Model-based unsupervised segmentation of birdcalls from field recordings

Anshul Thakur
School of Computing and Electrical Engineering
Indian Institute of Technology Mandi
Himachal Pradesh, India
Email: anshul_thakur@students.iitmandi.ac.in

Padmanabhan Rajan
School of Computing and Electrical Engineering
Indian Institute of Technology Mandi
Himachal Pradesh, India
Email: padman@iitmandi.ac.in

***Abstract*—In this paper, we describe an unsupervised, species independent method to segment birdcalls from the background in bioacoustic recordings. The method follows a two-pass approach. An initial segmentation is performed utilizing K-means clustering. This provides labels to train Gaussian mixture acoustic models, which are built using Mel frequency cepstral coefficients. Using the acoustic models, the segmentation is refined further to classify each short-time frame as belonging either to the background or to call-activity. Different features, namely short-time energy, Fourier transform phase-based entropy and inverse spectral flatness (ISF) are evaluated within the framework of the proposed method. Our experiments with real field recordings on two datasets reveal that the ISF reliably provides better segmentation performance when compared to the other two features.**

I. Introduction

In passive bioacoustic monitoring, recording devices collect audio signals continuously from the environment. Such automated recording devices can be programmed to capture audio signals throughout the day and night, or can be programmed to record at specific times. The data collected in this manner can be used by ecologists and conservationists to detect, determine the range, population size, etc. of sound-emitting animals. Since several of these devices can be deployed in the region of interest, a large amount of audio data can be collected relatively easily. The volume of data collected is usually too large for human experts to analyze. Therefore, such recordings are usually processed by automatic or semi-automatic techniques. In this paper, we describe an automatic method to process recordings primarily consisting of birdcalls.

The first step in processing such recordings is to distinguish (or segment) birdcalls from the background. In real-world field recordings, there may be a variety of background sounds, including the sound of the wind, the rushing flow of a jungle river and the sound of the rain. The quality of the recordings may also vary considerably. The technique proposed in this paper is shown to work reliably in a variety of environments and outperforms earlier techniques utilized for birdcall segmentation.

Some studies on bird species identification using bioacoustic data have used manual segmentation [1] [2] [3]. Time-domain segmentation based on energy is used in some studies [4] [5] [6]. However in low signal-to-noise ratio (SNR) conditions, the performance of energy-based segmentation is likely to suffer. In [7], a time-frequency based segmentation procedure using a random forest classifier is proposed to segment the bird vocalizations in low SNR conditions. In [8], a KL-divergence based technique is used for segmentation. The KL-divergence between the normalized power spectral density of an audio frame and the uniform distribution is computed. The local minima of this KL-divergence correspond to the boundaries of the bird vocalizations. In [9], an entropy-based bird phrase segmentation technique using the spectrogram is described. Birdcalls have more structure or have higher correlation in comparison to the background; hence call regions exhibit lesser entropy than the background. This difference in entropy has been used to distinguish the background from the birdcalls. A similar technique using the entropy estimated from the phase of the Fourier transform has been proposed in [10]. The entropy-based segmentation techniques requires prior knowledge of the average length of the segments between the calls. For optimal performance, this information has to be species specific, and hence needs to be modified when applied to the field recordings containing calls from multiple species.

In this work, we propose an unsupervised, species independent birdcall segmentation method, with no assumptions about the call duration or the recording environment. The technique is a two-pass process: an initial birdcall segmentation is performed using K-means clustering on the inverse spectral flatness (ISF). This provides training labels to build Gaussian mixture acoustic models, using Mel frequency cepstral coefficients (MFCCs). Then, Bayes rule with a recording-specific prior is used to refine the segmentation decisions for each short-time frame. The initial segmentation, acoustic model building and segmentation refinement are all done using the audio recording under consideration, and require no other data.

Various studies [11] [12] [13] have used unsupervised, two-pass approaches for distinguishing active and inactive regions in human speech signals. These approaches utilize an automatic initial segmentation which is then used to build acoustic models for speech and non-speech. In all of the above studies, short-term energy based segmentation has been used to make reliable initial decisions. In [12], zero crossing rate (ZCR) is also used along with energy to get initial segmentation decisions. In [13], energy based segmentation with spectral subtraction is used to get reliable initial decisions in low SNR conditions. Maximum likelihood (ML) training is used in [11] and [13], while maximum a posteriori (MAP) training is used in [12] to build the acoustic models. The main highlight of the technique proposed in this work is the use of inverse spectral flatness (ISF) to get reliable training labels to build the acoustic models. The ISF is estimated from the linear prediction (LP) residual [14] and is shown to reliably separate the background and call-activity regions even in low SNR conditions. Additionally, the proposed method uses a recording-specific prior to refine the final segmentation decisions.

