ARCHETYPAL ANALYSIS FOR AUDIO DICTIONARY LEARNING

Aleksandr Diment, Tuomas Virtanen

Tampere University of Technology
Department of Signal Processing
Korkeakoulunkatu 1, 33720, Tampere, Finland
firstname.lastname@tut.fi

ABSTRACT

This paper proposes dictionary learning with archetypes for audio processing. Archetypes refer to so-called pure types, which are a combination of a few data points and which can be combined to obtain a data point. The concept has been found useful in various problems, but it has not yet been applied for audio analysis. The algorithm performs archetypal analysis that minimises the generalised Kullback-Leibler divergence, shown suitable for audio, between an observation and the model. The methodology is evaluated in a source separation scenario (mixtures of speech) and shows results, which are comparable to the state-of-the-art, with perceptual measures indicating its superiority over all of the competing methods in the case of medium-size dictionaries.

Index Terms— archetypes, audio analysis, non-negative matrix factorisation, sparse representation

1. INTRODUCTION

The increasing amount of audio in the surrounding digital world motivates a need for developing methods for its analysis and content-based processing. Methods that address the problem of characterising meaningful audio units as non-negative linear combinations of dictionary atoms have been attracting attention over the past years.

Non-negativity plays an important role in creating an intuitive representation of audio: it enables the so-called compositional models, which essentially assume sounds to be of compositional nature. These models attempt to explain the audio as non-negative linear combinations of the dictionary atoms and are able to model sound mixtures, which are naturally constructive: purely additive, cancellation takes place only intentionally. For acquiring a non-negative decomposition of a mixed signal, such methods as non-negative matrix factorisation (NMF) [1] and non-negative sparse representations [2] have been proposed and applied for various problems. In audio analysis, they have been used to obtain state-of-the-art results in multitude tasks, including noise-robust automatic speech recognition [3, 4], voice conversion [5], music transcription [6] and speech separation [7].

NMF obtains models of the components in a form of fixed magnitude or power spectra, which are, naturally, non-negative. The procedure is to minimise the reconstruction error between an observed magnitude or power spectrum and its reconstruction from the model, while enforcing non-negativity on the matrix entries. Compared to vector quantisation (VQ) and principle component analysis (PCA), NMF learns parts-based representations, which allow only additive combinations [8].

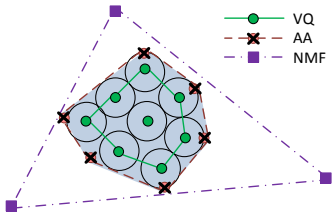

Figure 1: A schematic illustration of differences between dictionaries learnt with NMF, VQ, and archetypal analysis (AA). The observed data points are within the shaded region, and the large circles stand for groups of points, whose centroids appear in VQ. The lines correspond to the convex hulls spanned by respective learnt factors. AA learns a convex hull that compactly represents the observed data.

Archetypes are a novel, relatively recently emerging dictionary learning approach, which conceptually stands for representing extreme data values in such a way that the data points in turn can be represented using archetypes. Archetypal analysis performs discovery of latent factors in high-dimensional data. It has been proposed [9] as an alternative to PCA, and compared to the latter, factors learnt by archetypal analysis (archetypes) are to be convex combinations of data points. Archetypal analysis is strongly connected to NMF (it provides non-negative factorisation given non-negative data) and sparse representations (the approximation of the data points by convex combination of a few archetypes is sparse) [10]. The exploration of the concept started [9] on examples of modelling head dimensions with extreme types, resulting in faces of real individuals, represented as a mixture of the archetypes. Other early examples include air pollution data, used to represent a few "prototypical days". Later on, a variation of archetypal analysis, the so-called moving archetypes, was proposed for separating the travelling and stationary parts of moving structures, such as travelling waves [11].

The main benefit of archetypal analysis is the intuitivity and interpretability of the obtained models. The convex combinational association between the learnt factors and the data points enables a simple way of looking at multivariate data and providing information about its structure. Until very recently the method attracted a limited research interest because of its computational complexity. Recently, Chen *et al.* [10] proposed a computationally efficient optimisation scheme and developed its open-source implementation. Additionally, they showed the efficiency of the technique for such machine

T. Virtanen has been funded by the Academy of Finland, grant number 258708.

learning problems as computer vision by codebook learning and visualisation of the requests to a large image collection. In the latter, archetypal analysis provided a set of both intuitive and less expected images that constitute the typical outputs of a query by a keyword.

