APE: ARCHETYPAL-PROTOTYPAL EMBEDDINGS FOR AUDIO CLASSIFICATION

Arshdeep Singh, Anshul Thakur, Padmanabhan Rajan

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

ABSTRACT

Archetypal analysis deals with representing data points using archetypes, which capture the boundary of the data. Prototypal analysis deals with representing data points using prototypes, which capture the average behaviour of the data. Utilising these two complementary representations, we propose a simple, fixed-length representation for audio signals. We employ a well-studied method for determining archetypes, and utilise Gaussian mixture modelling to represent prototypes. Archetypes and prototypes are concatenated and utilised within a dictionary learning framework. Time-frequency representations of audio signals are projected on these dictionaries, under simplex constraints, to obtain the proposed archetypal prototypal embedding or APE. Experimental results on the tasks of bioacoustic classification and acoustic scene classification demonstrate the effectiveness of the APE representation for audio classification.

Index Terms— Archetypal analysis, bioacoustic classification, acoustic scene classification

1. INTRODUCTION

Archetypal analysis (AA) [1] provides an alternate viewpoint to represent multivariate data using "pure types" or *archetypes*. AA represents each point in a dataset as a convex combination of archetypes. Archetypes themselves are convex combinations of the points in the dataset. The constraints employed during the construction of the archetypes make them fall on the boundary (the convex hull) of the data points. Moreover, the representation of an archetype in terms of data points is also sparse. The previous two properties enable a useful interpretation of archetypes: they are points around the boundary of the data points, and are a combination of a few data points. They also provide information about the geometry or overall structure of the dataset as well [2]. In addition, archetypes are unique as there is no rotational ambiguity, invariance to scaling and affine transformations [3]. AA has found several applications in genetics and phytomedicine [4], market research and marketing [5], computer vision [6] and a few in audio analysis as well [7] [8] [9].

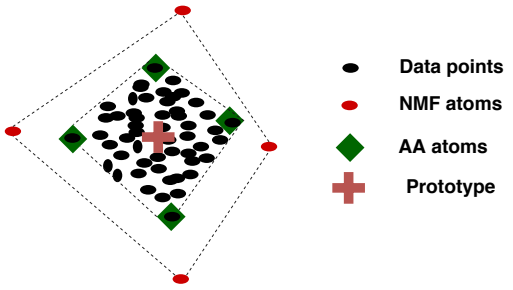

Fig. 1. Schematic illustration showing the atoms learned using NMF, AA and prototypal analysis [7]

Under non-negative data conditions, AA can also be considered as a non-negative matrix factorization (NMF) method. NMF has been widely used in audio analysis to obtain state-of-the-art results for a multitude tasks, including noise-robust automatic speech recognition [10], music transcription [11] and speech separation [12]. The representation learned using NMF covers the simplicial cone and there is no simplex constraint during the learning process [13]. This may result in the representation learnt using NMF to reside outside the given subspace [7]. Fig.1 schematically illustrates the atoms learned using AA and NMF, given a data space.

As mentioned earlier, AA models the convex hull or the boundary of the data. This can be related to extreme value theory (EVT) [14] in statistics. EVT deals with the modelling of extremal data points (the data points that are far away from average points and having low frequency of occurrence). Extreme data points play an important role in several applications such as structural engineering, finance, earth sciences, traffic prediction, and geological engineering [14].

Knowing the extreme points can help data analysis in two ways: (i) it can give structural behaviour of the data without assuming any prior distribution of the data, (ii) it can provide information about rare events which are far apart from the most commonly occurring events [14]. In many situations, rare events corresponding to the tail of the distribution might provide important cues which can help in discriminating events of interest.

978-1-5386-5477-4/18/$31.00

However, considering only the extremal or structural behaviour can ignore the average behaviour of the data. In this regard, many studies model the average or representative behaviour of the data points by using *prototypes*. The prototype vectors can capture the central trends where most of the data lies [15]. A commonly used prototype representation is by using mean vectors. For example, for multimodal data, Gaussian mixture models (GMM) can be used to interpret each data point in terms of prototypes, the prototypes being the mean vectors of each mixture component. Fig. 1 also illustrates the prototype of the data space.

In this paper, we propose to interpret data points by combining extremal data vectors (archetypes) and average data vectors (prototypes.) By utilising both these representations within a dictionary learning framework, we are able to capture both structural and average behaviour of data points. This results in data representations which are better able to discriminate individual acoustic events in two audio classification tasks. The representation is termed archetypal-prototypal embedding or APE. We demonstrate that the combination of both structural behaviour and average behaviour for representation provides a significant improvement over either of them alone.

