2018 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2018, AALBORG, DENMARK

CONVEX LIKELIHOOD ALIGNMENTS FOR BIOACOUSTIC CLASSIFICATION

Anshul Thakur, Arshdeep Singh and Padmanabhan Rajan

School of Computing and Electrical Engineering
Indian Institute of Technology, Mandi
Email: {anshul_thakur, d16006}@students.iitmandi.ac.in, padman@iitmandi.ac.in

ABSTRACT

In this work, we propose a bioacoustic classification framework based on Gaussian mixture models (GMM) and archetypal analysis (AA). The framework utilizes acoustic topic modelling to obtain an intermediate symbolic representation where the discrimination between target classes is more evident than in the input feature space. The proposed framework utilizes the GMM as an acoustic topic model and weighted likelihoods obtained from this GMM are utilized as the intermediate symbolic representation. Class-specific archetypal dictionaries are used to obtain the proposed feature representation, designated as convex likelihood alignments (CLAs), from this intermediate representation. Class-specific signatures are highly evident in these CLAs making them an ideal representation for various bioacoustic classification tasks. Through experiments on two available datasets, it is shown that the proposed CLAs yield comparable or better results than state-of-art approaches.

Index Terms— bioacoustics, archetypal analysis, bird species classification.

1. INTRODUCTION

The advent of programmable automatic audio recording devices has made the collection of bioacoustic data convenient. These devices can record the ecological audio scene for days, and hence, are capable of collecting large amount of audio data. This recorded data often contains important clues about the biodiversity such as population trends and migratory patterns, which can be helpful in surveying and passive monitoring of a target region or an ecosystem [1, 2]. The manual processing of this recorded data is not feasible due to its sheer volume. Hence, automatic techniques to identify or classify the targeted bioacoustic signals is of great utility.

In this work, we propose a bioacoustic signal classification framework that combines the generative abilities of Gaussian mixture modelling [3] and the extremal/convex-hull modelling abilities of archetypal analysis (AA) [4, 5]. The framework produces a discriminative feature representation, designated as convex likelihood alignment (CLA), which are effective for bioacoustic classification. The proposed framework maps the input segments of the target bioacoustic signals to a particular pseudo-acoustic topic. These pseudo-topics emerge when the segments extracted from signals of all the target classes are clustered using a GMM and each GMM component is considered to be a particular pseudo-topic. Thus, each segment of any bioacoustic signal, mapping to a particular GMM component, is considered to be an exemplar of the corresponding pseudo-topic.

The temporal-frequency modulations present in a bioacoustic signal have class-specific signatures [6]. Hence, the short-term modulations present in the segments of bioacoustics signals also exhibit class-specific characteristics. Working on these segments, rather than the whole target signals (such as complete birdcalls), provide generative characteristics to the framework. The proposed framework parameterizes an input segment by a symbolic representation, designated as the weighted likelihood alignment (WLA) vector, using pseudo-topic modelling. For an input segment, a WLA vector is obtained by calculating its weighted likelihood with respect to each pseudo-topic. This projects the segments of all the target classes in the WLA space where the discrimination between the target classes is more evident. This is due to the fact that ideally each class is composed of a particular set of pseudo-topics, which correspond to different components of the GMM. As a result, the weighted likelihoods obtained from different GMM components show distinct divergence for different classes. Hence, the WLA vectors obtained from the segments of different classes show different behaviour.

In the intermediate WLA space, the extremal modelling capability of AA is applied to further increase the discrimination between the classes. AA is applied on WLAs of a particular class to obtain an archetypal dictionary that models the extremal or convex hull of the respective class. These archetypal dictionaries are utilized to obtain the proposed convex likelihood alignments (CLA), from WLAs, using the simplex projection [6]. If the data points of different classes overlap, then the correlation among their archetypal dictionaries will be high. As a result, the convex representations obtained using these dictionaries will not be that discriminative. As discussed earlier, the nature of pseudo-topic modelling forces

