Deep Convex Representations: Feature Representations for Bioacoustics Classification

Anshul Thakur[1], *Vinayak Abrol*[2], *Pulkit Sharma*[1] *& Padmanabhan Rajan*[1]

[1]IIT Mandi, India
[2]Idiap Research Institute, Martigny, Switzerland

{anshul_thakur, pulkit_s}@students.iitmandi.ac.in, vinayak.abrol@idiap.ch, padman@iitmandi.ac.in

Abstract

In this paper, a deep convex matrix factorization framework is proposed for bioacoustics classification. Archetypal analysis, a form of convex non-negative matrix factorization, is used for acoustic modelling at each level of this framework. At first level, the input feature matrix is factorized into an archetypal dictionary and corresponding convex representations. The representation matrix obtained at the first level is further factorized into a dictionary and convex representations at second level. This hierarchical factorization continues until a desired depth is achieved. We observe that the dictionaries at different levels model complimentary information present in the data. The atoms of the dictionary learned at the first layer lie on convex hull of the data, thus try to model the extremal behaviour. On the contrary, atoms of the deeper dictionaries lie on the convex hull as well as inside the convex hull. Hence, these dictionaries can simultaneously model the extremal and average behaviour of the data. The convex representations obtained from these deeper dictionaries are referred as deep convex representations. Due to inherent sparsity, they result in efficient classification performance. Through experiments on two available bioacoustics datasets, we show that the proposed approach yield comparable or better results than state-of-art approaches.

Index Terms: deep convex representation, archetypal analysis, bioacoustics classification

1. Introduction

Acoustic monitoring provides a convenient and passive way to survey and monitor the animal and avian diversity of a particular habitat of interest [1, 2]. Acoustic monitoring makes it possible to remotely monitor various ecosystems such as swamps, marshes, remote islands or even aquatic ecosystems, where manual monitoring is difficult or not feasible. Bioacoustic signal classification is an important module in any acoustic monitoring system and can play a significant role in facilitating the analysis of population trends of different animal and bird species. Hence, bioacoustics classification can help in boosting conservation efforts for species under the threat of population decline or extinction.

Learned feature representations obtained by matrix factorization on spectrograms or Mel frequency cepstral coefficients (MFCC) have been employed successfully for various acoustic and bioacoustic classification tasks such as acoustic scene classification [3], acoustic event detection [4], speech recognition [5], bird audio detection [6] and bird species classification [7]. The intent of matrix factorization is to decompose a spectrogram or multivariate feature matrix, $\mathbf{X}$, into a dictionary $\mathbf{D}$ and representation $\mathbf{A}$ such that $\mathbf{X} \approx \mathbf{DA}$ with certain constraints on both $\mathbf{D}$ and $\mathbf{A}$. These representations can model the characteristics present in the data more effectively than hand-crafted feature representations [8]. Recently, there is an influx of deep matrix factorization techniques [8, 9, 10] to obtain learned representations. These techniques factorize a matrix $\mathbf{X}$ into multiple factors as $\mathbf{X} \approx \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3 \ldots \mathbf{D}_k\mathbf{A}_k$ where k is the depth of factorization. This deep factorization helps in emphasizing various latent attributes which can be helpful in analysing the data better.

Motivated by the success of deep matrix factorization in speech recognition [8], we propose a deep convex framework targeting the task of bioacoustic classification. Archetypal analysis (AA) [11] form the crux of the proposed framework. AA decomposes a matrix $\mathbf{X}$ as: $\mathbf{X} \approx \mathbf{DA}$, where $\mathbf{D}$ is a dictionary and $\mathbf{A}$ is convex-sparse representation. The dictionary, $\mathbf{D}$, is composed of archetypes or the extremal points. These archetypes lie on convex hull of the data and are constrained to be the convex combination of data points i.e., $\mathbf{D} = \mathbf{XB}$. The archetypes provide compact and meaningful representation of the data as they model the extremal or convex hull [12]. Although convex-sparse representations, obtained using AA, have been effectively used in various bioacoustics classification tasks such as bird activity detection [12] and bird species classification [7], AA suffers from a major disadvantage. It cannot model the average behaviour of the data which is often defined by prototypes such as the mean, the median or the medoid of the data. Hence, it can be hypothesized that combining the extremal modelling properties of AA with the behaviour defined by the prototypes can provide better modelling capabilities.