The rest of this paper is organized as follows. In section 2, we briefly review archetypal analysis. Subsequently, the proposed method is described in section 3. Performance evaluation and conclusions are included in sections 4 and 5 respectively.

2. BACKGROUND

2.1. Archetypal analysis

Archetypal analysis (AA) [1], proposed by Cutler and Breiman, is a simple and intuitive way to understand multivariate data. AA represents every point in the data set as a convex combination of archetypes. The archetypes themselves are imposed to be convex combinations of individual points in the data set. Under non-negative data conditions, AA can be considered as a form of NMF. AA approximates the data points by a convex combination of a few archetypes and produces sparse representations [9].

Given a set of n data points $\{\mathbf{x_j}\}$ in d-dimensional space, AA aims to find a subset of p archetypes $\{\mathbf{z_k}, 1 \leq k \leq p < n\}$. The archetypes are defined in such a way that (a) $\{\mathbf{x_j}\}$ can be approximated as a convex combination of $\{\mathbf{z_k}\}$ as given in equation 1, (b) each archetype can also be represented as a convex combination of data points as given in equation 2.

$$\mathbf{x_j} \approx \sum_{k=1}^{p} \alpha_{kj}\mathbf{z_k} \text{ s.t } \alpha_{kj} \geq 0, \sum_{k=1}^{p} \alpha_{kj} = 1 \tag{1}$$

$$\mathbf{z_k} \approx \sum_{j=1}^{n} \beta_{kj}\mathbf{x_j} \text{ s.t } \beta_{kj} \geq 0 \, , \sum_{j=1}^{n} \beta_{kj} = 1 \tag{2}$$

With the above definition of archetypes, a set of p archetypes $\{\mathbf{z_k}\}$ can be obtained which minimize the residual sum of square $RSS(p)$ as given in equation 3.

$$RSS(p) = \sum_{j=1}^{n} ||\mathbf{x_j} - \sum_{k=1}^{p} \alpha_{jk}\mathbf{z_k}||^2 \tag{3}$$

The above equation can also be written as a matrix factorization problem:

$$\min_{\substack{\boldsymbol{\alpha}_j \in \triangle_p \text{ for } 1 \leq j \leq n \\ \boldsymbol{\beta}_k \in \triangle_n \text{ for } 1 \leq k \leq p}} ||\mathbf{X} - \mathbf{ZA}||_F^2 \tag{4}$$

$$\triangle_p := \left\{ \boldsymbol{\alpha} \in \mathbb{R}^p \text{ s.t } \boldsymbol{\alpha} \geq 0 \text{ and } \sum_{k=1}^{p} \alpha[k] = 1 \right\} \tag{5}$$

$$\triangle_n := \left\{ \boldsymbol{\beta} \in \mathbb{R}^n \text{ s.t } \boldsymbol{\beta} \geq 0 \text{ and } \sum_{j=1}^{n} \beta[j] = 1 \right\} \tag{6}$$

In the above, $\mathbf{X} \in \mathbb{R}^{d \times n}$ is data matrix , $\mathbf{Z} = \mathbf{XB}$ is the archetypal dictionary of p archetypes $\{\mathbf{z_k}\}$, $\mathbf{B} = \{\boldsymbol{\beta}_k; \boldsymbol{\beta}_k \in \triangle_n\}$, $\mathbf{A} = \left\{\boldsymbol{\alpha}_j; \boldsymbol{\alpha}_j \in \triangle_p\right\}$, and $||.||_F$ denotes the Frobenius norm. Equation 4 is a constrained non-linear least square optimization problem with simplex constraints and squared loss function.

