Unsupervised birdcall activity detection using source and system features

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

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

***Abstract*—In this paper, we describe an unsupervised method to segment birdcalls from the background in bioacoustic recordings. The method utilizes information derived from both source features as well as system features. Three types of source features are extracted from the linear prediction residual signal, and Mel frequency cepstral coefficients are extracted from the system features. The source features are used to generate automatic labels, which are then used to train acoustic models for distinguishing birdcall frames from the background. In the context of a technique proposed earlier, our study demonstrates the improvements brought about by the inclusion of additional source features.**

I. Introduction

Acoustic monitoring of habitats holds potential for various ecological studies, including the detection of a specific species, or in estimating the biodiversity of a given region. Automatic processing of such bioacoustic recordings also have utilities in archiving, indexing, and in search and retrieval. Sophisticated, weather-proof bioacoustic recording devices [1] are available, which when deployed in the field, produce a large amount of acoustic data. Processing this data usually happens offline. For most applications, the first step in this is to identify portions of interest (i. e. active regions) in the recording. In this paper, we perform activity detection (or segmentation) of birdcalls in a given bioacoustic recording.

Previous approaches to automatic birdcall segmentation have used signal energy [2] [3] [4], time-frequency analysis [5], KL-divergence [6], template-matching [7] and spectral entropy [8]. Energy-based techniques, though simple, are prone to noise. The time-frequency representation based method proposed in [5] classifies each time-frequency unit as birdcall activity or as background. Hence it is capable of segmenting the birdcall portions which overlap in time. The KL divergence based method in [6] measures the deviation of the signal spectrum from a flat spectrum. Birdcalls have more correlation when compared to the background, and hence this can be used to distinguish between the two. Template-matching based approaches, such as the one in [7], are typically applied when the structure of calls are known apriori (i. e. they are species specific.) Entropy-based techniques are prone to lower performance when there are calls from other birds in the background. In this scenario, for an application that processes calls from a specific species, entropy-based segmentation may produce several false alarms.

In the methods listed above, the energy based methods in [2] [3] [4], the spectral entropy based technique in [8] and the KL-divergence based method in [6] are unsupervised techniques. On the other hand, the time-frequency based method [5] is supervised and requires prior training of a random forest classifier on bird and non-bird sounds. In a practical setting, supervised approaches for segmenting birdcalls from the background are of limited use. In field conditions, it is usually not known which species (or individuals within species) are vocalizing at a given instant. Also, the nature of the background is unpredictable. Hence, unsupervised approaches for segmentation are attractive.

Our prior work in this topic proposed a two-stage, unsupervised method to segment birdcalls from the background [9]. In this work, we propose improvements in both stages, by utilizing information from the excitation signal driving the avian vocal tract. The source-system model for human speech is well studied. This model has been applied to avian vocalizations as well [9], [10], [11]. These works have mostly focused on using the information present in the system component, which represents the vocal tract. In this paper, we explore the use of the source component, which represents the excitation input to the vocal tract. Motivated by the use of such source features for the processing of human speech, we apply the same to segment bird vocalizations. Our studies reveal that inclusion of source features considerably improves segmentation performance. However, in an unsupervised setting, the problem is nontrivial.

In the processing of human speech, source features have been shown to contain useful information. This has been exploited for tasks including speaker recognition [12], [13], [14], pitch tracking [15], voice activity detection [16], speech enhancement [17] and audio clip classification [18]. Conventional system-based features such as Mel frequency cepstral coefficients (MFCCs) are prone to a large amount of variation, based on how the speech is uttered. On the other hand, features derived from the source are less prone to degradation from channels and microphones [19]. Although, by themselves, these features may not perform as well as the system-based

978-1-5090-5356-8/17/$31.00 © 2017 IEEE

features, the combination of source and system features have shown improvements when compared to the system-based features alone [13], [14], [16]. These studies demonstrate the complimentary information present in the source-based features.

Motivated by the above applications of source features, we investigate the possible improvements brought about by them in distinguishing birdcall activity from the background in bioacoustic recordings. We utilize source features in the framework of the unsupervised segmentation algorithm proposed in [9]. The rest of this paper is organized as follows. In section II, we summaries the algorithm proposed in [9]. The residual features used in the present work are described in section III. In section IV, the improvements to the work in [9] are explained in detail. Performance analysis, discussion and conclusion are in sections V, VI and VII respectively.