The principal difference between the mechanisms behind the popular approaches for dictionary learning, such as NMF, vector quantisation (VQ) and archetypal analysis, is schematically depicted in Figure 1. NMF will learn dictionary atoms that are outside observation subspace. VQ learns atoms that are inside it. Archetypal analysis learns atoms that are on the boundaries and therefore leads to a more compact convex hull spanned by the atoms, potentially providing a more accurate representation. Considering a special case of the number of archetypes set equal to the number of data points, the approach yields archetypes being the data points, in which sense it becomes similar to the sparse non-parametric representation proposed for audio source separation in [12].

The standard archetypal analysis [9] minimises the Euclidean distance between observations and the model. In the case of audio analysis, minimising the Kullback-Leibler (KL) divergence (within the NMF framework) has been shown more effective due to its capability to capture the non-linear nature of the contributions of frequency bands [13, 4]. Archetypal analysis, only recently revisited, has not yet been applied for audio analysis. The expected benefit of analysing audio by means of this method is, similarly to other applications, within the intuitivity of the obtained models, as well as potentially higher accuracy.

This paper investigates the relevance of archetypes for audio analysis. We propose to perform the archetypal analysis that minimises KL divergence. The performance of the method is evaluated in a task of representing spectra of mixtures of speech. The primary contribution is in showing usefulness of archetypal analysis for audio and its competitiveness against the state-of-the-art, including the conventional NMF, exemplar-based representation and VQ.

Firstly, a brief overview of the concept of archetypes and the properties of archetypal analysis are presented in Section 2. The algorithm for audio analysis is proposed in Section 3 and evaluated in terms of its capability to represent mixtures of speech in Section 4, which also discusses the obtained results. Finally, the conclusions on the applicability of the proposed methodology along with future research suggestions are drawn in Section 5.

2. ARCHETYPAL ANALYSIS

Archetypal analysis produces sparse representation of N data points $\mathbf{x}_i$, $i = 1, \ldots, N$ of the dataset $\{\mathbf{x}_i\}$ by approximating them with convex combinations of extreme types, or archetypes $\mathbf{z}_1, \ldots, \mathbf{z}_P$. Archetypes, in turn, are defined as a convex combination of the data points. The requirement of the archetypes to be representable by a mixture of data points and data points as a mixture of archetypes is what makes this approach different from principle components.

Archetypes are located on the boundary of data points that minimises residual sum of squares (RSS). In the case the number of archetypes P is set to one, it is located at the sample mean, and in the case P equals the sample size, the archetypes are the data points.

The algorithm for computing archetypes is referred to as the archetype algorithm. It attempts to find archetypes $\mathbf{z}_1, \ldots, \mathbf{z}_P$ as the linear combination of data points weighted by β_{pj} as

$$\mathbf{z}_p = \sum_{j=1}^{N} \beta_{pj} \mathbf{x}_j \tag{1}$$

and $\{\alpha_{ip}\}$ to minimise the residual sum of squares

$$\mathrm{RSS} = \sum_{i=1}^{N} \left\| \mathbf{x}_i - \sum_{p=1}^{P} \alpha_{ip} \mathbf{z}_p \right\|^2 \tag{2}$$

subject to the following constraints of α_{ip} and β_{pj}:

$$\alpha_{ip} \geq 0 \quad \text{for} \quad i = 1, \ldots, N,\ p = 1, \ldots, P \tag{3}$$

$$\textstyle\sum_{p=1}^{P} \alpha_{ip} = 1 \quad \text{for} \quad i = 1, \ldots, N \tag{4}$$

$$\beta_{pj} \geq 0 \quad \text{for} \quad p = 1, \ldots, P,\ j = 1, \ldots, N \tag{5}$$

$$\textstyle\sum_{j=1}^{N} \beta_{pj} = 1 \quad \text{for} \quad p = 1, \ldots, P. \tag{6}$$

