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 (for eg. the SongMeter series from Wildlife Acoustics Inc. [1]), 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 (also called segmentation.) An entropy-based bird phrase segmentation technique was developed in [2]. 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 [3] [4] [5]. The same property is beneficial in the processing of bird vocalizations. In [2], 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 performed to separate the birdcalls from the background.

Many studies on birdcalls have used manual segmentation [6] [7] [8]. Time domain segmentation using energy has been used in many studies [9] [10] [11] . 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 [12]. KL-divergence between normalized power spectral density of a frame and uniform distribution is computed. The more KL-divergence corresponds to less entropy and vice-versa. Local minima of KL-divergence act as change points for bird vocalizations. In [13], time-frequency based segmentation using a random forest classifier is proposed to segment the bird vocalizations from a noisy audio recording.

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. The group delay function $\tau(\omega)$ can be derived as [14]

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \qquad (1)$$

where $x(n) \leftrightarrow X(\omega)$ and $y(n) \leftrightarrow Y(\omega)$ are Fourier transform pairs and $y(n) = nx(n)$. However, $\tau(\omega)$, as computed above 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. At spectral nulls, the value of the denominator in equation 1 tends to zero, leading to spurious peaks which mask the formant structure [3], [15], [16]. Several methods have been proposed in the literature to overcome the effects due to these artifacts [5], [17], [16]. Another technique to overcome this difficulty is to model the vocal tract as an all-pole filter, hence

avoiding the zeros altogether. Such a technique derived using linear prediction (LP) analysis was used in the detection of formants in human speech [18].

The vocal tract is represented in the LP model as

$$H(\omega) = \frac{G}{1 - \sum_{k=1}^{P} a(k) e^{-j\omega k}}, \tag{2}$$

where the predictor model order is P, G represents the gain and $a(k)$ are the predictor coefficients [20]. 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 all-pole group delay function (APGDF) is defined as

$$\tau_A(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|H(\omega)|^2}, \tag{3}$$

which is obtained by replacing the power spectrum in the denominator of equation 1 with the all-pole power spectrum of equation 2. This avoids spectral nulls in the denominator of equation 3, thus retaining the formant structure in the group delay function.

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.

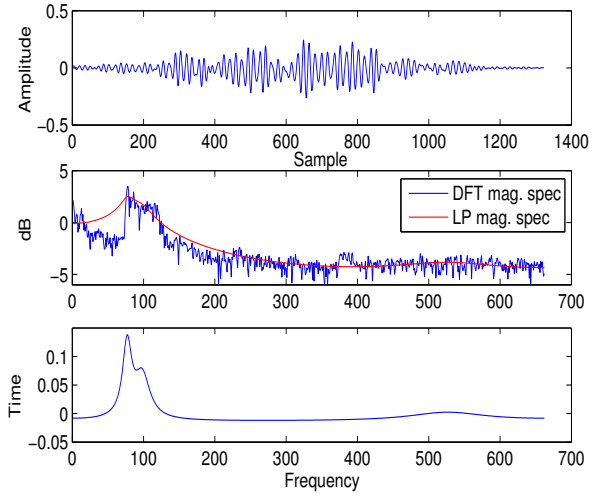

Figure 1: A frame of audio signal (top panel), corresponding LP magnitude spectrum superimposed on DFT magnitude spectrum (middle panel) and all-pole group delay function (bottom panel)

Recently, feature vectors derived from such a representation were used in speech [22] and speaker recognition [4].

There is strong evidence that birds use their vocal tract as a selective filter to modify the final sound [19]. Given this, the source-filter model developed for analysing human speech can be applied to bird vocalizations as well. LP analysis of human speech signals models the vocal tract spectrum as an all-pole filter [20] excited by a single source. When applied to birdcalls, this is a simplification of the 'two-voice' theory of avian vocalization [19], 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 [21], where LP analysis has been applied in analysing the song of the greater racket-tailed drongo.