II. Unsupervised Framework

The model-based unsupervised framework proposed in [9] uses a two-pass process. This framework uses the input recording itself to train acoustic models and no separate training data is required. In the first pass, labels for training the acoustic models are created automatically. In the second pass, these training labels are used to build acoustic models and final classification decisions are taken using these acoustic models. Figure 1 depicts a block diagram of the method proposed in [9].

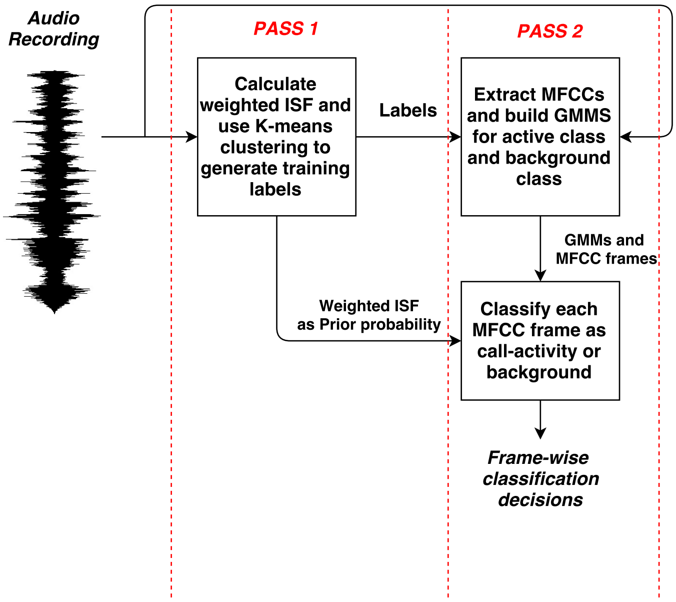
Audio Recording
PASS 1
PASS 2
Calculate weighted ISF and use K-means clustering to generate training labels
Labels
Extract MFCCs and build GMMS for active class and background class
GMMs and MFCC frames
Weighted ISF as Prior probability
Classify each MFCC frame as call-activity or background
Frame-wise classification decisions

Fig. 1. Block diagram showing the two-pass unsupervised framework.

During the first pass, inverse spectral flatness (ISF) is used to distinguish birdcall activity and the background regions in the input recording. ISF is defined as the ratio of the energy of a short segment of the input audio signal to the energy of the linear prediction (LP) residual of the corresponding segment [17], [20]. The ISF exhibits high values for the birdcall activity regions while it is close to unity for the background regions. Figure 2(b) shows the behavior of ISF for call activity and background regions for the recording in Figure 2(a). To give approximately equal values to the ISF of all birdcall activity regions and to ignore low energy background sounds, the ISF is post-processed using a $tanh$ based smooth weighting function defined in equation 1 [17].

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

Here ζ_k is the ISF value of kth frame. ζ_{max} and ζ_{min} define the range of output values. α_g is a constant scaling factor and α_0 defines the slope of the $\tanh$ function.

K-means clustering with $K = 2$ is applied on the weighted ISF values (corresponding to each frame) to get two clusters. One cluster is likely to contain majority of birdcall activity frames while the other mostly contains background frames. The upper 50% of the frames from birdcall activity cluster are labeled as activity. Similarly, the lower 50% of the frames from background cluster are labeled as background. These labels are given as inputs to the second pass. In pass 2, Mel frequency cepstral coefficients (MFCC) are extracted for each frame of the input audio recording. Gaussian mixture models (GMM) for both classes are built using the MFCCs and the training labels generated in pass 1. The final frame-wise classification decision is done using Bayes' rule. The likelihood for both classes is obtained using the GMMs while the weighted ISF of a frame from pass 1 is used as prior probability estimate.

In this work, we propose changes to both pass 1 and pass 2 by including source features. These changes are discussed in section IV.