The rest of this paper is organized as: In section II, three different features i.e short term energy, entropy and inverse spectral flatness are explored for applying the initial segmentation and generating training labels for building the acoustic models. In section III, the proposed unsupervised method for refining the initial segmentation is explained. Section IV and V describes performance analysis and conclusion respectively.

II. Features for preliminary decisions

Given an audio recording, we first get preliminary frame-wise labels, each frame being classified as belonging to the background or call-activity. Each short-time frame is of size 20 ms with an overlap of 25%. For this task, we examine short term energy (STE), phase-based entropy [10] and inverse spectral flatness (ISF) [14]. Also, a post-processing technique is applied on these features to further increase the contrast between background and call-activity in the feature domain.

A. Short term energy

Short term energy (STE) is defined as the sum of squares of samples in a frame. In ideal conditions, the STE associated with a background frame is less than the STE of a call-activity frame. This difference in STE between the background frames and the call-activity frames can be used to distinguish between them. Figure 1(b) shows the STE for a given recording containing two birdcall segments. The ability of STE to discriminate between the background and the call-activity becomes poorer as the SNR decreases.

B. Entropy from the Fourier transform phase

In the spectral domain, birdcalls have more structure (e.g. harmonics may be present) or are more correlated in comparison to the background. The higher correlation leads to lower entropy and vice-versa. Thus, the call-activity regions in a recording will exhibit lower entropy than the background. This entropy difference can be used to segment birdcalls. In [9], the entropy is estimated from a time-frequency window in the spectrogram, and a Bayesian change point detection scheme is applied to obtain the start and end points of bird phrases. The width of the time-frequency window (along the time axis) should be greater than smallest distance between two consecutive phrases in the audio recording, while the height of window depends on the frequency range of sounds produced by the species under consideration. In our earlier study [10], we proposed to use the entropy estimated from the phase spectrum of the Fourier transform, rather than from the magnitude spectrum. The entropy estimated from the phase spectrum demonstrated better segmentation performance when compared to that from the magnitude spectrum. In this paper, we explore this phase-based entropy to get the initial segmentation. Figure 1(c) depicts the behaviour of the phase-based entropy for the background and call-activity.

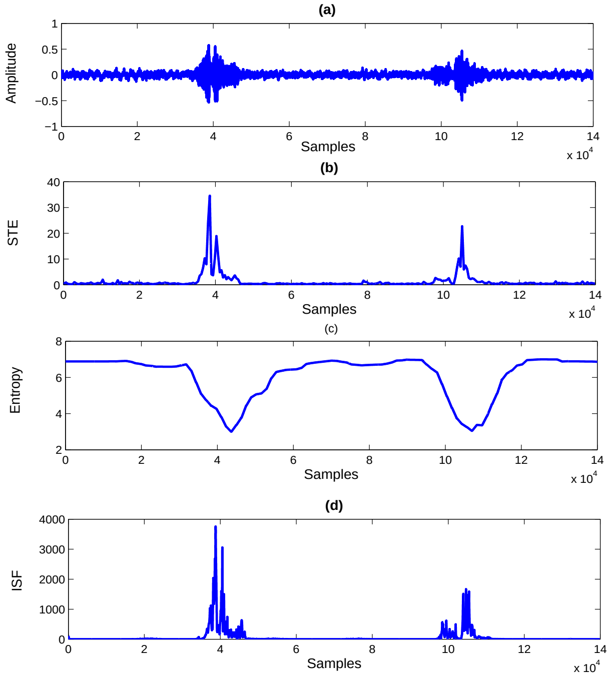

Fig. 1. **(a)** Audio segment containing two calls from Cassin's vireo. **(b)** Short term energy (STE) calculated using 20 ms frames. **(c)** Entropy calculated from whitened phase spectrum using the frequency range of 1500 Hz to 7000 Hz and time-frequency window of 138 ms. **(d)** Inverse spectral flatness (ISF) of the samples in 1(a), calculated using 2 ms analysis frames and LP model order of 5.