AA is used in the proposed deep convex framework to factorize the matrices at each level. At first level, an input matrix is factorized into an archetypal dictionary and the convex-sparse representation. The representation matrix obtained at the first level is further factorized into a dictionary and a convex-sparse representation at the second level. This hierarchical process is continued till the desired level of the factorization. Fig. 1 illustrates the proposed deep convex framework. By the definition of AA, the atoms of the dictionary obtained at the first level of the framework lie on the convex hull or extremal. However, the dictionaries obtain at the deeper levels show different modelling capability as compared to the archetypal dictionary of the first level. Some of the atoms of these deeper dictionaries lie on the convex hull while others lie inside the convex hull or boundary. The atoms lying on the convex hull model the extremal while atoms lying inside the boundary can be seen as the representatives of the average behaviour. Hence, these deeper dictionaries have better data modelling capabilities than the typical archetypal dictionaries. The convex representations obtained using these deeper dictionaries, designated as deep convex representations, are used for bioacoustics classification in this work.

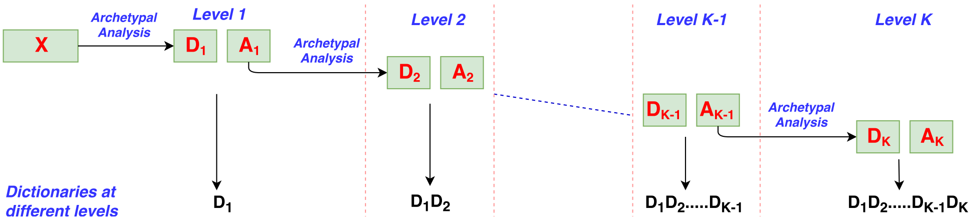

Figure 1: *Illustration of the archetypal analysis based deep matrix factorization. Here a matrix* $\mathbf{X}$ *is factorized into* $K+1$ *factors as:* $\mathbf{X} \approx \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3 \ldots \mathbf{D}_K\mathbf{A}_K$.

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

2. Related Work

Many studies have targeted the problems of bioacoustics classification. In [13], a low resource (computation and memory) framework, which utilizes mutual singular spectrum analysis to obtain the bases defining the subspaces of the bioacoustics classes, has been proposed. Canonical angles are used to measure the similarity between the subspaces to classify any test audio signal. The low resource utilization of this framework makes it suitable for acoustic monitoring applications. However, the reported classification performance of this framework is not up-to the desired standards. In [14], a convolutional neural network (CNN) that simultaneously segments and classifies bird vocalizations is proposed. This network bypasses the tedious task of segmentation but requires pixel-wise labelling of spectrograms which is often not available. Quin *et. al* proposed to use kernel-based extreme learning machines for classifying bird vocalizations [15]. Extreme learning machine [16] is a feed-forward neural network with random weights that does not require back-propagation based training. Hence, in comparison to deep neural networks (DNN), the amount of data required to train these networks is usually less. Apart from neural networks, SVM powered by dynamic kernels have also been utilized for bird species classification [17] and bird activity detection [18]. In our earlier studies, we successfully utilized archetypal analysis to obtain learned feature representations for the tasks of bird species classification [7] and bird activity detection [12].

3. Proposed Framework

This section describes the process to obtain deep convex representations using the AA based deep matrix factorization. Firstly, we briefly explain the compressed super-frame (CSF) feature, a representation shown to be suitable for bioacoustics classification [7]. Then, the process to learn dictionaries in a deep AA framework is explained in detailed, followed by the process to obtain deep convex representations (using the learned dictionaries) for the testing phase.

3.1. Obtaining compressed super-frames

The proposed framework factorizes a collection of CSFs [7] for obtaining deep convex representations. These CSFs are derived from the spectrogram by concatenating W neighbouring frames. This concatenation process helps in effectively capturing the frequency and temporal modulations which characterize the vocalizations of different animal and bird species. However, the concatenation also produces a high-dimensional (Wm-dimensional, m is the number of frequency bins in a frame) representation. In order to avoid the high dimensionality, these concatenated frames are compressed using random projections (since they are sparse) resulting in CSFs (Z-dimensional such that $Z < Wm$). More details about CSF representation can be found in [7].