III. Source Features Used

In this work, along with ISF we have used three additional features derived from the LP residual which models the source information. These are: summation of residual harmonics (SRH) [15], prediction gain (PG) [21] and power difference of spectra in sub-bands (PDSS) [22]. ISF, SRH and PG are used to generate training labels in pass 1 while PDSS along with MFCC is used for building the acoustic models in pass 2.

A. Summation of residual harmonics

Summation of residual harmonics (SRH) is a robust pitch tracker proposed in [15] and is used for voice activity detection in [16]. SRH exploits the strength of harmonics and sub-harmonics present in the spectrum of the LP residual signal to estimate pitch and to make voice/non-voice decision. SRH exhibits high values in the presence of voicing due to strong harmonic components present in the residual spectrum. On the other hand, SRH values are lower for the background due to the lack of harmonic components in the residual spectrum. In [15], SRH is shown to be robust in presence of different additive noises at low SNRs. For a frame, SRH can be calculated using the following steps [16]:

- Calculate the LP residual $r(n)$ by inverse filtering the speech signal.
- Calculate the power spectrum $R(f)$ of the residual.
- Calculate SRH using the equation below:

$$\mathrm{SRH} = \underset{f}{\operatorname{argmax}} \left(R(f) + \sum_{k=2}^{\mathrm{N_{harm}}} [R(k.f) - R((k - \frac{1}{2}).f)] \right) \tag{2}$$

where k is the kth harmonic and $\mathrm{N_{harm}}$ is the total number of harmonics. $\mathrm{N_{harm}}$ is fixed to 5 [15]. The f is varied in a possible pitch frequency range.

With the assumption that the LP residual of many birdcalls also exhibit harmonic structure, SRH can be used to distinguish call-activity from the background. The behavior of SRH for call-activity and background is depicted in Figure 2(c).

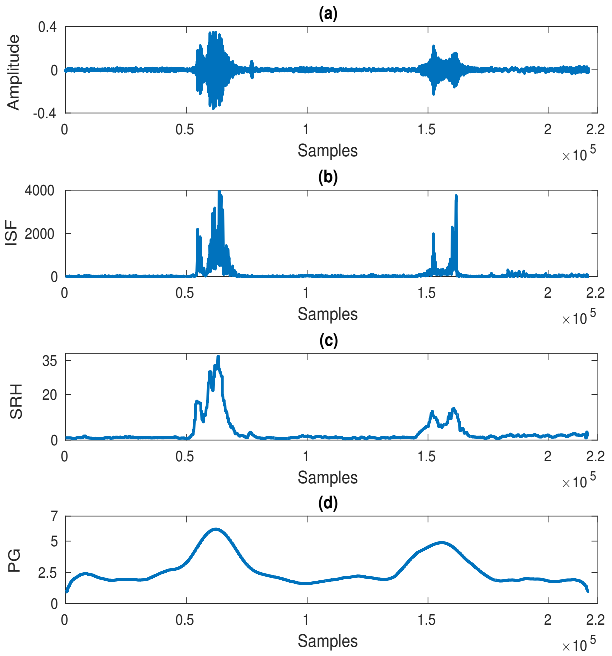

Fig. 2. **(a)** Audio segment containing two calls from Cassin's vireo. **(b)** ISF calculated from audio segment shown in (a). **(b)** SRH calculated from audio segment shown in (a) using pitch range of 200 Hz to 700 Hz. SRH is post-processed using a median filter of order 21. **(d)** Prediction gain calculated from audio segment shown in (a). PG is also post-processed using a median filter of order 21. The LP model order is fixed to 10 for calculating all the three features.

B. Prediction gain

Prediction gain (PG) is similar to the ISF and is defined as the ratio of the signal energy to the LP residual signal energy [21]. The signal energy is obtained from the autocorrelation at zero lag. For the nth frame, PG can be calculated as [21]:

$$G(n) = log(R_{xx}(n, 0)/\epsilon(n)) \tag{3}$$

Here $\epsilon(n)$ is error in the last step of the Levinson-Durbin recursion. The higher correlation in samples of call regions leads to low prediction error. Hence the denominator of equation 3 becomes small for calls and $G(n)$ attains higher values. Figure 2(d) shows the behavior of PG in presence and absence of birdcall activity.