One variant for the AA algorithm is proposed by [9], to make archetypal analysis robust against outliers as follows:

$$\min_{\substack{\boldsymbol{\alpha}_j \in \triangle_p \text{ for } 1 \leq j \leq n \\ \boldsymbol{\beta}_k \in \triangle_n \text{ for } 1 \leq k \leq p}} \sum_{j=1}^{n} h(||\mathbf{x_j} - \mathbf{Z}\boldsymbol{\alpha_j}||_2), \tag{7}$$

Here, h is a Huber loss function, and is used as compared to the squared loss function as mentioned in equation 4. As given in equation 8, $h : \mathbb{R} \rightarrow \mathbb{R}$, is defined for a scalar u in $\mathbb{R}$ and ϵ, a positive constant. The loss function penalizes outliers which differs significantly from the rest of the data.

$$h(u) = \begin{cases} \frac{u^2}{2\epsilon} + \frac{\epsilon}{2}, if \ |u| \leq \epsilon \\ |u|, \text{otherwise} \end{cases} \tag{8}$$

2.2. Prototypal analysis

Prototypal analysis aims to learn one or more representative (or prototype) vectors from a set of data points. Various measures of central tendency including the mean, median or mode could serve as prototypes. A simple interpretation of prototypes for multimodal data can be made using Gaussian mixture models. For acoustic data, the mean vectors of GMMs can also be viewed as the centriods of various acoustic clusters. By utilising the standard expectation-maximization algorithm, the GMM components can be estimated, and the component mean vectors serve as prototypes.

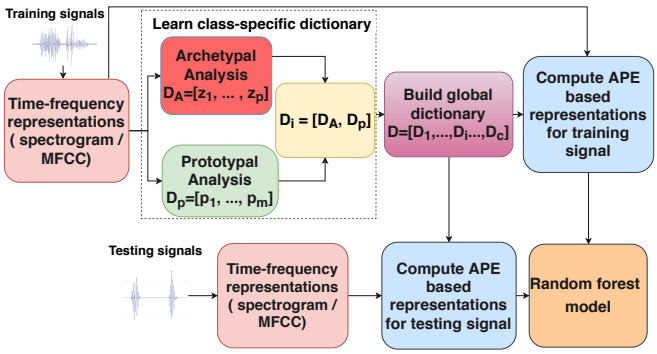

Fig. 2. Schematic illustration of the training and testing modules of APE-based proposed framework for audio classification. After learning c class-specific dictionaries, a global dictionary D is obtained.

3. PROPOSED METHOD

In this section, we describe the steps involved in converting an audio recording into the archetypal-prototypal embeddings (APE) representation. The overall flow of the proposed framework is shown in Fig. 2.

The time-frequency representations of an audio signal are extracted $\in \mathbb{R}^{d \times N_i}$, where d corresponds to the number of frequency bins and N_i is the number of time frames for the i^{th} audio signal. A class-specific data matrix $\mathbf{X}$ is constructed by concatenating the time-frequency representations of class-specific audio signals. Once the data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ (n is the total number of time frames of from all examples from a given class) is formed, archetypal analysis and prototypal analysis is performed on the data matrix to learn archetypes and prototypes.

3.1. Training: Class-specific dictionary learning

A set of p archetypes are obtained by solving the non-convex optimization problem mentioned in equation 7 with simplex constraints using the robust archetypal analysis algorithm proposed in [9]. This utilizes an active-set method to update $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$. The algorithm gives $\mathbf{A}$ and $\mathbf{B}$ decomposition matrices. The archetypes are computed using $\mathbf{Z} = \mathbf{XB}$.

To obtain the prototypes, GMM with m mixtures is trained using the class-specific data matrix $\mathbf{X}$. This gives m class-specific representative prototype vectors. Once the archetypes and prototypes are learned, a class-specific dictionary $\mathbf{D_i} = [\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_p}, \mathbf{p_1}, ..., \mathbf{p_m}] \in \mathbb{R}^{d \times (p+m)}$ is obtained by concatenating the class-specific archetypes and prototypes.

To visualize the archetypes and prototypes, principal component analysis is performed on class-specific data matrices. Fig. 3-5 shows the proposed dictionary atoms for three acoustic scene classes namely "tram", "library" and "beach" respectively [16].

3.2. Computing APE, feature representation and classification

After learning the class-specific dictionary $\mathbf{D_i}$, a global dictionary $\mathbf{D} = [\mathbf{D_1}, \mathbf{D_2}, ..., \mathbf{D_i}, ..., \mathbf{D_c}] \in \mathbb{R}^{d \times c(p+m)}$, where c is the number of classes, is obtained. Given the learned dictionary $\mathbf{D}$ and the frame-wise time-frequency representation ($\mathbf{x_j} \in \mathbb{R}^d$) of an audio signal, archetypal-prototypal embeddings $\boldsymbol{\gamma}_j \in \mathbb{R}^{c(p+m)}$ are computed by solving the following optimization problem under simplex constraints [9].

$$\min \sum_{j=1}^{n} ||\mathbf{x_j} - \mathbf{D}\boldsymbol{\gamma}_j||_2^2 \; s.t \; ||\boldsymbol{\gamma}_j||_1 = 1 \; , \boldsymbol{\gamma}_j \geq 0 \qquad (9)$$

The APE representation $\in \mathbb{R}^{c(p+m)}$) of the complete audio signal is computed by averaging frame-wise APE representations. Finally, the APE representation is used to train and evaluate a classifier.

