This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



# Deep Archetypal Analysis Based Intermediate Matching Kernel For Bioacoustic Classification

Anshul Thakur and Padmanabhan Rajan

***Abstract*—We introduce a new classification framework that combines the characteristics of matrix factorization with the discriminative capabilities of kernel methods. Short-time analysis of audio signals having different durations result in sets of feature vectors having different cardinalities. Support vector machines handle such varying-length feature sets using dynamic kernels, such as the intermediate matching kernel (IMK). IMK works by utilising the so-called virtual vectors which select pairs of feature vectors to learn discrimination between classes. Existing formulations of IMK choose virtual vectors from the most information-bearing regions of classes, such as cluster means. This form of IMK completely ignores the feature vector pairs that lie around the class boundaries. To overcome this limitation, we propose an alternative formulation of IMK based on archetypal analysis (AA) and deep archetypal analysis (DAA). AA represents the data in terms of boundary elements, whereas DAA represents data in terms of both boundary and average elements. The proposed AA and DAA based intermediate matching kernel (AA/DAA-IMK) utilizes the elements generated from AA and DAA as the virtual feature vectors. Experimental evaluation on four different bioacoustic datasets show that the introduction of AA and DAA into the IMK framework leads to a noticeable improvement in classification accuracy.**

***Index Terms*—deep archetypal analysis based IMK, bioacoustic classification, bird species classification, kernel methods**

## I. Introduction

For the past many years, there is an upsurge in efforts for conserving various avian and amphibian species [1], [2]. These conservation tasks often include monitoring and surveying different species in their natural habitats. Automated acoustic monitoring provides a convenient and passive way to monitor target species effectively [3]. It is cost-effective and requires less human intervention compared to manual field studies, which are often tedious, expensive and require experienced ecologists [4], [5]. Bioacoustic signal classification is an important module in any automated acoustic monitoring system [6] such as bird species classification, bird activity detection [7], [8] and frog species classification. Most of these bioacoustic classification tasks suffer from the lack of labeled data. Due to this, there is a hindrance in using state-of-the-art data-intensive frameworks such as deep neural networks (DNN) and convolutional neural networks (CNN). Thus, there is a need for techniques which could provide effective bioacoustic classification under low training data conditions.

This being said, the literature does contain some bioacoustic studies that utilize convolutional neural networks [9], [10], [11], [12]. However, these deep learning frameworks require a sufficient amount of data to provide effective generalization and may not be effective for many bioacoustic classification tasks such as species identification and vocalization segmentation where labeled data is scarce. Some studies such as [13], [14] have overcome the limitation of labeled data by utilizing transfer learning. However, these techniques have only been tested on a small number of classes and their scalability to a large-scale bioacoustic classification task is yet unproven.

Classical methods such as sinusoidal modeling of bird syllables for species identification have successfully been used in a few early studies [15], [16]. Hidden Markov models (HMM) have been effectively used to model the temporal arrangements of syllables for birdsong [17] and species classification [18]. However, in field conditions, many birds and other animals vocalize at the same time. Hence, obtaining a completely unaltered sequence of syllables or phrases for temporal modelling can be difficult. Apart from syllables and birdsong modelling, many studies have shown the effectiveness of data-driven feature representations for bioacoustic classification tasks, such as bird and frog species identification. Based on spherical K-means, Stowell and Plumbley proposed an unsupervised feature learning method for large-scale bird species classification [19]. A random forest classifier was used to highlight the discriminative abilities of these unsupervised features. In [20] and [21], convex representations obtained from dictionary learning frameworks are used as the feature representations. These representations exhibit good class-specific characteristics and are shown to be effective for bird species classification. A major disadvantage of these methods is that if the correlation between class-specific dictionaries is large, the discriminative characteristics of convex representations are significantly affected.

Along with the aforementioned data-driven feature representations, kernel methods based classification frameworks have also been successfully exploited for bird activity detection [7] and species classification. In [22], Quin *et al.* utilized kernel-based extreme learning machines [23] for classifying bird vocalizations. Chakraborty *et al.* [24] successfully used support vector machines (SVM) powered with different dynamic kernels [25] including probabilistic sequence kernels (PSK) and intermediate matching kernel (IMK) [26] to classify calls of 26 bird species. In [8] and [27], the variants of PSK incorporated in an SVM framework are used for bird activity detection. Although the performances of both data-driven feature representations and kernel-based classification frameworks have been effective, there is still a room for improvement. In this paper, the authors explore this possibility by utilizing novel data-driven approaches to improve existing

Anshul Thakur and Padmanabhan Rajan are with School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, India. E-mail: anshul_thakur@students.iitmandi.ac.in, padman@iitmandi.ac.in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



dynamic kernels.

Dynamic kernels [28], [29] are of particular interest for bioacoustic (or acoustic) classification as they empower classification frameworks such as SVMs to handle feature sets with varying cardinalities. These varying cardinalities arise from the short-term analysis of audio signals to extract representations such as mel spectrograms or mel frequency cepstral coefficients (MFCC). After short-time analysis, the audio signal is represented as a set of feature vectors. The cardinality of this set depends on the duration of the signal. Thus, audio recordings of different durations are represented by feature sets of different cardinalities. Dynamic kernels are used to determine the similarity between two feature sets irrespective of their cardinality. The similarity between the feature sets is used by the SVM to learn the discrimination between classes. Dynamic kernels are of two types: explicit mapping kernels and matching-based kernels. Explicit mapping kernels utilize generative models such as Gaussian mixture models to map feature sets of different cardinalities to a fixed-length representation. These fixed length representations can then be processed by the SVM to learn the discrimination between classes. Matching-based kernels bypass the problem of varying cardinalities by computing similarities between pairs of feature vectors from different feature sets. The individual similarities are then accumulated to calculate the overall similarity between the feature sets. The feature set similarities between training examples can be used to construct a kernel gram matrix that is utilized by the SVM for learning the discrimination between classes. More details about dynamic kernels and their working can be found in [25].

Intermediate matching kernel (IMK) [26] is a type of matching-based dynamic kernel that has been successfully utilized for various audio classification tasks such as speaker identification [30], speech recognition[29], [31] and birdcall classification [24]. IMK is characterized by the utilization of a set of *virtual vectors* $\mathcal{V}$ to calculate the kernel gram matrix between feature sets of audio examples. Each virtual vector is used to select a pair of feature vectors from the feature sets (one from each set). These selected feature vectors are referred as *local feature vectors* (LV) and are used to estimate similarity between the feature sets. Let $\mathcal{X}_m = \{\mathbf{x}_m^1, \mathbf{x}_m^2, \ldots, \mathbf{x}_m^y\}$ and $\mathcal{X}_n = \{\mathbf{x}_n^1, \mathbf{x}_n^2, \ldots, \mathbf{x}_n^z\}$ be two feature sets from two classes, having cardinalities $y$ and $z$ respectively. To compute IMK between $\mathcal{X}_m$ and $\mathcal{X}_n$, first $\mathcal{V} = \{\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^d\}$ is used to chose pairs of feature vectors from $\mathcal{X}_m$ and $\mathcal{X}_n$ as:

$$\mathbf{x}_m^{j*} = \underset{\mathbf{x}_m^k}{\operatorname{argmin}} \|\mathbf{x}_m^k - \mathbf{v}^j\|_2, \text{where } k = 1 \ldots y \tag{1}$$

$$\mathbf{x}_n^{j*} = \underset{\mathbf{x}_n^k}{\operatorname{argmin}} \|\mathbf{x}_n^k - \mathbf{v}^j\|_2, \text{where } k = 1 \ldots z \tag{2}$$

where $\mathbf{v}^j \in \mathcal{V}$ is the $j$th virtual vector. $\mathbf{x}_m^{j*}$ and $\mathbf{x}_n^{j*}$ are a pair of the local feature vectors selected from $\mathcal{X}_m$ and $\mathcal{X}_n$ using $\mathbf{v}^j$. Thus, $d$ pairs of feature vectors (one for each virtual vector) are chosen. A base kernel, generally a Gaussian kernel [32], is calculated between each chosen pair as:

$$K_{base}(\mathbf{x}_m^{j*}, \mathbf{x}_n^{j*}) = \exp(-\delta \|\mathbf{x}_m^{j*} - \mathbf{x}_n^{j*}\|_2^2), \tag{3}$$

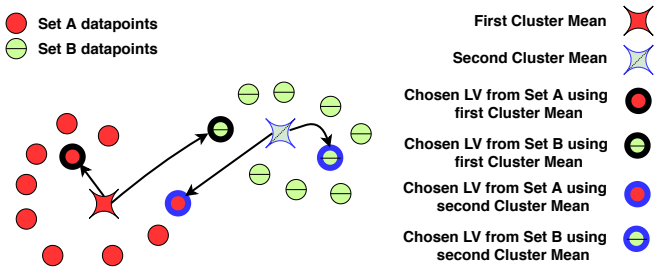



Fig. 1: Illustration of the process of selecting pairs of local vectors (LV) from two sets (Set A and Set B), sampled from two different classes, using cluster means. A pair of the local vectors (one from each set) which are closest to a cluster centre are chosen for calculating a base kernel.

where $\delta$ denotes the width parameter of Gaussian kernel and is fine-tuned empirically. These base kernels are aggregated to obtain the IMK between $\mathcal{X}_m$ and $\mathcal{X}_n$ as:

$$K_{imk}(\mathcal{X}_m, \mathcal{X}_n) = \sum_{j=1}^{d} K_{base}(\mathbf{x}_m^{j*}, \mathbf{x}_n^{j*}) \tag{4}$$

Since the base Gaussian kernels are positive definite, IMK is bound to be positive definite [26].

