2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

RÉNYI ENTROPY BASED MUTUAL INFORMATION FOR SEMI-SUPERVISED BIRD VOCALIZATION SEGMENTATION

Anshul Thakur, Vinayak Abrol, Pulkit Sharma and Padmanabhan Rajan

School of Computing and Electrical Engineering
Indian Institute of Technology, Mandi
Email: {anshul_thakur, vinayak_abrol, pulkit_s}@students.iitmandi.ac.in, padman@iitmandi.ac.in

ABSTRACT

In this paper we describe a semi-supervised algorithm to segment bird vocalizations using matrix factorization and Rényi entropy based mutual information. Singular value decomposition (SVD) is applied on pooled time-frequency representations of bird vocalizations to learn basis vectors. By utilizing only a few of the bases, a compact feature representation is obtained for input test data. Rényi entropy based mutual information is calculated between feature representations of consecutive frames. After some simple post-processing, a threshold is used to reliably distinguish bird vocalizations from other sounds. The algorithm is evaluated on the field recordings of different bird species and different SNR conditions. The results highlight the effectiveness of the proposed method in all SNR conditions, improvements over other methods, and its generality.

Index Terms— Bird call segmentation, feature learning using PCA, Rényi entropy

1. INTRODUCTION

Birds play important roles in maintaining the balance of ecosystems. They are present at various steps of the food chain, help in pollination and in seed dispersal. Many bird species are under the threat of population decline due to habitat destruction. Surveying and monitoring are essential for their conservation. Acoustic monitoring provides a convenient and passive way to monitor bird populations in their natural habitats. With the advent of automated recording devices (for eg. the SongMeter series from Wildlife Acoustics Inc.), acoustic monitoring has become easier. These sophisticated devices can collect large amounts of bioacoustic data. By analyzing audio recordings containing bird vocalizations collected in this manner, it is possible to perform tasks such as species identification, tracking of migrant species or examining the avian biodiversity of a given region. Typically, the collected data is processed off-line. In this process, the first step is usually to determine regions of interest in the recording, also termed as segmenting the vocalizations/calls from the background. The segmentation task becomes challenging due to the presence of various background sounds such as rain, insects, other animals, passing traffic and other sounds. In field conditions, background sounds are unpredictable which makes acoustic modeling of the background difficult. There is a need for unsupervised or semi-supervised segmentation techniques in which modeling the background is not required.

In this work, we propose a semi-supervised bird vocalization segmentation algorithm which can work in different recording environments and is not influenced by background disturbances to a large extent. A dictionary of basis vectors modeling bird vocalizations is learnt from a small amount of labeled training data. The proposed method uses singular value decomposition (SVD) to learn these basis vectors from the time-frequency representation (spectrogram) of bird vocalizations. By projecting the time-frequency representation of a test audio recording onto these bases, a compact representation is obtained for each short-time frame. By estimating the Rényi entropy based mutual information between each pair of consecutive frames, and with some simple post-processing, bird vocalizations can be easily distinguished from other sounds. This is an advantage over some other methods, which are prone to false alarms by being unable to distinguish naturally occurring background sounds like the sounds of insects. Our experimental evaluation on the vocalization of passerine birds demonstrates the generic nature of the method.

The rest of this paper is organized as follows. In section 2, we discuss some of the methods proposed in the literature targeting the bird vocalization segmentation along with their drawbacks. We also describe how the proposed algorithm overcome these drawbacks. In section 3, the proposed algorithm is described in detail. Performance analysis and conclusion are in sections 4 and 5 respectively.

2. COMPARISONS TO PRIOR WORK

Most methods for segmenting bird vocalizations have either utilized time-domain representations, or have used spectrogram-based representations. We briefly categorize some of these below.

Manual segmentation: A few bioacoustic studies dealing with bird vocalizations such as species identification [1] [2] have used manually segmented bird vocalizations. Manual segmentation can be tedious and unfeasible if the amount of data to be processed is large.

Energy or entropy-based methods: Energy calculated in the time domain has been used to segment the bird vocalizations in [3] [4] [5]. Spectral entropy and KL-divergence based segmentation methods are proposed in [6] and [7] respectively. The regions containing bird vocalizations exhibit low entropy in comparison to the background regions. In [7], KL-divergence between normalized power spectral density of an audio frame and the uniform distribution is computed. This KL-divergence is a measure of entropy; higher KL-divergence corresponds to less entropy and vice-versa. These methods are unsupervised in nature which is desirable. However, these methods are not able to distinguish bird vocalizations from any other sound event. Also, energy-based segmentation is affected by the presence of background noise.