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 $f_{\min}$ to $f_{\max}$ is considered over the APGDF vectors. As in [2], the entropy of this window is estimated using the expression

$$h_k = -\sum_{n=kT+1}^{kT+w} \sum_{f=f_{\min}}^{f_{\max}} \tau_N(n,f) \ln \tau_N(n,f), \tag{4}$$

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

$$\tau_N(n,f) = \frac{\tau_A(n,f)}{\sum_{n=kT+1}^{kT+w} \sum_{f=fmin}^{fmax} \tau_A(n,f)}, \tag{5}$$

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

The entropy calculated from the time-frequency window is less susceptible to sudden changes in the background as compared to the entropy calculated at each time instance [2]. The entropy of the time-frequency window containing birdcalls is lower than the one only containing the background. As the window moves to the call region from the background, there is a gradual drop in entropy.

3.1. Whitening the 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 the entropy during a call period. To mitigate this problem to some extent, as in [2], the APGDF is whitened before calculating the entropy. The entropy calculated from the whitened APGDF is almost constant for the background but dips enough to mark the presence of bird vocalizations. Due to this, small changes in entropy can be detected fairly reliably.

To whiten the APGDF (τ_A), 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 (τ_W) is calculated as

$$\tau_W = \text{diag}\left(\frac{1}{\sqrt{\text{diag}(S) + \epsilon}}\right) \times U^T \times \tau_A \tag{6}$$

Here ϵ is a negligible positive number added to prevent division by zero. The difference between entropy calculated from the original APGDF and the whitened APGDF is evident in figure 2.

3.2. Detecting change points using thresholding

To detect change points, extrema-based thresholding is used. Local minima and maxima are estimated on the entropy, and a

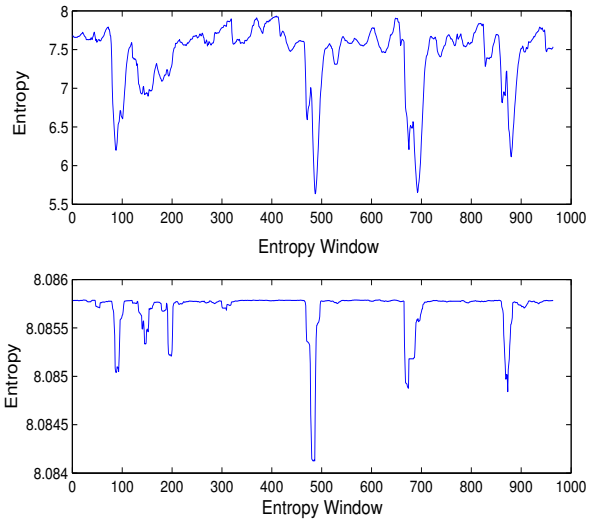

Figure 2: Entropy calculated from APGDF (Uppen Panel) and entropy calculated using white APGDF (Lower Panel).

threshold is applied on the difference between consecutive minima and maxima. 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 the recording.

4. Performance analysis

The proposed algorithm is evaluated on two datasets. The first consists of recordings of a single species, the Cassin's vireo; the second consists of recordings from several species, and is a subset of the MLSP 2013 bird classification challenge. In the Cassin's vireo dataset [23], the total duration of recordings is about 45 minutes, out of which about 5 minutes correspond to the calls. These recordings are collected over two months and are fairly clean. Phrase annotations are provided with the dataset, the start and end of which are used as the ground truth for segmentation.

The MLSP 2013 dataset [24] is much nosier and includes wind and rain in the background. The recordings are done for over two years at 13 different locations. Twenty five files of ten seconds length each are selected from this data to evaluate the proposed method. Manual annotations are done to mark the presence of bird vocalizations in these audio files. The audio files contain about 60 bird vocalizations, which occupy 13.35% of the total time of recordings.

The proposed algorithm is compared with a modified version of technique proposed in [2]. Entropy estimated from the whitened power spectrum is used to determine the change points. Similar to the proposed algorithm, extrema-based thresholding is used to determine the change points. Thus, in the experiments, entropy estimated from the power spectrum is compared to the entropy estimated from the APGDF.