The analysis of the aforementioned formulation of IMK shows that the choice of virtual vectors can have a significant effect on classification performance. Earlier studies on IMK have either used clustering [26] or Gaussian mixture modelling [29] to obtain virtual vectors. In [26], cluster centres, obtained by clustering the training data, are used as the virtual vectors. A pair of local feature vectors (one from each feature set) which exhibit minimum distance from a cluster centre are selected for learning a base kernel. This process of selecting pairs of the feature vectors using cluster means is illustrated in Fig. 1. In [29], [31], the authors proposed to use Gaussian mixture models (GMM) to choose these local feature vector pairs. For each GMM component, a feature vector from each set, having maximum affinity to the respective GMM component is chosen. The GMM-based IMK has provided better classification than the cluster-centre based approach [29]. This can be attributed to the fact that along with mean vectors, GMM-based IMK also utilizes covariance and weight of each component in choosing the feature vectors. These earlier studies utilize the most informative regions of the data (cluster centres or GMM components) as the virtual vectors. Hence, pairs of feature vectors chosen for calculating the base kernels are easily classifiable. Whereas, some of the most confusing (hard to classify) feature vector pairs are ignored. These pairs lie around the class boundaries and provide important cues about the separation between classes.

Archetypal analysis (AA) [33] is a matrix decomposition method that factorizes an input matrix into a dictionary of archetypes and convex-sparse representations. Archetypes lie on the convex hull or boundary of the data spread and hence, model the extremal behaviour of the data. In this work, a new classification framework is proposed by combining the data modelling capabilities of AA-based matrix factorization with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



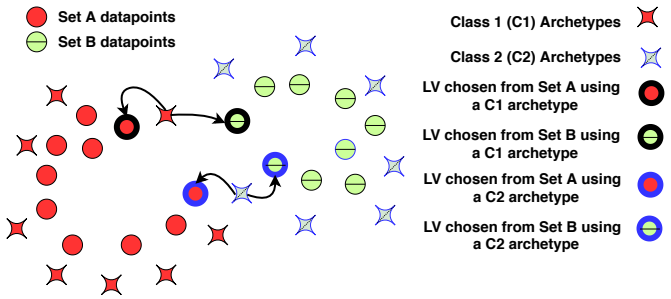



Fig. 2: Illustration of the process of selecting pairs of local vectors (LV) from two sets (Set A and Set B), from two different classes, using class-specific archetypes in AA-IMK framework. For illustration, only one class-specific archetype is shown to be used for the selection of local vectors. In the proposed method, all archetypes are used for learning AA-IMK.

the implicit discriminative abilities of kernel methods. A new formulation of IMK that targets the limitations existing in both the classical IMK and standard AA based dictionary learning frameworks is proposed. The inclusion of AA improves classical IMK by helping in choosing pairs of feature vectors that may lie around the class boundaries and are difficult to classify. The incorporation of such feature vector pairs in the kernel computation helps in learning a better classifier. On the other hand, the combination of kernel methods with an AA based dictionary learning framework helps in increasing the inter-class discrimination. AA are data-modeling methods and no external efforts are made to increase the inter-class separation as in the kernel methods. Thus, if two classes are overlapping, the dictionaries learnt using AA exhibit high correlation, leading to a less discriminative convex-sparse representation [20]. Taking advantage of kernel methods, AA based IMK learns the class-separation boundaries in an implicit higher dimensional space where the two-overlapping classes may be linearly separable.

In the proposed AA based formulation of IMK (AA-IMK), class-specific archetypes learned from the training data are utilized as the virtual vectors. Since archetypes lie on the convex hull or boundary of a class, they help in choosing pairs of local feature vectors that lie on or near the class boundaries. This behaviour of AA-IMK is illustrated in Fig. 2. As discussed earlier, these pairs of local vectors are hard to classify and including them in the training process helps in learning a better classifier. Though AA shows good extremal modeling capabilities, it lacks the ability to model the average or prototypical behaviour of the data. To overcome this problem, deep archetypal analysis (DAA) [21], [34] was proposed in our recent studies. In the DAA framework, the convex-sparse representation matrix obtained from AA is further factorized. This chain of factorization is continued up to a desired depth. In [21], it has been observed that atoms of the deeper dictionaries can model the extremal as well as the prototypical behaviour, i.e. some atoms can lie on or near the boundary while others can exist inside the boundary (More details about AA and DAA are in Section II). These deeper dictionary atoms can help in choosing local feature vectors from the most informative regions (similar to classical IMK) as well as from the class boundaries (similar to AA). Thus, utilization of DAA in the IMK framework can help in combining the properties of both IMK and AA-IMK.

The main contributions of this paper are as follows:

- A new classification framework that embeds the properties of matrix factorization in a kernel method framework.
- Based on AA and DAA, two alternative formulations of the classical IMK are proposed for bioacoustic classification. The data modeling capabilities of AA and DAA are exploited to choose pairs of feature vectors for learning the base kernels.
- Two variants to choose local feature vectors in AA/DAA-IMK framework are proposed. In first variant, nearest neighbour approach is followed to select local feature vector pairs. In the second approach, simplex decomposition [33] on class-specific archetypal dictionaries is utilized to choose the local feature vectors (see Section II).

The rest of this paper is organized as follows. In Section II, the proposed AA/DAA-IMK are described in detail. Experimental setup is described in Section III. Results and Discussion are in Section IV. Section V concludes this paper.

## II. Proposed Framework

In this section, first, we describe archetypal analysis (AA) and deep archetypal analysis (DAA) in detail. Then, we describe the proposed AA/DAA based intermediate matching kernel (AA/DAA-IMK).

### A. Archetypal and deep archetypal analysis

Archetypal analysis (AA) [35] decomposes a matrix containing $l$ feature vectors of $K$ dimensions, $\mathbf{X} \in \mathbb{R}^{K\times l}$, as $\mathbf{X} \approx \mathbf{DA}$. $\mathbf{D} \in \mathbb{R}^{K\times d}$ contains $d$ archetypes, which lie on the convex hull of the data and are forced to be convex combinations of the input features i.e., $\mathbf{D} = \mathbf{XB}$ where $\mathbf{B} \in \mathbb{R}^{l\times d}$ and $\mathbf{A} \in \mathbb{R}^{d\times l}$ are convex representation matrices. The archetypal dictionary $\mathbf{D}$ can be obtained by solving the following optimization problem [33]:

$$\underset{\substack{\mathbf{B},\mathbf{A}\\ \mathbf{b}_j\in\Delta_l,\mathbf{a}_i\in\Delta_d}}{\operatorname{argmin}} \quad \|\mathbf{X}-\mathbf{XBA}\|_F^2,$$

$$\Delta_l \triangleq [\mathbf{b}_j \succeq 0, \|\mathbf{b}_j\|_1 = 1], \Delta_d \triangleq [\mathbf{a}_i \succeq 0, \|\mathbf{a}_i\|_1 = 1]. \tag{5}$$

Here $\mathbf{a}_i$ and $\mathbf{b}_j$ are columns of $\mathbf{A} \in \mathbb{R}^{d\times l}$ and $\mathbf{B} \in \mathbb{R}^{l\times d}$, respectively. The objective defined in Equation 5 is non-convex; however, it is convex in terms of $\mathbf{A}$ if $\mathbf{B}$ is fixed or vice-versa. Thus, the optimization objective in Equation 5 can be easily solved using the block-coordinate descent scheme. More details about the implementation of AA can be found in [33].