3.2. Learning dictionaries using AA based deep factorization

To learn class-specific dictionaries, CSFs of the vocalizations of a particular class are pooled together to form a matrix $\mathbf{X} \in \mathbb{R}^{Z \times L}$, where Z is the dimensionality of the CSFs and L is the number of pooled CSFs. The feature matrix $\mathbf{X}$ is fed to the proposed deep framework (as illustrated in Fig. 1) to obtain a discriminative feature representation. In this work, we have used a robust AA which utilizes a weighting function to decrease the effect of outliers in the process of learning the archetypes. More details about robust AA can be found in [7, 11].

At the first level, $\mathbf{X}$ is factorized using AA to obtain an archetypal dictionary, $\mathbf{D}_1 \in \mathbb{R}^{Z \times n_1}$ (with n_1 number of archetypes) and a convex-sparse representation matrix $\mathbf{A}_1 \in \mathbb{R}^{n_1 \times L}$ as $\mathbf{X} \approx \mathbf{D}_1\mathbf{A}_1$. This representation, $\mathbf{A}_1$, is passed to the next level for further factorization. Again, at the second level, $\mathbf{A}_1$ is factorized using robust AA to obtain dictionary $\mathbf{D}_2 \in \mathbb{R}^{n_1 \times n_2}$ (an under-complete dictionary with n_2 archetypes such that $n_2 < n_1$) and convex-sparse representations $\mathbf{A}_2 \in \mathbb{R}^{n_2 \times L}$. At second level, we have the data decomposition as: $\mathbf{X} \approx \mathbf{D}_1\mathbf{A}_1 \approx \mathbf{D}_1\mathbf{D}_2\mathbf{A}_2 = \mathbf{D}_{l2}\mathbf{A}_2$, where $\mathbf{D}_{l2} \in \mathbb{R}^{Z \times n_2}$ is the global dictionary at this level. Again $\mathbf{A}_2$ is passed to next level for further factorization and this hierarchical process continues till a desired depth. By generalization, at any level K, $\mathbf{X}$ is factorized as: $\mathbf{X} \approx \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3 \ldots \mathbf{D}_K\mathbf{A}_K = \mathbf{D}_{lK}\mathbf{A}_K$, where $\mathbf{D}_{lK} \in \mathbb{R}^{Z \times n_K}$ (with n_K number of archetypes) is the global dictionary and $\mathbf{A}_K \in \mathbb{R}^{n_K \times L}$ is the convex-sparse representation matrix obtained at the Kth level of the framework.

3.2.1. Geometric interpretation and visualization

The geometric interpretation of archetypes is well known. They are the convex combination of the data and lie on the convex hull. Thus, the dictionaries $\mathbf{D}_1, \mathbf{D}_2 \ldots \mathbf{D}_K$ at each layer lie on the convex hull of the data from which they are learned i.e., $\mathbf{X}, \mathbf{A}_1 \ldots \mathbf{A}_{K-1}$ respectively. However, there is no convexity constraint present on the atoms of $\mathbf{D}_{lK}$, as they are just the linear combination of atoms from the global dictionary obtained at $K-1$th level. This suggests that the atoms of the deeper level global dictionary may lie inside the convex hull or in other words, they can behave as *prototypes* instead of archetypes.

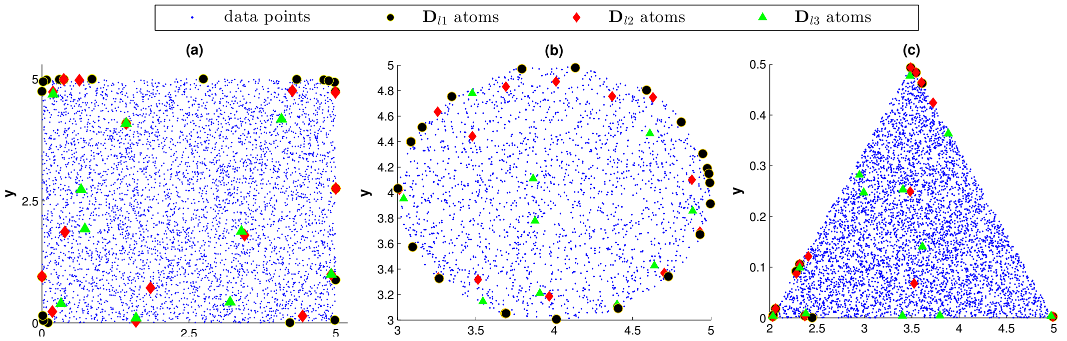

Figure 2: *Behaviour of atoms of the dictionaries, learned at three different levels of the proposed framework by factorizing data points, uniformly and randomly sampled from a square (a), circle (b) and triangle (c).*