C. Inverse spectral flatness

The inverse spectral flatness (ISF) is defined as the ratio of the energy of a small segment of the audio signal (1-2 ms) to the energy of the corresponding segment of its linear prediction (LP) residual signal [14] [15]. The LP residual is obtained by inverse filtering the input audio signal with the LP filter. Since the samples of the LP residual are uncorrelated, a small analysis window (1-2 ms) can be used to compute the ISF. For each small window, the ratio of the energy of the input signal and that of the residual signal is indicative of the reduction of correlation among the samples of the respective signals. This is analogous to comparing the flatness of the input signal spectrum and the residual spectrum. For noise-like regions, the input signal already exhibits a flat spectrum. Hence, for these regions, the ratio of the energies of the input signal and the residual signal is close to unity [14]. Whereas, for call-activity regions, the ratio will be higher. Moreover, the use of the small analysis window increases the temporal resolution. Figure 1(d) depicts the behavior of ISF for the background and the call-activity regions.

The ISF is calculated from the pre-emphasized audio signal using the following steps [15]:

- Calculate the LP residual signal from the pre-emphasized input signal using LP analysis with 20 ms frames that overlap by 50%.
- Calculate the ISF by taking the ratio of the energy of the input signal to the ratio of the energy of the LP residual signal for every 2 ms non-overlapping frame. This gives one ISF value per frame.
- Repeat each ISF value l times, here l is the number of samples in a 2 ms frame (882 samples for a sampling rate of 44.1 kHz). This is done to bring sample-wise synchronization between the ISF and the input signal.
- Smooth the ISF vector using a Hamming window of length $l + 1$.

D. Post-processing

The presence of background sounds like rain, wind etc. can affect the segmentation performance. All the features discussed earlier are susceptible to these kind of sounds. To mitigate this problem, a $\tanh$ based mapping function (equation 1) [15] can be used.

$$\zeta_k = \left(\frac{\zeta_{max} - \zeta_{min}}{2}\right) \tanh\left(\alpha_g \pi \left(\zeta_k - \alpha_0\right)\right) \left(\frac{\zeta_{max} + \zeta_{min}}{2}\right) \tag{1}$$

Here ζ_k is the feature value associated with kth frame. ζ_{max} and ζ_{min} define the range of output values. In this work, we have considered ζ_{max} and ζ_{min} to be 0.8 and 0.2 respectively. α_g is a constant scaling factor and α_0 defines the slope of the $\tanh$ function. More details about these parameters is given in section IV-C.

This function normalizes the features by smoothly mapping them to a user-specified range. The features with low amplitude are mapped to the lowest values and those with larger amplitude are mapped to the highest values in the specified range (ζ_{min} to ζ_{max}). This increases the contrast between background regions exhibiting low amplitude features and call-activity regions having high amplitude features. The slope (α_0) of this $\tanh$ function defines the extent to which low feature values corresponding to the background disturbances are removed. Figure 2 shows the function corresponding to equation 1.

Unlike the STE or the ISF, equation 1 cannot be directly applied to the entropy values due to its reverse behaviour (low entropy implies calls and high implies background). Therefore, to post-process the entropy, the function defined in equation 1 is applied with a negative value of α_g. This function with negative α_g is capable of smoothly mapping the higher entropy values to lower values and lower entropy values to higher values in the user-specified range. Thus, after post-processing the entropy, the lower mapped values will correspond to the background and the higher values will correspond to the call-activity. The results of applying the post-processing on STE, entropy and ISF given in figure 1, are depicted in figure 3.

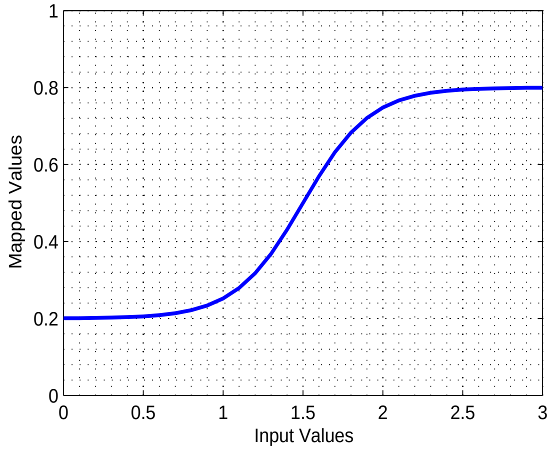

Fig. 2. $\tanh$ function with α_g=0.75, α_0=50% of the range of input values, ζ_{min}=0.2 and ζ_{max}=0.8

