COMPRESSED CONVEX SPECTRAL EMBEDDING FOR BIRD SPECIES CLASSIFICATION

A. Thakur, V. Abrol, P. Sharma and P. Rajan

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

ABSTRACT

This paper focuses on the problem of bird species identification using audio recordings. Following recent developments in deep learning, we propose a multi-layer alternating *sparse-dense* framework for bird species identification. Temporal and frequency modulations in bird vocalizations are captured by concatenating frames of spectrograms, resulting in a high dimensional super-frame based representation. These super-frame representations are highly sparse. Hence, we propose to use random projections to compress these super-frames. This is followed by class-specific archetypal analysis, employed on these compressed super-frames for acoustic modeling, to obtain a convex-sparse representation. These convex-sparse representations are referred as compressed convex spectral embeddings (CCSE). It is observed that these representations efficiently capture species-specific discriminative information. Experimental results show compelling evidence that the proposed approach shows performance comparable to existing methods such as deep neural networks (DNN) and dynamic kernel based SVMs.

***Index Terms*—** bioacoustics, bird species identification, archetypal analysis

1. INTRODUCTION

Habitat destruction due to human activities and global climate change has led to a rapid decline in populations of various avian species. In recent times, various conservation efforts have been started to protect these species. Most of these efforts start with surveying and monitoring birds in their natural habitats, for which acoustic monitoring is a convenient and passive method [1]. It can be utilized in habitats where manual monitoring is difficult such as in swamps, marshes and remote islands. In this work, we target the problem of species identification from audio recordings, which is a major task in acoustic monitoring. One of the major steps in designing such a system is the choice of acoustic features with species-specific signatures. This work is motivated by the success of learnt features obtained by processing the spectrogram using matrix factorization (MF) approaches [2]. MF has been shown to perform well for various tasks in bioacoustics such as segmentation [3] and bird audio detection [2, 4].

In contrast to the direct factorization of the spectrogram, we propose a supervised multi-layer alternating dense-sparse framework for bird species identification. The proposed multi-layer framework (illustrated in Fig. 1) helps in modeling various latent hidden acoustic attributes. To the best knowledge of the authors, such a framework has not been applied for bioacoustics, although a few works exists in context of natural images and speech/audio signals [5, 6]. First, a given recorded audio signal ($dense$) is converted into a magnitude spectrogram (*sparse*). The notion of sparsity comes from the fact that bird vocalizations acquire only a few frequency bins in the spectrogram [7]. Employing a conventional MF approach for feature learning at this stage, will not capture the species-specific signatures efficiently, which arises due to the time-frequency modulations present in bird vocalizations . To address this, context information is embedded around each frame of the spectrogram by concatenating a fixed number of frames (see Fig. 2). However, this results in a high dimensional ($sparse$) super-frame representation which is not suitable for acoustic modeling due to high computational complexity. This issue is addressed by compressing the super-frames to a relatively low dimensional representation ($dense$) using random projections [8]. The motivation for such relatively low-dimensional embeddings comes from the pairwise distance preserving property of random projections, following the Johnson-Lindenstrauss (JL) lemma [9]. Next, acoustic modeling on the compressed representations is performed using restricted robust archetypal analysis (AA) for each class/bird species. Compared to its counterparts such as NMF or sparse dictionary learning [10], AA results in a more compact, probabilistic and interpretable representation [11]. The learned archetypal dictionaries are used to obtain a sparse-convex representation for compressed super-frames. In this work, these sparse representations are referred to as compressed convex spectral embeddings (CCSE). These CCSE are used for species classification using a minimum reconstruction error based classifier [12].

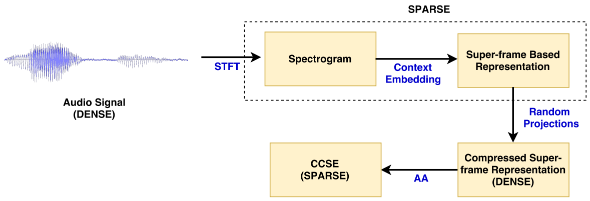

Fig. 1. Proposed pipeline for obtaining CCSE from audio signal.