Also, $\mathbf{D}_{l2}$ (global dictionary obtained at second layer) is obtained by the multiplication of $\mathbf{D}_1$, and $\mathbf{D}_2$, where $\mathbf{D}_1$ is modelling the convex hull of $\mathbf{X}$. Thus, due to the presence of $\mathbf{D}_1$ in the calculation, some of the atoms of $\mathbf{D}_{l2}$ can lie on or around the convex hull of the data. This can be interpolated to the Kth level and we can deduce that some of the atoms of the global dictionary $\mathbf{D}_{lK}$ also lies on convex hull due to the the presence of $\mathbf{D}_1$ in calculation of $\mathbf{D}_{lK}$.

To illustrate this, we factorize 3 different randomly generated two-dimensional datasets using the proposed deep convex factorization framework. For analysis, we employed only three level factorizations, with $n_1 = 25$, $n_2 = 15$ and $n_3 = 10$ being the number of archetypes learned at different levels. Fig. 2 illustrates the differences in the modelling capabilities of different dictionaries learned using the proposed framework. As expected, Fig. 2 demonstrates that the atoms of $\mathbf{D}_{l1}$, are lying on the convex hull thereby modelling the extremal behaviour of the data points. Whereas, the atoms of dictionaries $\mathbf{D}_{l2}$ and $\mathbf{D}_{l3}$ of the second and third level, lie on the convex hull as well as inside the convex hull. It can be inferred that the atoms lying inside the hull are the representations of the prototypical behaviour of the data. The prototypes such as mean, median and medoids are the representatives of the average behaviour of the data points. Hence, deeper dictionaries combine the extremal modelling behaviour of AA with the average behaviour of the data. Thus, these deeper dictionaries have better data modelling capability than the conventional AA factorization.

3.3. Classification: deep convex representations as features

Training: The class-specific global dictionaries learned at the Kth level are concatenated to get the final dictionary, $\mathbf{D}_f = [\mathbf{D}^1_{lk}\mathbf{D}^2_{lk}\ldots\mathbf{D}^q_{lk}]$ where $\mathbf{D}^q_{lk} \in \mathbb{R}^{Z\times n_k}$ is the Kth level global dictionary of the qth class. To obtain deep convex representations for a given CSF $\mathbf{x}_i$, it is projected on a simplex defined by the atoms of $\mathbf{D}_f \in \mathbb{R}^{Z\times qn_k}$ (with qn_k number of atoms):

$$\mathbf{a}_i = \underset{\substack{\mathbf{a}_i \\ \mathbf{a}_i \in \Delta_{qn_k}}}{\operatorname{argmin}} \|\mathbf{x}_i - \mathbf{D}_f\mathbf{a}_i\|_2^2, \tag{1}$$

such that $\Delta_{qn_k} \triangleq [\mathbf{a}_i \succeq 0, \|\mathbf{a}_i\|_1 = 1]$. The convex representations obtained using equation 1 are inherently sparse [11]. In this work, we have used the active-set algorithm, proposed in [11], to solve equation 1. During the training process, deep convex representations are obtained for all training CSFs using equation 1 and a multi-class classifier such as support vector machines (SVM) or random forest is trained using these representations.

Testing: A test vocalization is represented as a set of CSFs, $\mathcal{X} = [\mathbf{x}^t_1\mathbf{x}^t_2\ldots\mathbf{x}^t_n]$, where $\mathbf{x}^t_n$ is the nth CSF in the test vocalization. A deep convex representation is obtained for each CSF in $\mathcal{X}$ using $\mathbf{D}_f$. These convex representations are fed into a trained classifier to get the CSF level decisions. Finally, a voting rule is applied on these CSF level decisions to classify $\mathcal{X}$ or the corresponding animal or bird vocalization.

4. Performance Analysis

4.1. Dataset used

The proposed framework is evaluated on two datasets containing audio recording of bird and frog species. The first dataset containing audio recordings of 50 different bird species is obtained from three different sources. The recordings of 264 bird species were obtained from the Great Himalayan national park (GHNP). The recordings of 7 bird species were obtained from the bird audio database maintained by the Art & Science centre, UCLA [19]. The remaining 174 bird species audio recordings were obtained from the Macaulay Library [20]. The recording used here are 16-bit mono, sampled at 44.1 kHz and are of durations varying from 15 seconds to 3 minutes. 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 second dataset contains audio recordings of 10 different frog species used in [13] for bioacoustic classification and is available at `http://goo.gl/FFBzbb`. This set of recordings are 16-bit mono and are sampled of 44.1 kHz.