E. Comparison of STE, Entropy and ISF

To determine the feature that should be used to get initial training labels, we apply simple thresholding based segmentation on clean and noisy field recordings of the Cassin's vireo [16]. The recordings are at different SNR levels, 0, 5, 10, 15 and 20 dB, in three different recording environments (rain, river and waterfall). The ground truth is provided with the audio recordings. For more details about the dataset, please refer to section 4. The F$_1$-score is used as a metric to compare the segmentation performances of the three features discussed earlier. The F$_1$-score is the harmonic mean of precision and recall, and is defined as:

$$F_1\text{-}score = 2 \times \left(\frac{precision \times recall}{precision + recall}\right). \tag{2}$$

Figure 4 depicts the performance of STE, entropy and phase based entropy for the initial segmentation task.

It is clear from figure 4 that ISF performs better than other features across all the considered conditions. Therefore, we use ISF to perform the initial segmentation.

III. Refining the initial segmentation

The previous section established that the ISF post-processed by the $\tanh$ mapping function is a good candidate to perform the preliminary segmentation (ie. the first pass). The ISF values are generated for every sampling instance of the original signal. To synchronize the sample wise ISF values to short-time frames of 20 ms, the mean ISF value in a 20 ms window is used.

K-means clustering is then performed on these mean ISF values to form two clusters (K=2), one corresponding to the background, and the other corresponding to the call-activity regions. Once all the frames are clustered, 50% of the frames from the background cluster and 50% of the frames from

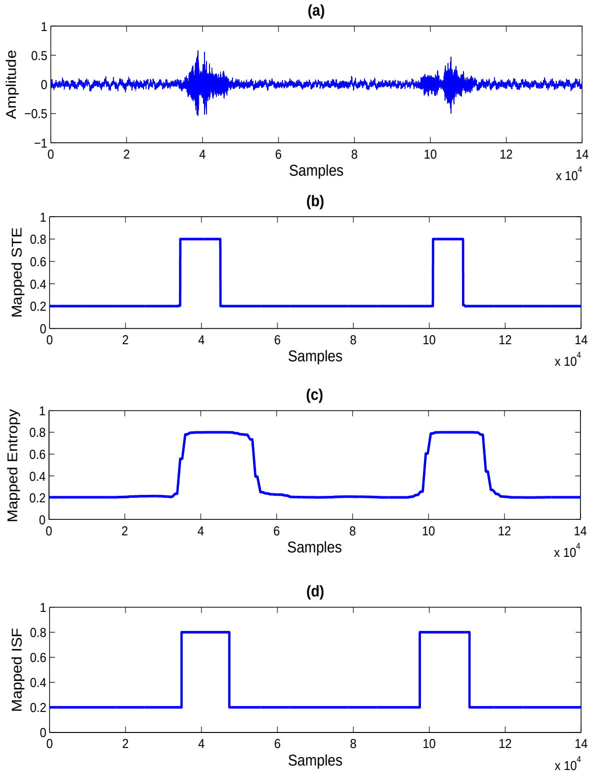

Fig. 3. **(a)** Audio segment containing two calls from Cassin's vireo. **(b)** Mapped STE after applying equation 1 on STE shown in figure 1(b) using α_g=0.75 and α_0= 25% of difference between maximum STE and minimum STE. **(c)** Mapped entropy after applying equation 1 on entropy shown in figure 1(c) using α_g=-1.5 and α_0= 50% of difference between maximum entropy and minimum entropy. **(d)** Mapped ISF calculated using α_g=0.75 and α_0= 40% of difference between maximum ISF and minimum ISF.

call-activity cluster serve as labels to train acoustic models. We utilize Gaussian mixture models (GMMs) built with Mel frequency cepstral coefficients (MFCCs) with energy, delta and acceleration coefficients as acoustic models. The GMM λ_0 refers to the background acoustic model and λ_1 refers to the call-activity acoustic model.