Template-based method: A noise-robust template matching based method is proposed in [8]. This methods uses dynamic time warping (DTW) and high-energy regions of spectrograms to build noise-robust templates for each type of vocalization. Each vocalization template is built using various examples of that vocalization. This method is effective for most of the background conditions. However, the disadvantage is that we must know beforehand what vocalizations we wish to segment. Hence, this method may not be scalable in real-world scenarios.

Other methods: In [9], time-frequency based segmentation using a random forest classifier is proposed to segment bird vocalizations in noisy conditions. This method requires a large amount of training examples. A spherical K-means based feature learning method [10] is proposed to model the bird vocalizations of different species. In [11], an unsupervised two-pass segmentation method is proposed. In the first pass, training labels are generated using inverse spectral flatness (ISF) from the input recording itself. ISF is used to distinguish vocalizations and background regions from the input recordings. These vocalizations and background regions are used to build Gaussian mixture models which are used in the second pass to classify each input frame as the background or the bird vocalization. However, like energy and entropy, the ISF used in first pass is also unable to distinguish bird sounds from non-bird sounds.

2.1. Advantages of the proposed algorithm

Ability to discriminate other background sounds: Using only a few of the basis vectors helps in retaining the information corresponding only to the bird vocalizations and not to the other background sounds. Hence, unlike some of the earlier methods, the proposed method can discriminate between bird vocalizations and non-bird sounds (see section 3).

Better precision: Mutual information criteria based on Rényi entropy is calculated for each pair of consecutive frames i.e. between the current frame under process and the previous frame, providing more precision as compared to the time-frequency window based entropy calculated in [6]. Entropy calculated from a time-frequency window several frames long will exhibit the presence of bird vocalization even if the vocalization is present in the first or last few frames of the window, leading to a decrease in segmentation precision.

Generalization: Since the learnt dictionary is a generative model, a bird vocalization not used in learning the basis vectors can still be approximated by the learned basis vectors. This makes the proposed algorithm more generic as compared to the template based technique described in [8]. This behavior is analyzed in detail in section 4.

3. PROPOSED METHOD

3.1. Dictionary learning

To learn the basis vectors, the time-frequency representations of bird vocalizations are extracted using a small amount of labeled training data. The training labels provide information about the start and end time of the vocalizations. These extracted vocalizations are pooled together to form a matrix $\mathbf{M}$ of dimensions $D \times N$. Here N is the number of pooled frames and D represents the number of FFT bins used in the spectrogram. This matrix, $\mathbf{M}$, is factorized using singular value decomposition (SVD): $\mathbf{M} = \mathbf{U} \times \mathbf{\Sigma} \times \mathbf{V}^*$. Here $\mathbf{U}$ is a $D \times D$ unitary matrix whose columns contains the left singular vectors, $\mathbf{\Sigma}$ is a $D \times N$ diagonal matrix containing singular values and $\mathbf{V}$ is $N \times N$ unitary matrix whose columns contains the right singular vectors. The columns of U are used as the basis vectors of the subspace on which the time-frequency representation of the audio recording is projected to get the features in the testing stage.

Typically vocalizations of many song birds occupy only few frequency bins at any given time. Hence the information regarding bird vocalization regions is mostly consolidated in the first few columns of $\mathbf{U}$ which correspond to the directions of highest singular values and hence highest variances. Hence to retain only the bird vocalization information in the feature domain, the input test audio recording is projected on the first K columns of $\mathbf{U}$ as: $\mathbf{F} = \mathbf{B}^T \times \mathbf{P}$. Here $\mathbf{B}$ is matrix of $D \times K$ dimensions whose columns are the first K columns of $\mathbf{U}$. $\mathbf{P}$ is the time-frequency representation of the test audio signal having $D \times M$ dimensions, M is the number of frames and D is the number of frequency bins. $\mathbf{F}$ is the matrix of dimension $K \times M$ whose columns contain the feature representations of each input frame. The value of K is determined experimentally.