978-1-5386-5477-4/18/$31.00 ©2018 IEEE

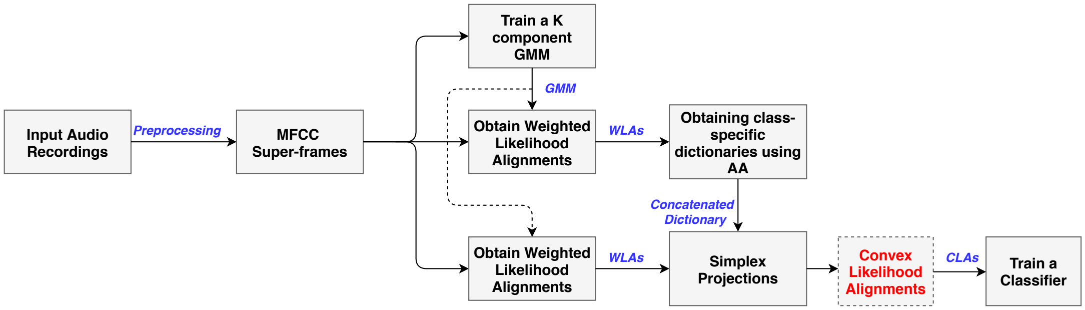

Fig. 1. A systemic diagram of training procedure of the proposed framework. Different audio recordings are used for obtaining the GMM, the archetypal dictionaries and CLAs.

WLAs of different classes to be different. Hence, the class-specific archetypal dictionaries obtained from WLAs show less or no correlation among each other. The CLAs obtained using these uncorrelated archetypal dictionaries exhibit inherent group-sparse structure (see Section 3.2), making them an effective representation for classification.

The rest of this paper is organised as follows. In Section 2, we discuss some of the methods proposed in the literature for bioacoustic classification. In Section 3, the proposed framework is described in detail. Performance analysis and conclusion are in Sections 4 and 5, respectively.

2. BACKGROUND

In the literature, many studies have addressed the task of bioacoustic classification. In [5], the convex representations obtained using AA are used in a SVM framework for the task of bird activity detection. Quin *et. al* proposed to use kernel-based extreme learning machines [7] for classifying bird vocalizations [8]. In comparison to deep neural networks (DNN), these extreme learning machines can be trained effectively, even in low data conditions. In [9], a computationally efficient framework has been proposed for bioacoustic classification. This framework utilises mutual singular spectrum analysis (MSSA) for obtaining bases defining the subspaces of the target bioacoustic classes. Improvising over this MSSA framework, Souza *et. al* proposed to represent a bioacoustic class by a set of subspaces mapped on a Grassmann manifold [10]. A discriminant mechanism is applied on this Grassmann manifold to increase the separability between target classes. Owing to their low computation requirements, both MSSA and its Grassmann variant are ideal for bioacoustic classification in field conditions. However, the documented classification performances of these frameworks can be improved. In other studies, SVM powered by dynamic kernels have also been utilised for birdcall classification [11] and bird activity detection [12]. Apart from dynamic kernels, deep neural networks (DNN) have also been used for birdcall classification in [11, 13]. In a recent study [6], an AA based framework is successfully utilised to obtain convex representations for the tasks of bird species classification. This framework assumes that all the bird vocalizations of a particular class lie on the same subspace in the feature domain which is normally not the case in real-life situations [10]. Apart from these studies, convolutional neural networks (CNN) have also been used in many bioacoustic studies [14, 15]. However, CNN usually are data-intensive frameworks and their effective utilisation in bioacoustics is often deterred by the unavailability of the labelled data. It must be noted that the documented classification performances of the above mentioned methods cannot be directly compared due to differences in the datasets used in their respective studies.

3. PROPOSED FRAMEWORK

In this section, we explain the proposed framework in detail. First, we describe the feature representation to parameterize bioacoustic signals. Then, we explain the process to obtain GMM, weighted likelihood alignments (WLAs), class-specific archetypal dictionaries and convex likelihood alignments (CLAs). Finally, the procedure to classify any input test audio recording is described.

3.1. Feature Representation: MFCC super-frames

Short-term analysis is applied to parameterize each input bioacoustic signal by a sequence of Mel-frequency cepstral coefficients (MFCC), with delta and acceleration coefficients (d-dimensional). The frequency-temporal modulations present in bioacoustic signals cannot be modelled effectively due to the short-term nature of these MFCC vectors. To overcome this issue, W neighbouring MFCC vectors are concatenated to form a Wd-dimensional feature representation, designated as MFCC super-frame. This representation is given as input to the proposed framework and has more context information than one MFCC vector and hence, can model the temporal-frequency modulations in a better manner.

