Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 81 (2016) 129–137

www.elsevier.com/locate/specom

Under-determined reverberant audio source separation using Bayesian Non-negative Matrix Factorization

Sayeh Mirzaei[a,*], Hugo Van Hamme[a], Yaser Norouzi[b]

[a] *Department of Electrical Engineering-ESAT, KULeuven, Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven, Belgium*
[b] *Department of Electrical Engineering, Amirkabir University of Technology, 424 Hafez Ave, Tehran, Iran*

Received 28 May 2015; received in revised form 11 January 2016; accepted 13 January 2016
Available online 3 February 2016

Abstract

In this paper, we address the task of audio source separation for a stereo reverberant mixture of audio signals. We use a full-rank model for the spatial covariance matrix. Bayesian Non-negative Matrix Factorization(NMF)frameworks are introduced for factorizing the time-frequency variance matrix of each source into basis components and time activations. We also propose to incorporate the temporal dependencies in the Bayesian model through (1) recursively updating the prior hyperparameters or (2) applying a prior with Markov chain structure to favor the smoothness of the solution and we compare the performance of these two schemes. The EM algorithm is applied to derive the update relations of the unknown parameters. The separation performance improvement over the non-Bayesian standard NMF method as well as the conventional full-rank unconstrained method are investigated by calculating objective separation evaluation metrics.

Keywords: Blind Source Separation (BSS); Bayesian Non-negative Matrix Factorization(BNMF); Spatial covariance model; Temporal dependencies; Reverberant mixture.

1. Introduction

We often deal with a mixture of sounds coming from different acoustic sources. Separation of these audio signals and extracting the individual source signals is required in many applications including speaker diarization, meeting transcription systems, hearing aids, polyphonic music transcription, etc.

When no prior information of the sources or channel mixing system is available, the task is called Blind Source Separation (BSS). The multichannel mixture signal $\mathbf{x}(t) \in \mathbb{R}^M$ can be expressed as

$$\mathbf{x}(t) = \sum_{n=1}^{N} \mathbf{y}_n(t) \tag{1}$$

where $\mathbf{y}_n(t), \quad n = 1 \dots N$ is the nth source spatial image vector over M channels. The mixing process consists of a linear time-invariant filtering of the source signals as:

$$\mathbf{y}_n(t) = \sum_{l=0}^{L-1} \mathbf{h}_n(l) s_n(t-l) \tag{2}$$

where $s_n(t)$ is the nth source signal, $\mathbf{h}_n(l) \in \mathbb{R}^M$ is the mixing filter vector which denotes the acoustic path from source n to the M microphones and L is the filter length. Most of the proposed BSS methods are based on the assumption that the mixing process at each frequency bin can be approximated by complex-valued multiplication:

$$\mathbf{Y}_n(f,t) \approx \mathbf{H}_n(f) S_n(f,t) \tag{3}$$

where $\mathbf{Y}_n(f, t)$ is the spatial image of source n in the Short Time Fourier Transform Domain(STFT) domain, $s_n(f, t)$ denotes the source STFT and $\mathbf{H}_n(f)$ specifies the Fourier transform of the mixing filter $\mathbf{h}_n(t)$.

If $S_n(f, t)$ is a zero-mean variable with variance $v_n(f, t)$, the covariance of $\mathbf{Y}_n(f, t)$ can be written as:

$$\mathbf{R}_{\mathbf{Y}_n}(f,t) = v_n(f,t) \mathbf{R}_n(f) \tag{4}$$

[*] Corresponding author. Tel.: +989126850714.
E-mail address: sayehm62@gmail.com (S. Mirzaei).

http://dx.doi.org/10.1016/j.specom.2016.01.003

The assumption in (3) implies that the spatial covariance matrix of each source, $\mathbf{R}_n(f)$, has rank 1. Assuming the rank-1 model, BSS can be achieved using time-frequency (TF) masking techniques (Yilmaz and Rickard, 2004) or MAP estimation assuming sparse prior distributions (Winter et al., 2007), or modeling the source variances with Non-negative Matrix Factorization (NMF) (Févotte et al., 2009; Ozerov and Févotte, 2010). The rank 1 assumption is only valid when the filter length L is sufficiently small with respect to the STFT window length. This is violated in most realistic scenarios where reverberation exists. A full-rank spatial covariance matrix model is proposed in Duong et al. (2009) to provide better approximation in reverberant environments. The Maximum Likelihood (ML) solution is then found in an oracle context where both the spatial covariance matrix, $\mathbf{R}_n(f)$, and the scalar variance of the sources, $v_n(f, t)$, are known and also in a semi-blind context where the spatial covariance matrix is estimated from single-source training data. In Duong et al. (2010a), the EM algorithm was used for blindly estimating both of the above parameters. The source permutation problem which arises when the unknown parameters are independently estimated at each frequency bin, has also been solved in Duong et al. (2010a).

