Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 746–762

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation☆

Emad M. Grais*, Hakan Erdogan

Faculty of Engineering and Natural Sciences, Sabanci University, Orhanli Tuzla, 34956 Istanbul, Turkey

Received 30 December 2011; received in revised form 1 August 2012; accepted 10 September 2012
Available online 19 September 2012

Abstract

We introduce a new regularized nonnegative matrix factorization (NMF) method for supervised single-channel source separation (SCSS). We propose a new multi-objective cost function which includes the conventional divergence term for the NMF together with a prior likelihood term. The first term measures the divergence between the observed data and the multiplication of basis and gains matrices. The novel second term encourages the log-normalized gain vectors of the NMF solution to increase their likelihood under a prior Gaussian mixture model (GMM) which is used to encourage the gains to follow certain patterns. In this model, the parameters to be estimated are the basis vectors, the gain vectors and the parameters of the GMM prior. We introduce two different ways to train the model parameters, sequential training and joint training. In sequential training, after finding the basis and gains matrices, the gains matrix is then used to train the prior GMM in a separate step. In joint training, within each NMF iteration the basis matrix, the gains matrix and the prior GMM parameters are updated jointly using the proposed regularized NMF. The normalization of the gains makes the prior models energy independent, which is an advantage as compared to earlier proposals. In addition, GMM is a much richer prior than the previously considered alternatives such as conjugate priors which may not represent the distribution of the gains in the best possible way. In the separation stage after observing the mixed signal, we use the proposed regularized cost function with a combined basis and the GMM priors for all sources that were learned from training data for each source. Only the gain vectors are estimated from the mixed data by minimizing the joint cost function. We introduce novel update rules that solve the optimization problem efficiently for the new regularized NMF problem. This optimization is challenging due to using energy normalization and GMM for prior modeling, which makes the problem highly nonlinear and non-convex. The experimental results show that the introduced methods improve the performance of single channel source separation for speech separation and speech–music separation with different NMF divergence functions. The experimental results also show that, using the GMM prior gives better separation results than using the conjugate prior.

Keywords: Nonnegative matrix factorization; Single-channel source separation; Gaussian mixture models; Trained prior models

☆ This paper has been recommended for acceptance by Jon Barker.

* Corresponding author. Tel.: +90 216 483 9607.
E-mail addresses: grais@sabanciuniv.edu (E.M. Grais), haerdogan@sabanciuniv.edu (H. Erdogan).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.09.002

1. Introduction

1.1. Motivation and literature review

Nonnegative matrix factorization (Lee and Seung, 2001) is extensively used in source separation applications, especially when only a single observation of the mixed signal is available. The observed mixed signal can be a mixture of: multiple speaker signals (Schmidt and Olsson, 2006), multiple musical instrument signals (Helén and Virtanen, 2005; Wang and Plumbley, 2006), speech and music signals (Grais and Erdogan, 2011a,c,d; Nakano et al., 2011; Raj et al., 2010), and speech and noise signals (Wilson et al., 2008b). In NMF based single-channel source separation, NMF uses the training data for each source to train a set of nonnegative basis vectors. After observing the mixed signal, NMF is used again to decompose the mixed signal as a weighted nonnegative linear combination of the trained basis vectors of all sources together. The estimate for each source is found by summing the decomposition terms that include the corresponding trained basis vectors.

To improve the performance of NMF, there have been many works that aim to encourage the NMF decomposition weights to satisfy certain characteristics of the nature of the source signals to be estimated. In Virtanen (2007), continuity and sparsity priors were placed on the decomposition weights. In Bertin et al. (2009), and Bertin et al. (2010), harmonicity and smoothness were enforced in Bayesian NMF and applied to music transcription. In Wilson et al. (2008b), regularized NMF was used to impose statistical structure for each audio frame. It was also used in Wilson et al. (2008a) in addition to modeling frame-to-frame temporal structure. In Cichocki et al. (2006), and Chen et al. (2006), different constrained NMF algorithms were used for different applications. In Fevotte et al. (2009), regularized NMF with Itakura-Saito (IS-NMF) divergence was introduced with Markov chain prior models for smoothness within a Bayesian framework. The conjugate prior distributions on the NMF weights and basis matrices solutions with the Poisson observation model within Bayesian framework was introduced in Virtanen et al. (2008). In Virtanen et al. (2008), the Gamma distribution was used as a prior for the basis matrix and the Gamma Markov chain (Cemgil and Dikmen, 2007) was used as a prior for the weights/gains matrix. In Virtanen and Cemgil (2009), a mixture of Gamma prior model was used as a prior for the basis matrix.