A frame length of 20 ms and increment of 5 ms is used to perform short-time processing. APGDF computed on each frame is whitened before computing the entropy as given in equation 6. A time-frequency window of length w=138.8 ms is used along with a shift of T = 15 ms to estimate the entropy (see equation 4). The frequency range of the time-frequency window is set from $f_{\min}$=1.5 kHz to $f_{\max}$=7 kHz [2]. Although these parameters are possibly species specific, we use the same parameters for both datasets.

Receiver operating characteristics (ROC) curves are used to analyze the performance of both techniques. True positive rate (TPR) and false alarm rate (FAR) are calculated as:

$$\text{TPR}(\%) = \frac{\text{Correctly Classified Call Frames}}{\text{Total Call Frames}} \times 100 \quad (7)$$

$$\text{FAR}(\%) = \frac{\text{Wrongly Classified Call Frames}}{\text{Total Background Frames}} \times 100 \quad (8)$$

Figures 3 and 4 depict ROC curves comparing performance of the two methods on the respective datasets. It is clear from figures 3 and 4 that the APGDF-based technique outperforms the power spectrum based method at most of the operating points.

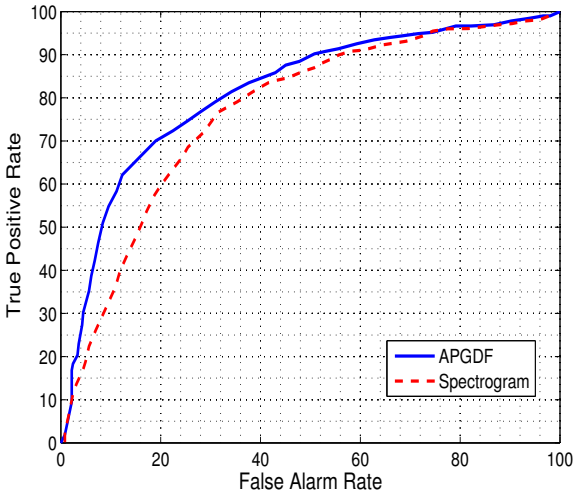

Figure 3: ROC curves comparing performance of methods based on APGDF and spectrogram on Cassin's Vireo dataset.

It is to be noted that the method in [2] utilises a sophisticated Bayesian change-point detection technique. When compared to the simple thresholding technique used in our evaluation, their method acheives a better true positive rate at 20% false alarm rate (see figure 2 in [2]). Replacing thresholding with a more relaiable change-point detection technique is expected to bring improvements while using the APGDF-based technique.

4.1. Determining the model order for LP analysis

To correctly model the all-pole filter while estimating the APGDF, one requires to know the optimum number of poles to be used. In [20], the Akaike information criterion (AIC) is used to determine the optimal model order for the LP filter. Assuming a Gaussian probability distribution for the signal, the criterion approximated as

$$I(p) = \log V_p + \quad (9)$$

where V_p is XXX fill this up, and mention what is N_e.

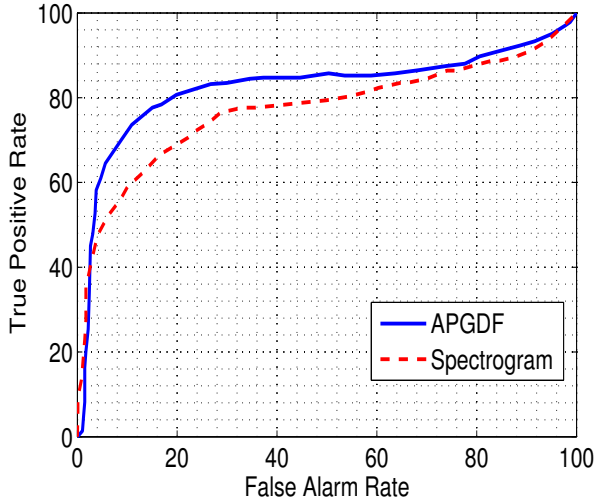

Figure 4: ROC curves comparing performance of methods based on APGDF and spectrogram on MLSP dataset.