In Arberet et al. (2010), the source variances $v_n(f, t)$ are modeled by NMF and the EM algorithm is used for blindly estimating the parameters similar to what is done in Duong et al. (2010a). In Duong et al. (2010c), the use of a non-uniform TF representation on the auditory-motivated equivalent rectangular bandwidth(ERB) scale is investigated. It has been shown that this representation is beneficial for multichannel convolutive source separation provided that the full-rank covariance model is used. This has also been investigated in Burred and Sikora (2006) for instantaneous mixtures and (Vincent, 2006) for convolutive mixtures.

In Duong et al. (2010b), four specific covariance models including the rank-1 anechoic model, the rank-1 convolutive model, the full-rank direct+diffuse model and the full-rank unconstrained model are considered. A hierarchical clustering-based method is used to initialize the parameters. Also, a Direction of Arrival (DoA) based approach is proposed to align the order of the estimated sources across all frequency bins.

In Duong et al. (2013) some spatial location prior distributions consistent with the theory of statistical room acoustics are proposed for application to the spatial covariance matrices and EM algorithms are derived for Maximum a Posteriori (MAP) estimation. In Nikunen and Virtanen (2014), a spatial covariance matrix model is proposed which consists of a weighted sum of Direction of Arrival kernels. This covariance model is combined with the Complex NMF (CNMF) framework proposed in Sawada et al. (2013) and the update relations for finding the unknown parameters are subsequently derived.

In Arberet et al. (2010), the nth source variance matrix $\mathbf{V}_n(F \times T)$ consisting of the above variance elements, $v_n(f, t)$, is approximated as a product of two non-negative matrices $\mathbf{W}_n(F \times K)$ and $\mathbf{H}_n(K \times T)$ which specify the basis components and time activation matrices respectively. It is assumed that the number of the components K required for modeling each source is known in advance. However this may not be a suitable presumption when the goal is to blindly separate the individual source signals and there is no prior information about the source types. Here, we propose a Bayesian NMF framework to automatically infer the number of basis vectors for each source. In our first approach, we develop a Bayesian framework assuming that the time activation matrix elements H_n are random variables with a Gamma prior distribution. An EM algorithm is developed for deriving the update equations. The update relations given in Arberet et al. (2010) are replaced with the newly derived relations for the factors of the source variance matrices which are obtained through MAP estimation. We have also modeled the temporal dependencies through imposing constraints to the prior distribution of the temporal activations. A procedure inspired from Mohammadiha et al. (2012) is used for updating the scale parameters of the prior distributions of the time activations.

In the second approach, we favor the smoothness of the results through applying an inverse-Gamma chain prior distribution inspired from Févotte et al. (2009).

In Smaragdis et al. (2014), a comprehensive study of the NMF methods which model the temporal statistics is done. One flexible approach for considering the actual temporal dependencies is to impose constraints on the model activations (Essid and Févotte, 2013; Févotte, 2011; Févotte et al., 2009; Mohammadiha et al., 2013; 2012; Virtanen, 2007; Wilson et al., 2008). These approaches are called dynamic or smooth NMF. They differ by the used penalty term in non-probabilistic settings or by the choice of the observation model and prior structure in the Bayesian frameworks. In Virtanen (2007), temporal continuity and sparseness constraints are applied to the activation coefficients. Temporal continuity is favored by using a cost term which is the sum of squared differences between the activations in adjacent frames, and sparseness is favored by penalizing nonzero activations. A Non-negative Dynamical System (NDS) is introduced in Févotte et al. (2013) for modeling speech spectra. It can be regarded as an extension of NMF to support Markovian dynamics. Non-negativity preserving Gamma or inverse-Gamma Markov chain priors are considered in Févotte (2011); Févotte et al. (2009); Mohammadiha et al. (2013, 2012) and Markov random fields in Kim and Smaragdis (2013). In Nakano et al. (2010), the spectrogram of music signals is modeled as the combination of Markov-chained spectral patterns.

The approaches proposed in this paper can be regarded as Bayesian extensions of the method proposed in Arberet et al. (2010) accentuating the smoothness of the estimates. The Gamma prior model has been chosen for its effectiveness in modeling sparse parameters. Meanwhile, Gamma and inverse-Gamma prior distributions are preferred because we are going to model non-negative elements of the activation matrix, thus other sparse prior distributions such as Laplace cannot be useful here. The novel aspects of our proposed approaches can be summarized as follows: