Applied Acoustics 129 (2018) 277–283

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Incremental basis estimation adopting global k-means algorithm for NMF-based noise reduction

Kisoo Kwon [a], Jong Won Shin [b,*], Nam Soo Kim [a]

[a] Department of Electrical and Computer Engineering and INMC, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea

[b] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 123 Cheomdan-gwagiro, Buk-gu, Gwangju, Republic of Korea

ARTICLE INFO

Article history:
Received 29 September 2016
Received in revised form 25 July 2017
Accepted 7 August 2017
Available online 17 August 2017

Keywords:
Nonnegative matrix factorization
Basis training
Incremental approach
Global k-means
Source separation

ABSTRACT

Nonnegative matrix factorization (NMF) is a data decomposition technique enabling to discover meaningful latent nonnegative components. Since, however, the objective function of NMF is non-convex, the performance of the source separation can degrade when the iterative update of the basis matrix in the training procedure is stuck to a poor local minimum. In most of the previous studies, the whole basis matrix for a specific source is iteratively updated to minimize a certain objective function with random initialization although a few approaches have been proposed for the systematic initialization of the basis matrix such as the singular value decomposition and k-means clustering. In this paper, we propose an approach to robust bases estimation in which an incremental strategy is adopted. Based on an analogy between clustering and NMF analysis, we estimate the NMF bases in a similar way to the global k-means algorithm popular in the data clustering area. Experiments on audio separation from noise showed that the proposed methods outperformed the conventional NMF technique using random initialization by about 2.04 dB and 2.34 dB in signal-to-distortion ratio when the target source was speech and violin, respectively.

1. Introduction

Over the recent years, nonnegative matrix factorization (NMF) has been widely applied to many tasks, and it has shown impressive performances particularly in image and audio signal processing [1–22]. NMF is a sort of latent factor analysis technique for which unsupervised learning algorithms are used to discover underlying part-based representations for the given nonnegative data. NMF is conceptually similar to other well-known matrix factorization or even data clustering techniques which can be expressed in a unified formulation [16,23,24]. NMF has shown certain benefits compared with other factorization schemes such as the independent component analysis and principal component analysis (PCA) [1,2], in the areas including source separation [3,4] and document classification [5]. Since the publication of [1], a number of attempts have been made to improve NMF under some specific conditions, which include Itakura-Saito NMF [2], sparse NMF [6–8], convolutive NMF [10], and discriminative NMF [11–13].

Though NMF shows an impressive performance in several fields, one of its weakness is that the final result is sensitive to the initial values of the bases [24]. Because the objective function of NMF is not convex, the optimized solution obtained from the iterative updates of the basis matrix can be stuck to a local minimum, which implies that the overall performance may significantly depend on the initial parameter values. For this reason, several previous works attempt to provide systematic ways to initialize the basis and encoding matrices such as the centroids of k-means clustering, PCA, and singular value decomposition (SVD)-based methods [25–31]. Though some of these methods show lower reconstruction errors and faster convergence speeds than the random value initialization, the source separation performance are not of primary concern. Moreover, the SVD- and PCA-based methods cannot support over-complete bases in which the number of bases is larger than the dimension of the input vector. Recently, an incremental approach inspired by Linde-Buzo-Gray (LBG) algorithm is proposed which doubles the number of bases in each step [9]. It shows promising source separation performance, but the number of basis is strictly restricted to be a power of 2.

The conventional vector quantization task can be interpreted as a special case of the matrix factorization where each basis vector corresponds to a codeword and only a single basis is activated at each time [23]. This analogy implies that the data clustering techniques can provide some useful cues for the initialization of the NMF bases. Unfortunately, however, conventional codebook training approaches such as the k-means clustering can only guarantee

* Corresponding author.
E-mail address: jwshin@gist.ac.kr (J.W. Shin).

http://dx.doi.org/10.1016/j.apacoust.2017.08.008
0003-682X/

suboptimal solutions similar to the case of NMF bases estimation and the final centroids are sensitive to the initialization of the code vectors. In order to alleviate this difficulty, several modified k-means algorithms have been developed [32–34]. The core idea of these algorithms is to increase the number of code vectors gradually while optimizing a certain criterion so that the final result can be less dependent on the initial parameter values.

