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 relaiable 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 [1]. 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 [2] [3] [4]. The same property is beneficial in the processing of bird vocalizations. In [1], 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 [3]. 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 [5] [6] [7]. Time domain segmentation using energy has been used in many studies [8] [9] [10] . The energy based segmentation method is highly influenced by background noise and will where bird calls have low energy in comparison to the background. A KL-divergence based segmentation method is proposed in [11]. 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 [12] , 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 [13]. 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 [4, 14]. 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 [15].

There is strong evidence that birds use their vocal tract as a selective filter to modify the final sound [16]. 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 [17] excited by a single source. When applied to birdcalls, this is a simplification of the ‘two-voice’ theory of avian vocalization [16], 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 [18], 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

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

where the predictor model order is P, G represents the gain and $a(k)$ are the predictor coefficientses [17]. 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 1 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.

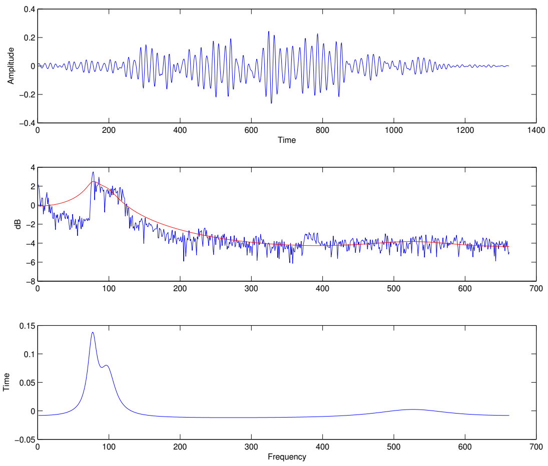

Figure 1: based on all pole model (in the bottom panel).

Recently, feature vectors derived from such a representation were used in speech [19] and speaker recognition [3].

3. Entropy-based segmentation of birdsong

Computing the APGDF for every frame enables good time-frequency resolution of 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 [1], the entropy of this window is estimated using the expression

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

where T is time-frequency window shift and $\tau_N(n,f)$ is normalized APGDF phase spectrum.

$\tau_N(n,f)$ is estimated using the expression [1]

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

where $\tau(n,f)$ is APGDF phase spectrum at time n and frequency f.

The entropy of the time-frequency windows containing bird call is lower than the one containing the background. The entropy calculated from block is less susceptible to the sudden changes in background as compared to the entropy calculated at each time instance [1].

3.1. Whitening APGDF before Entropy Calculation

The drop in entropy at the presence of bird vocalizations makes it possible to detect call periods. This drop can only be detected accurately if background entropy is not changing drastically. However this is not always the case in raw sound recordings. 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 [1]. To overcome this problem, whole APGDF spectrum is whitened before calculating entropy. The entropy calculated from whitened APGDF spectrum is almost constant for background but dips enough to mark the presence of bird vocalizations. Due to this nature of background entropy, even a small change in entropy can be detected easily.

To whiten the APGDF spectrum ($\tau(n,f)$), the covariance matrix of mean subtracted APGDF spectrum is calculated. The eigen values matrix (S) and eigen vectors matrix (U) of this covariance matrix are calculated. The white APGDF spectrum matrix ($\tau_W(n,f)$) is calculated using the following expression:

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

The difference between entropy calculated from normal phase spectrum and whitened phase spectrum is evident in Figure 2.

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 maximum and local minimum to find out 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 propose algorithm is evaluated on two datasets: Cassins Vireo [20] and MLSP Bird Classification Challaenge 2013 dataset [21]. 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 APGDF spectrum, 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 [1].Phrase annotations are available with the dataset [20].