Compared to conventional matrix factorization, AA can be considered as a deep model with three factors, the first being the data itself ($\mathbf{X} \approx \mathbf{XBA}$). Inspired by this observation, AA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



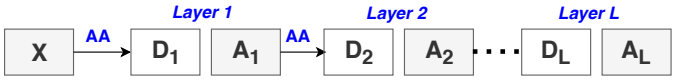



Fig. 3: Illustration of the deep archetypal analysis (DAA) based matrix factorization framework. Here a matrix $\mathbf{X}$ is factorized into $L+1$ factors as: $\mathbf{X} \approx \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3 \dots \mathbf{D}_L\mathbf{A}_L$.

is used in a deep matrix factorization framework to uncover the hidden attributes of the data by further decomposing the convex-sparse representations [21], [34]. This deep AA (DAA) framework is a layered architecture which begins with factorizing the input matrix $\mathbf{X}$ into an archetypal dictionary $\mathbf{D}_1$ and convex-sparse representation matrix $\mathbf{A}_1$ using AA as discussed earlier. $\mathbf{A}_1$ is passed to the second layer and is again factorized using AA to obtain dictionary $\mathbf{D}_2$ and the convex-sparse representations $\mathbf{A}_2$. At this layer, the overall data decomposition is: $\mathbf{X} \approx \mathbf{D}_1\mathbf{A}_1 \approx \mathbf{D}_1\mathbf{D}_2\mathbf{A}_2 = \mathbf{D}_{L2}\mathbf{A}_2$, here $\mathbf{D}_{L2}$ is the DAA dictionary obtained at the second layer of DAA framework. This process is followed till the desired levels of factorization. Fig.3 illustrates the factorizations at each layer of DAA. Thus, in this deep variant of AA, $\mathbf{X}$ is factorized into $L+1$ factors as:

$$\mathbf{X} \approx \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3 \dots \mathbf{D}_L\mathbf{A}_L. \tag{6}$$

At each layer of DAA framework, the factorizations can be unfolded as:

$$\begin{aligned} \mathbf{X} \approx \mathbf{D}_1\mathbf{A}_1 &= \mathbf{X}\mathbf{B}_1\mathbf{A}_1 \\ \mathbf{A}_1 &\approx \mathbf{A}_1\mathbf{B}_2\mathbf{A}_2 \\ \mathbf{A}_2 &\approx \mathbf{A}_2\mathbf{B}_3\mathbf{A}_3 \\ &\vdots \\ \mathbf{A}_{L-1} &\approx \mathbf{A}_{L-1}\mathbf{B}_L\mathbf{A}_L \\ \mathbf{A}_L &\approx \mathbf{A}_L\mathbf{B}_{L+1}\mathbf{A}_{L+1}. \end{aligned} \tag{7}$$

Note that the factorizations at each layer of the DAA framework are obtained in a greedy fashion, i.e. the optimization objective solved at each layer is independent of decompositions done at other layers.

*Analyzing geometric properties of AA/DAA atoms:* The geometric properties of archetypes are well studied [35]. As discussed earlier, archetypes lie on the convex hull and model the geometry or extremal of the data. Fig. 4(A), (D) and (G) exhibit the geometric modeling capabilities of archetypes on three different datasets. The 2-dimensional data points in Fig. 4(A) and Fig. 4(D) are randomly generated from a uniform distribution and a Gaussian distribution respectively, whereas data points in Fig. 4(G) are two-dimensional t-SNE [36] representations of 39-dimensional MFCC feature vectors obtained from the song phrases of Cassin's vireo, a North American song bird.

Unlike archetypes, the DAA dictionary ($\mathbf{D}_{Li}$, where $i > 1$) atoms are observed to lie on or near the boundary as well as inside the data spread. Fig. 4 illustrates this behaviour of the DAA atoms. Fig. 4 (B), (E) and (H) show the nature of the dictionary atoms obtained at the second layer of DAA framework. Similarly, Fig. 4 (C), (F) and (I) illustrate the geometric properties of the third layer DAA dictionary ($\mathbf{D}_{L3}$) atoms. To analyze this behaviour, we consider first two layers of the DAA framework. At first layer, a matrix $\mathbf{X}$ is factorized into an archetypal dictionary $\mathbf{D}_1$ and convex representations $\mathbf{A}_1$. At the second layer, $\mathbf{A}_1$ is factorized to $\mathbf{D}_2$ and $\mathbf{A}_2$. At this point, the input matrix $\mathbf{X}$ can be represented as: $\mathbf{X} = \mathbf{D}_1\mathbf{D}_2\mathbf{A}_2 = \mathbf{D}_{L2}\mathbf{A}_2$. By the definition of AA, it is known that $\mathbf{D}_2$ is the convex combination of columns of $\mathbf{A}_1$ and $\mathbf{A}_1$ is the convex combination of columns of $\mathbf{D}_2$. Since $\mathbf{A}_1$ is already a convex representation matrix, the convex combinations of its columns also exhibit the properties of convex representations. As a result, each column $\mathbf{d}$ of $\mathbf{D}_2$ is: $\mathbf{d} \succeq 0, \|\mathbf{d}\|_1 = 1$. Thus, the second layer DAA dictionary, $\mathbf{D}_{L2} = \mathbf{D}_1\mathbf{D}_2$, can be seen as the convex combination of archetypes ($\mathbf{D}_1$) of $\mathbf{X}$, where $\mathbf{D}_2$ is a convex representation matrix. Hence, the atoms of $\mathbf{D}_{L2}$ can lie anywhere in the space spanned by convex combination of archetypes. A DAA dictionary atom lies near the boundary if the contribution of an archetype in defining this atom is significantly greater than the contribution of other archetypes. On the other hand, DAA atoms lie inside the data spread (away from boundary) when multiple archetypes have significant contributions in the definition of these atoms. This allows the DAA atoms to tessellate the entire data spread (as shown in Fig. 4) and equips the DAA dictionaries with better data modeling capabilities. The atoms lying around or near the boundary model the extremal behaviour while the atoms lying inside the boundary exhibit average or prototypical behaviour of the data.

### *B. Archetypal/deep archetypal analysis (AA/DAA) based IMK*

Characteristically, kernel methods can only be used for binary classification and IMK is no exception. However, these methods can be extended to multiple classes using the *one-vs-rest* approach where a binary classifier such as SVM is trained to discriminate examples of one class from the other remaining classes. Thus, all the other classes are regarded as one non-target class and multiple binary classifiers are trained to solve a given multi-class classification problem.

In this subsection, we describe the proposed AA/DAA-IMK for binary classification. The proposed formulations can be easily extended to the multi-class setting using *one-vs-rest* approach. To calculate the proposed AA/DAA-IMK between examples of two possible classes, the class specific DAA dictionaries i.e. $\mathbf{D}^1_{Li}$ and $\mathbf{D}^2_{Li}$ are obtained using the DAA framework as explained earlier. Here $\mathbf{D}^C_{Li}$ represents the $C$th class dictionary obtained from the $i$th layer of DAA framework. Each atom of these class-specific dictionaries are used as a virtual vector for choosing a pair of the local feature vectors. The proposed formulations utilize two different methods to select pairs of local feature vectors from $\mathcal{X}_m = \{\mathbf{x}^1_m, \mathbf{x}^2_m, \dots, \mathbf{x}^y_m\}$ and $\mathcal{X}_n = \{\mathbf{x}^1_n, \mathbf{x}^2_n, \dots, \mathbf{x}^z_n\}$:

*1. Nearest neighbour approach:* The $j$th atom, $\mathbf{d}^C_j$, of $\mathbf{D}^C_{Li}$ is used to choose a pair of the local feature vectors (one from each $\mathcal{X}_m$ and $\mathcal{X}_n$) as:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