3.2. Training Procedure

The training procedure of the proposed framework is illustrated in Fig. 1. Each training bioacoustic signal is processed to obtain the MFCC super-frames as discussed earlier. An audio recording often contains unimportant background acoustic events along with the target signals. These background events are removed from the training procedure by applying the segmentation method proposed in [16]. Only the target signals, corresponding to these segmented regions, are used in the training process.

3.2.1. Pseudo-topic modelling: obtaining WLAs

To find the pseudo-topics present in the training data of all the target classes, the MFCC super-frames of these classes are clustered together using a K component GMM. The parameters of this GMM are estimated using the expectation-maximization algorithm [3]. Each component of the GMM corresponds to a particular pseudo-topic. The term 'pseudo' refers to the fact that we are assuming each pseudo-topic to be an independent acoustic entity (which they are not).

The GMM (pseudo-topic model) is used to project the input MFCC super-frames into the WLA space. Suppose $\mathcal{M} = [\mathbf{m}_1\mathbf{m}_2\ldots\mathbf{m}_n]$ be a set of MFCC super-frames (not involved in building the GMM). For each $\mathbf{m}_i$, a weighted likelihood alignment vector, $\psi(\mathbf{m}_i) = [\gamma_1(\mathbf{m}_i)\gamma_2(\mathbf{m}_i)\ldots\gamma_K(\mathbf{m}_i)]^T$ is obtained. Here K is the number of GMM components and $\gamma_q(\mathbf{m}_i)$ represents the relative likelihood of $\mathbf{m}_i$ being generated from the qth component of GMM, and is calculated as:

$$\gamma_q(\mathbf{m}_i) = \frac{w_q\mathcal{N}(\mathbf{m}_i|\mu_q,\mathbf{\Sigma}_q)}{\sum_{j=1}^{K} w_j\mathcal{N}(\mathbf{m}_i|\mu_q,\mathbf{\Sigma}_q)}. \qquad (1)$$

Here w_q, μ_q and $\mathbf{\Sigma}_q$ represent weight, mean and covariance of the qth component of the GMM. Using this procedure, all the training MFCC super-frames are converted into WLAs: $\mathcal{M} = [\psi(\mathbf{m}_1)\psi(\mathbf{m}_2)\ldots\psi(\mathbf{m}_n)]$.

3.2.2. Learning class-specific dictionaries using robust AA

To learn class-specific dictionaries, WLAs obtained for the MFCC super-frames of a particular class are pooled together to form a matrix $\mathbf{X} \in \mathbb{R}^{K\times N}$ (K is the dimensionality of a WLA vector and N is the number of pooled WLAs) and AA is applied on $\mathbf{X}$. AA is a form of matrix factorization that decomposes $\mathbf{X}$ as $\mathbf{X} = \mathbf{DA}$ where $\mathbf{D}$ is the required dictionary. $\mathbf{D}$ contains extremal points or archetypes, which lie on the convex hull of the data and are forced to be the convex combination of the data points i.e., $\mathbf{D} = \mathbf{XB}, \mathbf{D} \in \mathbb{R}^{K\times d}$ (d is the number of archetypes) [4]. Since archetypes model the convex hull, AA can be seen as a geometric modelling of the data. The presence of outliers can alter the geometry of the data spread and affect the performance of AA. In the proposed framework, these outliers can arise due to the errors in segmentation process.

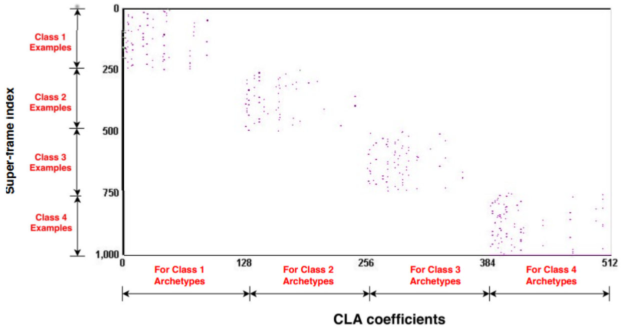

Fig. 2. Inherent block sparse behaviour of CLAs obtained for pseudo-topics of 4 different bird species. Each row represents a 512-dimensional CLA calculated for a MFCC super-frame. Magenta represents the high magnitude coefficients while white represents the low coefficient values.