Once the acoustic models are built, the input recording is segmented again using the models (the second pass). By applying Bayes rule, the posterior probability of each class is estimated for every MFCC frame. The posterior probabilities of the frame $\mathbf{x}_t$ belonging to the background and the call-activity are given by equation 3 and equation 4 respectively.

$$p(\lambda_0|\mathbf{x}_t) = \frac{p(\mathbf{x}_t|\lambda_0)p(\lambda_0)}{p(\mathbf{x}_t)} \tag{3}$$

$$p(\lambda_1|\mathbf{x}_t) = \frac{p(\mathbf{x}_t|\lambda_1)p(\lambda_1)}{p(\mathbf{x}_t)} \tag{4}$$

For the tth frame, the likelihood of each class can be estimated from the acoustic model. The prior probability, p(λ_0) or p(λ_1), depends on the chance of the frame $\mathbf{x}_t$ being from the background or from the active region. Most portions of typical recordings consist of the background, with calls occupying only a small percentage of the total duration. Hence it is reasonable to assign a large value for p(λ_0) for all the frames in a recording. However, this does not take into consideration the correlation between frames. A more useful value for the prior probabilities, p(λ_0) and p(λ_1), can be approximated by the mapped feature value obtained in the first pass. The mapping function ensures that the range of these prior probabilities lies between 0.2 and 0.8. This prevents the assignment of a very high or very low value to the prior probability. The frame $\mathbf{x}_t$, is assigned to the background class if $p(\lambda_0|\mathbf{x}_t) > p(\lambda_1|\mathbf{x}_t)$ or vice-versa. The proposed method for birdcall segmentation is described in algorithm 1.

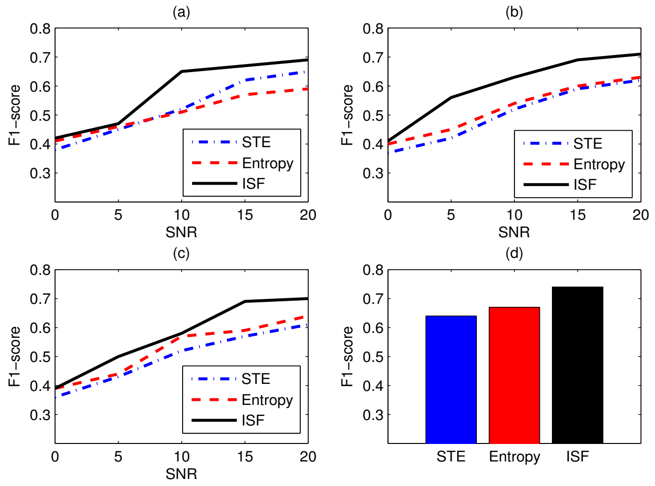

Fig. 4. Comparison of segmentation performances of STE, entropy and ISF on different noise types i.e. (a) rain , (b) river, (c) waterfall and (d) clean data, using thresholding.

IV. Performance Analysis

A. Datasets used

We perform experimental validation of the proposed algorithm on two datasets. The first dataset consists of the recordings of Cassin's vireo, a North American song bird. The other dataset has the recordings of the California thrasher, also a song bird which is found in California and Baja California. In the Cassin's vireo dataset [16], the total duration of recordings is about 45 minutes, out of which about 5 minutes correspond to approximately 800 calls. These recordings are collected over two months and are fairly clean. The California thrasher dataset [17] contains recordings having a total duration of 50 minutes. 20% of the total recordings consist of 3646 calls of the California thrasher. These recordings have been collected from mixed coniferous forests and chaparrals in the United States of America. Each recording in this datset has a quality rating, ranging from 2 (poor quality) to 5 (excellent quality). The excellent quality recordings have no background or microphone disturbances while these disturbances are overwhelming in the poor quality recordings with the other quality ratings in between. The labels (start and end time of each call) are provided with both datasets. These labels are used as the

Algorithm 1: Proposed method for birdcall segmentation

Computation of ISF and post-processing[15]

- Calculate inverse spectral flatness (ISF) from the pre-emphasized input signal as described in II-C. The number of elements in ISF vector is same as the number of samples in the input audio signal.
- Apply post-processing on the ISF as explained in section II-D.

Determine training examples for background and call activity models (Pass 1)

- Divide the ISF vector, calculated in the previous step, into 20 ms frames with an overlap of 25%.
- Find the mean ISF value for each frame calculated in the previous step.
- Apply K-means clustering on these mean ISF values to get two clusters corresponding to the background and the call-activity.
- Frames corresponding to the lower 50% of ISF values in the background cluster and the frames corresponding to the higher 50% ISF values in call-activity cluster serve as labels to build acoustic models in the second pass.

