Segmentation of birdsong

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

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

Abstract

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

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 propreties, 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. These differences in entropy levels enable effective distinction between birdcalls and the background. Applying a similar technique with the group delay function, rather than the spectrogram, provides increased frequency resolution, and hence more effective entropy computation. Spectral whitening is applied to smooth the entropy estimates and simple thresholding is applied to separate the birdcalls from the background.

Other methods for segmentation of birdsong include

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. 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. Such a technique derived using linear prediction analysis was used in the detection of formants in human speech [**?**]. More recently, feature vectors derived from such a representation was used in speech [**?**] and speaker recognition [**?**].

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. 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. In [**?**], LP analysis has been applied in analysing the song of the greater racket-tailed drongo. Thus, modeling the avain vocal tract as an all-pole filter is a reasonable assumption.

3. Entropy-based segmentation of birdsong

This section describes how to calculate entropy from spectrogram and phase spectrum, also how to use entropy to identify bird vocalizations. In [**?**], the entropy of spectrograms from recordings can be effectively used for distinguishing between background and bird vocalizations [**?**]. Spectrogram of single bird song is generally sparse i.e. high power components acquire only a small portion of time-frequency bins and the background noise of spectrogram is relatively white. Hence the entropy of sliding time frequency block over spectrogram is low when block contains a signal and is high when only background is present in that block. It is known that phase spectrum has more information than magnitude spectrum. So, phase spectrum can also be used to calculate entropy. Group delay functions from all pole models are used to calculate the phase spectrum.

3.1. Entropy Calculation

Entropy is calculated for each time-frequency block on power spectrum. This time-frequency block of time length ***w*** and having F frequency bins ranging from ***f1*** to ***fn*** is moved horizontally from beginning to the end of power spectrum. The frequency range is different for each target species. ***p(n,f)*** is power spec-

trum at time ***n*** and frequency ***f***. The entropy is calculated using following equation:

$$h_k = \sum_{n=kT+1}^{kT+w} \sum_{f=f1}^{fn} z(n,f) \ln z(n,f)$$

Here ***T*** is time-frequency block shift and ***z(n,f)*** is normalized power spectrum.

$$z(n,f) = \frac{p(n,f)}{\sum_{n=kT+1}^{kT+w} \sum_{f=f1}^{fn} p(n,f)}$$

The entropy calculated from block is less susceptible to the bursts of background noise in comparison to the entropy calculated at each time instance.

3.2. Whitening Spectrogram before Entropy Calculation

To identify the bird vocalizations using entropy, there should be a clear distinction between the entropy of background and call period. However this is not always the case in raw sound recordings. Depending on the interference level, entropy at a quiet period can even be higher than the entropy at a call period or bird call activity [1]. To overcome this problem, whole spectrogram or power spectrum is whitened using PCA before calculating entropy. The entropy calculated from whitened spectrogram is almost constant for background but dips enough to mark the presence of bird vocalizations. Even low energy bird vocalizations can be detected accurately using this method. The missed detection rate is decreased if whitened spectrogram is used for entropy calculation.

To whiten the spectrogram (*PS*), the covariance matrix of mean subtracted spectrogram is calculated. The Eigen values matrix (*S*) and Eigen vectors matrix (*U*) of this covariance matrix are calculated. The spectrogram matrix is whitened using the following equation:

$$WhitePS = diag(\frac{1}{\sqrt{diag(S)+\epsilon}}) * U' * PS$$

Figure 1 depicts the difference between entropy calculated from normal spectrogram and whitened spectrogram:

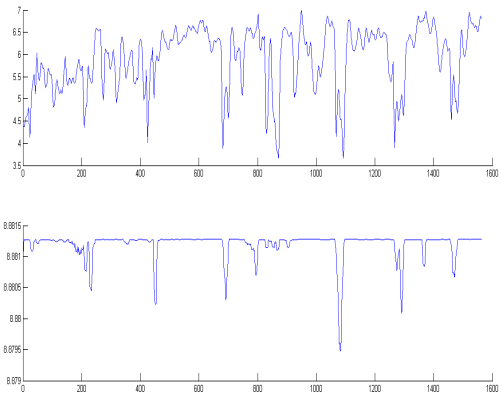

Figure 1: Entropy calculated from normal spectrogram and whitened spectrogram

It becomes evident that identifying bird vocalizations using entropy calculated from whitened spectrogram is easier.

3.3. Entropy Calculation from phase spectrum

3.4. Detecting change points using thresholding

To detect the change points, thresholding is used. The local minimums and local maximums are calculated on the entropy. The difference between consecutive local minimums and local maximums is calculated. If this difference is greater than predefined threshold, then corresponding local maximum is considered as the start of a bird vocalization or a change point. Then the difference between corresponding local minima and next local maxima is calculated. If this difference is greater than the threshold, local maxima is considered as the end of bird vocalization or another change point. Hence 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. Figure 2 shows the local maximums and local minimums on an entropy plot along with change points calculated from thresholding.

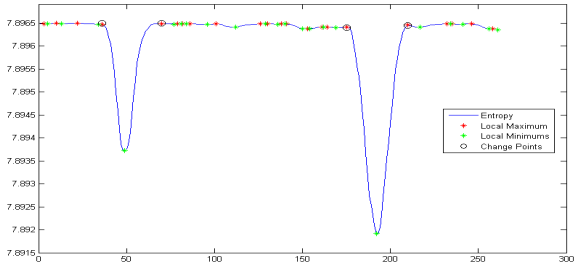

Figure 2: Change points generated by thresholding and Extrema calculated on entropy

4. Experimentation and Performance Analysis

For experimentation, the labeled recordings of Cassins Vireo (*Vireo cassinii*) and single species MLSP Bird Classification Challaenge 2013 datasets are used [**?**].