To mitigate the affect of outliers, a robust version of AA, proposed in [4], is used in this work. This robust version of AA utilises a weighting mechanism to reduce the impact of outliers on the archetypal estimation. The archetypal dictionary, $\mathbf{D}$, is obtained by optimizing the following objective function [4]:

$$\begin{aligned} &\underset{\substack{\mathbf{B},\mathbf{A}\\ \mathbf{b}_j\in\Delta_N,\mathbf{a}_i\in\Delta_d}}{\operatorname{argmin}} \sum_i h(\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2) \\ &\qquad = \frac{1}{2}\sum_i \frac{1}{w_i}\|\mathbf{x}_i - \mathbf{X}\mathbf{B}\mathbf{a}_i\|_2^2 + w_i, \\ &\Delta_N \triangleq [\mathbf{b}_j \succeq 0, \|\mathbf{b}_j\|_1 = 1], \Delta_d \triangleq [\mathbf{a}_i \succeq 0, \|\mathbf{a}_i\|_1 = 1], \\ &\quad w_i \geq \epsilon \end{aligned} \qquad (2)$$

where $\mathbf{x}_i$, $\mathbf{a}_i$ and $\mathbf{b}_j$ are the columns of $\mathbf{X} \in \mathbb{R}^{K\times N}$, $\mathbf{A} \in \mathbb{R}^{d\times N}$ and $\mathbf{B} \in \mathbb{R}^{N\times d}$, respectively. For any scalar u and constant ϵ, the Huber function is defined as, $h(u) = 1/2 \min_{w\geq\epsilon}[u^2/w + w]$. A weight, $w_i = max(\|\mathbf{x}_i - \mathbf{XBa}_i\|_2, \epsilon)$, is assigned for each $\mathbf{x}_i$. During optimization, the value of w_i becomes greater for outliers and reduces their importance in calculating the archetypes. More details about robust AA and its implementation can be found in [4].

3.2.3. Obtaining convex likelihood alignments (CLAs)

All the class-specific archetypal dictionaries obtained during training are concatenated to form a final dictionary, $\mathbf{D}_f = [\mathbf{D}_1\mathbf{D}_2\ldots\mathbf{D}_z]$, where $\mathbf{D}_z$ is the robust archetypal dictionary of the zth class. To obtain the CLA $\mathbf{y}_i$ for any MFCC super-frame $\mathbf{m}_i$, its WLA representation, $\psi(\mathbf{m}_i)$, is projected on the simplex whose vertices are defined by the columns of $\mathbf{D}_f \in \mathbb{R}^{K\times zd}$ (zd is the number of archetypes in the concatenated dictionary $\mathbf{D}_f$):

$$\mathbf{y}_i = \underset{\substack{\mathbf{y}_i\\ \mathbf{y}_i\in\Delta_{zd}}}{\operatorname{argmin}} \|\psi(\mathbf{m}_i) - \mathbf{D}_f\mathbf{y}_i\|_2^2 \qquad (3)$$

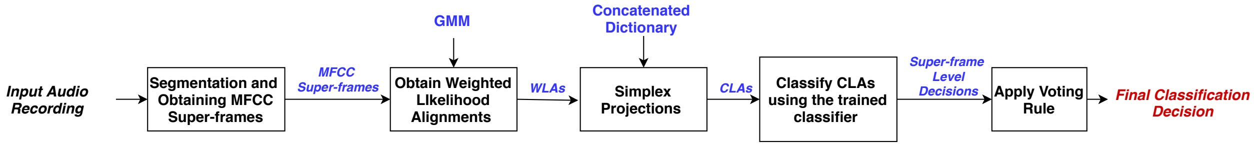
GMM
Concatenated Dictionary
Input Audio Recording
Segmentation and Obtaining MFCC Super-frames
MFCC Super-frames
Obtain Weighted LIkelihood Alignments
WLAs
Simplex Projections
CLAs
Classify CLAs using the trained classifier
Super-frame Level Decisions
Apply Voting Rule
Final Classification Decision

Fig. 3. A systemic block diagram illustrating the testing procedure used in the proposed framework.

where $\Delta_{zd} \triangleq [\mathbf{y}_i \succeq 0, \|\mathbf{y}_i\|_1 = 1]$. In this work, we have used the active-set algorithm proposed in [4] to solve equation 3.