C. Power difference of spectra in sub-bands of residual (PDSS)

PDSS is the spectral flatness computed for various sub-bands. It uses the differences in spectral flatness of sub-bands to find the peaks and dips in the power spectrum. The larger power differences between these peaks and dips indicates the higher periodicity [22]. Although PDSS was originally used for speaker identification task in [14], [22], its ability to capture the information present in the harmonic structure in the residual spectrum makes it a suitable feature representation for detecting the presence/absence of birdcalls. For a given frame, PDSS can be calculated using the following steps [14]:

- Calculate the LP residual, $r(n)$.
- Calculate the power spectrum of the residual, $R(f)$.
- Group the power spectrum into M sub-bands.
- Calculate the ratio of geometric mean and arithmetic mean of power spectrum for sub-band, i and subtract it from 1.

$$\mathrm{PDSS}(i) = 1 - \frac{(\pi_{k=L_i}^{H_i} R(k))^{1/N_i}}{\sum_{k=L_i}^{H_i} R(k)/N_i} \tag{4}$$

Here L_i and H_i are lower and higher frequency limits of ith sub-band. Also N_i is the number of frequency samples in the ith sub-band. The value of PDSS is close to 1 if LP residual spectrum exhibits higher periodicity. On the other hand, this value is close to zero for low periodicity [22].

ISF, SRH and PG are 1-dimensional features whereas PDSS is a M-dimensional feature vector where M corresponds to the number of sub-bands in equation 4.

IV. Proposed Improvements

In this section we describe the improvements to the method discussed in section II. First we discuss the changes proposed for automatically generating training labels from the input recording during pass 1. Then we describe the changes proposed for pass 2.

A. Improvements in pass 1

Along with ISF, SRH and PG are used to get reliable training labels. For each frame of the input audio signal, these features are extracted as discussed in section III. The SRH and PG values are smoothed using a median filter before further processing. Then, the ISF, SRH and PG are post-processed separately using equation 1. These post-processed features lie in the range specified by ζ_{min} and ζ_{max}. We have kept ζ_{min} and ζ_{max} to be 0.2 and 0.8 respectively as described in [17].

K-medoids clustering is a variation of K-means such that each centroid is one among the data points. These data points are called medoids and are at minimum distance to all the other data points in a cluster. The mean is more influenced by outliers in comparison to the medoid, making K-medoids clustering more robust against noise and outliers [23]. Hence we have applied K-medoids clustering instead of K-means on ISF, SRH and PG individually to get three sets of two clusters. One cluster corresponds to the background and other corresponds to the birdcall activity. Hence for each frame, one label is generated by each of ISF, SRH and PG. The following voting rule is applied to decide final labels of the frames: two out of three labels generated by these features have to agree on any labeling decision. If this condition is not met, then that frame is treated as unlabeled.

The pass 1 also outputs prior probabilities values for each frame along with labels to be used in pass 2. Instead of using the ISF as a prior for any frame, the mean of the post-processed ISF, SRH and PG is used as prior probability. Since each feature is between 0.2 and 0.8 as discussed earlier, the mean of these features for any frame also lies in this range. This range prevents the assignment of a low or a very high prior to any frame. Figure 3 depicts the pass 1 as a block diagram.