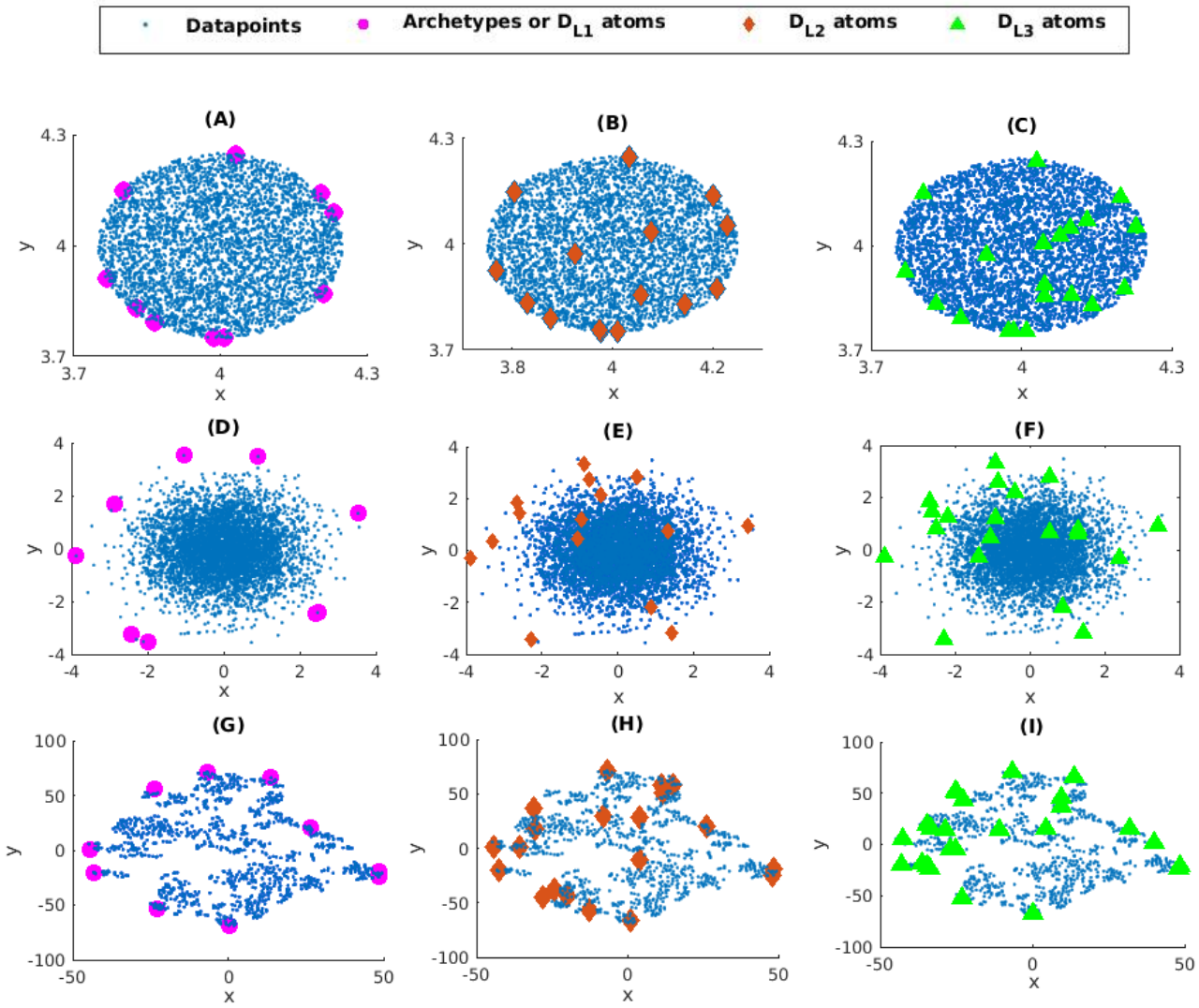



Fig. 4: Illustration of data modeling abilities of deep archetypal dictionaries ($\mathbf{D}_{L1}$, $\mathbf{D}_{L2}$ and $\mathbf{D}_{L3}$ i.e. first, second and third level dictionaries respectively) on two randomly generated data (in first and second row) and on two dimensional t-SNE representation of MFCC vectors obtained from Cassin's vireo song phrases (in last row). The number of dictionary atoms are the same in each layer. Note that some of atoms of the lower layer dictionaries are falling in same position.

$$\mathbf{x}_m^{j*} = \underset{\mathbf{x}_m^k}{\operatorname{argmin}} \|\mathbf{x}_m^k - \mathbf{d}_j^C\|_2 \tag{8}$$

$$\mathbf{x}_n^{j*} = \underset{\mathbf{x}_n^k}{\operatorname{argmin}} \|\mathbf{x}_n^k - \mathbf{d}_j^C\|_2, \tag{9}$$

where $\mathbf{x}_m^k \in \mathcal{X}_m$, $\mathbf{x}_n^k \in \mathcal{X}_n$ and $\mathbf{D}_{Li}^C$ is the $C$th class dictionary ($C = 1, 2$) obtained from the $i$th layer. Using equations 8 and 9, $2d$ pairs of local feature vectors (assuming $d$ atoms in each class-specific dictionary) are selected for computing IMK.

*2. Simplex decomposition approach:* In this method, the simplex projections of feature vectors on the class-specific DAA dictionaries are used to choose the local feature vector pairs for calculating the IMK. This approach is based on the contribution of the dictionary atoms in defining a feature vector rather than the Euclidean distance between dictionary atoms and the feature vectors. This contribution is defined by the magnitude of the coefficient corresponding to a dictionary atom in the convex-sparse representation of the given vector. A vector $\mathbf{x}_m^k$ is projected on a simplex whose vertices correspond with atoms of the class-specific DAA dictionary $\mathbf{D}_{Li}^C$ (having $d$ atoms) to obtain the convex-sparse representation:

$$\begin{aligned} \mathbf{r}_m^{*k} &= \underset{\substack{\mathbf{r}_m^k \\ \mathbf{r}_m^k \in \Delta_d}}{\operatorname{argmin}} \|\mathbf{x}_m^k - \mathbf{D}_{Li}^C \mathbf{r}_m^k\|_2^2 \\ \Delta_d &\triangleq [\mathbf{r}_m^k \succeq 0, \|\mathbf{r}_m^k\|_1 = 1], \end{aligned} \tag{10}$$

where $\mathbf{r}_m^k \in \mathbb{R}^d$ is the convex-sparse representation that defines the contribution of each atom of $\mathbf{D}_{Li}^C$ in representing $\mathbf{x}_m^k$. Using Equation 10, the convex-sparse representations for all $y$ and $z$ vectors in $\mathcal{X}_m$ and $\mathcal{X}_n$ respectively are calculated and are pooled in matrices $\mathbf{C}_m \in \mathbb{R}^{y \times d}$ and $\mathbf{C}_n \in \mathbb{R}^{z \times d}$. Here $\mathbf{C}_m(o, j)$ represents the coefficient corresponding to the $\mathbf{d}_j^C$ ($j$th atom) in the convex-sparse representation of the $o$th

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



feature vector of $\mathcal{X}_m$. For an atom $\mathbf{d}_j^C$, a pair of local feature vectors is chosen as:

$$ind_m^j = \underset{o}{\operatorname{argmax}}\, \mathbf{C}_m(o,j) \tag{11}$$

$$ind_n^j = \underset{o}{\operatorname{argmax}}\, \mathbf{C}_n(o,j). \tag{12}$$

Here $ind_m^j$ and $ind_n^j$ represent indices of the selected local feature vectors ($\mathbf{x}_m^{j*}$ and $\mathbf{x}_n^{j*}$) chosen from $\mathcal{X}_m$ and $\mathcal{X}_n$ using $j$th atom ($\mathbf{d}_j^C$) of $\mathbf{D}_{Li}^C$. In this way, $2d$ pairs of the local feature vectors ($d$ dictionary atoms per class-specific dictionary) are chosen for calculating the proposed AA/DAA-IMK.

Once these $2d$ pairs of local feature vectors are chosen (using either simplex decomposition or nearest neighbour approach), a Gaussian base kernel is computed between each pair using Equation 3. These $2d$ base kernels are aggregated to obtain the proposed AA/DAA-IMK between $\mathcal{X}_m$ and $\mathcal{X}_n$:

$$K_{aa/daa-imk}(\mathcal{X}_m, \mathcal{X}_n) = \sum_{j=1}^{2d} K_{base}(\mathbf{x}_m^{j*}, \mathbf{x}_n^{j*}) \tag{13}$$

Note that when dictionaries obtained from the first layer of DAA framework are used to calculate IMK, the proposed kernel is dubbed as AA-IMK. Similarly when deeper dictionaries are used, the proposed kernel is referred as DAA-IMK.

These kernel gram matrices can be incorporated in support vector machine (SVM) framework to classify the feature sets of varying cardinalities. During training, the proposed AA/DAA-IMK (Equation 13) is calculated between each pair of training examples and these kernel values are stored in a kernel-gram or similarity matrix. This matrix is passed to the SVM to learn the discrimination between classes. The overall procedure to train SVM with AA/DAA-IMK is summarized in Algorithm 1. To test an example, a test kernel-gram matrix is calculated between this test example and each training example. Then, the test kernel-gram matrix is passed to the trained SVM to obtain the prediction. Algorithm 2 contains the pseudo-code explaining the testing procedure.

**Algorithm 1:** Training SVM with AA/DAA-IMK for binary classification

**input :** $\mathcal{X} = \{\mathcal{X}_1\mathcal{X}_2....\mathcal{X}_M\}$: Examples of $i$th class
$\mathcal{Y} = \{\mathcal{Y}_1\mathcal{Y}_2....\mathcal{Y}_N\}$: Examples of $j$th class
$\mathbf{D}_L^i$: $L$th layer DAA dictionary of $i$th class
$\mathbf{D}_L^j$: $L$th layer DAA dictionary of $j$th class

**output:** SVM: Trained support vector machine

$K = [\,]$ // Empty matrix of size $M \times N$ to store kernel values
$Labels = [repeat(1, M); repeat(-1, N)]$ // Creating M "1" labels for $i$th class and N "-1" labels for $j$th class.
**for** $x \leftarrow 1$ **to** $M$ **do**
  **for** $y \leftarrow 1$ **to** $N$ **do**
    $\mathcal{I}$,$\mathcal{J}$=get_Local_Vectors($\mathcal{X}[x]$, $\mathcal{Y}[y]$, $\mathbf{D}_L^i$,$\mathbf{D}_L^j$) // Selecting 2d local vector pairs w.r.t. $\mathbf{D}_L^i$ and $\mathbf{D}_L^i$ either by nearest neighbours (using Equations 8-9) or simplex decomposition (using Equations 10-12)
    $kernel = 0$
    **for** $z \leftarrow 1$ **to** $2d$ **do**
      $base$=get_Base_Kernel($\mathcal{I}[z]$,$\mathcal{J}[z]$) // computing base kernel using Equation 3
      $kernel = kernel + base$ // Aggregating base kernels (Equation 13)
    **end**
    $K[x, y] = kernel$ // Kernel-gram matrix
  **end**
**end**
SVM=train_SVM($K, Labels$) // Training SVM

## III. Experimental Setup

In this section, the experiments designed to evaluate the classification performance of the proposed AA/DAA-IMK are described. The datasets, train-test data distribution, comparative methods and parameter settings used for experimentation are also discussed here.