AA involves approximating the convex-hull of the data. The estimation of convex hull is computationally demanding [13]. Hence, in order to obtain a better estimate and speed up the process of finding archetypes, we propose to use restricted AA. In other words, AA is restricted to the data points around the boundary/convex hull only. Moreover, AA is performed individually for each class and no separate effort is made to increase the inter-class discrimination. Hence, there can be high correlation between inter-class dictionary atoms which may decrease the discrimination ability. One way to overcome this problem is to learn dictionaries in a supervised manner as done in approaches like label-consistent K-SVD [14]. However, they have large computational complexity (time and space). To address this, we propose an efficient procedure to choose atoms from each dictionary such that the gross correlation over all dictionaries is reduced. This helps in decreasing the size of dictionaries as well as may boost up the classification performance. The major contributions of this work are listed below:

- A compressed super-frame based representation that captures time-frequency bird vocalization modulations.
- Restricted robust archetype analysis for modeling bird vocalizations.
- An iterative pruning procedure to choose a subset from the atoms of each dictionary such that the gross-correlation between inter and intra dictionary atoms is reduced.

The rest of this paper is organized as follows. In Section 2, we discuss some of the methods proposed in the literature for bird species identification using acoustic data. In Section 3, the proposed framework is described in detail. Performance analysis and conclusion are in sections 4 and 5, respectively.

2. RELATED WORKS

Many studies have targeted the task of bird species identification. In [15], syllables modeled as frequency and amplitude modulated sinusoidal pulses are used for species identification. This method works only for species producing tonal sounds. Lee *et. al* proposed to use two-dimensional cepstral coefficients for bird species identification in [16]. Each class is modeled by a set of prototype vectors which are centroids of vector quantization or means vectors of Gaussian mixture models (GMM). Further, linear discriminant analysis (LDA) is applied on these prototypes to increase the inter-class variations. The final classification is done using a k-NN classifier. In [17], SVM with various dynamic kernels such as probabilistic sequence kernel (PSK), GMM supervector (GMMSV) kernel, GMM-UBM mean interval (GUMI) kernel, GMM-based intermediate matching kernel (GMM-IMK) and GMM-based pyramid match kernel (PMK) are used for bird species classification. Apart from this, a deep neural network (DNN) has also been proposed for species identification. For large-scale species identification, an unsupervised feature learning method based on spherical K-means has been proposed in [18]. The discriminative qualities of these features are verified using a random forest based classification framework. In [19], a convolutive neural network (CNN) based framework is used for species identification. It uses Segnet [20] to segment vocalizations from the spectrogram and simultaneously classifies these bird vocalizations. Hence, this framework bypasses the need to segment the bird vocalizations before classification.

3. PROPOSED FRAMEWORK

In this section, we describe the steps involved in converting an audio recording into the compressed super-frame based representation (see Fig. 1). Further, we explain restricted robust AA to learn class-specific dictionaries from the training audio recordings. This is followed by the proposed pruning procedure to reduce the inter-class correlation between dictionary atoms. Finally, we describe the procedure for computing CCSE during the testing/classification phase.

3.1. Computing compressed super-frames

The short-time Fourier transform (STFT) is used to convert each input audio recording into a magnitude spectrogram, $\mathbf{S}$ ($m \times N$, m is the number of frequency bins, N is the number of the frames). Context information is embedded to the current frame of the spectrogram by concatenating W previous and W next frames around the current frame. This gives rise to a high $2Wm + m$ dimensional representation (see Fig. 2). After this concatenation process, the pooled spectrograms of all the examples of a particular class, $\hat{\mathbf{S}}$, are transformed to a super-frame based representation matrix, $\mathbf{F} \in \mathbb{R}^{(2Wm+m)\times l}$, where l is the number of pooled super-frames. Working with these high-dimensional super-frames is computationally expensive. However, such super-frame representations are highly sparse. Hence, exploiting the JL lemma [9], we propose to use random projections to compress these super-frames. It is conjectured in the literature that with high probability Gaussian random matrices satisfy Johnson-Lindenstrauss lemma i.e., they preserve the pair-wise distance between super-frames in the transformed domain [21]. In particular, the transformation, $\phi : \mathbb{R}^{2Wm+m} \rightarrow \mathbb{R}^{K}$ is applied by using a random Gaussian matrix ($\mathbf{G}$ of dimensions $K \times 2Wm + m$) to compress the super-frames. This compressed transformed representation, $\mathbf{X} = \mathbf{GF}, \mathbf{X} \in \mathbb{R}^{K\times l}$, is used for both learning the dictionaries (for acoustic modeling) in the training phase and obtaining the CCSEs during the testing phase.