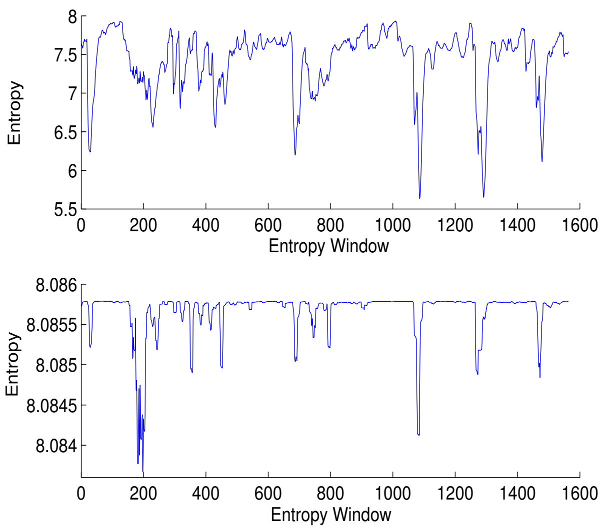

Figure 2: Entropy calculated from phase spectrum and whitened phase spectrum respecively.

These annotations are used as ground truth for evaluation of the proposed approach. For performance analysis, receiver operating characteristic (ROC) curves are used. A ROC curve is a plot between false alarm rate (FAR) and true positive rate (TPR). True positive rate and false alarm rate are calculated using following equations:

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

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

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 APGDF method is outperforming the technique in [1] using whitened spectrogram and thresholding.

The second dataset used is a subset of MLSP bird classification challenge 2013 dataset [21]. The data is recorded in more challenging conditions and recordings have low signal to noise ratio. Twenty Five files of ten seconds length are used to evaluate the proposed method. Figure 4 shows ROC curves comparing performance of methods based on entropy calculated from whitened APDGF spectrum and whitened power spectrum. Here group delay phase spectrum based method is outperforming spectrogram based method for almost all the operating points.

4.1. Model Order vs AIC

To establish the optimum model order, Akaike Information criteria (AIC) [17] is used. Figure 5 depicts the AIC for different model orders for Cassin's Vireo phrases.

Figure 6 depicts the AIC for different model orders for three different bird species i.e. Cuckoo, Great Barbet and Laughing Thresh.

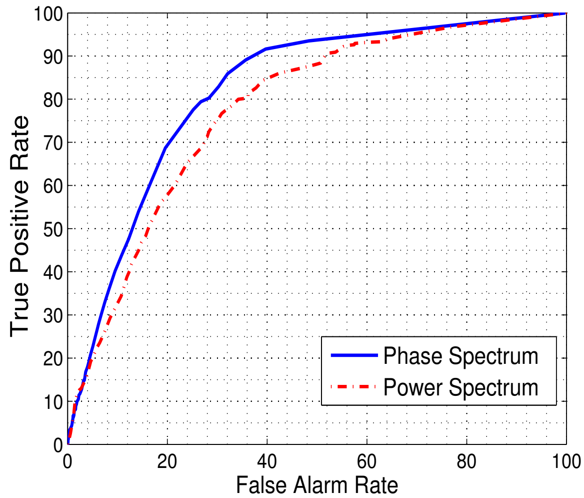

Figure 3: ROC curves comparing performance of methods based on white phase spectrum and white power spectrum on Cassin's Vireo dataset.

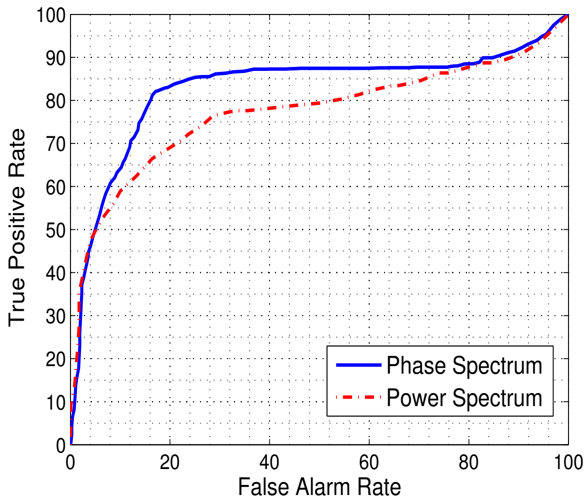

Figure 4: ROC curves comparing performance of methods based on white phase spectrum and white power spectrum on MLSP 2013 single species dataset

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: AIC vs model order for Cassin's Vireo