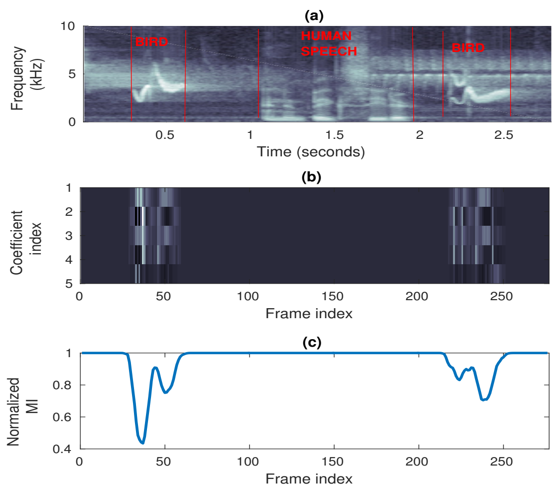

Fig. 1: (a) Spectrogram of an audio recording containing human speech, background noise and vocalization of Cassin's vireo (b) Feature representation of the above spectrogram (c) Normalized MI extracted from the feature representations depicted in (b).

Figure 1(b) depicts the feature representation learned from the time-frequency representation of an audio recording shown in Figure 1(a). This feature representation is obtained by projecting the time-frequency representation on the first 5 columns of $\mathbf{U}$. This $\mathbf{U}$ is learned by factorizing the pooled time-frequency representations of the vocalizations of Cassin's vireo, a North American song bird, using SVD. The audio corresponding to the spectrogram shown in figure 1(a) contains human speech and two Cassin's vireo vocalizations. By analyzing figure 1(b), it is clear that information corresponding to the human speech and other background disturbances is not reflected in the feature domain. Each of the coefficients calculated for any non-bird frame has magnitude close to zero. Hence, the variance of coefficients of any non-bird frame is low. On the other hand, each coefficient calculated for any bird frame has larger magnitudes in comparison to the non-bird frame. The variance of coefficients within a bird sound frame is also high. This is due to the fact that none of the learned basis vectors have any contribution in defining the background frames. However, a bird vocalization frame can be represented as a combination of the scaled versions of the learned basis vectors [12]. The contribution of some basis vectors in defining the input bird frame is higher than the others. This leads to the difference in magnitudes of the coefficients calculated for bird frames.

This behavior is highlighted in figure 2. The box plots of coefficients of 100 human speech frames, 100 background frames and 100 bird vocalization frames are shown in Figure 2. From this figure it is evident that the magnitude of coefficients for human speech and background frames is almost constant. However, a significant amount of variation is present for the bird vocalizations.

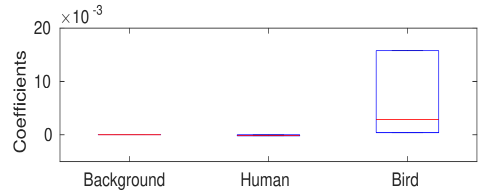

Fig. 2: Box plots of the coefficients calculated for (a) 100 background frames (b) 100 human speech frames and (c) 100 bird sound frames.

3.2. Rényi entropy based mutual information

The feature representation of nth frame, $\mathbf{x}_n$, is converted into a normalized vector using the softmax function: $(\mathbf{x}_n)_j = \frac{e^{\mathbf{x}_{n_j}}}{\sum_{k=1}^{K} e^{\mathbf{x}_{n_k}}}$, for $j = 1, 2, .., K$.

Since the feature coefficients of each non-bird frame approach zero, each coefficient of the frame becomes almost equal after applying the softmax function. However, for any bird vocalization frame, some coefficients exhibit higher values than others. Hence, the normalized feature representations for all the non-bird frames are nearly similar and more variation occurs for the bird vocalization frames. Considering these feature vectors as sampled random vectors in $\mathbb{R}^K$, this can behavior can be discriminated by using mutual information.

Mutual information (MI) between normalized feature representations of each pair of the consecutive frames (i.e. between n_{th} and $(n-1)_{th}$ frames) is calculated. This serves the purpose of considering the previous frame along with the current frame in making the segmentation decisions. MI of a random vector with itself is highest, therefore MI between two almost similar feature representations will be higher than between two representations which are different. Hence for non-bird regions, MI will be high as compared to the regions having vocalizations as depicted in Figure 1(c). Also, since the feature representations for the frames of non-bird regions are almost similar, MI across all these regions is almost constant.