The 2-dimensional t-SNE plot for the three acoustic scene classes shown earlier is given in Fig. 6. APE representations of 2400 dimensions were reduced to 2 dimensions for visualization.

4. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed APE based representations for two audio classification tasks. The first task is the classification of bird species from their calls, and the second is the classification of acoustic scenes.

4.1. Bird Species Classification

4.1.1. Dataset Used

Audio recordings containing vocalizations of 50 different bird species are used for performance evaluation. These audio recordings are obtained from the Great Himalayan national park (GHNP), in north India, the UCLA Art & Science center [17] and the Macaulay Library of Cornell University. All the recordings available are 16-bit WAV files having a sampling rate of 44.1 kHz, with the duration ranging from 18 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`.

4.1.2. Experimental setup

The compressed super-frame (CSF) representation, proposed in [18], is used for parameterization of bird vocalizations. These CSFs are obtained by concatenating the neighbouring frames of spectrograms and compressing the resultant vector using random projections. The spectrogram is obtained by utilizing a frame size of 20 ms with a 50% overlap, and has

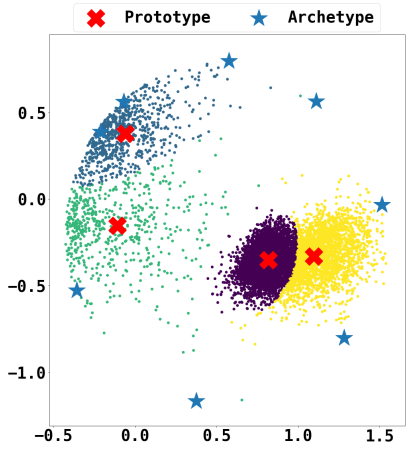

Fig. 3. Tram class: archetypes and prototypes (different color indicates different clusters)

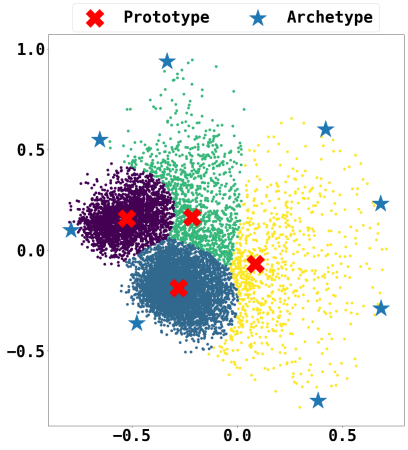

Fig. 4. Library class: archetypes and prototypes (different color indicates different clusters)

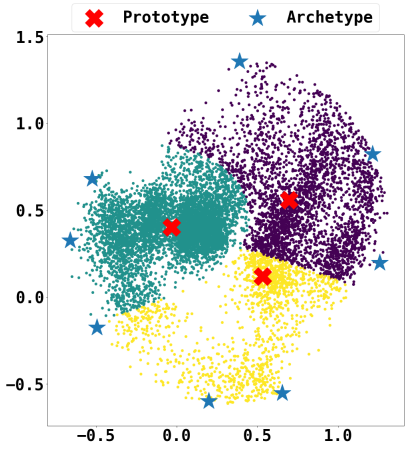

Fig. 5. Beach class: archetypes and prototypes (different color indicates different clusters)

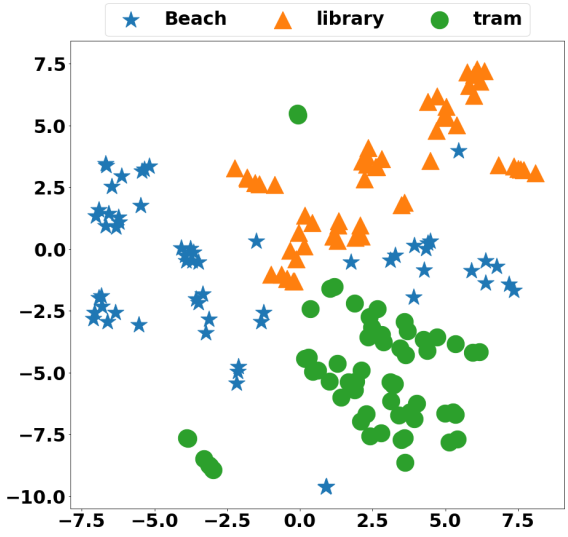

Fig. 6. t-SNE plot showing global representations for acoustic scene classes (a) Beach (b) Library (c) Tram