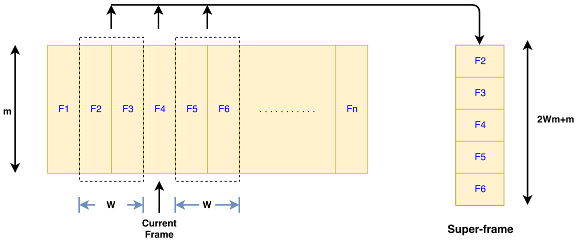

Fig. 2. Procedure of obtaining super-frames from the spectrogram.

3.2. Training: Learning dictionaries using restricted robust AA

For acoustic modeling, we have employed AA, which is a form of non-negative MF technique where the matrix containing compressed super-frames, $\mathbf{X}$, is decomposed as $\mathbf{X} = \mathbf{DA}$. The dictionary, $\mathbf{D}$, consists of the extremal points or archetypes, which lie on the convex hull of the data and are restricted to be the convex combination of the individual data points i.e., $\mathbf{D} = \mathbf{XB}, \mathbf{D} \in \mathbb{R}^{K\times d}$. Since archetypes model the convex hull, they provide more compact and meaningful representation of the data [22, 4]. To learn efficient dictionaries, the super-frame representation, $\mathbf{F}$, is only obtained from those audio regions where vocalizations are present. The vocalizations are extracted using a semi-supervised method proposed in our earlier study [3]. The presence of outliers can affect the performance of AA [11]. In the proposed framework, these outliers may arise due to noise, segmentation errors or wrong labels. To tackle this issue, we propose to use robust AA which effectively handles the outliers present in the data. In particular, the archetypal dictionary, $\mathbf{D}$, is learned by optimizing the following function [11]:

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

where $\mathbf{a}_i$ and $\mathbf{b}_j$ are the columns of $\mathbf{A} \in \mathbb{R}^{d \times l}$ and $\mathbf{B} \in \mathbb{R}^{l \times d}$, respectively. For any scalar u and constant ϵ, the Huber function is defined as, $h(u) = 1/2 \min_{w \geq \epsilon}[u^2/w + w]$. The weight w_i for $\mathbf{x}_i$ is defined as: $w_i = max(\|\mathbf{x}_i - \mathbf{X}\mathbf{B}\mathbf{a}_i\|_2, \epsilon)$. During optimization, w_i becomes larger for outliers and reduces their importance while finding archetypes. More details about robust AA can be found in [11].

Restricted AA: MF, in general, has large time/memory requirements. However, since archetypes lie on the boundary of the convex hull, it is possible to restrict the search of the archetypes to the set of data points, indexed by $\mathcal{B}$, that lie around the boundary. To achieve this speed up, we first minimize the following objective:

$$\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 \text{ s.t. } diag(\mathbf{C}) = 0,\ \mathbf{c}_i \succeq 0, \text{ and } \|\mathbf{c}_i\|_1 = 1, \quad (2)$$

where, $diag(.)$ denotes the diagonal elements. The solution $\mathbf{C} \in \mathbb{R}^{l \times l}$ (having columns $\mathbf{c}_i$), can be seen as the coefficient matrix for representing each exemplar ($\mathbf{x}_i$) in $\mathbf{X}$ as a linear combination of other exemplars [13]. The significant values of the solution refer to the points $\mathbf{x}_k$ ($k \in \mathcal{B}$) around the convex hull, due to the fact that they are selected iteratively by maximizing the negative gradient of the objective function defined in Eq. (2) with respect to each $\mathbf{c}_i$,

$$\underset{k}{\operatorname{argmax}}\ -\nabla = \mathbf{X}^T(\mathbf{x}_i - \mathbf{X}\mathbf{c}_i) = \langle \mathbf{x}_i, \mathbf{x}_k \rangle - \sum_l c_i^l \langle \mathbf{x}_l, \mathbf{x}_k \rangle, \quad (3)$$

where c_i^l is the lth coefficient of $\mathbf{c}_i$. By property of convex geometry, inner product between two points is maximized when one of the points is an extremal point [23]. As a result, the solution to Eq. (2) ensures that the union of the indicies of significant elements (having high magnitude) of each $\mathbf{c}_i$ refers to data points around the boundary. Note that problem in Eq. (2) can be solved using a fast quadratic programming solver, and our experiments reveal that performing AA on the reduced training set $\mathbf{X}[:, \mathcal{B}]$, is approximately $5\times$ faster than on the full $\mathbf{X}$.