The CLAs obtained for different classes exhibit different behaviour, making them an ideal representation for classification. To highlight this behaviour, we obtained CLAs for the super-frames extracted from the vocalizations of 4 different bird species. These CLAs are obtained using a 32-component GMM for pseudo-topic modelling and class-specific archetypal dictionaries having 128 atoms. The MFCC super-frames are obtained by concatenating $W = 10$ neighbouring MFCC vectors. Fig. 2 depicts the discriminative nature of these CLAs. This figure shows the 512-dimensional CLAs obtained for 1000 pseudo-topics (250 of each class). The analysis of Fig. 2 makes it clear that CLA exhibits an inherent group sparse structural behaviour i.e. the CLA coefficients obtained for the super-frames of a particular class exhibit high magnitude for the atoms of that class only. It must be noted that this group sparse behaviour is due to the nature of simplex projections and no specific sparsity constraints are applied on equation 3. During training phase, a classifier such as random forest or SVM is trained using these CLAs. This classifier is used during testing for classifying super-frames extracted from a test example. These super-frame level decisions are used for assigning a class label to the test example.

3.3. Classification Procedure

An input audio recording is segmented to isolate the target bioacoustic signal from the background. These segmented signals are processed to obtain MFCC super-frames. As a result, this input audio recording is represented by a set of MFCC super-frames, $\mathcal{E} = [\mathbf{e}_1\mathbf{e}_2 \dots \mathbf{e}_n]$. The GMM trained during the training phase is used to obtain the WLA vector, $\psi(\mathbf{e}_i)$ for each $\mathbf{e}_i \in \mathcal{E}$. These WLAs are projected on the concatenated archetypal dictionary $\mathbf{D}_f$, under simplex constraints, to obtain a set of CLAs ($\hat{\mathcal{E}} = [\mathbf{y}_1\mathbf{y}_2 \dots \mathbf{y}_n]$) using equation 3. Each $\mathbf{y}_i \in \hat{\mathcal{E}}$ is categorized using a trained classification model to obtain super-frame level classification decisions. A voting rule is applied on these decisions to obtain the final label for $\hat{\mathcal{E}}$. The overall testing procedure is illustrated in Fig. 3.

4. PERFORMANCE EVALUATION

4.1. Datasets Used

The classification performance of the proposed framework is evaluated on two different datasets: bird species classification dataset and frog species classification dataset. The bird dataset contains audio recordings of 50 bird species and are obtained from three different sources. The recordings of 26 bird species were obtained from the Great Himalayan national park (GHNP), in north India. These recordings are used in [11] for evaluating the classification performance. The recordings of 7 bird species were obtained from the bird audio database maintained by the Art & Science center, UCLA [17]. The audio recordings of the remaining 17 bird species were obtained from the Macaulay Library [18]. All these recordings are 16-bit mono wav files having a sampling rate of 44.1 kHz. The information about these 50 species along with the total number of recordings and vocalizations per species is available at http://goo.gl/cAu4Q1.

The frog dataset containing audio recordings of 10 different species, used in [9, 10], is also used here. These recordings are also 16-bit mono, wav files, sampled at 44.1 kHz and are available at http://goo.gl/FFBzbb.

4.2. Experimental Setup

Parameter setting: A frame size of 20 ms with a 50% overlap is used to obtain a 39-dimensional MFCC representation. 10 neighbouring ($W = 10$) MFCC vectors are concatenated to obtain 390-dimensional pseudo-topics. A GMM with 512 components is trained to obtain 512-dimensional WLAs. For each class, an archetypal dictionary with 128-atoms is obtained. A random forest classifier with 100 decision trees is used for categorizing CLAs. For both the datasets, the above mentioned parameters are used. The optimal number of GMM components are chosen using Akaike information criterion (AIC). All the other parameters are tuned empirically.