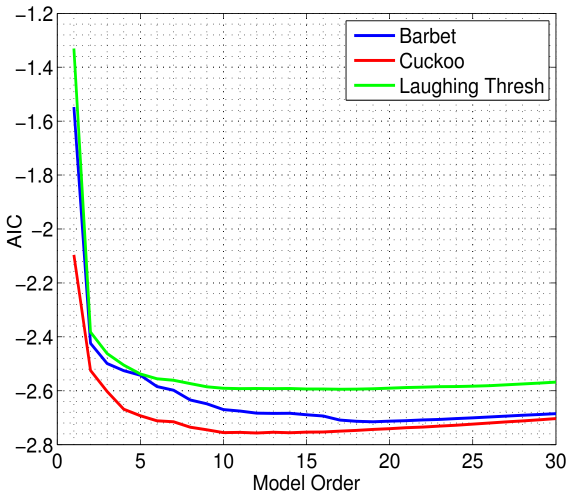

Figure 6: AIC vs model order for three different species

6. References

[1] N. C. Wang, R. E. Hudson, L. N. Tan, C. E. Taylor, A. Alwan, and K. Yao, "Bird phrase segmentation by entropy-driven change point detection," in *Proc. Int. Conf. Acoust. Speech, Signal Process*, 2013, pp. 773–777.

[2] H. A. Murthy and V. R. R. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 2003, pp. 68–71.

[3] P. Rajan, T. Kinnunen, C. Hanili, J. Pohjalainen, and P. Alku, "Using group delay functions from all-pole models for speaker recognition," in *Proc. Interspeech*, 2013, pp. 2489–2493.

[4] R. M. Hegde, H. A. Murthy, and V. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 190–202, Jan. 2007.

[5] V. M. Trifa, A. N. G. Kirschel, C. E. Taylor, and E. E. Vallejo, "Automated species recognition of antbirds in a mexican rainforest using hidden markov models," *Jnl. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2424–2431, Apr 2008.

[6] C. H. Lee, C. C. Han, and C. C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *IEEE Trans. Audio, Speech, Language Process*, vol. 16, no. 8, pp. 1541–1550, Nov 2008.

[7] K. Kaewtip, L. N. Tan, A. Alwan, and C. E. Taylor, "A robust automatic bird phrase classifier using dynamic time-warping with prominent region identification," in *Proc. Int. Conf. Acoust. Speech, Signal Process*, 2013, pp. 768–772.

[8] A. Harma and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *Proc. Int. Conf. Acoust. Speech, Signal Process*, vol. 5, 2004, pp. 701–704.

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

[10] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 64–64, Jan. 2007.

[11] B. Lakshminarayanan, R. Raich, and X. Fern, "A syllable-level probabilistic framework for bird species identification," in *Proc. Int. Conf. Mach. Learn. Applicat.*, 2009, pp. 53–59.

[12] L. Neal, F. Briggs, R. Raich, and X. Z. Fern, "Time-frequency segmentation of bird song in noisy acoustic environments," in *Proc. Int. Conf. Acoust. Speech, Signal Process*, 2011, pp. 2012–2015.

[13] H. Banno, J. Lu, S. Nakamura, K. Shikano, and H. Kawahara, "Efficient representation of short-time phase based on group delay," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 2, 1998, pp. 861–864.

[14] D. Zhu and K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 125–128.

[15] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *Jnl. Acoust. Soc. Amer.*, vol. 63, no. 5, pp. 1638–1640, 1978.

[16] C. K. Catchpole and P. J. Slater, *Bird song: biological themes and variations*. Cambridge university press, 2003.

[17] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[18] S. Agnihotri, P. Sundeep, C. S. Seelamantula, and R. Balakrishnan, "Quantifying vocal mimicry in the greater racket-tailed drongo: a comparison of automated methods and human assessment," *PloS one*, vol. 9, no. 3, p. e89540, 2014.

[19] E. Loweimi, S. M. Ahadi, and T. Drugman, "A new phase-based feature representation for robust speech recognition," in *Proc. Int. Conf. Acoust. Speech, Signal Process*, 2013.

[20] "Cassin's vireo recordings," http://taylor0.biology.ucla.edu/al/bioacoustics/, accessed: 2016-03-20.

[21] "Mlsp bird classification challenge 2013," https://www.kaggle.com/c/mlsp-2013-birds/data, accessed: 2016-03-20.