In this paper, we propose a novel approach to estimate the basis and encoding matrices for the NMF analysis. Exploiting the analogy between NMF analysis and data clustering, a systematic method for estimating the NMF basis matrix is proposed by combining the standard NMF basis training procedure and an efficient codebook learning algorithm. The proposed methods borrow an idea from the global k-means algorithm [34]. One of the prominent features of this algorithm is that it estimates the parameters incrementally, i.e., increases the number of bases by one at each step. Unlike our previous approach based on LBG algorithm [9], the proposed approach aligns with the geometric analysis of NMF and does not have a restriction on the number of basis vectors. In order to evaluate the performance of the proposed techniques, we carried out an experiment on audio separation from noise. In the experimental result, we can see that the proposed methods outperformed other bases estimation methods.

2. NMF-based audio source separation

When NMF is applied to audio source separation, it generally approximates the magnitude spectra of a given mixture $\mathbf{V} \in \mathbb{R}_+^{M\times N}$ as the product of a basis matrix $\mathbf{W} \in \mathbb{R}_+^{M\times R}$ and an encoding matrix $\mathbf{H} \in \mathbb{R}_+^{R\times N}$ ($\mathbf{V} \approx \mathbf{WH}$) where M, N, and R denote the number of frequency bins, short-time frames, and the number of basis vectors, respectively. In order to resolve the non-unique factorization problem, it is needed to impose some constraints on the structures of $\mathbf{W}$ or $\mathbf{H}$. In our work, all the column vectors of $\mathbf{W}$ are constrained to have a unit L_2-norm. The process of NMF-based source separation is given in Fig. 1. In this case, the basis matrix $\mathbf{W}$ is considered as a concatenation of the target and noise basis matrices, $\mathbf{W}_S \in \mathbb{R}_+^{M\times R_s}$ and $\mathbf{W}_N \in \mathbb{R}_+^{M\times R_n}$ where R_s and R_n indicate the number of target signal and noise basis vectors, respectively. $\mathbf{W}_S$ and $\mathbf{W}_N$ are usually trained separately with clean target signal and noise DBs, respectively. The objective function of NMF is given as the discrepancy between $\mathbf{V}$ and $\mathbf{WH}$, i.e.,

$$f(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) \tag{1}$$

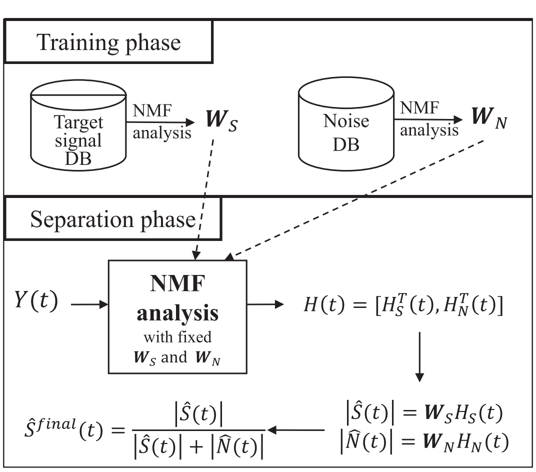

Fig. 1. The process of NMF-based audio source separation.