257 frequency bins. The 5 neighbouring frames of spectrogram are concatenated to obtain a high dimensional representation. This representation is projected on a 500-dimensional space to obtain CSFs, using a Gaussian random matrix.

A three fold cross validation is used for evaluation. For each fold, 33% of the bird vocalizations (from each class) are used for training while remaining are used for evaluation. Out of these 33%, 75% of the vocalizations used for learning the dictionary and the rest are used for obtaining the APE. For each class, 64 archetypes and a 5-component GMM is trained. This gives rise to a 69-atom class-specific dictionary. A random forest classifier with 100 trees is utilized on APE representations to obtain the final classification decisions.

The classification performance of APE based framework is compared with various existing methods such as dynamic kernel based support vector machine (SVM), deep neural networks (DNN) [19] and compressed convex spectral embeddings (CCSE) [18]. Dynamic kernels such as intermediate matching kernel (IMK), probabilistic sequence kernel (PSK) and GMM-UBM mean interval (GUMI) kernel are used. The CCSE framework utilizes archetypal analysis alone to obtain the dictionaries and can be seen as a subset of the proposed APE framework. The parameters used in both the proposed framework and the comparative methods are optimized empirically to obtain the best classification performance.

4.1.3. Results and Discussion

Fig. 7 depicts the performance of APE based representations and the comparative methods for the task of bird species classification. The analysis of this figure highlights that APE based representations show comparable performance to the existing bird species classification frameworks. As expected, APE outperforms the GMM baseline and CCSE by 10.51% and 1.58% respectively. This justifies the utilization of GMM means and archetypes in the proposed framework for capturing average and extremal behaviour simultaneously. Apart from that, APE based representations show better classification performance than dynamic kernels. However, DNN shows an improvement of 1.36% over APE.

4.2. Acoustic Scene Classification

4.2.1. Dataset Used

The TUT acoustic scene classification DCASE 2016 development dataset [16] consists of 15 acoustic scene classes. Each audio recording is of 30 seconds length and is recorded at a

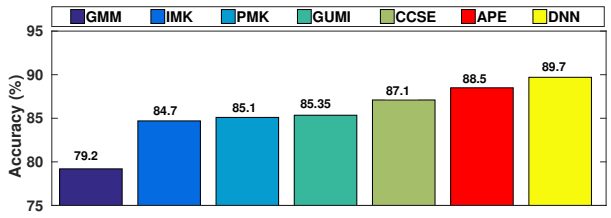

Fig. 7. Comparison of classification performance of different methods on 50 bird species

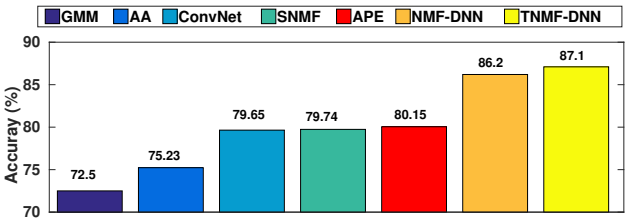

Fig. 8. Comparison of classification performance of different methods on DCASE'16 ASC development dataset

sampling rate of 44.1 kHz. The dataset consists of 1170 audio samples.

4.2.2. Experimental setup

The time-frequency representations of acoustic scene signals are computed using mel-frequency cepstral coefficients, delta and acceleration coefficients as proposed in [16].

The performance of the proposed APE framework is measured in terms of classification accuracy. A 4-fold cross-validation is performed as per the DCASE'16 protocol. For each class, 128 archetypes and 32-component GMM is learned. This gives rise to 160 class-specific dictionary atoms. The obtained APE representations are used to train a random forest classifier with 100 trees.

The proposed method is compared with the performances of the GMM baseline proposed in [16], archetypal analysis (AA), supervised NMF, convolution neural network (ConvNet) proposed in [20], NMF-DNN and TNMF-DNN as proposed in [21]. The SNMF and ConvNet use log mel-frequency representations whereas the NMF-DNN and the TNMF-DNN use constant-Q transform (CQT) representations.