Here, α_{ip} is the weight of an archetype $\mathbf{z}_p$ applied to obtain a data point $\mathbf{x}_i$, and β_{pj} is the weight of a data point $\mathbf{x}_j$ applied to obtain an archetype $\mathbf{z}_p$. The archetype problem is to find α's and β's to minimise the RSS subject to the constraints (3)–(6) [9]:

$$\alpha^\star, \beta^\star = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \left\| \mathbf{x}_i - \sum_{p=1}^{P} \alpha_{ip} \sum_{j=1}^{N} \beta_{pj} \mathbf{x}_j \right\|^2. \tag{7}$$

3. ARCHETYPAL ANALYSIS FOR AUDIO

Generalized Kullback-Leibler (KL) divergence is found better than the Euclidean distance in audio applications [13], with benchmarks performed *e.g.* for speech separation [4] and sparse coding using musical instrument templates [14]. The KL divergence between matrices $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathrm{KL}(\mathbf{X}, \mathbf{Y}) = \sum_{i,j} x_{i,j} \log \frac{x_{i,j}}{y_{i,j}} - x_{i,j} + y_{i,j}. \tag{8}$$

It is preferred over the Euclidean distance in audio analysis since it de-emphasises high-energy observations in the decomposition and it is able to distinguish between the noise floor and higher-energy components. We propose a novel algorithm for archetypal analysis minimising the KL divergence, presented in Algorithm 1.

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ (N–number of samples, M–sample dimensionality), number of archetypes P.
[Initialisation]:
Each entry β_{jp} of $\mathbf{B} \in \mathbb{R}^{N \times P}$ with random positive value.
Normalise so that $\boldsymbol{\beta}_p \in \Delta$.
Similarly for each entry α_{pi} of $\mathbf{A} \in \mathbb{R}^{P \times N}$.
repeat
$\mathbf{A} \leftarrow \mathbf{A}. * \dfrac{\mathbf{B^T} \cdot \mathbf{X^T} \cdot (\frac{\mathbf{X}}{\mathbf{XBA}})}{\mathbf{B^T} \cdot \mathbf{X^T} \cdot \mathbf{1}}$
Normalise each column of $\mathbf{A}$ to unity (L_1-norm).
$\mathbf{B} \leftarrow \mathbf{B}. * \dfrac{\mathbf{X^T} \cdot (\frac{\mathbf{X}}{\mathbf{XBA}}) \cdot \mathbf{A^T}}{\mathbf{X^T} \cdot \mathbf{1} \cdot \mathbf{A^T}}$
Normalise each column of $\mathbf{B}$ to unity (L_1-norm).
until *convergence, decided on by the relative change of* (8).
Output: Basis matrix $\mathbf{XB} \in \mathbb{R}^{P \times M}$.

Algorithm 1. *Archetypal analysis minimising KL divergence. Operation .* is entry-wise multiplication, and divisions are entry-wise.*

For the number of iterations, a threshold of 100 can be set. During preliminary experiments, a clear saturation of the value of divergence (8) was observed at the iteration count of 40–60.

4. EVALUATION

The proposed algorithm for archetypal analysis is evaluated against other dictionary learning methods in a supervised source separation task. Source separation refers to estimating the individual source signals of which a mixture is composed. The scenario addressed in the evaluation is representing mixtures of speech magnitude spectrograms. Methods that utilise dictionary-based representations currently give state-of-the-art results in supervised separation [15]. Dictionary learning is motivated by the idea that for each sound source there is a separate set of atoms. Given two dictionaries $\mathbf{D}^{(s)}$, $s = 1, 2$, an observed mixture vector $\mathbf{x}$ can be modelled as

$$\mathbf{x} \approx \mathbf{D}^{(1)}\mathbf{w}^{(1)} + \mathbf{D}^{(2)}\mathbf{w}^{(2)}, \tag{9}$$

where weight vectors $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ are constrained to be entry-wise non-negative. Reconstruction of individual source can be obtained as

$$\hat{\mathbf{x}}^{(s)} = \mathbf{D}^{(s)}\mathbf{w}^{(s)}, \tag{10}$$

or by the Wiener filtering approach described in Section 4.3.