MI between feature representations of two consecutive frames i.e. $\mathbf{x}_n$ and $\mathbf{x}_{n-1}$, each of dimensions $K \times 1$, can be calculated as

$$MI(\mathbf{x}_n, \mathbf{x}_{n-1}) = H(\mathbf{x}_n) + H(\mathbf{x}_{n-1}) - H(\mathbf{x}_n, \mathbf{x}_{n-1}) \quad (1)$$

Here $H()$ represents the entropy. Rényi entropy is used in this work to calculate the MI. Rényi entropy of the pth order for feature representation of nth frame,$\mathbf{x}_n$ can be calculated as [13]:

$$H(\mathbf{x}_n) = \frac{p}{1-p} \log(\|\mathbf{x}_n\|_p). \quad (2)$$

where p controls the sensitivity towards the shape of probability distribution of the coefficients of $\mathbf{x}_n$ [14]. The value of p ($0 < p < 1$) is determined experimentally.

3.3. Segmentation using thresholding

The nature of MI calculated from the feature representations makes the task of thresholding simple. Since the MI for background regions is almost constant, any drop in the value of MI signifies the presence of bird vocalization. The calculated MI is smoothed using a moving average filter and normalized to be between 0 and 1. This results in the MI to be close to 1 for background regions as can be seen in Figure 1(c). Thus, a threshold t, just below one, is able to reliably discriminate call regions from other regions.

4. PERFORMANCE ANALYSIS

4.1. Datasets used

Experimental validation of the proposed algorithm is performed on three datasets. Two datasets consists of the recordings of Cassins vireo, a North American songbird. The third dataset has the recordings of another song bird, California thrasher. The first Cassin's vireo dataset (CV1) contains twelve audio recordings and are available at [15]. These audio recordings were collected over two months in 2010 and contain almost 800 bird vocalizations or song phrases of 65 different types. The second Cassin's vireo dataset (CV2) and California thrasher recordings (CT) are available at [16]. Out of the available 459 recordings of Cassin's vireo, we have used only 100 recordings here. The recordings having longest durations and maximum number vocalizations are chosen. These recordings contain almost 25000 Cassin's vireo vocalizations of 123 different types. Similarly out of the available 698 California thrasher recordings, we have chosen 100 recordings having maximum durations and number of vocalizations. These 100 recordings contain about 15000 bird vocalizations. All the recordings from these three sources are field recordings and contain various types of background noise including human speech. These recordings are 16-bit mono WAV files having a sampling rate of 44.1 kHz.

To test the proposed algorithm in extreme conditions, background sounds are artificially added to the recordings of the CV1 dataset. Three different types of background sounds i.e. rain, waterfall, river and cicadas at 0 dB, 5 dB, 10 dB, 15 dB and 20 dB SNR are added using Filtering and Noise Adding Tool (FaNt) [17]. The sound files are downloaded from FreeSound [18].

4.2. Experiments

Two different experiments are performed to evaluate various aspects of the proposed algorithm. In the first experiment, we compare the performance of the proposed algorithm with existing unsupervised segmentation methods such as short-term energy (STE), spectral entropy (SE) [6], inverse spectral flatness (ISF) [19] and two-pass unsupervised method (US) [11]. Apart from these methods, the performance is also compared with the supervised template-based method (TM) in [8], and two variants of the proposed algorithm. The first variant uses non-negative matrix factorization (NMF) for learning the basis vectors instead of SVD. The second variant uses the normalized energy of the feature coefficients (CE) instead of Rényi entropy based mutual information. The second experiment is to demonstrate the general nature of the proposed method by testing on unseen vocalizations.

F_1 score, defined as the geometric mean of precision and recall, is used as a metric for evaluation, by comparing with the manually labeled ground truth. Both experiments use 10-fold cross-validation. During each fold, one audio recording was used for learning bases and the rest were used as test examples. The average results of these 10 folds are presented in Figure 3 and Table 1 for the first and second experiments respectively.

A frame length of 20 ms with a 10 ms overlap, Hamming window and 512 FFT points are used to compute the time-frequency representations of the input audio. For calculating the feature coefficients, we project the time-frequency representation of the test audio file on the top $K = 5$ left singular vectors. For calculating Rényi entropy, an order of $p = 0.7$ is used, and a moving average filter of length 10 is used to smooth the MI. A threshold $t = 0.9999$ is applied on MI to segment the bird vocalizations. These values of K, p and t are chosen experimentally using a validation set. Two recordings from CV1 having the shortest durations are chosen to form the validation set. These audio recordings are not used for either learning basis vectors or testing in any of the experiments. After validation, the same values of K, p and t are used for all the experiments (including the noisy cases.)