Training the GMM models and testing procedure (Pass 2)

- Calculate Mel frequency cepstral coefficients (MFCC) with energy, delta and acceleration coefficients from the input signal using frame size of 20 ms and an overlap of 25%.
- GMMs for the background and call-activity are trained using the labels generated in pass 1. MFCC features corresponding to these labels are used to build GMMs.
- The final classification decision for each 20 ms frame of the input signal is obtained using equations 3 and 4. The mean ISF m (computed in pass 1) of a frame, $\mathbf{x}_t$, and $1 - m$ are used as prior probabilities for the call-activity and the background class respectively. The frame $\mathbf{x}_t$ is assigned to the call-activity class if $p(\lambda_1|\mathbf{x}_t) > p(\lambda_0|\mathbf{x}_t)$ or vice-versa.

ground truth for evaluating the segmentation performance. The recordings in both the datasets are sampled at 44.1 kHz.

Since the Cassin's vireo dataset is fairly clean, noise is added to study the robustness of the proposed method. Three different types of background sounds i.e. rain, waterfall and river at 0 dB, 5 dB, 10dB, 15 dB and 20 dB SNR are added to the Cassin's vireo recordings. These sounds are considered for this study as they are more likely to influence a field recording obtained from a forest. Filtering and Noise Adding Tool (FaNt)[18] is used to add the sounds to the data. The sound files are downloaded from FreeSound [19].

B. Experimentation

A frame length of 20 ms with 10 ms increment and LP model order of 5 is used to calculate the LP residual signal from the pre-emphasized audio recording. The ISF is calculated using analysis frames of 2 ms length as described in secton II-C. The post-processing discussed in section II-D is applied on the ISF to obtain the mapped ISF values. The details about parameters of the post-processing function (equation 1) is given in next sub-section.

Once training labels are obtained for the two acoustic models using K-means clustering, GMM models are built using MFCC features. Since the variation in the background data is relatively less, a single mixture component is used for this class. Two mixture components are used for the call-activity class. Varying the number of mixtures did not improve the segmentation performance.

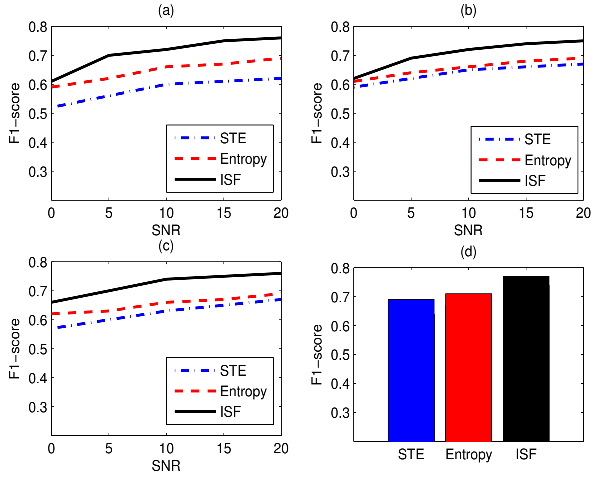

Fig. 5. Segmentation performances of the proposed method in terms of F1-scores on clean and noisy Cassin's vireo data using STE, phase-based entropy and ISF for generating labels during pass 1 and for estimating prior probabilities during pass 2. The segmentation performances are depicted for (a) rain, (b) river, (c) waterfall and (d) clean data.

TABLE I
SEGMENTATION PERFORMANCES OF THE PROPOSED METHOD IN TERMS OF F1-SCORES ON CALIFORNIA THRASHER DATA USING STE, PHASE-BASED ENTROPY AND ISF FOR GENERATING LABELS DURING PASS 1 AND FOR ESTIMATING PRIOR PROBABILITIES DURING PASS 2.

Quality	STE	Entropy	ISF
2 (poor)	0.39	0.41	**0.51**
3	0.51	0.54	**0.68**
4	0.6	0.62	**0.77**
5 (excellent)	0.61	0.62	**0.79**
Complete data	0.59	0.61	**0.72**

The segmentation performance of the proposed method is summarized in figure 5 (for the Cassin's vireo data) and in table I (for the California thrasher data). To make a fair comparison, the proposed method is also evaluated utilizing STE and phase-based entropy for the initial segmentation in the first pass. Acoustic models using MFCCs are built based on the labels generated by the respective features. The prior probabilities used in the segmentation refinement are provided by the mapped feature values (STE or phase-based entropy). The results clearly reveal that ISF provides better initial segmentation and more accurate prior probabilities for segmentation refinement, resulting in higher F1-scores.

