Biomedical Signal Processing and Control 36 (2017) 168–175

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Non-negative matrix factorization for speech/music separation using source dependent decomposition rank, temporal continuity term and filtering

S. Abdali, B. NaserSharif*

Department of Computer Engineering, K.N. Toosi University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 24 November 2016
Received in revised form 11 March 2017
Accepted 26 March 2017

Keywords:
Non-negative matrix factorization (NMF)
Cost function
Regularization term
Filter
Decomposition rank

ABSTRACT

Non-negative matrix factorization (NMF) is a recently well-known method for separating speech from music signal as a single channel source separation problem. In this approach, spectrogram of each source signal is factorized as a multiplication of two matrices known as basis and weight matrices. To obtain a good estimation of signal spectrogram, weight and basis matrices are updated based on a cost function, iteratively. In standard NMF, each frame of signal is considered as an independent observation and this assumption is a drawback for NMF. For overcoming this weakness, a regularization term is added to the cost function to consider spectral temporal continuity. Furthermore, in the standard NMF, the same decomposition rank is usually used for different sources. In this paper, in accompany with using a regularization term, we propose to apply a filter to the signals estimated by NMF. The filter is constructed by signals which are estimated using a regularized NMF method. Moreover, we propose to use different decomposition ranks for speech and music signals as different sources. Experimental results on one hour of speech and music signals show that the proposed method increases signal to inference ratio (SIR) values for speech and music signals in comparison to conventional NMF methods.

1. Introduction

Speech/Music separation is one of the single channel source separation methods that can be used as a pre-processing step for different applications such as in-car voice command systems or searching keywords in spoken news archives. It can be also used for improving the quality of hearing aid's output (as a medical equipment) when background music is considered as the background noise.

Many approaches have been proposed for single channel source separation problem. Most of these approaches need training data for each source signal. The training data can be modeled by probabilistic models such as Gaussian Mixture Model (GMM) or Hidden Markov Model (HMM).

These models are used in the separation stage under this assumption that appeared sources in the mixed signal have the same energy level as the training data. Resolving this limitation requires complicated computations [1–5].

Another technique for training data modeling is to train non-negative basis sets for each source. This method is more flexible where it does not need any assumption about the energy differences between the signals in the training and separation stages. The main drawback of this approach is that any non-negative linear combination of the trained vectors is not necessarily a valid estimation of the source signals. This may decrease separation quality [9]. Non-negative matrix factorization (NMF) is an example of such methods. Usually, NMF method is used for decomposing a mixed signal spectrogram. In standard NMF, each source signal is defined as the multiplication of two matrices with non-negative elements known as weight and basis matrices. NMF has two stages: training and test (decomposition) stages. In the training stage, training data are used for training basis vectors of each source's basis matrix. In this stage, a cost function is used to estimate the matrices. Then, in the decomposition stage, estimated basis vectors are used for frame-by-frame separation of mixed signal without considering any smooth transition between frames or other information that may exist in the subsequent frames. Until now, different cost functions such as Euclidean distance, Kullback–Leibler (KL) and Itakora–Saito (IS) divergences have been used [6–10]. A problem with Standard NMF is that each element of the basis matrix is considered as an independent observation. Many approaches have

* Corresponding author.
E-mail address: bnasersharif@kntu.ac.ir (B. NaserSharif).

http://dx.doi.org/10.1016/j.bspc.2017.03.010
1746-8094/© 2017 Elsevier Ltd. All rights reserved.

been proposed to solve this problem. One of these approaches is to add a regularization term to the cost function of NMF, which codes prior knowledge and shows temporal continuity of spectrum [11,12]. In fact, this term is used to penalize big differences of adjacent frames in the weight matrix [11,13]. To determine this term, different methods have been proposed based on used NMF cost function and its mathematical properties. For example, in [14,15], a term is considered to compensate sparseness of weight matrix besides temporal continuity term. In another approach, the temporal continuity is considered using a regularization term based on HMM and smoothness for rows of weight matrix [9]. In [16,17], this regularization term is obtained using statistical properties of spectrum's temporal continuity. Moreover, it has been proposed to apply post processing methods such as Wiener mask for enhancing source separation [18].

In this paper, for improving separation quality, besides using regularization term, we propose to apply a filter on separated output signals. In addition, we propose to use different decomposition ranks for music and speech signals.

The remainder of this article is organized as follows: Section 2 introduces the NMF method mathematically, then in Section 3, KL divergence, which broadly is used as the NMF cost function for speech processing, has been described. In Section 4, our proposed method and the motivations have been propounded and finally in Sections 5 and 6 the experimental results and conclusion are presented, respectively.