The parameter setting used in [11] are used for implementing STE, SE, ISF and US. Similarly the parameter values discussed in [8] are used for implementing TM. The parameter setting used in the proposed algorithm is also used for implementing NMF and CE. However, in the NMF variant, approach, 256 basis vectors are learned from the training data. This number is chosen experimentally.

4.2.1. Comparison of the proposed algorithm with other methods

The performance of the proposed algorithm is compared with other methods on dataset CV1 and on the artificially created noisy versions of CV1. In the proposed algorithm, the NMF variant and the CE variant, the basis vectors are learned from labeled bird vocalizations extracted from a single audio recording of CV1 and the rest of the recordings are used for testing during each of the 10 folds. For segmenting noisy versions of CV1, the basis learned from the clean audio recordings of CV1 are used. The segmentation performance of the proposed algorithm along with other methods is summarized in Figure 3.

By analyzing Figure 3, it can be concluded that the performance of the proposed algorithm is better than all the other methods except TM in both noisy and clean conditions. However, TM is a template based method which may not be scalable and requires prior knowledge of all the vocalizations we wish to segment, in the form of templates. Also, the performance of the proposed method is not affected vigorously as compared to performances of STE, ENT, ISF and US in low SNR conditions. The NMF and CE variants also outperforms these methods. The use of the top K left singular vectors in the proposed algorithm instead of the NMF based dictionary provided better segmentation in all conditions. The CE variant method gave good segmentation performance. This shows that a simple energy-based segmentation is good enough to segment the bird vocalizations learnt from basis vectors. But since Rényi entropy based MI uses context information in terms of the previous and current frames, it provides slightly better segmentation.

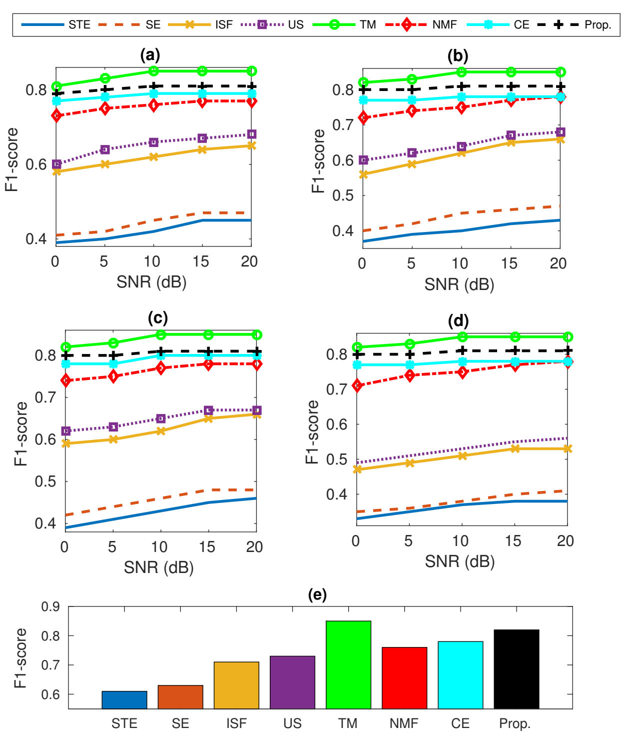

Fig. 3: **Results of experiment 1:** Comparison of segmentation performances of different segmentation methods on noisy variations of CV1 generated by adding noise types (a) rain, (b) river, (c) waterfall (d) cicada and (e) on CV1.

4.2.2. Generic nature of the proposed algorithm

The second experiment has two parts, and establishes the generic nature of the proposed algorithm. In the first part, we learn basis vectors from CV1 and segment the audio recordings of CV2. In the second part, we use the basis vectors learnt from CV1 to segment the recordings of CT i.e. we learn the basis vectors from the recordings of one species and segment the audio recordings of another. Again, 10 fold cross-validation was used. During each fold, one recording from CV1 was used to learn basis vectors and testing was performed on all the recordings of CV2 and CT. These are tabulated in Table 1. The analysis of Table 1 shows that the proposed algorithm is able to segment CV2 recordings having 123 different types of Cassin's vireo vocalizations using the basis vectors learned from a single audio recording of CV1 which has 10 to 25 different types of Cassin's vireo vocalizations (the number of vocalizations in the training recording depends on the fold). Hence, the proposed method is able to segment vocalizations which are not used in learning the basis vectors.