C. Sensitivity to the parameters of the mapping function

In this section, we note a few points regarding the parameters used in the mapping function (equation 1). In the proposed method, the mapping function serves two purposes. Firstly, it enhances the contrast between the feature values in the low SNR regions and high SNR regions in the first pass. Secondly, it gives an approximate value of the framewise prior probability used in the second pass. For the three features considered in the experimental evaluation, the values used for the parameters are shown in table II.

TABLE II
PARAMETER SETTING FOR THE POST-PROCESSING FUNCTION

Features	ζ_{min}	ζ_{max}	α_g	α_0
STE	0.2	0.8	0.75	25% of input range
Entropy	0.2	0.8	0.75	50% of input range
ISF	0.2	0.8	-1.5	40% of input range

To some extent, these parameters need to be tuned to the range of the input features given to the function. The parameter α_g controls the shape of the function, whereas the parameter α_0 defines the amount of contrast the mapping achieves. In [15], the parameters α_g and α_0 were set according to factors such as perceived distortion and quality of the enhanced speech. The parameter α_0 controls the slope of the $\tanh$ function. Lower values of α_0 enhances low-valued features; hence this will lead to more false-alarms in the initial segmentation. Similarly, higher values of α_0 will result in missing medium valued features, leading to more misses in the initial segmentation. Both of these cases may result in the GMM models being trained with impure data. Hence a conservative value of α_0 is around the median of the range of the input values (i. e. $\alpha_0 \approx 50\%$ of the input range). The value of α_g is negative for the entropy due to its reverse behavior (see section II-D), otherwise it is set following [15]. Minor variations in segmentation performance were observed by varying these parameters around these values. The values in Table II obtained the best results for each of the features considered.

V. CONCLUSION

In this work, we proposed an unsupervised method to segment birdcalls from the background in bioacoustic recordings. Initial segmentation using K-means clustering is refined using MFCC-based Gaussian mixture acoustic models. Short-time energy, Fourier transform phase-based entropy and inverse spectral flatness (ISF) were evaluated in the framework of the method. The range of values of these features between low SNR regions and high SNR regions was increased by using a mapping function. Compared to the other two features, ISF demonstrated better ability to segment the calls more accurately. By taking advantage of the uncorrelated nature of the LP residual, the ISF effectively discriminates the background regions from the call-activity regions present in the recording. Short-time energy is unable to take advantage of the spectral information, whereas the entropy-based feature has reduced temporal resolution when compared to the ISF.

Future work includes the evaluation of the proposed method on more diverse datasets, with recordings containing calls from many different species. Other classifiers like support vector machines or neural networks can also be explored.

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] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] A. Thakur and P. Rajan, “Entropy-based segmentation of birdcalls using Fourier transform phase,” 2016, (Under review).

[11] X. Anguera, C. Wooters, and J. M. Pardo, “Robust speaker diarization for meetings: ICSI RT06s meetings evaluation system,” in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 346–358.

[12] H. Sun, T. L. Nwe, B. Ma, and H. Li, “Speaker diarization for meeting room audio,” in *Proc. Interspeech*, 2009, pp. 900–903.

[13] T. Kinnunen and P. Rajan, “A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data,” in *Proc. Int. Conf. Acoust. Speech, Signal Process*, May 2013, pp. 7229–7233.

[14] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. Murthy, “Processing linear prediction residual for speech enhancement,” in *EUROSPEECH*, 1997.

[15] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, “Speech enhancement using linear prediction residual,” *Speech Commun.*, vol. 28, no. 1, pp. 25–42, 1999.

[16] “Cassin’s vireo recordings,” http://taylor0.biology.ucla.edu/al/bioacoustics/, accessed: 2016-03-20.

[17] “Art-sci center, University of California,” http://artsci.ucla.edu/birds/database.html/, accessed: 2016-07-10.

[18] “Filtering and noise adding tool,” http://dnt.kr.hs-niederrhein.de/, accessed: 2016-07-10.

[19] “Freesound,” http://freesound.org/, accessed: 2016-07-10.