Comparative Methods: The classification performance of the proposed framework is compared with various existing bioacoustics classification methods such as GMM, SVM with dynamic kernels (intermediate matching kernel (IMK) and pyramid matching kernel (PMK)) and a DNN based approach proposed in [11]. The performance of the proposed framework is also compared with compressed convex sparse representations (CCSE) proposed in [6]. MSSA [9] and Grassmann

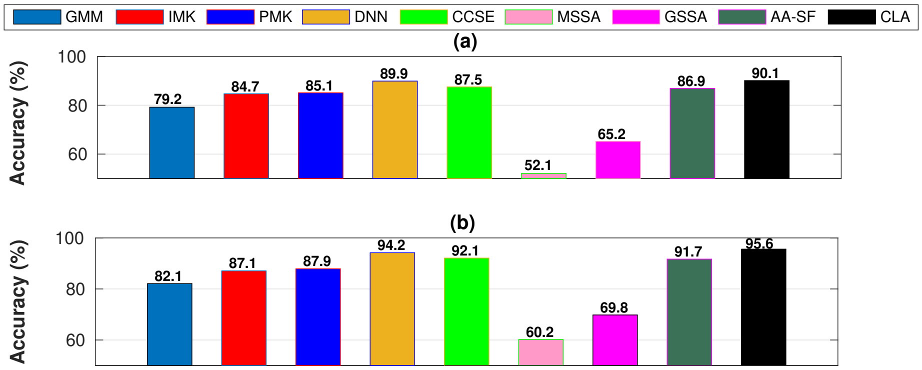

Fig. 4. Classification performance of different methods on (a) 50 bird species and (b) 10 frog species (averaged across three folds).

singular spectrum analysis (GSSA) [10] are also used for the performance comparison. In addition, a variant of the proposed framework is also used as a comparative method. This variant applies AA on MFCC super-frames (AA-SF) rather than the WLAs. It must be noted that all the parameters used in these comparative methods are tuned empirically to obtain the optimal classification performance.

Train/test data distribution: A three-fold cross-validation is used to compare the classification performance of the proposed framework and the comparative methods. 33.33% of the vocalizations/target signals present in each fold (per class) are used for training while the remaining vocalizations are used for testing. Out of these 33.33% training vocalizations, 50% are used for building the GMM and rest of the 50% are used for obtaining the archetypal dictionaries and CLAs. The results presented here are averaged across three folds.

4.3. Results and Discussion

4.3.1. Classification performance

The classification performances of the proposed framework and other comparative methods on bird and frog datasets are illustrated in Fig. 4(a) and Fig. 4(b) respectively. Following can be inferred from the analysis of these two figures:

- The classification performance of the proposed framework (CLA) is either significantly better or comparable to the existing methods over both the datasets.

- The classification performance of the proposed CLA based framework is comparable to the DNN over both the datasets. CLA shows a relative improvement of 1.32% and 1.46% over DNN on the bird and frog datasets respectively.

- CLA significantly outperforms GMM, SVM with dynamic kernels, MSSA, GSSA and CCSE over both the datasets. CLA shows a relative improvement of 13.06%, 7.03%, 6.5%, 42.1%, 28.38% and 3.95% over GMM, IMK, PMK, MSSA, GSSA and CCSE respectively on the bird dataset. Similarly, a relative improvement of 14.12%, 8.89%, 8.05%, 35.4%, 25.8% and 3.8% is observed on the the frog dataset.

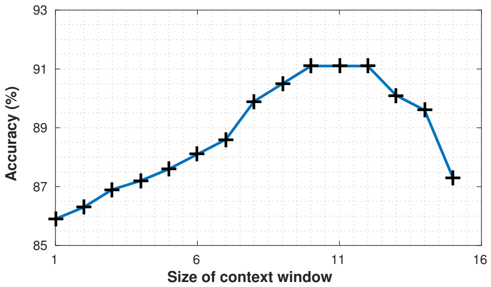

Fig. 5. Effect of context window size on the classification performance of the proposed framework on bird dataset.

- MSSA and GSSA are based on singular spectrum analysis which deals with the spectrum of the eigenvalues of the covariance matrix obtained from the input Hankel matrix of time-domain signal rather than the power spectrum of the input signal itself (as done in other methods). This nature of the singular spectrum analysis could be attributed to the low performances of both MSSA and GSSA.