Also, the proposed algorithm is able to segment recordings of California thresher using basis vectors learned from Cassin's vireo. The segmentation performance obtained in this cross-species experiment is also compared with the segmentation performance obtained by using the basis vectors learned from the vocalizations of California thrasher. No significant difference is observed in the performances which further supports the generic nature of the proposed algorithm. Table 1 also depicts the performance of other methods with the proposed method. By analyzing Table 1, it is clear that the proposed method outperforms the other methods except TM, which requires templates of the vocalizations.

Table 1: **Results of experiment 2:** Performance of the proposed method for various train-test conditions. (-) indicates that the method is unsupervised. The TM method uses templates of the vocalizations.

Method	Training Dataset	Testing on CV2	Testing on CT
STE	-	0.53	0.55
SE	-	0.56	0.57
ISF	-	0.6	0.61
US	-	0.64	0.63
TM	CV2, CT	0.79	0.78
NMF	CV1	0.72	0.71
NMF	CT	0.7	0.73
CE	CV1	0.74	0.71
CE	CT	0.71	0.73
Prop.	CV1	0.76	0.73
Prop.	CT	0.74	0.75

4.2.3. Discussion

The proposed method provides reliable segmentation performance only if the vocalizations represented by the bases are similar to the ones in the evaluation data. If the bases are learned from bird vocalizations which exhibit rapid frequency and temporal modulations but the target vocalizations are wideband in nature, the proposed algorithm will fail. For example, the basis vectors learned from CV1 are not able to segment the sounds of greater sooty owls and forest ravens which are wide-band in nature. On the other hand, sounds from other passerine birds like the Verditer flycatcher and blue magpies (found in the Indian subcontinent) are effectively segmented by the bases learnt from Cassin's vireo. Five recordings of the vocalizations of these species (downloaded from [20]) resulted in an F1-score of 0.32, 0.38, 0.85 and 0.79 respectively.

5. CONCLUSION

This paper presented an algorithm for segmenting birdcalls from the background using matrix factorization and Réyni entropy based MI. Experimental evaluation, including comparisons with six existing methods demonstrated the effectiveness and the generality of the proposed method. The results also indicate that the method can be utilized for segmenting the vocalizations of similar birds in other geographic regions, irrespective of data used in learning the basis vectors.

6. REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

[8] K. Kaewtip, L. N. Tan, C. E. Taylor, and A. Alwan, "Bird-phrase segmentation and verification: A noise-robust template-based approach," in *Proc. Int. Conf. Acoust. Speech, Signal Process*, 2015, pp. 758–762.

[9] 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.

[10] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, pp. e488, 2014.

[11] A. Thakur and P. Rajan, "Model-based unsupervised segmentation of birdcalls from field recordings," in *Proc. Int. Conf. Signal Process. Commun. Syst.*, 2016.

[12] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis Mach. Intel.*, vol. 31, no. 2, pp. 210–227, 2009.

[13] S. Dolinar, "Maximum-entropy probability distributions under l_p-norm constraints," in *TDA Progress Report, NASA, Communications Systems Research Section*, February 1991, pp. 74–87.

[14] T. Maszczyk and W. Duch, "Comparison of Shannon, Renyi and Tsallis entropy used in decision trees," *Art. Intel. Soft Comp.*, pp. 643–651, 2008.

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

[16] "Art-sci center, University of California," `http://artsci.ucla.edu/birds/database.html/`, Accessed: 2016-07-10.

[17] "Filtering and noise adding tool," `http://dnt.kr.hs-niederrhein.de/`, Accessed: 2016-11-14.

[18] "Freesound," `http://freesound.org/`, Accessed: 2017-3-13.

[19] B Yegnanarayana, Carlos Avendano, Hynek Hermansky, and P Satyanarayana Murthy, "Speech enhancement using linear prediction residual," *Speech Commun.*, vol. 28, no. 1, pp. 25–42, 1999.

[20] "Xeno-canto," `http://www.xeno-canto.org`, Accessed: 2017-3-14.