where $D(a|b)$ denotes the divergence between a and b. One of the popular choices for the divergence measure is Kullback-Leibler divergence (KLD) which is given as [1]

$$D(\mathbf{V}|\mathbf{WH}) = \sum_{m,n} \mathbf{V}_{m,n} \log \frac{\mathbf{V}_{m,n}}{(\mathbf{WH})_{m,n}} - \mathbf{V}_{m,n} + (\mathbf{WH})_{m,n} \tag{2}$$

where $\mathbf{A}_{m,n}$ denotes the m-th row and n-th column component of the matrix $\mathbf{A}$. When the multiplicative update rule (MuR) [1] is used, $\mathbf{W}$ and $\mathbf{H}$ are updated as follows:

$$\mathbf{H}_{r,n} \leftarrow \mathbf{H}_{r,n} \frac{\sum_{k=1}^{M} \frac{\mathbf{W}_{k,r}\mathbf{V}_{k,n}}{\sum_{f=1}^{R} \mathbf{W}_{k,f}\mathbf{H}_{f,n}}}{\sum_{k=1}^{M} \mathbf{W}_{k,r}}, \tag{3}$$

$$\mathbf{W}_{m,r} \leftarrow \mathbf{W}_{m,r} \frac{\sum_{p=1}^{N} \frac{\mathbf{H}_{r,p}\mathbf{V}_{m,p}}{\sum_{f=1}^{R} \mathbf{W}_{m,f}\mathbf{H}_{f,p}}}{\sum_{p=1}^{N} \mathbf{H}_{r,p}}. \tag{4}$$

The final estimate for $\mathbf{H}$ and $\mathbf{W}$ are obtained by iterative application of the update rules (3) and (4) for a fixed number of iterations. The MuR is an well-known approach to estimate $\mathbf{W}$ and $\mathbf{H}$ which is simple to implement and shown to yield good results. Each basis matrix, $\mathbf{W}_S$ and $\mathbf{W}_N$, is obtained separately by (3) and (4).

In the separation phase, a noisy magnitude spectrum $|Y(t)|$ is approximated as $|Y(t)| \approx \mathbf{W}H(t)$ for each frame with the fixed basis matrix $\mathbf{W} = [\mathbf{W}_S\, \mathbf{W}_N]$ obtained during the training phase where $H(t) = \left[H_S(t)^T\, H_N(t)^T\right]^T \in \mathbb{R}^{(R_s+R_n)\times 1}$ denotes the encoding vector of the mixed signal in the t-th frame, $Y(t)$ represents the short-time Fourier transform (STFT) coefficients of the noisy input, and $|\cdot|$ denotes taking element-wise magnitude. Keeping $\mathbf{W}$ fixed, $H(t)$ is computed by iterating (3) for a fixed number of times, in which $H_S(t)$ and $H_N(t)$ are initialized to nonnegative random numbers. After a fixed number of iterations, the magnitude spectra of the target and noise signals are estimated as follows:

$$|\hat{S}(t)| = \mathbf{W}_S H_S(t), \qquad |\hat{N}(t)| = \mathbf{W}_N H_N(t). \tag{5}$$

Instead of directly using the estimated magnitude spectra in (5), a spectral gain function similar to the Wiener filter is adopted in [2–4,20]. In this scheme, the gain function is given by

$$G(t) = \frac{|\hat{S}(t)|}{|\hat{S}(t)| + |\hat{N}(t)|} \tag{6}$$

where $\frac{A}{B}$ denotes element-wise division of vectors, and $G(t) \in \mathbb{R}^{M\times 1}$ is a gain vector obtained at the t-th frame. Finally, the STFT coefficients of the enhanced speech signal at the t-th frame are obtained according to $\hat{S}^{final}(t) = G(t) \otimes Y(t)$ where $\otimes$ indicates an elementary-wise multiplication.

3. Global k-means clustering-based NMF bases estimation

In this section, we propose a novel approach to estimate NMF bases, which is based on an analogy between the NMF basis training and the codebook design in vector quantization. If the encoding vector of the NMF analysis is allowed to have only one non-zero component, then each NMF basis can be viewed as a codeword vector and the reconstruction error can be treated as the distance between the input vector and its nearest codeword. Our approach to NMF bases estimation is motivated by the global k-means (GKM) clustering technique [34], which has demonstrated smaller clustering error than several other variants of the k-means clustering approach. In general, for data clustering, we need to find R codewords and a rule to map any M-dimensional input vector into one of the R codewords for the sake of minimizing the sum of the squared Euclidean distances between each input vector and the