4.1. Acoustic data

The evaluation setup is identical to [15]. The subset of the GRID corpus [16] used as the training set in the Speech Separation Challenge [17] served as the evaluation data. The dataset is comprised of 500 simple sentences from each of 34 speakers. An evaluation subset was generated consisting of 100 mixtures from two speakers, randomly selected for each mixture, pronouncing random sentences. The lengths of the sentences were equalised by zero-padding. Each speaker signal was normalised to equal RMS level prior to mixing.

Short-time spectrograms, being a standard tool employed for audio content analysis, were used to represent the data. A frame blocking and windowing were performed with the following parameters: Hanning window, frame length 60 ms, frame hop of 15 ms. From each frame, a magnitude spectrum was obtained by taking the absolute value of the discrete Fourier transform. This yields non-negative frame-wise feature vectors of length 751.

4.2. Dictionaries

The dictionaries were constructed in such a way, that for each speaker an individual set of atoms is acquired. The training data for each speaker consisted of all the sentences uttered by the speaker except the ones used for generating the evaluation mixtures.

Prior to the training, the magnitude spectra were normalised to sum to unity in order to make the dictionaries gain-independent. From the spectra the dictionaries were learnt with the evaluated methods. Dictionaries of nine different sizes were considered: 2, 3, 5, 10, 25, 50, 75, 100 and 1000 atoms per speaker.

The proposed technique of archetypal analysis for audio was evaluated in comparison to the state-of-the-art methodology applied in the field: vector quantisation (VQ) [15], NMF [7], as well as exemplar-based representations [3]. Each of the methods was used to obtain a different dictionary. The conceptional difference between the methods is in the way they select the training samples for acquiring dictionary atoms. In the case of VQ, groups of training samples are represented by their centroid points to obtain a dictionary that best represents the training data. NMF (the approach of minimising the KL divergence was used in this case) finds a dictionary that best represents the training data when atoms are linearly combined. The exemplar-based approach [18] randomly samples the training data.

Additionally, the proposed method was compared in the evaluation with another, existing implementation of the archetypal analysis, which is distributed as part of the SPAMS toolbox [10]. The SPAMS implementation was previously applied for problems, other than audio analysis; it is robust against outliers due to the Huber loss function, and operates on the Euclidean norm as in (7), in contrast to the proposed operation in terms of KL divergence.

With each method, two subcases were considered with the sparseness cost parameter λ values of 0 and 1. In the case of NMF, the dictionary size 1000 was omitted from the evaluation since regular NMF is not capable of meaningfully handling an overcomplete dictionary.

4.3. Source estimation and evaluation

The mixtures are composed of sentences uttered by two speakers. The weights $\mathbf{w}$ of dictionary atoms $\mathbf{B}$ are obtained by means of Active-set Newton algorithm for minimising KL divergence between an observation $\mathbf{x}$ and the model $\hat{\mathbf{x}}$ (see (10)) [15]. It estimates and updates a set of active atoms with non-zero weights. From the learnt atoms of individual speakers and from their estimated weights frequency-domain filters are composed for each speaker. They are then applied on the magnitude spectrum of the mixture signal (a "Wiener-style" reconstruction), and the phases are taken directly from the mixture spectrum. Thereupon, an inverse DFT is taken to obtain time-domain separated signal, followed by the overlap-add procedure.

Signal-to-distortion (SDR) ratio between the separated $\hat{s}(t)$ and the original source signal $s(t)$ was used:

$$\text{SDR}_{\text{dB}} = 10 \log_{10} \frac{\sum_t s(t)^2}{\sum_t (s(t) - \hat{s}(t))^2}. \tag{11}$$

The values of SDR were averaged across the speakers and evaluated methods. Additionally, source-to-interference ratio (SIR) and source-to-artifacts ratio (SAR) were computed using the implementation from the BSS Eval toolbox [19]. With their aid, estimation errors mostly dominated by either interferences or artifacts can be distinguished. Furthermore, evaluation with the overall perceptual score (OPS) was performed using the latest version of the PEASS Toolkit [20, 21]. The method is based on sub-band decomposition of the distortion signal into components, whose salience is assessed with auditory-based methods, and the measure correlates well with the human mean opinion score assessments of the separation quality.