In Cassin's Vireo dataset, the total duration of recordings is about 45 minutes. Out of 45 minutes, about 5 minutes of recordings correspond to the phrases of Cassin's Vireo. To calculate spectrogram, frame length of 20 ms and increment of 5 ms is used. The time-frequency window of 138.8 ms is used along with increment of 15 ms to calculate entropy. The frequency range of the block is from 1.5 kHz to 7 kHz.

The method is analyzed using manual annotations of bird vocalizations. For performance analysis, three metrics i.e. true positive rate, missed detection rate and false alarm rate are used. These metrics are calculated using following equations:

$$\text{True Positive rate } (\%) = \frac{\text{Frames correctly classified as calls}}{\text{Total frames contating call activity}} \times 100$$

$$\text{Missed Detection } (\%) = \frac{\text{Frames misclassified as background}}{\text{Total call activity frames}} \times 100$$

$$\text{False Alarms } (\%) = \frac{\text{Background frames classified as calls}}{\text{Total background frames}} \times 100$$

ROC curves are used for analyzing the method. Figure 3 depicts ROC curves comparing performance of methods based on entropy calculated from whitened spectrogram and entropy calculated from whitened group delay phase spectrum. It is clear

from ROC plot in Figure 3 that whitened Group delay phase spectrum method is outperforming whitened power spectrum method.

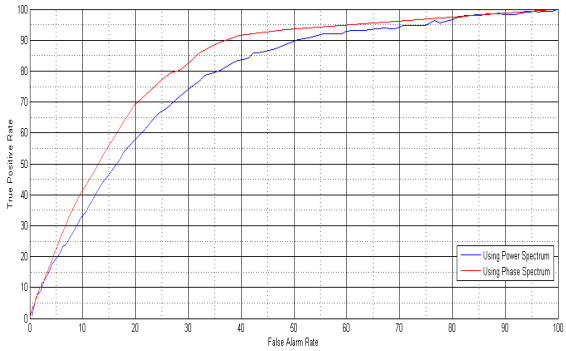

Figure 3: ROC curves comparing white phase spectrum and white power spectrum methods

The effect of whitening the phase spectrum before entropy calculation is evident from the ROC curves depicted in Figure 4 and Figure 5. The analysis of ROC in Figure 4 establishes that whitening the phase spectrum gives better performance than using phase spectrum which is not whitened.

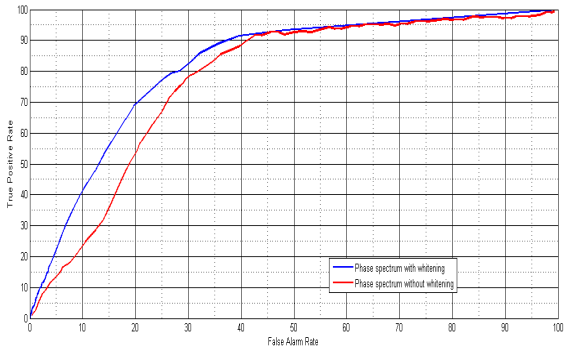

Figure 4: ROC curves comparing performance of methods based whitened phase spectrum and non white phase spectrum

Figure 5 shows comparison of ROC plots between methods based on white and non white power spectrum.

The method is also evaluated on an another dataset i.e. single species MLSP Bird Classification Challaenge 2013. The recordings have low signal to noise ratio. Figure 6 shows ROC curves comparing performance of methods based on entropy calculated from whitened spectrogram and entropy calculated from whitened group delay phase spectrum on single species MLSP data.

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.

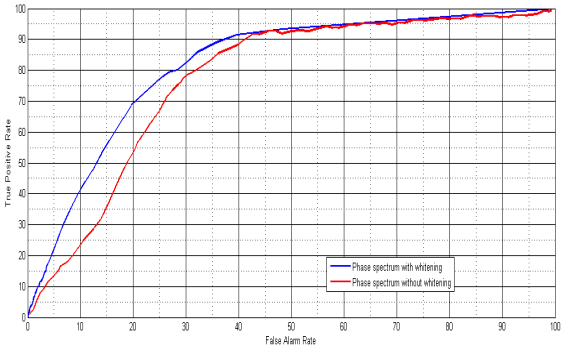

Figure 5: ROC curves comparing performance of methods based whitened power spectrum and non white power spectrum

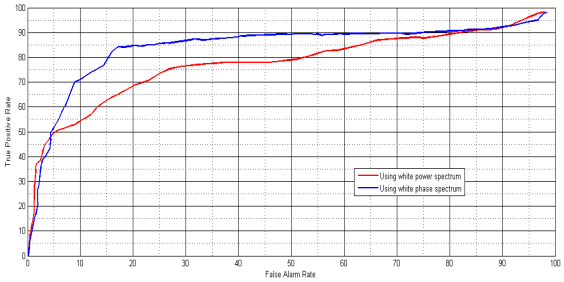

Figure 6: ROC curves comparing performance of methods based on white phase spectrum and white power spectrum on MLSP 2013 single species dataset