4.2. Experimental setup

Parameter setting: The vocalizations are segmented from audio recordings using the semi-supervised method proposed in [21]. Only these segmented vocalizations are used for training and testing. Each input audio recording is converted into a spectrogram using the STFT with 512 DFT points using a frame rate of 20 ms with 50% overlap. A window with context size $W = 5$ is used to obtain super-frame representations which results in 1285-dimensional ($1285 = (257 \times 5)$) representation. Random projections are used to compress these representations to obtain 500-dimensional CSFs. The depth K of the deep convex framework is set to 3. The orders of factorization i.e. n_1, n_2 and n_3 are 128, 64 and 32, respectively. A random forest classifier with 100 trees is used for obtaining CSF level decisions. All the parameters mentioned here are tuned empirically.

Comparative methods: The classification performance of

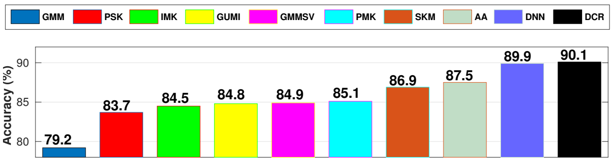

Figure 3: *Classification performance of different methods on 50 bird species (averaged across three folds)*

deep convex representations (DCR) is compared with various existing bioacoustics classification methods such as GMM, SVM with dynamic kernels (intermediate matching kernel (IMK), pyramid matching kernel (PMK), GMM-UBM mean interval (GUMI) kernel, GMM supervector kernel (GMMSV) and probabilistic sequence kernel (PSK)), spherical K-means (SKM) with random forest based framework proposed in [22] and a DNN based approach proposed in [17]. For SVM and DNN based classification schemes, MFCC using delta and acceleration coefficients are used as the feature representation. In addition, the classification performance of convex representations obtained using class-specific AA (first layer of the framework, $K = 1$) is also compared. Here, each class is modelled by a dictionary with 128 atoms and a random forest with 100 trees is used for classification.

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 present in each fold (per class) are used for training while the remaining are used for testing. The results presented here are averaged across three folds.

4.3. Results and Discussion

4.3.1. Classification performance

The classification performance of the proposed framework and other comparative methods on bird and frog datasets is shown in Fig. 3 and Fig. 4 respectively. The following inferences can be made from the analysis of these two figures:

- The classification performance of the proposed DCR is comparable to the DNN over both the datasets. DCR shows a minute relative improvement of 0.22% and 0.21% over the DNN on the bird and frog datasets respectively.
- DCR significantly outperforms the GMM, SVM powdered by dynamic kernels and SKM on both the datasets. DCR shows a relative improvement of $12.1\%, 7.65\%, 6.22\%, 5.88\%, 5.77\%, 5.55\%$ and 3.55% over the performances of GMM, PSK, GUMI, GMMSV, PMK, SKM and AA respectively on the bird dataset. Also, it exhibits a relative improvement of $13.7\%, 9.23\%, 7.87\%, 6.93\%, 6.51\%, 7.24\%$ and 4.83% using different approaches on the frog dataset.
- Classification performance of different dynamic kernels is almost similar. Also, these kernels are significantly outperformed by SKM, AA and DNN on both the datasets.
- DCR, convex representations obtained at the third level of the deep convex framework, performs better than AA (convex representations obtained at $K = 1$). DCR exhibits a relative improvement of 2.89% and 2.58% over AA on the bird and the frog datasets respectively. This highlights the importance of DCR over convex representations obtained using AA.