Decreasing correlation between inter class dictionary atoms: In the above framework, each class-specific dictionary is learned independently. This often leads to a high correlation among inter-class dictionary atoms which may degrade the classification accuracy. Also, the number of optimal archetypes to be learned for each class is unknown. Although, as discussed earlier, one can use supervised dictionary learning approaches such as label-consistent KSVD [14], these approaches are computationally expensive and the underlying objective may not converge, especially when the number of classes are large. In contrast, we propose a computationally efficient alternative, where the learned class-specific dictionaries are pruned to form a final concatenated dictionary.

We choose a set of atoms from each class-specific dictionary which decreases the correlation between the inter/intra-class dictionary atoms. Let us denote the pruned dictionary of the q^{th} class by $\mathbf{D}^{*^q}$. The algorithm starts by finding the independent atoms from the dictionary $\mathbf{D}^1$ of the first class iteratively using the following metric:

$$i = \max_{i \notin \mathcal{Z}} \|\mathbf{d}_i^1 - \mathbf{D}_{\mathcal{Z}}^1 \mathbf{D}_{\mathcal{Z}}^{1+} \mathbf{d}_i^1\|_2^2 \text{ s.t. } \mathbf{D}_{\mathcal{Z}}^{1^T} \mathbf{D}_{\mathcal{Z}}^1 \text{ is invertible.} \quad (4)$$

Here $\mathbf{d}_i^1$ is an atom of $\mathbf{D}^1$, $+$ denotes the pseudo-inverse and $\mathbf{D}_{\mathcal{Z}}^1 \subset \mathbf{D}^1$, denotes the current set of selected atoms. Eq. (4) computes the distance of an atom $\mathbf{d}_i^1$ to the space spanned by the atoms in $\mathbf{D}_{\mathcal{Z}}^1$, and selects the one which does not lie in the span of $\mathbf{D}_{\mathcal{Z}}^1$. To choose K atoms from $\mathbf{D}^1$, Eq. (4) is iterated K times. Hence, a pruned dictionary, $\mathbf{D}^{*^1} \subset \mathbf{D}^1$, is obtained. This whole procedure is repeated for the dictionary of each class to find the independent atoms with respect to the selected atoms of all the previous classes. Our implementation uses the fast exemplar selection (FES) algorithm (see [24] for more details) based on block matrix and incremental Cholesky factor updates. Thus, we obtain the final pruned dictionary as a concatenation of class-specific independent atoms as $\mathbf{D}^* = [\mathbf{D}^{*^1} \mathbf{D}^{*^2} \ldots \mathbf{D}^{*^q}]$.

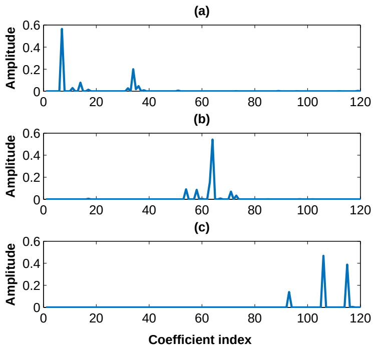

Fig. 3. CCSE obtained for a super-frame of bird species (a) black-throated tit, (b) black-yellow grosbeak and (c) black-crested tit.

3.3. Testing: Obtaining CCSE and classification

For classifying a given audio recording, bird vocalizations are segmented and the corresponding super-frames are converted into the compressed super-frame representation as described in Section 3.1. The random Gaussian matrix employed during training and testing is the same. For testing, we have used the generative classification approach, where a segmented test super-frame is classified to the class that gives the smallest reconstruction error. In particular, the final dictionary $\mathbf{D}^*$ is used to solve for the sparse/convex decomposition (CCSE), $\mathbf{a}_t = [\mathbf{a}_t^1 \ldots \mathbf{a}_t^q]$, for a given test compressed super-frame, $\mathbf{x}_t$, using an active-set QP solver from the SPAMS toolbox [25]. For each class, these CCSE are significantly different and can be used as features in any classification framework. This behavior is illustrated in Fig. 3. The figure depicts CCSE obtained for three super-frames of three different species. These CCSE are obtained using the pruned dictionary derived from the individual dictionaries of all three species (as described earlier). The pruned dictionary contains 40 atoms per class (the first 40 for black-throated tit, the next 40 for black-yellow grosbeak and the last 40 for black-crested tit) and in CCSE, these atoms exhibit higher amplitude for the super-frame belonging to their respective class. This provides a strong evidence of the presence of species-specific signatures.