### A. Experiments and datasets

The classification performances of AA/DAA-IMK are evaluated on four different tasks: bird species classification, frog species classification, bird activity detection and birdsong phrase classification. For bird species classification, a collection of audio recordings of 50 different species, obtained from three different sources, is used here. The recordings of 17 bird species were obtained from the Macaulay Library[1] and were provided on an academic license. The recordings of 7 bird species were downloaded from bird database maintained by Art & Science Centre, UCLA[2]. The recordings of 26 bird species were obtained from the Great Himalayan national park (GHNP) dataset[3] and were also used in [24]. The information about these 50 species along with the total number of recordings and vocalizations per species is available at https://goo.gl/z6UEQa. The publicly available Anuran dataset[4] is used for the frog species classification task. This dataset contains audio recordings of 10 different frog species, which are 16-bit mono and are sampled at 44.1 kHz.

For bird activity detection task, the publicly available development dataset provided for *BAD Challenge 2017*[5] is used here. This development dataset is composed of two datasets: *Freefield* and *Warblr*. A total of 16000 mono recordings (8000 birds and 8000 non-birds) having a sampling rate of 44.1 kHz are available in this dataset.

For song phrase classification, the thirteen Cassin's vireo audio recordings available at http://taylor0.biology.ucla.edu/

[1] http://www.macaulaylibrary.org

[2] http://artsci.ucla.edu/birds/database.html

[3] https://tinyurl.com/y9robcdy

[4] http://goo.gl/FFBzbb

[5] http://machine-listening.eecs.qmul.ac.uk/bird-audio-detection-challenge/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



**Algorithm 2:** Prediction with AA/DAA-IMK in SVM framework

**input :** $\mathcal{T} = \{\mathcal{T}_1\mathcal{T}_2....\mathcal{T}_t\}$: Test Examples
$\mathcal{Z} = \{\mathcal{X}_1\mathcal{X}_2....\mathcal{X}_M\mathcal{Y}_1....\mathcal{Y}_N\}$: Training Examples
SVM: Trained SVM
$\mathbf{D}_L^i$: $L$th layer DAA dictionary of $i$th class
$\mathbf{D}_L^j$: $L$th layer DAA dictionary of $j$th class

**output:** $\mathcal{P}$ : Predictions

$K = [\,]$ // Empty matrix of size $t \times MN$ to store test kernel gram matrix
**for** $x \leftarrow 1$ **to** $t$ **do**
  **for** $y \leftarrow 1$ **to** $MN$ **do**
    $\mathcal{I}$,$\mathcal{J}$=get_Local_Vectors($\mathcal{T}[x]$, $\mathcal{Z}[y]$, $\mathbf{D}_L^i$,$\mathbf{D}_L^j$)
      // Selecting 2d local vector pairs w.r.t. $\mathbf{D}_L^i$ and $\mathbf{D}_L^i$ either by nearest neighbours or (using Equations 8-9) or simplex decomposition (using Equations 10-12)
    $kernel = 0$
    **for** $z \leftarrow 1$ **to** $2d$ **do**
      $base$=get_Base_Kernel($\mathcal{I}[z]$,$\mathcal{J}[z]$)
        // computing base kernel using Equation 3
      $kernel = kernel + base$ // Aggregating base kernels (Equation 13)
    **end**
    $K[x, y] = kernel$ // Kernel-gram matrix
  **end**
**end**
$\mathcal{P}$=predict(SVM,$K$) // obtain predictions

are used. The recordings are segmented using the provided labels to obtain song phrases. The ten song phrase classes having maximum number of examples are used for the experimentation. These segmented song phrases are hosted at Figshare[6].

The classification accuracy is used as a performance metric for species and song phrase classification tasks, and, area under ROC curve (AUC) is used for bird activity detection.

### B. Train-test data distribution

For bird and frog species classification tasks, 10% of vocalizations from each class are used for validation. These vocalizations are not used for training or testing. The remaining 90% of vocalizations are used for three-fold cross-validation. In each fold, 33% of vocalizations are used for training while remaining vocalizations are used for testing. For bird species classification, in each fold, approximately 2970 vocalizations are used for training while approximately 5920 vocalizations are used for testing. For bird activity detection task, 50% of recordings (i.e. approx. 8000) are used for training, 10% for validation (i.e. approx. 1600) while the remaining (i.e. approx. 6400) are used for testing. For song phrase classification, five phrases of each class are used for training, five are used for validation while remaining phrases are used for testing.

[6]https://figshare.com/s/cfca142fcedd3f206b8b

### C. Comparative Methods

For the bird and frog species classification tasks, the performance of the proposed kernels is compared with six other methods. These include Gaussian mixture models (GMM), class-specific intermediate matching kernel (IMK) proposed in [31], probabilistic sequence kernels (PSK), a 3-layered fully-connected neural network proposed in [24] for bird species classification, an archetypal analysis based dictionary learning framework (CCSE) [20], deep convex representations (DCR) with random forest classifier [21] and spherical K-means based framework (SKM) proposed in [19].

For the task of bird activity detection, the performance of the proposed kernels is compared with some of the highest performing methods of *BAD Challenge 2017*. These include a CNN framework proposed in [12], a CNN-RNN hybrid network [11], a masked non-negative factorization framework [37], a probabilistic sequence kernel based method [8] and AA based convex sparse sequence kernel (AA-CSK) [27].

For the song phrase classification task, the performance of AA/DAA-IMK is compared against dynamic time warping (DTW), sparse representation based classifier (SR) and a two-pass framework fusing DTW and SR (DTW-SR-2Pass) [38]. In DTW-SR-2Pass, if there is a disagreement between DTW and SR classification decisions obtained in pass 1, then separate SR classifier is used to break this stand-off in pass 2. This separate SR classifier is a binary classifier trained on the examples of two classes predicted by DTW and SR during the first pass. The details about these methods can be found in [38]. Table I tabulates all the comparative methods along with their abbreviations used in this study.

### D. Parameter Settings

All the parameters used in the proposed frameworks and the comparative methods are fine-tuned on the respective validation datasets. For species classification, the bird and frog vocalizations are segmented from audio recordings using the semi-supervised method proposed in [39]. Only these segmented vocalizations are passed to the proposed framework and other comparative methods for both training and testing purposes. A feature representation derived from mel-frequency cepstral coefficients (MFCC) with delta and acceleration coefficients (39-dimensional) is used in the proposed framework. The frequency-temporal structures 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 $W \times 39$-dimensional feature representation. $W = 10$ is used here to obtain a 390-dimensional representation. This value of $W$ is determined using the validation dataset. To obtain MFCC, a frame length of 20 ms with 10 ms overlap is used in this work. The same feature representation is used in all the comparative studies considered for bird and frog species identification. During initial experimentation, the authors also tried linear frequency

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



TABLE I: Comparative methods used for the performance evaluation of the proposed AA/DAA-IMK for species classification, bird activity detection (BAD) and song phrase classification tasks.

| Method | Abbreviation | Nature | Task |
|---|---|---|---|
| SVM with Intermediate Matching Kernel [31] | IMK | Matching Dynamic Kernel | Species Classification |
| SVM with Probabilistic Sequence Kernel [8] | PSK | Mapping Dynamic Kernel | Species Classification |
| Compressed Convex Spectral Embeddings [20] | CCSE | Dictionary Learning | Species Classification |
| Deep Convex Representations [21] | DCR | Deep Dictionary Learning | Species Classification |
| Deep Neural Network [24] | NN | Multi-layer Perceptron | Species Classification |
| Spherical K-means with Random Forest [19] | SKM | Feature Learning | Species Classification |
| Masked Non-negative Matrix Factorization [37] | Masked-NMF | Non-negative Matrix Factorization | BAD |
| Convolutional-Recurrent Neural Network [11] | RCNN | Deep Learning | BAD |
| Convolutional Neural Network [12] | Bulbul | Deep Learning | BAD |
| Archetypal Analysis Based Convex Sequence Kernel [27] | AA-CSK | Dictionary Learning + Mapping Dynamic Kernel | BAD |
| Dynamic Time Warping | DTW | Sequence Matching | Phrase Classification |
| Sparse Representation based Classification [38] | SR | Exemplars based Dictionary Learning | Phrase Classification |
| DTW Followed by SR [38] | DTW-SR-2Pass | Combination of DTW and SR | Phrase Classification |
| AA based Intermediate Matching Kernel | AA-IMK | Dictionary Learning + Matching Dynamic Kernel | all |
| Deep AA based Intermediate Matching Kernel | DAA-IMK | Deep Dictionary Learning + Matching Dynamic Kernel | all |

cepstral coefficients (LFCC) as a feature representation. However, no significant difference was observed in classification performance and hence, the authors continued with MFCC features.

For the bird activity detection (BAD) experiments, 39-dimensional MFCC vectors are used as feature representation (no context embedding is used). A frame size of 20 ms with no overlap is used for feature extraction. These features are normalized and warped to have a normal distribution as explained in [8]. Since the comparative methods for BAD also include CNN based frameworks, using the same feature representation for all comparative methods is not possible. Hence, for this experiment, the feature representations used in the respective studies are used here.