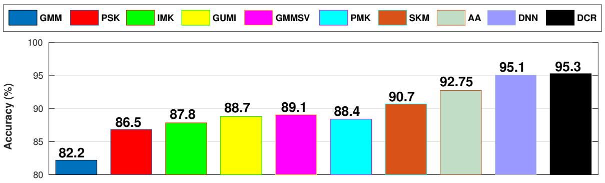

Figure 4: *Classification performance of different methods on 10 frog species (averaged across three folds)*

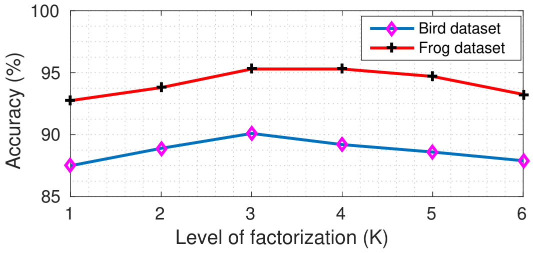

Figure 5: *Variation in classification accuracy with the level of factorization.*

4.3.2. Depth of factorization vs. classification performance

The extent of factorization required in the proposed deep convex framework depends on the dynamics of the data under processing and the optimal value of K is difficult to generalize. However, to establish the ideal depth of our deep framework for the two datasets used in this study, we analysed the classification performances with different number of levels ($K = 1, 2, 3, 4, 5$ and 6). The orders of factorization at first to sixth level of the proposed framework is set to $128, 64, 32, 24, 16$ and 8. Fig. 5 depicts the performance of convex representations as a function of depth. The analysis of this figure confirms that the maximum classification accuracy is observed at $K = 3$ for both the datasets. This justifies the use of three level factorization in the proposed framework. Also, as the level of factorization is increased, a small drop in classification accuracy is observed. This can be attributed to the decrease in the number of atoms of the dictionaries learned at 5th and 6th level of the framework. Each dictionary at 5th level has 16 atoms while dictionaries at 6th level has 8 atoms. These number of atoms may not be efficient to fully capture the variations present in the data, leading to a drop in accuracy.

5. Conclusions

In this paper, we have proposed deep convex representations obtained using a deep factorization framework for bioacoustics classification. Our experiments indicate that the deeper convex dictionaries has better data modelling capabilities than the dictionaries learned using single level archetypal factorization. The experimental observations highlight the advantages of these deep representations over the convex representations obtained using AA. Also, the classification performance of these deep convex representations is comparable to the state-of-art bioacoustics classification methods. Future work may include the application of deep convex representation for other audio classification tasks.

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," *Trans. Audio, Speech, Language Process*, vol. 16, no. 8, pp. 1541–1550, 2008.

[3] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, March, 2016, pp. 6445–6449.

[4] P. Giannoulis, G. Potamianos, P. Maragos, and A. Katsamanis, "Improved dictionary selection and detection schemes in sparse-CNMF-based overlapping acoustic event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 25–29.

[5] P. Sharma, V. Abrol, A. Dileep, and A. K. Sao, "Class specific GMM based sparse feature for speech units classification," in *Proc. Eusipco*, 2017, pp. 528–532.

[6] I. Sobieraj, Q. Kong, and M. Plumbley, "Masked non-negative matrix factorization for bird detection using weakly labelled data," in *Proc. Eusipco*, 2017, pp. 1819–1823.

[7] A. Thakur, V. Abrol, P. Sharma, and P. Rajan, "Compressed convex spectral embedding for bird species classification," in *Proceedings of Int. Conf. Acoust. Speech, Signal Process. (to appear)*, April, 2018.

[8] P. Sharma, V. Abrol, and A. K. Sao, "Deep-sparse-representation-based features for speech recognition," *Trans. Audio, Speech and Lang. Process.*, vol. 25, no. 11, pp. 2162–2175, November 2017.

[9] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller, "A deep matrix factorization method for learning attribute representations," *Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 417–429, 2017.

[10] H.-J. Xue, X.-Y. Dai, J. Zhang, S. Huang, and J. Chen, "Deep matrix factorization models for recommender systems," in *Proc. of IJCAI*, 2017.

[11] Y. Chen, J. Mairal, and Z. Harchaoui, "Fast and robust archetypal analysis for representation learning," in *Proceedings of Comp. Vis. Pattern Recog.*, June, 2014, pp. 1478–1485.

[12] 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 Eusipco*, Aug., 2017, pp. 4436–4440.

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

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

[15] 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.

[16] 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.

[17] 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 Int. Conf. Mach. Learn. App.*, Dec., 2016, pp. 280–285.

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

[19] "Art-sci center, University of California," http://artsci.ucla.edu/birds/database.html/, accessed: 2017-10-10.

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

[21] A. Thakur and P. Rajan, "Rényi entropy based mutual information for semi-supervised bird vocalization segmentation," in *Proceedings of MLSP*, Sept., 2017.

[22] 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.