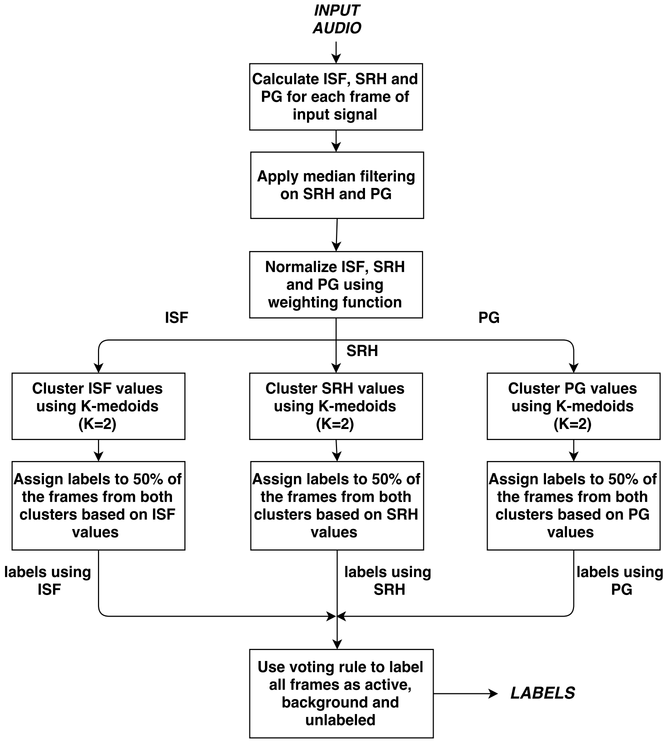
INPUT AUDIO
Calculate ISF, SRH and PG for each frame of input signal
Apply median filtering on SRH and PG
Normalize ISF, SRH and PG using weighting function
ISF
SRH
PG
Cluster ISF values using K-medoids (K=2)
Cluster SRH values using K-medoids (K=2)
Cluster PG values using K-medoids (K=2)
Assign labels to 50% of the frames from both clusters based on ISF values
Assign labels to 50% of the frames from both clusters based on SRH values
Assign labels to 50% of the frames from both clusters based on PG values
labels using ISF
labels using SRH
labels using PG
Use voting rule to label all frames as active, background and unlabeled
LABELS

Fig. 3. Block diagram of pass 1.

B. Improvements in pass 2

In pass 2, MFCC and PDSS features are calculated for each frame of the input signal. Using the labels generated in pass 1 and concatenation of MFCC and PDSS (early fusion) as features, GMMs for call-activity and background classes are built. The final classification decision for each frame is made using Bayes' rule. The posterior probability for each class is calculated. The likelihood is estimated using the GMMs and the prior probability (derived from the mean of ISF, SRH and PG) comes from pass 1. The frame is assigned to the class having maximum posterior probability.

V. Experimentation

The experimental setup is same as in [9]. The proposed approach is evaluated on audio recordings of the Cassin's vireo. These recordings are available at [24]. The recordings have total duration of 45 minutes, out of which 5 minutes correspond to the call-activity. There are approx. 800 bird vocalizations. These recordings are clean but contain low energy background bird vocalizations which are to be ignored. The audio files are sampled as 44.1 kHz. The ground truth is also provided with the dataset. To analyze the performance of the proposed approach in noisy conditions, we added three types of noise i.e. rain, river and waterfall to the audio recordings at 0 dB, 5 dB, 10 dB, 15 dB and 20 dB using Filtering and Noise Adding Tool (FaNt) [25]. The sounds of rain, river and waterfall are obtained from FreeSound [26].

For LP analysis, we use the frame size of 20 ms with an overlap of 50% and a model order of 10. The ISF is calculated using an analysis window of 2 ms as described in [17]. A possible pitch range of 200-500 Hz is used for calculating SRH. This is in accordance with the possible pitch range of Cassin's vireo. The parameter setting for post-processing the ISF, SRH and PG using equation 1 is same as the one used in [9] i.e. ζ_{max} and ζ_{min} is set to 0.8 and 0.2 respectively, $\alpha_g = 0.75$ and α_0 is 40% of the range of input feature values. In pass 2, 12 MFCCs with log energy, delta and acceleration coefficients are used. For calculating PDSS, the residual spectrum is divided into 10 sub-bands of 500 Hz. Different sizes and numbers of sub-bands did not bring about significant changes to the segmentation performance. Delta and acceleration coefficients are also added to PDSS. The removal of delta and acceleration coefficients from both MFCC and PDSS led to a drop in performance. The combination of MFCC and PDSS (MFCC+PDSS) is used for training the GMMs. The number of mixtures of call-activity and background GMM are 2 and 1 respectively. The number of mixtures are decided experimentally.