Using CCSE, the reconstruction error of a super-frame, $\mathbf{x}_t$ belonging to the qth class, is calculated as: $\mathbf{r}_t^q = \|\mathbf{x}_t - \mathbf{D}^{*^q} \mathbf{a}_t^q\|_2$. Finally, to classify the input test recording, a voting rule is applied on the segmented super-frames of the input test recording.

4. PERFORMANCE EVALUATION

4.1. Dataset used

The proposed framework is evaluated on a dataset containing audio recordings of bird species found in the lower Himalayan regions. This dataset was collected at the Great Himalayan National Park (GHNP) located in North India and contains audio recordings from 26 different bird species. Each recording has a sampling rate of 44.1 kHz and is labeled by experienced birdwatchers. The duration of these recordings vary from 15 to 86 seconds. We segmented bird

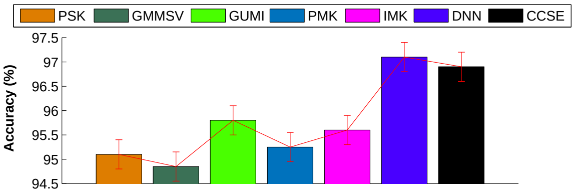

Fig. 4. Comparison of classification performance of different methods on GHNP dataset (averaged across three folds).

vocalizations from these recordings using the segmentation method proposed in [3]. These segmented vocalizations are used for training and testing.

4.2. Experimental setup and parameter selection

In all the experiments, each input recording is converted into a spectrogram by applying the STFT with 512 FFT points on 20 ms frames having 50% overlap. The parameter $W = 2$ is used to obtain super-frames, which are compressed via random projections to have a dimension of 500. For each class, 256 archetypes are learned using restricted robust AA. After applying the independent atom selection procedure, 40 atoms are chosen from each dictionary. All these parameters are obtained empirically. The performance of the proposed framework is measured in terms of classification accuracy and is compared with the performances of the DNN and dynamic kernels based SVM used in [17] for species identification. For comparison, five different dynamic kernels are used. These include : probabilistic sequence kernel (PSK), GMM supervector kernel (GMMSV), GMM-UBM mean interval kernel (GUMI), GMM-based pyramid match kernel (PMK) and GMM-based intermediate matching kernel (IMK). The DNN architecture (3 layers with 512 hidden units in each layer) proposed in [17] is also used here for comparison. The DNN and aforementioned kernels use Mel-frequency cepstral coefficients (MFCC) as the feature representation. The results reported here are with the optimal parameter settings for all the comparative methods. A three-fold cross-validation is used for experimentation and 33.33% of the vocalizations of each species are used for training and the rest are used for testing. The results reported here are averaged across all three folds.

4.3. Results and Discussion

Classification performance: The comparison of the classification accuracies of the proposed CCSE based framework with the DNN and SVM powered by dynamic kernels is depicted in Fig. 4. It can be observed that the proposed framework performs better than the SVM based approaches, whereas the DNN system performs slightly better (0.2% relative improvement) than the proposed framework. The relative improvements of 1.89%, 2.16%, 1.15%, 1.73% and 1.36% are observed by the proposed framework on PSK, GMMSV kernel, GUMI kernel, IMK and PMK, respectively. Although the performance of the DNN and the proposed framework is comparable, the proposed framework is computationally efficient as the number of trainable parameters are significantly less. Also, AA inherently requires less data to capture variations present in data [4]. This makes the proposed framework more suitable in conditions where the amount of labeled training data is small (as is common in bioacoustic problems).

Compression vs. classification trade-off: To establish the extent of compression that can be achieved in the super-frames, we experimented with different compression ratios and the results are shown in Fig. 5. It can be observed that one can achieve a 61% compression i.e. $K = 500$ from the original dimension of 1285 (obtained by using $W = 2$), without any decrease in the classification accuracy. This confirms the claim that given enough measurements of the sparse super-frame feature, it is possible to preserve the context information in bird vocalizations to a great extent. This is in contrast to existing approaches (such as in [18]) which embed the context information around a frame by averaging or max-pooling. The temporal and frequency modulations may not be fully modeled due to smearing out or loss of the detailed information which may degrade the classification performance. To establish this, experiments were also performed by averaging the super-frames (with no compression) followed by AA analysis for classification of bird species. A relative drop of 2.89% (96.9% to 94.1%) in accuracy was observed.

