerative clustering on the recording of Figure 2. Merged segment boundaries are indicated by vertical black lines. Cluster numbers are indicated within the segment. Here there are three unique clusters namely 0, 1 and 2. We correctly estimate the number of species present as 2.

7 shows a scatter plot of the number of species present in the recording versus the number of clusters determined via agglomerative clustering. From this figure a linear trend is clear. Of the 179 recordings, there are 76 recordings for which the number of species is correctly estimated. That is the accuracy is 42.5%.

By examining the spectrograms of the recordings, we see that there is significant noise in the spectral band below 2kHz. This noise is due to wind, rain and other unwanted sound sources. Also, there is little signal of interest in this band. This suggests that passing the signal through a high pass filter (HPF) will remove noise and possibly improve system performance. To test this we filtered the recordings using a 10th order Butterworth HPF with cut off frequency ranging from 500Hz to 4kHz. Table I shows the system accuracy as a function of the cut-off frequency. We see that the best performance of 52% is achieved with a cut-off frequency of 3kHz. As we increase the cut-off beyond this point the performance begins to deteriorate once again indicating that significant signal is present above 3kHz.

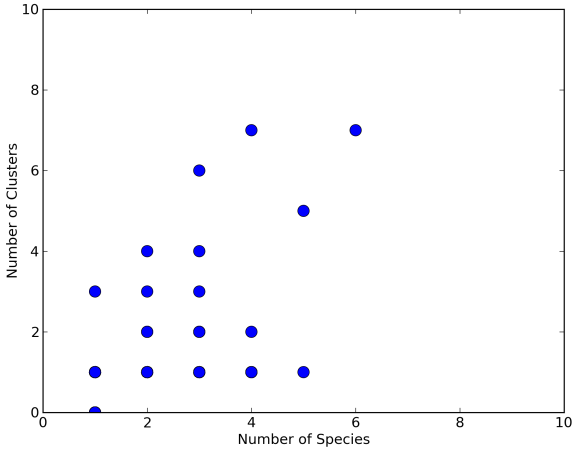

Fig. 7. Scatter plot of the number of species present in the recording versus the number of clusters determined via agglomerative clustering.

TABLE I. EFFECT OF PASSING THE SIGNAL THROUGH A HIGH PASS FILTER ON PERFORMANCE.

Cut-off Frequency (kHz)	Accuracy (%)
0.5	38.5
1	38.5
1.5	42.5
2	44.7
2.5	48.6
3	**52.0**
3.5	45.8
4	38.5

IV. CONCLUSION

In this work we have demonstrated the use of speaker diarization techniques in biodiversity assessment. The system works by estimating the number of bird species in a recording and using this as an estimate of the species richness in the area from which the recording was taken. We obtain satisfactory results using publicly available data and in future work we aim to test the methodology on recordings obtained from the wildlife conservancy located within Dedan Kimathi University of Technology.

REFERENCES

[1] N. Myers, R. A. Mittermeier, C. G. Mittermeier, G. A. Da Fonseca, and J. Kent, “Biodiversity hotspots for conservation priorities,” *Nature*, vol. 403, no. 6772, pp. 853–858, 2000.

[2] J. H. Lawton, D. Bignell, B. Bolton, G. Bloemers, P. Eggleton, P. Hammond, M. Hodda, R. Holt, T. Larsen, N. Mawdsley *et al.*, “Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest,” *Nature*, vol. 391, no. 6662, pp. 72–76, 1998.

[3] J. T. Kerr, A. Sugar, and L. Packer, “Indicator taxa, rapid biodiversity assessment, and nestedness in an endangered ecosystem,” *Conservation Biology*, vol. 14, no. 6, pp. 1726–1734, 2000.

[4] J. Sueur, S. Pavoine, O. Hamerlynck, and S. Duvail, “Rapid acoustic survey for biodiversity appraisal,” *PLoS One*, vol. 3, no. 12, p. e4065, 2008.

[5] M. Depraetere, S. Pavoine, F. Jiguet, A. Gasc, S. Duvail, and J. Sueur, “Monitoring animal diversity using acoustic indices: implementation in a temperate woodland,” *Ecological Indicators*, vol. 13, no. 1, pp. 46–54, 2012.

[6] K. Adi, M. T. Johnson, and T. S. Osiejuk, “Acoustic censusing using automatic vocalization classification and identity recognition,” *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 874–883, 2010.

[7] S. E. Tranter and D. A. Reynolds, “An overview of automatic speaker diarization systems,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557–1565, 2006.

[8] S. Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the bayesian information criterion,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*. Virginia, USA, 1998, p. 8.

[9] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, Z. Lei, W. Cukierski, S. F. Hadley, A. Hadley, M. Betts, X. Z. Fern *et al.*, “The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment,” in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–8.

[10] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, “Bob: a free signal processing and machine learning toolbox for researchers,” in *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*. ACM Press, Oct. 2012. [Online]. Available: http://publications.idiap.ch/downloads/papers/2012/Anjos_Bob_ACMMM12.pdf