For song phrase classification task, MFCC with temporal context of 10 frames is used as a feature representation in AA/DAA-IMK. DTW is applied on the mel-spectrogram representation for phrase matching. Whereas, for SR classifier, each song phrase is represented by a 128-dimensional compressed feature representation. This representation is obtained from the time warped mel-spectrograms by concatenating all frames and compressing the concatenated representation using PCA [38]. Across all methods, a frame length of 20 ms with 50% overlap is used for short-term analysis.

*Parameters in AA/DAA-IMK:* For BAD and species classification experiments, three layered DAA framework is used to obtain dictionaries. The order of factorization i.e. 128 is maintained at each layer to obtain $\mathbf{D}_{Li} \in \mathbb{R}^{390\times128}$ where $i$ ranges from 1 to 3. The number of layers in DAA framework and the order of factorizations are dependent on the nature of data and are fine-tuned using the validation datasets. For phrase classification, 16 atoms per dictionary and three layers of factorizations are used.

*Parameters in other methods:* The number of GMM components used in the class-specific GMM and in PSK are determined using the Akaike information criterion (AIC). In CCSE and AA-IMK framework, 128 archetypes per class-specific archetypal dictionary are used. In DCR, three layers of DAA are used. The order of factorization at each layer is 128. A random forest with 100 trees is used to classify convex representations obtained in DCR. In SKM, a spherical K-means with K=256 is used for feature learning and a random forest with 250 trees is used for both birds and frog species classification. The trade-off parameter, $C$ in the SVM and the width parameter of the Gaussian kernel are empirically chosen. All the aforementioned parameters are fine-tuned on the validation datasets. Libsvm[7] is used for building all SVM based classifiers. SPAMS[8] toolbox is used for learning AA and DAA dictionaries.

## IV. Results and Discussion

In this section, the classification performances of the proposed AA/DAA-IMK on four bioacoustic tasks are presented. Apart from that, the effect of increasing depth on the DAA atoms is also discussed here.

### A. Classification performance

***Species Classification:*** Box-plots depicting the performances of AA/DAA-IMK and other comparative methods for bird and frog species classification tasks are illustrated in Fig. 5a and Fig. 5b respectively. The performance trends of all classifiers including the proposed AA/DAA-IMK are similar between the two datasets. The following can be observed from the analysis of these figures:

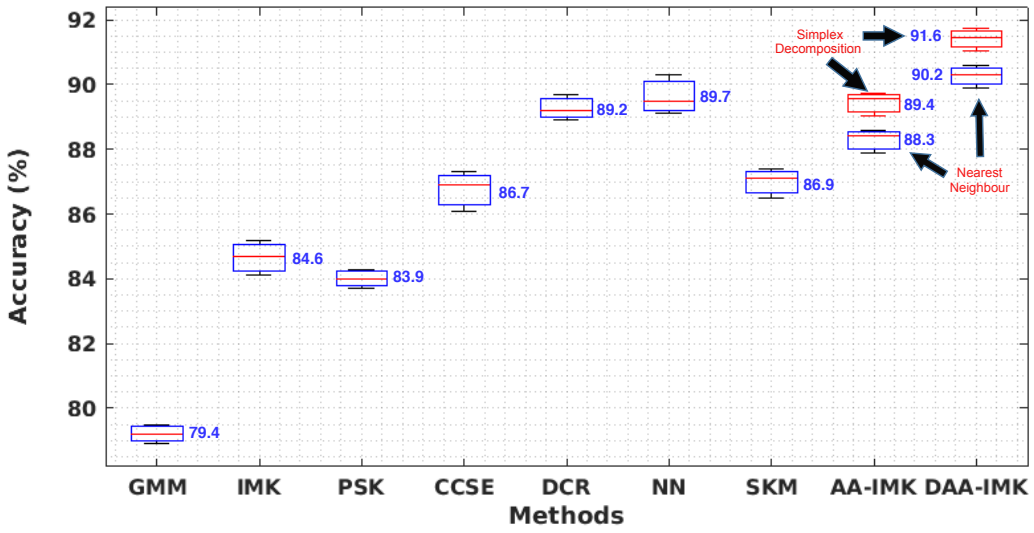
- The proposed AA-IMK and DAA-IMK provide better classification than the traditional kernels such as IMK and PSK across all three folds on both the datasets. The AA-IMK shows a relative improvement of $4.37\%$ and $5.2\%$ over the average classification accuracy achieved by IMK and PSK on the bird dataset. Similarly, relative improvements of $6.6\%$ and $7.1\%$ are shown by DAA-IMK over IMK and PSK respectively. It can be concluded that the introduction of AA and DAA in the kernel method framework led to a better performance.
- AA-IMK shows a relative improvement of $1.85\%$ and $2.27\%$ in average classification accuracy over CCSE on birds and frog dataset respectively. Also, the performances of DAA-IMK is better than DCR by $1.21\%$ and $0.77\%$ on bird and frog dataset respectively. This justifies the hypothesis that the introduction of kernel methods in matrix factorization frameworks can lead to better classification.

[7]https://www.csie.ntu.edu.tw/~cjlin/libsvm/

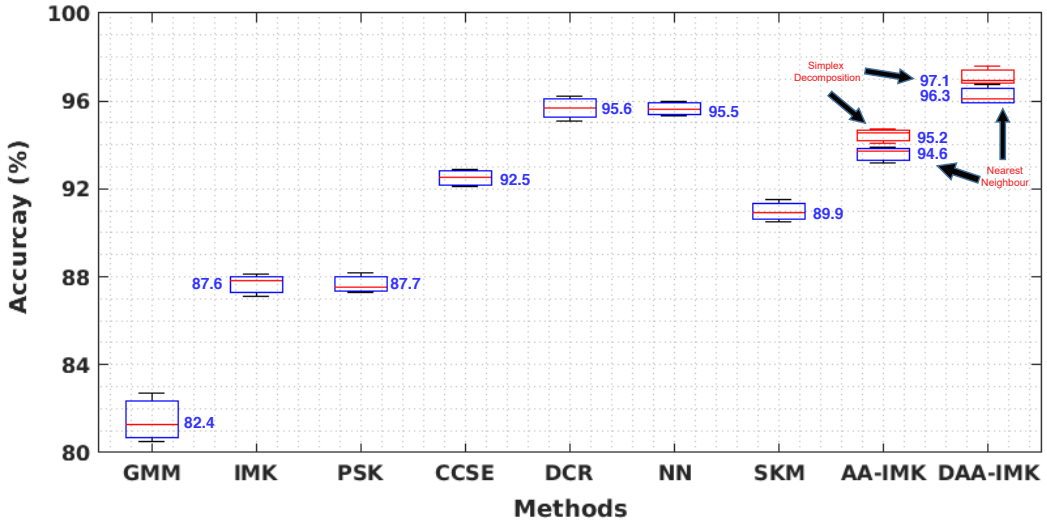
[8]http://spams-devel.gforge.inria.fr/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing






(a)




(b)

Fig. 5: Box plots depicting the classification performances of Gaussian mixture models (GMM), intermediate matching kernel (IMK), probabilistic sequence kernel (PSK), compressed convex spectral embeddings (CCSE), deep convex representations (DCR), deep neural network (NN), spherical K-means (SKM), AA-IMK and DAA-IMK on (a) 50 bird species and (b) 10 frog species (across three folds). Box-plots in red represent the performances of simplex-projection based variant of the proposed AA/DAA-IMK. The number next to each box-plot represents the average classification accuracy across three folds.

- The performance of DAA-IMK is better than all the other methods including AA-IMK on both the datasets. The better performance of DAA-IMK over AA-IMK can be attributed to the better data modeling capabilities of DAA atoms. As discussed earlier, DAA atoms can model both extremal and prototypical behaviour of the data. Thus, unlike AA-IMK or the classical IMK, DAA-IMK utilizes information about the whole data in learning the kernels.
- The simplex projection based DAA-IMK exhibits the best classification performance (across both datasets) among the methods chosen in this study. The simplex-projection based variants of DAA-IMK and AA-IMK (shown in red box-plots in Fig. 5a and Fig. 5b) show a relative improvement of $1.55\%$ and $1.25\%$ over the nearest neighbour variants on the bird dataset. Similarly, relative improvements of $0.83\%$ and $0.63\%$ are observed on the frog dataset. These improvements are small and the classification performances of both these variant are similar.

***Bird Activity Detection:*** Table II depicts the classification performances of different frameworks for the task of bird activity detection. The following inference can be drawn from the analysis of Table II.