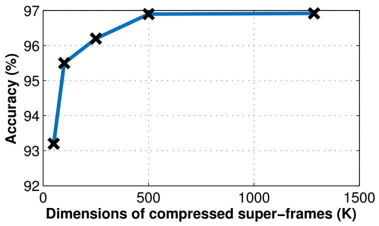

Fig. 5. Effect of compression on classification performance in the proposed framework.

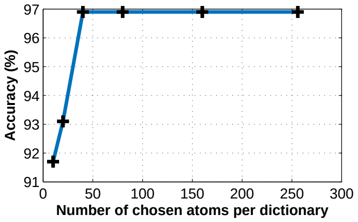

Fig. 6. Classification accuracy vs. the number of chosen atoms.

Size of pruned dictionary vs. classification trade-off: In this experiment, we analyzed the effect of different numbers of chosen atoms per dictionary (after pruning) on the classification performance while keeping other parameters fixed ($W = 2$ and $K = 500$). The results of this experiment are illustrated in Fig. 6, and it is confirmed that we are able to maintain the classification accuracy obtained in case of full dictionary (having 256 atoms), with just 40 atoms per dictionary. This justifies the use of pruning procedure in the proposed CCSE framework. This behavior can be beneficial when we are dealing with a large number of bird species.

5. CONCLUSION

In this work, we proposed CCSE based framework for bird species identification. The framework is based on a restricted version of robust AA which effectively models each bird species even in low data conditions. We also proposed an iterative procedure to choose atoms from each dictionary to decrease gross correlation among inter-dictionary atoms. This allowed us to decrease the size of dictionaries without degrading the classification performance. Future work may include to extend this framework for species classification on a large scale.

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] 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.

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

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

[5] Y. He, K. Kavukcuoglu, Y. Wang, A. Szlam, and Y. Qi, "Unsupervised feature learning by deep sparse coding," in *Proc. Int. Conf. Data Mining*, 2014, pp. 902–910.

[6] 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.

[7] N.-C. Wang, R. E. Hudson, L. N. Tan, C. E. Taylor, A. Alwan, and R. Yao, "Change point detection methodology used for segmenting bird songs," in *Proc. Int. Conf. Signal Info. Process.*, 2013, pp. 206–209.

[8] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *Trans. Info. Theory*, vol. 52, no. 9, pp. 4036–4048, September 2006.

[9] P. Frankl and H. Maehara, "The Johnson-Lindenstrauss lemma and the sphericity of some graphs," *Journal of Combinatorial Theory, Series B*, vol. 44, no. 3, pp. 355–362, 1988.

[10] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, March 2011.

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

[12] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and M. Yi, "Robust face recognition via sparse representation," *Trans. Pattern Anal. Machine Intel.*, vol. 31, no. 2, pp. 210–227, February 2009.

[13] V. Abrol, P. Sharma, and A. K. Sao, "Identifying archetypes by exploiting sparsity of convex representations," in *Workshop on The Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, June 2017.

[14] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Proc. Comp. Vis. Pattern Recog.*, 2011, pp. 1697–1704.

[15] A. Harma and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *Proc. Int. Conf. Acoust. Speech, Signal Process*, 2004, pp. 701–704.

[16] 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, and Lang. Process.*, vol. 16, no. 8, pp. 1541–1550, 2008.

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

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

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

[20] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *Trans. Pattern Anal. Mach. Intel.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[21] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[22] S. Seth and M. J. Eugster, "Archetypal analysis for nominal observations," *Trans. Pattern Anal. Mach. Intel*, vol. 38, no. 5, pp. 849–861, 2016.

[23] S. Mair, A. Boubekki, and U. Brefeld, "Frame-based data factorizations," in *Proc. Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 2305–2313.

[24] V. Abrol, P. Sharma, and A. K. Sao, "Fast exemplar selection algorithm for matrix approximation and representation: A variant oasis algorithm," in *Proc. Int. Conf. Acoust. Speech, Sig. Process.*, 2017, pp. 4436–4440.

[25] J. Mairal, "SPAMS toolbox," `http://spams-devel.gforge.inria.fr/doc/html/index.html`, Accessed: 2017-09-20.