4.4. Results

The results of the SDR, SIR, SAR and OPS evaluation of all the methods with $\lambda = 0$ and varied dictionary sizes are presented in Figure 2, where archetypal analysis is denoted by AA. The effect of the value of λ on the results is observed rather small. $\lambda = 0$ gave consistently better results in terms of SDR (a slight improvement within the range 0.2…1 db), therefore results with $\lambda = 1$ are not shown.

With smaller dictionary sizes, the SDR performance of the proposed archetypal analysis minimising KL divergence is superior to all the competing methods except NMF. Such behaviour can be explained by the fact that the convex hull estimated with methods other than NMF becomes too tight when there is not enough components. In the case of 10 atoms per speaker, the proposed method shows absolute superiority. With increasing dictionary size, the performance

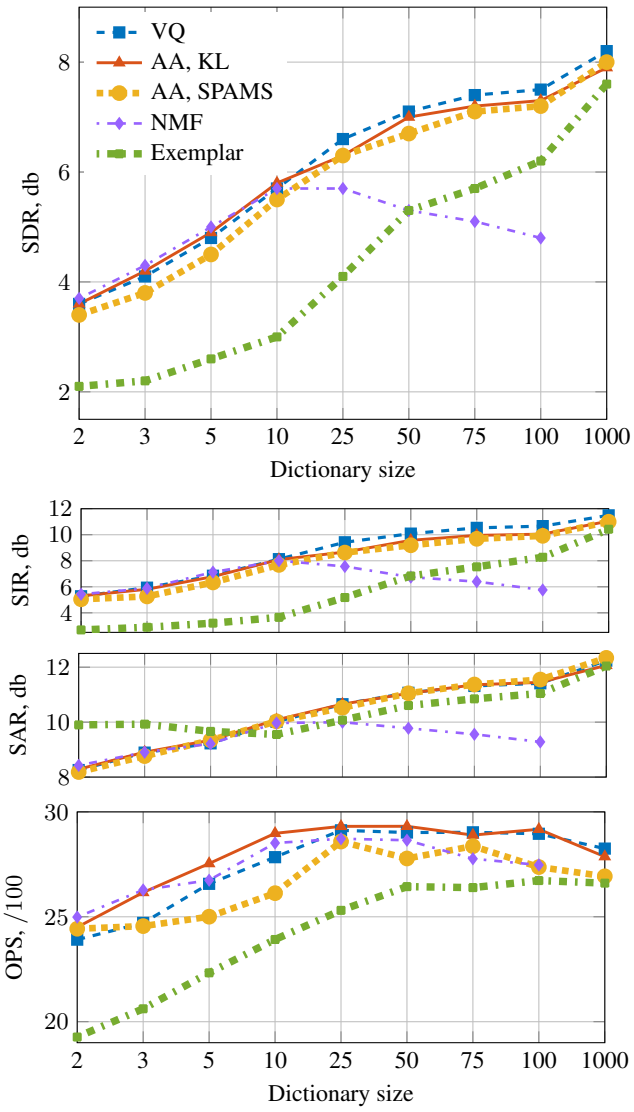

Figure 2: The source separation SDR, SIR, SAR, as well as OPS measures of the evaluated methods with varied dictionary size.

of NMF decreases. This is due to it being a parts-based representation: when the number of components becomes close to the rank of the observed data matrix, meaningful parts are not any more learnt. The performance of the proposed algorithm, however, follows the increasing trend, common to the other methods as well.

The evaluation in terms of SIR and SAR measures shows that the proposed method produces similar amount of artifacts as VQ, AA SPAMS and the exemplar-based representations with medium and large dictionaries, while interference from another source appears to mostly contribute to the overall performance difference among the methods. The results with the perceptually motivated measure shows that with most of the dictionary sizes, the proposed method achieves the best overall perceptual score among the evaluated methods.

