Entropy-based segmentation of birdcalls using Fourier transform phase

Author Name[1], *Co-author Name*[2]

[1]Author Affiliation
[2]Co-author Affiliation
author@university.edu, coauthor@company.com

Abstract

In this paper we describe an entropy-based algorithm for the segmentation of birdcalls from recordings. The entropy of time-frequency blocks are estimated from the phase of the Fourier transform. To overcome difficulties in processing the phase, the group delay function from an all-pole filter is utilised. The group delay function has good frequency resolution properties, and hence provides reliable estimates of the entropy. Furthermore, spectral whitening is performed to smooth the entropy estimate and the extremities are determined. A threshold is applied on the difference to distinguish the call periods from the background. The algorithm is evaluated on two different datasets, one of which is recorded in more challenging field conditions. When compared to entropy estimated from the power spectrum, the entropy from the group delay function provides better detection accuracy at almost all operating points. The choice of model order of the all-pole filter for different bird species is also briefly investigated.

Index Terms: bioacoustics, birdcall segmentation, Fourier transform phase

1. Introduction

With the advent of automated recording devices, the collection of large amounts of bioacoustic data has become relatively easy. By analysing birdcalls collected in this manner, it is possible to perform tasks such as the tracking of migrant species or examining the avian biodiversity of a given region. Typically, the collected data is processed offline. In this process, the first step is usually to determine regions of interest in the recording. An entropy-based bird phrase segmentation technique was developed in [**?**]. In this paper, we propose a modified version of that technique, by using information from the phase of the short-term Fourier transform (STFT). Most techniques for processing speech and audio signals have utilised the magnitude spectrum of the STFT. Although the phase spectrum of the STFT has useful information, its processing has remained difficult. A popular technique for exploiting information from the phase has been through group delay functions. In this work, we utilise information from group delay functions using parametric models, and apply it to segment birdcalls into active and inactive regions.

The group delay function has good frequency resolution properties, which enable it to be useful in tasks such as speech recognition and speaker recognition [**?**] [**?**] [**?**]. The same property is beneficial in the processing of bird vocalizations. In [**?**], the entropy within a sliding time-frequency window over the spectrogram has been effectively used for distinguishing active and inactive regions. The essential idea is that birdcalls have more structure (for eg. harmonics may be present), and thus have lower entropy when compared to background sounds, which have higher entropy. This difference in entropy levels enable effective distinction between birdcalls and the background. The high resolution property of group delay functions enable accurate tracking of time-frequency information [**?**]. In this work, the entropy of a sliding time-frequency window is estimated from the group delay representation. Spectral whitening is applied to smooth the entropy estimates and thresholding on differences of extrema is applied to separate the birdcalls from the background.

Most of the bioacoustic studies have used manually segmented bird calls [**?**] [**?**] [**?**]. Time domain segmentation using energy has been used in many studies [**?**] [**?**] [**?**] . The energy based segmentation method is highly influenced by background noise and will deteriorate where bird calls have low energy in comparison to the background. A KL-divergence based segmentation method is proposed in [**?**]. KL-divergence between normalized power spectral density of a frame and uniform distribution is computed. Local minima of KL divergence act as change points for bird vocalizations. In [**?**] , time-frequency domain based segmentation using random forest classifier is proposed to segment the syllables from noisy audio signal.

2. Utilising Fourier transform phase

Commonly used features for processing speech and audio signals are based on the magnitude spectrum of the short-term Fourier transform. The phase spectrum has received relatively lesser attention due to signal processing difficulties, one of them being the need to unwrap the phase spectrum. The unwrapping problem can be bypassed by utilising the group delay function, which is the negative derivative of the phase spectrum. The group delay function can be computed using properties of the Fourier transform, and hence avoids the need for explicit computation of the phase spectrum [**?**]. However, this method can produce artifacts in the form of spurious peaks at spectral nulls. These nulls correspond to zeros close to the unit circle when the vocal tract transfer function is represented in the Z domain. Several methods have been proposed in the literature to overcome the effects due to these artifacts [**?**] [**?**]. Another technique to overcome this difficulty is to model the vocal tract as an all-pole filter, hence avoiding the nulls altogether. Such a technique derived using linear prediction analysis was used in the detection of formants in human speech [**?**].

