Signal Processing 142 (2018) 137–148

Contents lists available at ScienceDirect

Signal Processing

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

Audio signal separation through complex tensor factorization: Utilizing modulation frequency and phase information

Shogo Masaya

[a] *Faculty of Applied Sciences, Delft University of Technology, Lorentzweg 1, 2628 CJ, Delft, The Netherlands*
[b] *INPEX Corporation, 5-3-1, Akasaka, Minato-ku, Tokyo, Japan*

ARTICLE INFO

Article history:
Received 18 December 2016
Revised 11 June 2017
Accepted 17 July 2017
Available online 18 July 2017

Keywords:
Audio-source separation
Tensor factorization
Complex non-negative matrix factorization
Modulation frequency domain
Phase

ABSTRACT

I propose a complex-valued tensor factorization algorithm for audio-source separation to exploit not only amplitude but phase information of audio signals in the modulation frequency (MF) domain. The proposed algorithm is extended from complex non-negative matrix factorization, which is capable of decomposing an arbitrary complex matrix such as the complex spectrum in the acoustic frequency domain. The proposed method enables us to factorize an arbitrary complex tensor of order 3. The detailed performance of the proposed algorithm for single-channel source separation is investigated through numerical experiments. I examine the quantitative contributions of the MF domain and phase information examined by additionally presenting three tensor factorization algorithms and using five objective indices for source separation.

1. Introduction

Blind source separation is a problem that involves separating outsource signals from mixture signals without any information or with limited information about the sources and/or the mixing processes. The separation techniques have potential applications in various fields such as speech enhancement, polyphonic music transcription, the front-end of automatic speech recognition, and hearing aids. However, the source separation still remains an open-ended and challenging research topic. A lot of attempts to use the mixture data, training data, and/or physical properties of target sources have been presented on the basis of the assumption that the target sources cannot be directly known.

One of commonly used methods for source separation is non-negative matrix factorization (NMF), which is capable of factorizing an arbitrary non-negative matrix into a product of two non-negative matrices [1]. The approaches based on NMF has been suggested in the field of audio-signal processing since NMF had been applied to polyphonic music transcription [2] by using the amplitude spectrum in the acoustic frequency (AF) domain, which is calculated via the short-time Fourier transforms (STFTs). Numerous NMF-based schemes have been also studied for source separation by applying NMF to the amplitude spectrum in the AF domain. For instance, it was shown that the NMF with Itakura–Saito (IS) divergence was suitable for the representation of music signals compared with the NMF with Euclidean or Kullback–Leibler divergence [3]. However, no complex spectrum including the phase spectrum in the AF domain can be decomposed by NMF due to its nonnegativity limitation. Thus, the phase information of signals is neglected in the NMF-based schemes, although prior studies have indicated that not only amplitude but also phase information has played a significant role in speech enhancement [4–8]. A complex NMF framework [9] was presented by Kameoka et al. for the purpose of addressing the mentioned limitation in NMF and taking account of the phase spectrum in the AF domain. King and Atlas simply called it “complex matrix factorization (CMF)” [10] and investigated its detailed performance with their additionally proposed algorithms for single-channel source separation [10–12]. Moreover, the extension of multichannel for CMF is reported in [13,14].

Greenberg and Kingsbury proposed the spectrogram of a modulation envelope, which was termed a “modulation spectrogram”, and it was demonstrated to be beneficial in representing speech signals [15]. Previous researches have proved that spectral processing like spectral subtraction in the modulation frequency (MF) domain could be more effective than the AF-domain schemes for speech enhancement [15–19]. Non-negative tensor factorization (NTF) [20], which is extended from NMF, is one of the approaches utilized the MF domain for multichannel source separation [21]. In recent years, NTF-based algorithms with the modulation spectrogram have been applied to single-channel audio source separation [22–24]. They showed that the modulation spectrogram was able to effectively represent redundant patterns across frequencies with similar features. The use of not only amplitude information but

E-mail address: S.Masaya@tudelft.nl

http://dx.doi.org/10.1016/j.sigpro.2017.07.013
0165-1684/© 2017 Elsevier B.V. All rights reserved.

also phase information in the MF domain methods would be desired to exploit more information on signals. However, NTF-based methods do not take the phase information of signals into account because they cannot factorize any complex-valued tensor such as complex spectrum including phase spectrum. To address this issue, we recently proposed a novel algorithm "complex tensor factorization (CTF)" and showed the preliminary results with simple experiments to demonstrate its validity for single-channel source separation [25]. CTF enables us to factorize an arbitrary complex tensor of order 3. Therefore, the separation approach based CTF can take account of not only the amplitude information but also the phase information in the MF domain. Furthermore, since the signal representation in the MF domain is calculated from two STFTs in the approach, it allows us to synthesize individual separated signals in the time domain via to two inverse STFTs (ISTFTs) without any loss of signal information through the MF and AF domains.

In this paper, in order to investigate the detailed performance of CTF, the contributions of the MF domain and phase information to speech enhancement and source separation are evaluated by newly presenting three tensor factorization algorithms. Here, five indices for assessing the performance of source separation and speech enhancement with four types of noises are used to more accurately prove the effectiveness of CTF in numerical experiments,

The organization of this paper is as follows. Section 2 introduces the background of existing source separation methods using matrix factorizations and the MF domain. Section 3 describes our proposed methods for source separation. Numerical experiments on our evaluations and a discussion are presented in Sections 4 and 5. Finally, the conclusions I drew on the basis of our experiments are given in Section 6.

2. Background

2.1. Non-negative matrix factorization

NMF is an approximation algorithm to factorize an arbitrary non-negative matrix into a product of two non-negative matrices and often provides the sparse solutions. The algorithm comes down to solving the following optimization problem (e.g. see the notation in Bronson and Depalle [26]):

$$\text{Given}: \boldsymbol{X} \in \mathbb{R}^{\geq 0, L\times M},\ K \in \mathbb{N}^{+}, \tag{1}$$

$$\text{Factorize}: \boldsymbol{X} \simeq \boldsymbol{BW}, \tag{2}$$

$$\text{Minimize}: \sum_{l,m}\left|X_{l,m} - \sum_{k} B_{l,k}W_{k,m}\right|^{2}, \tag{3}$$

$$\text{Subject to}: \boldsymbol{B} \in \mathbb{R}^{\geq 0, L\times K},\ \boldsymbol{W} \in \mathbb{R}^{\geq 0, K\times M}, \tag{4}$$

where $\boldsymbol{X}$ represents an arbitrary non-negative matrix corresponding to the input data to be factorized by NMF. K, $\boldsymbol{B}$, and $\boldsymbol{W}$ are the number of bases, the base matrix, and the weight matrix. Multiplicative update method [1] is a well-known iterative algorithm to solve the optimization problem. The method provides the update rules for $\boldsymbol{B}$ and $\boldsymbol{W}$ via the iteration. The way that the number of bases K is determined remains one of the challenges in NMF. In general, K is empirically given by taking into account each data feature, although the automatic relevance determination for its bases was suggested [27,28].

The objective function shown in Eq. (3) is described by the Euclidean distance, which is the simplest distance in NMF. However, as mentioned previously in Section 1, it is known that IS distance in NMF is effective for the representation of audio signals. Therefore, the NMF with the following IS distance is used in the evaluation for the speech enhancement and source separation in this paper:

$$\sum_{l,m}\left(\frac{X_{l,m}}{\sum_k B_{l,k}W_{k,m}} - \log\frac{X_{l,m}}{\sum_k B_{l,k}W_{k,m}} - 1\right). \tag{5}$$

Next, the application of NMF to single-channel source separation is mentioned. Only mixture data $y(t) \in \mathbb{R}^{N_{mix}}$, where $y(t) = y_1(t) + y_2(t)$, is assumed to be observed. Here, $y_1(t)$ is a clean signal, and $y_2(t)$ is noise or another clean signal at time t. N_{mix} represents the number of time samples in the mixture data. The complex spectrum $\boldsymbol{Y}(\omega_l, \tau_l)$ in the AF domain for the mixture data $y(t)$ is given by STFT:

$$\boldsymbol{Y}(\omega_l, \tau_l) = \int_{-\infty}^{\infty} y(t)w_l(t-\tau_l)e^{-j\omega_l t}dt, \tag{6}$$

where ω_l, τ_l, and $w_l(t)$ indicate the frequency, frame indices, and window function in the STFT. The complex spectrum $\boldsymbol{Y}(\omega_l, \tau_l)(=|\boldsymbol{Y}(\omega_l, \tau_l)|e^{j\arg\boldsymbol{Y}(\omega_l,\tau_l)})$ consists of its amplitude spectrum $|\boldsymbol{Y}(\omega_l, \tau_l)|$ and phase spectrum $\arg\boldsymbol{Y}(\omega_l, \tau_l)$. The amplitude spectrum $|\boldsymbol{Y}(\omega_1, \tau_1)|$ of mixture data is commonly used as the input data, which corresponds to the non-negative matrix $\boldsymbol{X}$ described in Eq. (1), for the source separation based on NMF [2]. Therefore, the amplitude spectrum is factorized into two non-negative matrices by NMF.

For the NMF-based source separation, we assume the existence of training data for the mentioned two sources y_1 and y_2. Let the two training data be $y_1(t_1) \in \mathbb{R}^{N_{s1}}$ and noise $y_2(t_2) \in \mathbb{R}^{N_{s2}}$ with their time samples N_{s1} and N_{s2}, respectively. Thus, the complex spectra $\boldsymbol{Y}_i(\omega_l, \tau_{l,i})$ in the AF domain for the two training data $(i = 1, 2)$ are given by:

$$\boldsymbol{Y}_i(\omega_l, \tau_{l,i}) = \int_{-\infty}^{\infty} y_i(t_i)w_l(t_i-\tau_{l,i})e^{-j\omega_l t_i}dt_i, \tag{7}$$

where $\tau_{l,\ i}$ represents the frame indices for each source in the STFT. As Eq. (2) has shown, the amplitude spectrum of the training data can be factorized via NMF:

$$\boldsymbol{X}_i \equiv |\boldsymbol{Y}_i(\omega_l, \tau_{l,i})| \simeq \boldsymbol{B}_i\boldsymbol{W}_i, \tag{8}$$

where $\boldsymbol{B}_i$ and $\boldsymbol{W}_i$ represent the base matrices and weight matrices factorized by NMF for the training data. Similarly, NMF is applied to the amplitude spectrum of the mixture data:

$$\boldsymbol{X} \equiv |\boldsymbol{Y}(\omega_l, \tau_l)| \simeq \begin{bmatrix}\hat{\boldsymbol{B}}_1 & \hat{\boldsymbol{B}}_2\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{W}}_1 \\ \hat{\boldsymbol{W}}_2\end{bmatrix} = \begin{bmatrix}\hat{\boldsymbol{X}}_1 \\ \hat{\boldsymbol{X}}_2\end{bmatrix}. \tag{9}$$

Here, assuming that $\hat{\boldsymbol{B}}_i = \boldsymbol{B}_i$ and the initial elements in the weight matrices $\hat{\boldsymbol{W}}_i$ are random in the experiment of Section 4, we can calculate the complex spectra of two separated signals with $\hat{\boldsymbol{Y}}_i(\omega_l, \tau_l) = \hat{\boldsymbol{X}}_i(\omega_l, \tau_l)e^{j\arg\boldsymbol{Y}(\omega_l,\tau_l)}$ by using separated amplitude spectra $\hat{\boldsymbol{X}}_i(\omega_l, \tau_l)$. Note that the phase spectra of the two signals are assumed to be the same as those of the mixture data. Finally, the two separated signals, $\hat{y}_1(t)$ and $\hat{y}_2(t)$, can be estimated by carrying out the ISTFT of the complex spectra $\hat{\boldsymbol{Y}}_i(\omega_l, \tau_l)$. Since NMF has a non-negative limitation for its input matrix, the phase spectrum which is a complex matrix in the AF domain cannot be estimated in NMF.

2.2. Complex non-negative matrix factorization

CMF has been proposed as an algorithm to overcome the limitation in NMF and factorizes an arbitrary complex matrix into two non-negative matrices and a complex-valued tensor of order 3 [9]. CMF is defined as the following optimization problem:

$$\text{Given}: \boldsymbol{Y} \in \mathbb{C}^{L\times M},\ K \in \mathbb{N}^{+},\ \lambda \in \mathbb{R},$$