For the Cassin's vireo, the AIC is computed for various model orders. The AIC curve reaches a minimum around p=10 and then increases with a gentle slope. In practice, a model order between 5 and 10 is well suited for this species. To investigate if this is species specific, a similar study is performed for four species from four different families. These are: Indian nightjar (*Caprimulgus asiaticus asiaticus* from the family Caprimulgidae, emerald dove (*Chalcophaps indica*) from the family Columbidae, abd XXX XXX. The model order versus AIC plots given in figure XXX also indicate that the same value of p (between 5 and 10) is suitable accross multiple species.

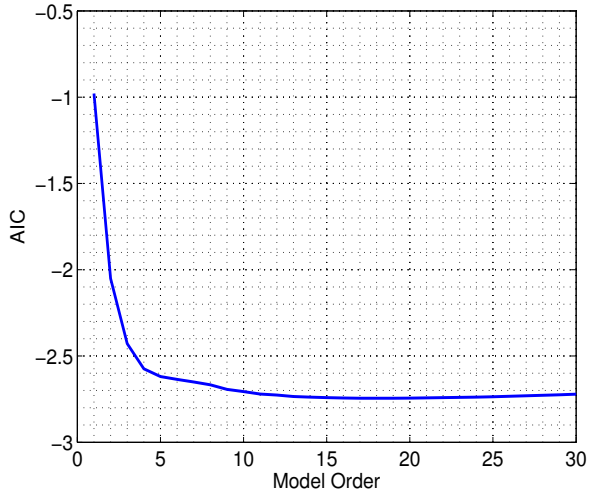

Figure 5: AIC vs model order for Cassin's Vireo

Figure 6 depicts the AIC for different model orders for four different bird species i.e. Indian Nightjar, Emerald Dove, Canary Flycatcher and Sarus Crane .

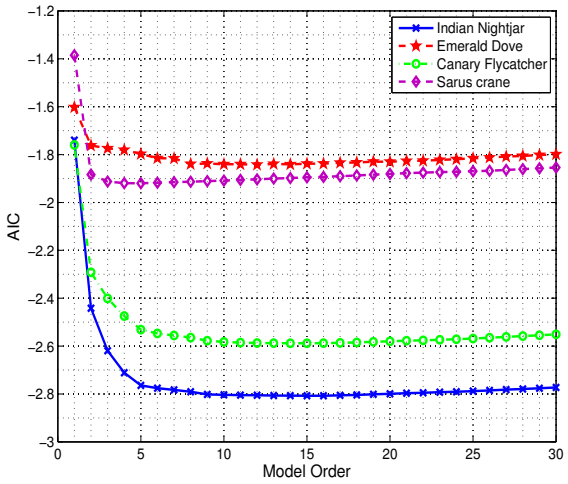

Figure 6: AIC vs model order for four different species

5. Conclusion

In this paper, we explored an entropy-based birdcall segmentation technique. The method utilises information from the phase spectrum of the short term Fourier transform. To avoid the difficulty of unwrapping the phase, the group delay function is utilised. An all-pole filter using LP analysis is used to model the vocal tract spectrum while estimating the group delay function, resulting in the APGDF represntation. The APGDF is whitened before estimating the entropy, and extrema-based thresholding is performed to distinguish call periods from the background. When compared to the entropy estimated from the power spectrum, the entropy from the APGDF resulted in better segmentation accuracy.

Studies on model order for the LP filter using the Akaike information criteria showed that a value of p between 5 and 10 is sufficient for species from different families. Future work will include applying the LP model to perform tasks such as phrase detection and species identification.

6. References

[1] Song meter sm4. [Online]. Available: http://wildlifeacoustics.com/products/song-meter-sm4

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] H. A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," *Signal Processing*, vol. 22, no. 3, pp. 259–267, 1991.

[16] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Commun.*, vol. 49, pp. 159–176, 2007.

[17] 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.

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

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

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

[21] 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.

[22] 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.

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

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