F$_1$-score is used as a metric to evaluate the performance of the proposed method. The F$_1$-score is the harmonic mean of precision and recall, and can be calculated as:

$$F_1\text{-}score = 2 \times \left(\frac{precision \times recall}{precision + recall}\right). \tag{5}$$

The performance of the approach proposed in this work is compared with the method proposed in [9]. This comparison is depicted in Figure 4. By analyzing figure 4, it is clear that the proposed approach outperforms the earlier method for almost all the cases. The average relative improvements of 1.43%, 1.41%, 2.82% and 2.6% in the F$_1$ scores across all SNRs for rain, river, waterfall and clean data respectively are observed. The improvement in performance is higher especially at low SNRs. Moreover, the method described in [9] outperformed other segmentation techniques based on energy and entropy.

The proposed improvements hence outperform those methods as well.

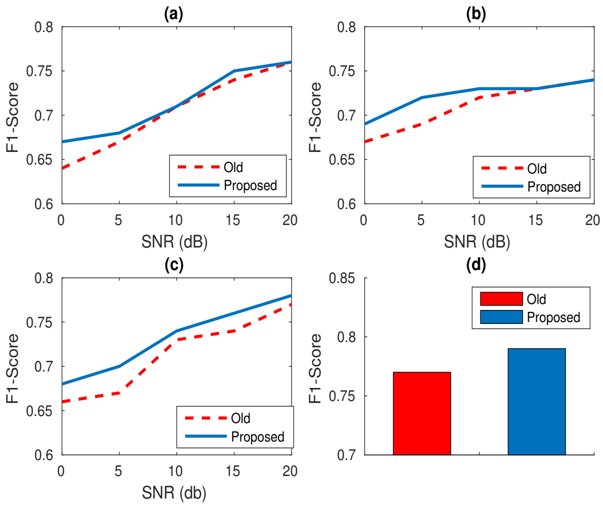

Fig. 4. Comparison of segmentation performances of the method proposed in [9] and the improved approach proposed in this work on different noise types i.e. (a) rain , (b) river, (c) waterfall and (d) clean data.

VI. Discussion

We analyse the improvements brought about by the proposed method in more detail. We first study the improvements brought about in pass 1 due to the inclusion of the SRH and PG features for automatic label generation.

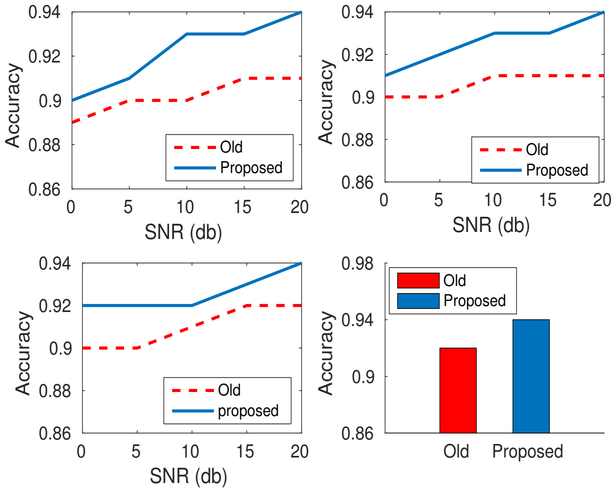

Fig. 5. Comparison of label accuracy generated by the method proposed in [9] and the improved approach on different noise types i.e. (a) rain , (b) river, (c) waterfall and (d) clean data.

The labels generated in pass 1 proposed here are more accurate in comparison to the labels generated by pass 1 of the earlier method in [9]. The comparison of the training label accuracies generated by these two approaches is depicted in figure 5. Here, the accuracy is the ratio of the number of correct labels generated with respect to the ground truth. It is clear from the figure 5 that the proposed improvements in pass 1 has led to an increase in accuracy of the generated training labels. The average relative improvements of 2.22%, 2.23%, 1.76% and 2.17% in label accuracies across all SNRs for rain, river, waterfall and clean data respectively are observed.