- As expected, deep neural network based frameworks such as RCNN and Bulbul outperforms all other frameworks. How-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



TABLE II: Classification performance of different methods for the task of bird activity detection on *BAD 2017 challenge* dataset.

| Framework | AUC (%) |
|---|---|
| Probabilistic Sequence Kernel (PSK) [8] | 83.2 |
| Masked-NMF [37] | 84.25 |
| Recurrent-convolutional neural network (RCNN) [11] | 88.2 |
| Bulbul [12] | 88.91 |
| AA based convex sequence kernel (AA-CSK) [27] | 84.1 |
| Intermediate Matching Kernel (IMK) | 83.5 |
| AA-IMK | 85.1 |
| AA-IMK with simplex projections | 85.9 |
| DAA-IMK | 86.3 |
| DAA-IMK with simplex projections | 86.95 |

TABLE III: Performance of various comparative methods on song phrase classification task.

| Method | Accuracy (%) |
|---|---|
| Sparse Representation based Classifier (SR) | 92.7 |
| Dynamic Time Warping (DTW) | 93.6 |
| DTW-SR-2Pass | 96.9 |
| AA-IMK | 93.1 |
| AA-IMK with simplex projections | 93.15 |
| DAA-IMK | 94.5 |
| DAA-IMK with simplex projections | 94.52 |

ever, the performances of DAA-IMK and DAA-IMK with simplex projections are comparable to RCNN and Bulbul. RCNN shows a relative improvement of 2.15% and 1.42% in AUC scores over DAA-IMK and its simplex projection variant respectively. Similarly Bulbul also shows a relative improvement of 2.92% and 2.1% in AUC scores over DAA-IMK and its simplex projection variant respectively.

- AA/DAA-IMK outperforms PSK, AA-CSK, IMK and Masked-NMF. AA-IMK shows a relative improvement of 2.24%, 1.18%, 1.88% and 1% over PSK, AA-CSK, IMK and Masked-NMF in AUC scores respectively. While DAA-IMK shows a relative improvement of 3.59%, 2.55%, 3.24% and 2.35% over PSK, AA-CSK, IMK and Masked-NMF respectively.
- Simplex projection variants of AA/DAA-IMK exhibit a small improvement in AUC scores over nearest local vector variants. AA-IMK with simplex projections shows a relative improvement of 0.94% over AA-IMK while DAA-IMK with simplex projections exhibits a relative improvement of 0.75% over DAA-IMK.

***Song Phrase Classification:*** The performances of AA/DAA-IMK and other comparative methods are documented in Table III. The analysis of this table highlights the following:

- An effective phrase classification is shown by AA/DAA-IMK. However, both these frameworks are outperformed by DTW-SR-2Pass. This can be attributed to the generative nature of AA/DAA. Although, the data requirements for AA/DAA are significantly lesser than deep learning frameworks, they still require a sufficient number of examples to provide effective generalization.
- DTW-SR-2Pass performs better than DTW and SR. This shows that combining the properties of template matching (DTW) and exemplar based generative modelling (SR) results in better phrase classification under limited training data conditions (5 examples per class).
- No statistically significant difference is observed in the simplex projection and nearest neighbour variants of AA/DAA-IMK.

The analysis of results obtained from all four experiments highlights the significance of AA/DAA-IMK. It is strongly suggestive that the classification performance of AA/DAA-IMK is better than the existing formulations of IMK. These results also demonstrate the power of matrix factorization combined with kernel methods.

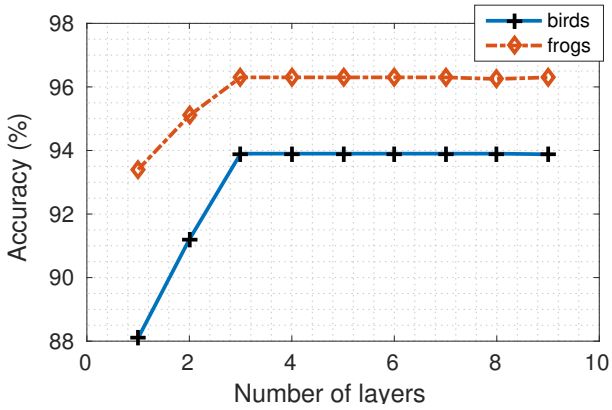



Fig. 6: Classification accuracy of DAA-IMK, on the validation sets, as a function of the number of layers of DAA framework.

### *B. Effect of depth on classification performance*

To choose the number of layers in DAA-IMK framework, the classification accuracy as a function of the number of layers was observed on the validation datasets for bird and frog species classification. This observation is illustrated in Fig. 6. The analysis of this figure shows that the classification accuracy improved as the number of layers are increased from one to three. However, using more than three layers does not account in any increment in the classification performance and up to nine layers the classification performance is almost constant. This behaviour can be attributed to the reasoning that there are small changes in modelling capabilities of DAA dictionary atoms after third layer and these changes do not effect the classification performance of DAA-IMK. To corroborate this claim, we chose a 2-dimensional randomly sampled dataset and factorized it up to nine layers to obtain the DAA dictionaries. The number of dictionary atoms i.e. 20 is kept same at all layers. Fig. 7 depicts the behaviour of DAA dictionaries obtained at first, third, sixth and ninth layer. This figure shows that though there are small differences in modeling capabilities of sixth and ninth layer dictionary atoms, both these dictionaries are modeling the average and extremal behaviour effectively. In an IMK framework, these small differences may or not have any effect on the classification performance. Hence, it can be inferred that after a particular depth, the modeling capabilities of deeper dictionaries start stagnating and going further deep may not improve the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



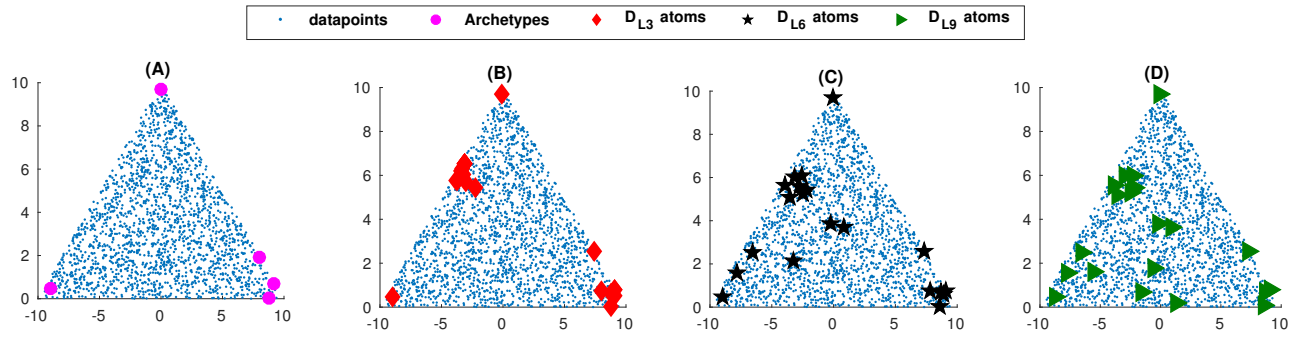



Fig. 7: Behaviour of DAA dictionary atoms obtained at (A) first layer, (B) third layer, (C) sixth layer and (D) ninth layer.

performance of DAA-IMK. This observation is in accordance with the constant classification performances observed after the third layer on the validation datasets (in Fig. 6).

Note that the optimum depth of DAA in the IMK framework is dependent on the geometric properties of data and can only be determined experimentally.

### C. Nearest neighbour search vs simplex decomposition

In Section II, two possible methods to select pairs of the local feature vectors are discussed. From the analysis of classification performances obtained for species classification, bird activity detection and phrase classification, it can be inferred that both nearest neighbour and simplex projection variants exhibit similar performance (less than 1% difference in all tasks.) This implies that mostly same local feature vector pairs are selected for kernel computation in both these variants. It can be attributed to the reason that in most cases, both nearest neighbour and simplex decomposition variants behave in a similar way. The contribution of an atom in representing a data point (by convex combination) is maximum if this data point exhibits maximum spatial proximity to the dictionary atom. As a result, the simplex decomposition coefficient exhibiting the highest value corresponds to the atom that lie at the minimum distance from the data point.

The difference in nature of nearest neighbour and simplex decomposition is highlighted when more than one vector exhibit the same minimum distance from a dictionary atom. In the nearest neighbour variant, any one of these nearest vectors can be selected with respect to the dictionary atom. However, depending on the ideal convex combination (solution of Equation 10), the simplex coefficients corresponding to this dictionary atom can be different in the convex-sparse representations obtained for these vectors. As a result, there can be a difference in local feature pairs selected using nearest neighbour and simplex decomposition. The small performance gain observed for the simplex decomposition in our experiments can be ascribed to selection of the local vector pairs that provided better discrimination.

It is worth mentioning that in many cases, multiple solutions are available for the simplex decomposition problem (Equation 10). In such cases, the local feature vector pairs selected for calculating IMK are dependent on the implementation of simplex decomposition.

### D. Amount of training data vs classification performance

To analyze the effect of training data on the proposed AA/DAA-IMK and other comparative methods, we conducted the bird activity detection (BAD) experiment with varying amount i.e. 10%, 25%, 50% and 75% of the training data. All the parameters and feature representations described in Section III-D are used here. The results of this experiment are depicted in Fig. 8. The analysis of this figure highlights the fact that deep learning frameworks i.e. RCNN and Bulbul significantly outperforms other methods when 50% and 75% of the data is used for training. However, as expected, their performaces significantly deteriorate at 10% and 25% training data configurations. The classification performances of DAA-IMK is better than other methods at 10% and 25% training data configurations. This shows that as desired, the proposed DAA-IMK can provide effective classification in low-training data conditions.