2. Non-negative matrix factorization

2.1. Basic definition

Mathematically, the NMF is formulated as follows. Let $V \in R_+^{M\times N}$ be a non-negative matrix, mean all the coefficients of which are positive or null, of size $M \times N$ (in music applications, V will very often the amplitude spectrogram). Non-negative matrix factorization approximates V by $\tilde{V}$ as follows:

$$\tilde{V} = BW \tag{1}$$

where $B \in R_+^{M\times K}$ and $W \in R_+^{K\times N}$, and where K is the factorization rank, generally chosen such that $K(M+N) \le MN$. The matrix B is called basis or the codebook. The matrix W is called the weight or activation matrix. Each column vector in matrix V is estimated by a weighted linear combination of basis vectors which are the same B columns. Weights for basis vectors appear in corresponding columns in matrix W. To approximate data in V as a non-negative linear combination of its component vectors, the non-negative basis vectors in matrix B are optimized. The following figure shows an example of how to decompose a piece of music [9] (Fig. 1):

The matrices B and W are estimated by solving following optimization problem [18]:

$$\min C(V||BW) \quad \text{where } B, W \ge 0 \tag{2}$$

where C is a cost function which estimates distance between V and BW. Different cost functions lead to different kinds of NMF. One of the well-known cost functions is KL divergence.

2.2. Training of speech and music

With two sets of training data for speech and music signals, the Fast Fourier Transform (FFT) is computed for each signal to obtain magnitude spectrogram of speech and music signals. Then, NMF is used for decomposing speech and music spectrograms into base and weight matrices. In other word, the aim of using NMF is to model the training data as a set of basis vectors to represent the spectral characteristics for each source signal [18].

$$S_{\text{train}} \approx B_{\text{speech}} W_{\text{speech}} \tag{3}$$

$$M_{\text{train}} \approx B_{\text{music}} W_{\text{music}} \tag{4}$$

Based on the used cost function, B and W are updated iteratively. B_{speech} and B_{music} have normalized columns, and after each iteration their columns should be normalized again. The initial values for B and W are random positive values [18].

2.3. Decomposition of the mixed signal

In the decomposition stage, NMF should be used again to decompose the signal X spectrogram. In this step, basis matrix is obtained from the training phase matrices as follows [18]:

$$X \approx \left[B_{\text{speech}} B_{\text{music}}\right] W \tag{5}$$

$$\tilde{S} \approx B_{\text{speech}} W_{\text{s}} \tag{6}$$

$$\tilde{M} \approx B_{\text{music}} W_{\text{m}} \tag{7}$$

where W_s and W_m are sub matrices in matrix W which correspond to the speech and music components respectively. $\tilde{M}$ and $\tilde{S}$ matrices contain estimations for the spectral magnitude of the music and speech signals [18].

3. NMF cost functions and regularization term

3.1. KL based cost function

Cost function based on KL divergence is defined as [7,8]:

$$D_{KL}(V||BW) = \sum_{i,j} \left(V_{ij} \log \frac{V_{i,j}}{(BW)_{i,j}} - V_{i,j} + (BW)_{i,j} \right) \tag{8}$$

To minimize the KL based cost function, B and W matrices can be computed through the following iterative updates:

$$B \longleftarrow B \otimes \frac{(V/BW)W^{\mathrm{T}}}{1W^{\mathrm{T}}} \tag{9}$$

$$W \longleftarrow W \otimes \frac{B^{\mathrm{T}}(V/BW)}{B^{\mathrm{T}}1} \tag{10}$$

where 1 is a matrix of ones with the same size of V, all multiplication and divisions are element wise and $\otimes$ indicates element wise multiplication [6,7].

As mentioned before, standard NMF does not consider any probabilistic assumption and so basis vector elements are considered independent. This tends to poor basis and weight vectors estimation. Some approaches use prior knowledge of basis vectors in the standard NMF training step [14].

A well-known method for considering the prior knowledge is based on adding a regularization term to the NMF cost function. This regularization term considers temporal continuity of components to penalize large changes between the weights of adjacent frames. This term can be computed as the sum of the squared differences between the weights as follows [14]:

$$C_{\text{TC}} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2} \sum_{t=2}^{T} \left(w_{t,j} - w_{t-1,j}\right)^2 \tag{11}$$

$$\sigma_j = \sqrt{\frac{1}{T}\sum_{t=1}^{T} w_{t,j}^2} \tag{12}$$