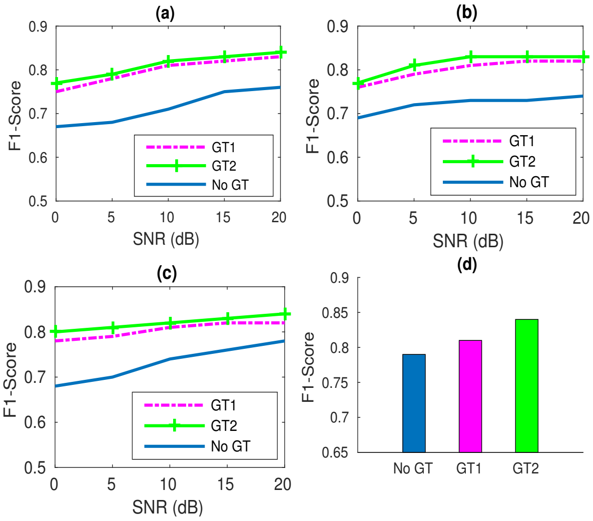

Fig. 6. Comparison of segmentation performances using ground truth as labels for building GMMs with MFCC (GT1) and MFCC+PDSS (GT2) on different noise types i.e. (a) rain, (b) river, (c) waterfall and (d) clean data. The segmentation performance of the proposed unsupervised approach which does not use ground truth for labeling (No GT) is also depicted here.

Figure 6 compares the performance of the proposed method (indicated by solid lines) with the improvements brought about after utilizing ground truth labels in pass 2 (indicated by dotted-dashed lines.) In other words, instead of using the labels generated in pass 1, if the ground truth labels were utilized, there is an average relative improvement of 10.8% percent in segmentation accuracy. This indicates that most of the segmentation errors produced by the proposed method is due to incorrect clustering in pass 1. Figure 6 also shows the performance of using only MFCC features in pass 2 (i.e. not using the PDSS features), while using the ground truth labels. The figures indicates that the PDSS features bring about improvements, but these are quite modest. These plots indicate the scope for improvements in accurate label generation (equivalent to generating clusters of high purity) in pass 1.

The early fusion of ISF, SRH, and PG to PDSS combined with MFCC in pass 2 does not result in major performance improvements. One possible reason for this is that the PDSS provides spectral flatness information for subbands. This incorporates the information provided by the other spectral flatness measures.

VII. Conclusion

In this work, we proposed improvements to our earlier unsupervised birdcall segmentation algorithm. We proposed the use of summation of residual harmonics and prediction gain along with inverse spectral flatness to automatically generate training labels. A voting rule is used on individual label decisions of each feature to get the final training labels. The use of this method decreased the amount of impure training data for building the acoustic models. Additionally, power difference in spectral sub-bands along with MFCC was utilised in building GMM acoustic models for the call class and the background class. This resulted in further improvement of segmentation performance.

References

[1] "Song Meter SM4," https://www.wildlifeacoustics.com/products/song-meter-sm4, accessed: 2016-11-14.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] S. Selouani, M. Kardouchi, E. Hervet, and D. Roy, "Automatic birdsong recognition based on autoregressive time-delay neural networks," in *Proc. Cong. Comput. Intell. Methods Applicant.*, 2005, pp. 6–pp.

[12] S. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Commun.*, vol. 48, no. 10, pp. 1243–1261, 2006.

[13] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, 2006.

[14] M. Chetouani, M. Faundez-Zanuy, B. Gas, and J. Zarader, "Investigation on LP-residual representations for speaker identification," *Pattern Recognition*, vol. 42, no. 3, pp. 487–494, 2009.

[15] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics." in *Proc. Interspeech*, 2011, pp. 1973–1976.

[16] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 252–256, 2016.

[17] 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.

[18] A. Bajpai and B. Yegnanarayana, "Exploring features for audio clip classification using LP residual and aann models," in *Proc. Int. Conf. Intell. Sensing Info. Process.*, 2004, pp. 305–310.

[19] B. Yegnanarayana, S. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. Int. Conf. Acoust. Speech, Signal Process*, 2002, pp. I–541.

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

[21] S. O. Sadjadi and J. H. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 197–200, 2013.

[22] S. Hayakawa, K. Takeda, and F. Itakura, "Speaker identification using harmonic structure of LP-residual spectrum," in *Proc. Int. Conf. on Audio Video-Based Biomet. Person Authent.*, 1997, pp. 253–260.

[23] X. Jin and J. Han, "K-medoids clustering," in *Encyclopedia of Machine Learning*. Springer, 2011, pp. 564–565.

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

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

[26] "Freesound," http://freesound.org/, accessed: 2016-07-10.