The proposed archetypal analysis in the current implementation is the least computationally efficient among the evaluated methods due to no particular attempt for increasing the efficiency made at this

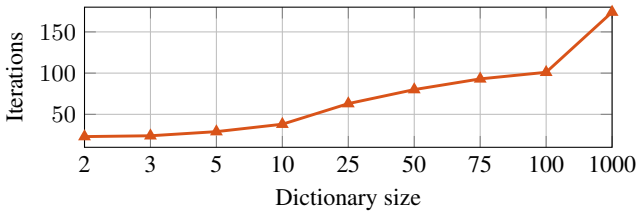

Figure 3: The number of iterations required for the learning with the proposed method to reach convergence decided upon a certain relative change of KL divergence.

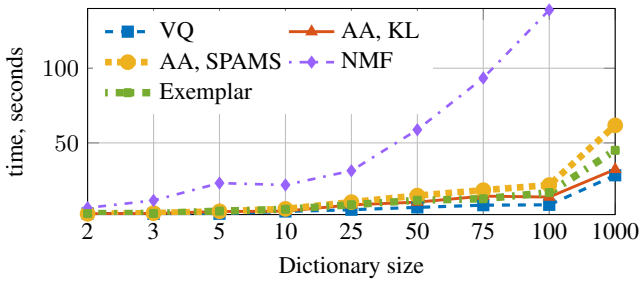

Figure 4: Time required for a source estimation from one test mixture on an average modern desktop machine.

stage. With regard to the dictionary size, the convergence decided upon a sufficiently small relative change of the KL divergence (an empirical value of 2.5×10^{-9} in this case) is achieved after fewer iterations with smaller dictionaries (see Figure 3). When setting the termination criterion to a certain number of iterations, the required processing time grows steadily, but moderately with increasing dictionary size. The processing time required for source estimation from one test mixture using the bases learnt with the evaluated methods with regard to dictionary size is depicted in Figure 4, showing a competitive performance of the proposed method.

A small set of demo signals separated with all the presented methods and the best-performing dictionary size was generated. The samples are available at the demonstration webpage[1].

5. CONCLUSIONS

This paper proposed archetypal analysis minimising the generalised KL divergence and showed its applicability for audio processing. The evaluation performed within the source separation problem demonstrated comparable results to other dictionary learning methods, including NMF, VQ and exemplar-based learning. In the case of medium-size dictionaries, the best performance across the evaluated methods was achieved with the proposed method. Thus, the approach was shown to be a valid alternative to the state-of-the-art, with an additional inherent benefit of the inuitivity of the learnt dictionaries.

The proposed methodology is expected span beyond the presented source separation scenario to other audio analysis areas. Archetypal analysis has been previously used also for classification problems [10]. An investigation of the applicability of the method for audio classification, *e.g.* of acoustic events, appears one of the valid future research directions.

[1] `http://www.cs.tut.fi/~diment/aademo.html`

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 556–562.

[2] S. A. Abdallah and M. D. Plumbley, "Polyphonic music transcription by non-negative sparse coding of power spectra," in *5th International Conference on Music Information Retrieval (ISMIR)*, 2004, pp. 318–325.

[3] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2067–2080, 2011.

[4] F. Weninger, M. Wollmer, J. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. 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.

[5] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 10, pp. 1506–1521, 2014.

[6] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* IEEE, 2003, pp. 177–180.

[7] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Interspeech — International Conference on Spoken Language Proceesing*, 2006.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

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

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

[11] A. Cutler and E. Stone, "Moving archetypes," *Physica D: Nonlinear Phenomena*, vol. 107, no. 1, pp. 1–16, 1997.

[12] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse nonparametric approach for single channel separation of known sounds," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 1705–1713.

[13] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, 2007.

[14] J. J. Carabias-Orti, F. J. Rodriguez-Serrano, P. Vera-Candeas, F. J. Cañadas-Quesada, and N. Ruiz-Reyes, "Constrained nonnegative sparse coding using learnt instrument templates for realtime music transcription," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 7, pp. 1671–1680, 2013.

[15] T. Virtanen, J. Gemmeke, and B. Raj, "Active-set newton algorithm for overcomplete non-negative representations of audio," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 11, pp. 2277–2289, Nov 2013.

[16] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[17] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

[18] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4546–4549.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.

[20] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2046–2057, 2011.

[21] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 430–437.