4.2.3. Results and Discussion

Fig. 8 depicts the performance of APE based representation and the comparative methods for the task of acoustic scene classification. As expected, it can be observed that APE based proposed framework outperforms the GMM baseline and AA by 10.34% and 3.45% respectively. This shows the utility of the proposed framework to incorporate both average and extremal behaviour simultaneously. Apart from this, APE based representations performs comparable as that of SNMF and ConvNet even without incorporating class-specific information while learning features. However, the NMF-DNN and TNMF-DNN shows an improvement of around 8.6% over APE. This is possibly due to the better time-frequency representation afforded by CQT, rather than MFCC, which was developed for human speech.

5. CONCLUSION

In this paper, we proposed the archetypal-prototypal embeddings within a dictionary learning framework for audio classification. As shown in the experimental evaluations, the proposed representations incorporate the structural and average behaviour of multivariate data effectively. The proposed framework is simple and intuitive in terms of representation and interpretability of the learnt dictionary atoms.

Future work will look at incorporating improved time-frequency representations before learning the archetypes or prototypes. The sensitivity of the method to various hyper parameters also need to be evaluated.

6. REFERENCES

[1] Adele Cutler and Leo Breiman, "Archetypal analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, 1994.

[2] Sohan Seth and Manuel JA Eugster, "Archetypal analysis for nominal observations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 849–861, 2016.

[3] Morten Mørup and Lars Kai Hansen, "Archetypal analysis for machine learning," in *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*. IEEE, 2010, pp. 172–177.

[4] Juliane Charlotte Thøgersen, Morten Mørup, Søren Damkiær, Søren Molin, and Lars Jelsbak, "Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways," *BMC bioinformatics*, vol. 14, no. 1, pp. 279, 2013.

[5] Christian Bauckhage and Kasra Manshaei, "Kernel archetypal analysis for clustering web search frequency time series," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 1544–1549.

[6] Shahzad Cheema, Abdalrahman Eweiwi, Christian Thurau, and Christian Bauckhage, "Action recognition by

learning discriminative key poses," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1302–1309.

[7] Aleksandr Diment and Tuomas Virtanen, "Archetypal analysis for audio dictionary learning," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*. IEEE, 2015, pp. 1–5.

[8] V Abrol, P Sharma, A Thakur, P Rajan, AD Dileep, and Anil K Sao, "Archetypal analysis based sparse convex sequence kernel for bird activity detection," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 1774–1778.

[9] Yuansi Chen, Julien Mairal, and Zaid Harchaoui, "Fast and robust archetypal analysis for representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1478–1485.

[10] Felix Weninger, Martin Wöllmer, Jürgen Geiger, Björn Schuller, Jort F Gemmeke, Antti Hurmalainen, Tuomas Virtanen, and Gerhard Rigoll, "Non-negative matrix factorization for highly noise-robust asr: To enhance or to recognize?," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4681–4684.

[11] Nancy Bertin, Roland Badeau, and Emmanuel Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 538–549, 2010.

[12] Mikkel N Schmidt and Rasmus K Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Ninth International Conference on Spoken Language Processing*, 2006.

[13] Tongliang Liu, Mingming Gong, and Dacheng Tao, "Large-cone nonnegative matrix factorization," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 9, pp. 2129–2142, 2017.

[14] Kshitij Sharma, Valérie Chavez-Demoulin, and Pierre Dillenbourg, "An application of extreme value theory to learning analytics: Predicting collaboration outcome from eye-tracking data," *Journal of Learning Analytics*, vol. 4, no. 3, pp. 140–164, 2017.

[15] Arnulf BA Graf, Olivier Bousquet, Gunnar Rätsch, and Bernhard Schölkopf, "Prototype classification: Insights from machine learning," *Neural computation*, vol. 21, no. 1, pp. 272–300, 2009.

[16] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1128–1132.

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

[18] 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.*, April, 2018.

[19] 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.

[20] Alain Rakotomamonjy and Alain Rakotomamonjy, "Supervised representation learning for audio scene classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 6, pp. 1253–1265, 2017.

[21] Victor Bisot, Romain Serizel, Slim Essid, and Gaël Richard, "Leveraging deep neural networks with non-negative representations for improved environmental sound classification," in *IEEE International Workshop on Machine Learning for Signal Processing MLSP*, 2017.