There is strong evidence that birds use their vocal tract as a selective filter to modify the final sound [**?**]. Given this, the source-filter model developed for analysing human speech can be applied to bird vocalizations as well. Linear predictive (LP) analysis of human speech signals models the vocal tract spectrum as an all-pole filter [**?**] excited by a single source. When applied to birdcalls, this is a simplification of the 'two-voice' theory of avian vocalization [**?**], in that there is assumed to be

only one source, rather than two. Nevertheless, this reasonable assumption is followed in this work. A similar assumption has been made in [**?**], where LP analysis has been applied in analysing the song of the greater racket-tailed drongo.

The vocal tract is represented in the LP model as

$$H(\omega) = \frac{G}{1 - \sum_{k=1}^{P} a(k) e^{-j\omega k}}, \quad (1)$$

where the predictor model order is P, G represents the gain and $a(k)$ are the predictor coefficientses [**?**]. The filter reprsented by $H(\omega)$ is an all-pole filter, and its group delay function does not suffer from the artifacts mentioned earlier. The group delay function computed in this manner is termed as all-pole group delay function (APGDF.) Figure **??** shows the magnitude spectrum, LP spectrum and APGDF derived from a 20 ms call of Cassins vireo (*Vireo cassinii.*) As can be seen, the APGDF emphasises the formants, as compared to the DFT magnitude spectrum or the LP magnitude spectrum. Two peaks which are merged in the magnitude spectra around the 100th frequency bin appear distinctly in the APGDF.

Figure 1: A frame of audio signal (top panel), corresponding LP magnitude spectrum superimposed on DFT magnitude spectrum (middle panel) and all-pole group delay function (bottom panel)

Recently, feature vectors derived from such a representation were used in speech [**?**] and speaker recognition [**?**].

3. Entropy-based segmentation of birdsong

Computing the APGDF for every frame enables good time-frequency resolution of the audio recording. Unlike the power spectrum, the APGDF can be negative. Since here we are interested only in the magnitudes and locations of the frequency components, the sign of the APGDF is ignored by taking the absolute value. Henceforth, APGDF means positive APGDF. A sliding time-frequency window of width w frames and frequency range $fmin$ to $fmax$ is considered over the APGDF vectors. As in [**?**], the entropy of this window is estimated using the expression

$$h_k = - \sum_{n=kT+1}^{kT+w} \sum_{f=fmin}^{fmax} \tau_N(n,f) \ln \tau_N(n,f), \quad (2)$$

where T is the time-frequency window shift and $\tau_N(n,f)$ is normalized APGDF. $\tau_N(n,f)$ is estimated using the expression

$$\tau_N(n,f) = \frac{\tau(n,f)}{\sum_{n=kT+1}^{kT+w} \sum_{f=fmin}^{fmax} \tau(n,f)}, \quad (3)$$

where $\tau(n,f)$ is the positive APGDF at frame n and frequency *f*.

The entropy of the time-frequency window containing bird call is lower than the one only containing the background. As the time-frequency window starts moving to the call region from background, entropy starts dipping. The entropy remains low if time-frequency window completely overlaps with the call period. The entropy starts increasing as the window moves from call region to background. The entropy calculated from time-frequency window is less susceptible to sudden changes in background as compared to the entropy calculated at each time instance [**?**].

3.1. Whitening APGDF before Entropy Calculation

The drop in entropy during the presence of bird vocalizations makes it possible to detect call periods. Due to the presence of various background sounds e.g. rain, thunder, other animals etc., entropy of the background can vary rapidly making it difficult to separate it from entropy at a call period. To mitigate this problem to some extent, APGDF is whitened before calculating entropy. The entropy calculated from whitened APGDF is almost constant for background but dips enough to mark the presence of bird vocalizations. Due to this nature of background entropy, small change in entropy can be detected reliably.

To whiten the APGDF ($\tau(n,f)$), the covariance matrix C of the mean-subtracted APGDF is estimated. The eigenvalue matrix S and the eigenvector matrix U of C are determined. The whitened APGDF ($\tau_W(n,f)$) is calculated as:

$$\tau_W(n,f) = \text{diag}\left(\frac{1}{\sqrt{\text{diag(S)} + \epsilon}}\right) \times \text{U' } \times \tau(n,f), \quad (4)$$

Henceforth, the whitened APGDF is used to calculate entropy. The difference between entropy calculated from normal phase spectrum and whitened phase spectrum is evident in Figure **??**.

Figure 2: Entropy calculated from APGDF (Uppen Panel) and entropy calculated using white APGDF (Lower Panel).

3.2. Detecting change points using thresholding

To detect the change points, extrema-based thresholding is used. Local minima and maxima are estimated on the entropy. The thresholding is applied on the difference of consecutive local maxima and local minima to determine the change point. Two contiguous change points correspond to the start and end of a bird vocalization. These change points can be tracked back to get the start and end time of the vocalization in sound recording.

4. Experimentation and Performance Analysis

The proposed algorithm is evaluated on two datasets. The first consists of recordings of a single species, the Cassin's vireo, and second is a subset of the MLSP 2013 bird classification challenge, consisting of recordings from several species. In the Cassin's vireo dataset [**?**], the total duration of recordings is about 45 minutes, out of which about 5 minutes correspond to the calls. These recordings are fairly clean and have less background noise. Phrase annotations are provided with the dataset. The start and end of these annotations are used as the ground truth.

The MLSP 2013 dataset [**?**] is much nosier and includes wind and rain in the background.The dataset is collected in the H. J. Andrews (HJA) long-term experimental research forest in Oregon. The recordings are done for over two years at 13 different locations. Twenty Five files of ten seconds length are taken from the data to evaluate the proposed method. Manual annotations are done to mark the presence of bird vocalizations in these audio files.

The proposed algorithm is compared with a modified version of technique proposed in [**?**]. The entropy estimated from whitened power spectrum is used to determine the change

points. Similar to the proposed algorithm, extrema-based thresholding is referred to determine the change points. Thus, entropy from power spectrum is compared to the entropy from APGDF.

The frame length of 20 ms and increment of 5 ms is used to divide the audio signal into the frames. For each frame, APGDF vector is calculated. The whitening is applied to APGDF before calculating entropy. A time-frequency window of length w=138.8 ms is used along with increment of T = 15 ms to estimate the entropy. The frequency range of the time-frequency window is set from f_{min}=1.5 kHz to f_{max}=7 kHz [**?**].

Receiver operating characteristic (ROC) curves are used to analyze the performance of both the techniques. True positive rate and false alarm rate are calculated using following equations:

$$\text{TPR}(\%) = \frac{\text{Correctly Classified Call Frames}}{\text{Total Call Frames}} \times 100 \quad (5)$$

$$\text{FAR}(\%) = \frac{\text{Wrongly Classified As Call Frames}}{\text{Total Background Frames}} \times 100 \quad (6)$$

Figure **??** depicts ROC curves comparing performance of the methods based on power spectrum and APGDF on Cassin's Vireo dataset. It is clear from figure **??** that the APGDF method outperforms the power spectrum method.

Figure 3: ROC curves comparing performance of methods based on APGDF and spectrogram on Cassin's Vireo dataset.

Figure **??** shows ROC curves comparing performance of methods based on entropy calculated from APDGF and spectrogram on MLSP dataset. Here too, APGDF based method outperforms spectrogram based method for almost all the operating points.

Figure 4: ROC curves comparing performance of methods based on APGDF and spectrogram on MLSP dataset

4.1. Model Order vs AIC

To establish the optimum model order, Akaike Information criteria (AIC) [**?**] is used. Figure **??** depicts the AIC for different model orders for Cassin's Vireo phrases.

Figure 5: AIC vs model order for Cassin's Vireo

Figure **??** depicts the AIC for different model orders for three different bird species i.e. Cuckoo, Great Barbet and Laughing Thresh.

Figure 6: AIC vs model order for three different species

5. Conclusion

We propose an entropy based bird vocalization segmentation method where entropy is calculated from Group Delay phase spectrum. It is also established that whitening the power or phase spectrum before entropy calculation improves the performance of entropy based segmentation. From experimentation,it is clear that the proposed group delay based method outperforms the power spectrum based method.