## V. Conclusion

In this paper, we introduced a new classification framework that combines properties of deep matrix factorization with kernel methods. The modeling capabilities of deep archetypal analysis (DAA) are analyzed to propose a variant of the traditional intermediate matching kernel (IMK). The proposed kernel utilizes DAA, a deep matrix factorization framework, for choosing pairs of local feature vectors for learning the base kernels. The nature of DAA helps in choosing local feature vector pairs which lie on or around the class boundaries, in addition to those in the interior. The utilization of these confusing pairs in training process helps in learning a better classifier. Experimental results on four different bioacoustic tasks show that the proposed AA/DAA-IMK outperforms the traditional IMK and matrix factorization based classification frameworks. Future work may include introducing DAA frameworks in other kernels such as PSK and AA-CSK.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2906465, IEEE Journal of Selected Topics in Signal Processing



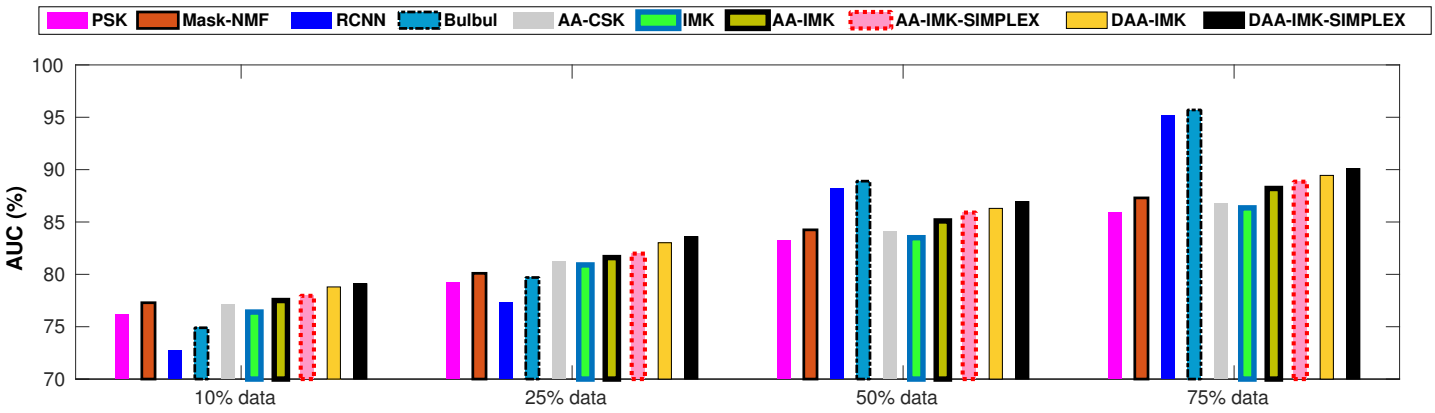



Fig. 8: Bar plots depicting the classification performances of probabilistic sequence kernels (PSK), masked NMF, recurrent convolutional neural network (RCNN), Bulbul, AA based convex sequence kernel (AA-CSK), intermediate matching kernel (IMK), AA-IMK, AA-IMK with simplex decomposition, DAA-IMK and DAA-IMK with simplex decomposition (from left to right in each bar group) at 10%, 25%, 50% and 75% of the training data.

## REFERENCES


[1] F. van Bommel, "Birds in Europe: Population estimates, trends and conservation status," *British Birds*, vol. 98, pp. 269–271, 2005.

[2] S. A. Cushman, "Effects of habitat loss and fragmentation on amphibians: a review and prospectus," *Biological Conservation*, vol. 128, no. 2, pp. 231–240, 2006.

[3] 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.

[4] A. L. Borker, M. W. McKown, J. T. Ackerman, C. A. Eagles-Smith, B. R. Tershy, and D. A. Croll, "Vocal activity as a low cost and scalable index of seabird colony size," *Conservation Biology*, vol. 28, no. 4, pp. 1100–1108, 2014.

[5] B. J. Furnas and R. L. Callas, "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," *The Journal of Wildlife Management*, vol. 79, no. 2, pp. 325–337, 2015.

[6] B. Gatto, J. Colonna, E. M. dos Santos, and E. F. Nakamura, "Mutual singular spectrum analysis for bioacoustics classification," in *Proceedings of International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept., 2017.

[7] "BAD challenge," http://machine-listening.eecs.qmul.ac.uk/bird-audio-detection-challenge/, accessed: 2017-2-1.

[8] A. Thakur, R. Jyothi, P. Rajan, and A. Dileep, "Rapid bird activity detection using probabilistic sequence kernels," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1754–1758.

[9] B. P. Tóth and B. Czeba, "Convolutional neural networks for large-scale bird song classification in noisy environment." in *Proceedings of Conference and Labs of the Evaluation Forum (CLEF)*, 2016, pp. 560–568.

[10] R. Narasimhan, X. Z. Fern, and R. Raich, "Simultaneous segmentation and classification of bird song using CNN," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) (ICASSP)*, 2017, pp. 146–150.

[11] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1744–1748.

[12] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1764–1768.

[13] S. Ntalampiras, "Bird species identification via transfer learning from music genres," *Ecological Informatics*, vol. 44, pp. 76–81, 2018.

[14] J. Xie, C. Ding, W. Li, and C. Cai, "Audio-only bird species automated identification method with limited training data based on multi-channel deep convolutional neural networks," *arXiv preprint arXiv:1803.01107*, 2018.

[15] A. Harma and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. 701–704.

[16] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Transactions on Audio, Speech, Language Processing*, vol. 14, no. 6, pp. 2252–2263, Nov 2006.

[17] W. Chu and D. T. Blumstein, "Noise robust bird song detection using syllable pattern-based hidden Markov models," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 345–348.

[18] V. M. Trifa, A. N. Kirschel, C. E. Taylor, and E. E. Vallejo, "Automated species recognition of antbirds in a mexican rainforest using hidden Markov models," *The Journal of Acoustical Society of America*, vol. 123, no. 4, pp. 2424–2431, 2008.

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

[20] A. Thakur, V. Abrol, P. Sharma, and P. Rajan, "Compressed convex spectral embedding for bird species classification," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April, 2018.

[21] ——, "Deep convex representations: Feature representations for bioacoustics classification," in *Proceedings of Interspeech*, 2018.

[22] K. Qian, Z. Zhang, A. Baird, and B. Schuller, "Active learning for bird sound classification via a kernel-based extreme learning machine," *The Journal of Acoustical Society of America*, vol. 142, no. 4, pp. 1796–1804, 2017.

[23] 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.

[24] D. Chakraborty, P. Mukker, P. Rajan, and A. Dileep, "Bird call identification using dynamic kernel based support vector machines and deep neural networks," in *Proceedings of International Conference on Machine Learning Applications*, 2016.

[25] P. S. Kumar and P. Amita, *Pattern Recognition And Big Data*. World Scientific, 2016.

[26] S. Boughorbel, J. P. Tarel, and N. Boujemaa, "The intermediate matching kernel for image local features," in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, July 2005, pp. 889–894.

[27] 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 *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2017, pp. 4436–4440.

[28] V. Wan and S. Renals, "Evaluation of kernel methods for speaker verification and identification," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.

[29] A. D. Dileep and C. C. Sekhar, "GMM-based intermediate matching kernel for classification of varying length patterns of long duration

speech using support vector machines," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1421–1432, 2014.

[30] A. Dileep and C. C. Sekhar, "Speaker identification using intermediate matching kernel-based support vector machines," in *Forensic Speaker Recognition*. Springer, 2012, pp. 389–424.

[31] A. D. Dileep and C. C. Sekhar, "Class-specific GMM based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," *Speech Communication*, vol. 57, pp. 126–143, 2014.

[32] W. Wang, Z. Xu, W. Lu, and X. Zhang, "Determination of the spread parameter in the gaussian kernel for classification and regression," *Neurocomputing*, vol. 55, no. 3-4, pp. 643–663, 2003.

[33] Y. Chen, J. Mairal, and Z. Harchaoui, "Fast and robust archetypal analysis for representation learning," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1478–1485.

[34] P. Sharma, V. Abrol, and A. Thakur, "Ase: Acoustic scene embedding using deep archetypal analysis and GMM," in *Proceedings of Interspeech*, 2018.

[35] A. Cutler and L. Breiman, "Archetypal analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, 1994.

[36] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[37] I. Sobieraj, Q. Kong, and M. Plumbley, "Masked non-negative matrix factorization for bird detection using weakly labelled data," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1819–1823.

[38] L. N. Tan, A. Alwan, G. Kossan, M. L. Cody, and C. E. Taylor, "Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1069–1080, 2015.

[39] A. Thakur, V. Abrol, P. Sharma, and P. Rajan, "Rényi entropy based mutual information for semi-supervised bird vocalization segmentation," in *Proceedings of International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017.