- CLA outperforms its AA-SF variant by 4.61% and 4.08% on bird and frog dataset respectively. This result justifies the use of pseudo-topic modelling and intermediate symbolic representation (WLA) in the proposed framework. Applying AA on WLAs rather the MFCC super-frames leads to an increment in classification performance. This affirms the claim that the class discrimination is more evident in WLA space rather than the input MFCC super-frame space.

4.3.2. *Effect of context size (W) on classification*

Using a small value of W leads to MFCC super-frames having less context information, while using a very large value can create super-frames which are templates of the target signal. Hence, an optimal value of W must be chosen. Fig. 5 depicts the classification accuracy as the function of W. The analysis of this figure illustrates that embedding the temporal context helps in improving the classification. The maximum accuracy is obtained at $W = 10, 11$ and 12. However, since the lowest dimensionality of super-frames is obtained for $W = 10$, this can be considered as the optimal value of W. A very large value of W ($W \geq 13$) can affect the generative nature of the proposed framework leading to the deterioration in classification performance as evident in Fig. 5.

5. CONCLUSION

In this work, we propose a framework for bioacoustic classification which combines the statistical modelling properties of GMM and geometric modelling behaviour of AA. The framework maps the segments of bioacoustic signals to a weighted likelihood space using a GMM as the pseudo-acoustic topic model. AA is applied on these likelihood alignments to obtain the proposed CLAs. These CLAs are shown to be an effective feature representation for bioacoustic classification. Future work may include the utilisation of more discriminative intermediate representations for further improving the results.

6. REFERENCES

[1] T. S. Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International*, vol. 18, no. S1, pp. S163–S173, 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] D. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, pp. 827–832, 2015.

[4] Y. Chen, J. Mairal, and Z. Harchaoui, "Fast and robust archetypal analysis for representation learning," in *Proc. Comp. Vision Pattern Recog.*, 2014, pp. 1478–1485.

[5] V. Abrol, P. Sharma, A. Thakur, P. Rajan, A. D. Dileep, and A. K. Sao, "Archetypal analysis based sparse convex sequence kernel for bird activity detection," in *Proc. Eusipco*, 2017, pp. 4436–4440.

[6] A. Thakur, V. Abrol, P. Sharma, and P. Rajan, "Compressed convex spectral embedding for bird species classification," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, April, 2018.

[7] S. Ding, H. Zhao, Y. Zhang, X. Xu, and R. Nie, "Extreme learning machine: algorithm, theory and applications," *Artificial Intelligence Review*, vol. 44, no. 1, pp. 103–115, 2015.

[8] K. Qian, Z. Zhang, A. Baird, and B. Schuller, "Active learning for bird sound classification via a kernel-based extreme learning machine," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1796–1804, 2017.

[9] B. Gatto, J. Colonna, E. M. dos Santos, and E. F. Nakamura, "Mutual singular spectrum analysis for bioacoustics classification," in *Proc. MLSP*, Sept., 2017.

[10] L. S. Souza, B. Gatto, and K. Fukui, "Grassmann singular spectrum analysis for bioacoustics classification," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, April, 2018.

[11] D. Chakraborty, P. Mukker, P. Rajan, and A.D. Dileep, "Bird call identification using dynamic kernel based support vector machines and deep neural networks," in *Proc. Int. Conf. Mach. Learn. App.*, 2016, pp. 280–285.

[12] A. Thakur, R. Jyothi, P. Rajan, and A. D. Dileep, "Rapid bird activity detection using probabilistic sequence kernels," in *Proc. Eusipco*, Aug., 2017, pp. 1754–1758.

[13] E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio based bird species identification using deep learning techniques.," in *CLEF (Working Notes)*, 2016, pp. 547–559.

[14] R. Narasimhan, X. Z. Fern, and R. Raich, "Simultaneous segmentation and classification of bird song using CNN," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, 2017, pp. 146–150.

[15] B. P. Tóth and B. Czeba, "Convolutional neural networks for large-scale bird song classification in noisy environment.," in *CLEF (Working Notes)*, 2016, pp. 560–568.

[16] A. Thakur, V. Abrol, P. Sharma, and P. Rajan, "Rényi entropy based mutual information for semi-supervised bird vocalization segmentation," in *Proc. Mach. Learn. Sig. Process.*, 2017.

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

[18] "Macaulay library," http://www.macaulaylibrary.org/, Accessed: 2017-11-14.