1.2. Overview of the proposed algorithm

In this work, we propose a new regularized NMF algorithm that incorporates the statistical characteristics of the source signals to steer the optimal solution of the NMF cost function during the separation process. The new algorithm makes better use of the available training data of the source signals to improve the separation performance. We define a new regularized cost function where the regularization term is the log-likelihood of the log-normalized gain vector under a prior Gaussian mixture model.

In supervised source separation problems, there is a training stage and a separation stage. In the training stage, we assume we have training data for each source which we use to learn the characteristics of each source first. In the separation (testing) stage, we observe a mixed signal and try to estimate each source in the mixed signal. Conventional use of NMF in supervised source separation is to decompose the magnitude or power spectra of the training data of each source into a trained basis matrix and a trained gains or "weights" matrix. In previous works, the columns of the trained basis matrix are usually used as the only representative model for the training source signals and the trained gains matrix was usually ignored. The columns of the trained gains matrix represent the valid weight combination patterns that the columns in the basis matrix can jointly receive for a specific type of source signal. A prior distribution can represent the statistical distribution of the gains vector in each column of the gains matrix and model the correlation between the entries. The prior model guides the NMF solution to prefer valid gain patterns during the separation stage. We use a multivariate Gaussian mixture model (GMM) as a prior model for the gains vector for each frame of each source. The GMM is a rich model for capturing the statistics and the correlations of the valid gain combinations for a certain type of source signal. GMMs are used extensively in speech recognition and speaker verification to model the multi-modal nature in speech feature vectors due to phonetic differences, gender, speaking styles, accents (Rabiner and Juang, 1993) and we conjecture that the gains vector can be considered as a feature extracted from the audio signal in a frame so that it can be modeled well with a GMM. The columns of the trained gains matrix for each source are normalized by the ℓ^2 norm, and their logarithm is taken and used in the prior GMM. In the proposed method, the

trained basis matrix and its corresponding gains prior GMM are jointly used as a representative model for the training data for each source.

The training can be performed either in two steps sequentially, or all the parameters can be learned using joint training. In sequential training, we first learn the basis and gains matrices using conventional NMF for each source from the corresponding training data and then fit a GMM to the log-normalized gains vectors obtained in the previous step. In joint training, we learn both the NMF matrices and the GMM parameters using coordinate descent (or alternating minimization) on the proposed regularized cost function directly. Jointly training the NMF and the prior models simultaneously is a novel idea introduced in this paper. In joint training, trained basis matrix is also changed since the gains matrix is enforced to satisfy the NMF equation guided by the GMM prior, so that the trained models are more consistent with the GMM prior assumption. For this reason, we use sequential training for initialization of the model parameters, but eventually use joint training of the model parameters in this work.

In the separation stage, regularized NMF is used to decompose the magnitude or power spectra of the observed mixed signal as a weighted linear combination of the columns of trained bases matrices for all source signals that appear in the mixed signal. The decomposition weights are encouraged to increase their log-likelihood with their corresponding trained prior GMMs using the regularized cost function.

1.3. Comparison with earlier work

Previous studies mostly focused on enforcing temporal continuity and sparsity of the gains matrices (Schmidt and Olsson, 2006; Virtanen, 2007; Bertin et al., 2010; Fevotte et al., 2009). One other study used a Gaussian prior for the gains in NMF (Wilson et al., 2008b). Wilson et al. (2008b) analyzed a limited case where the energy level for each source in the mixed signal is the same as the energy level with its corresponding training data. Moreover, Wilson et al. (2008b) assumed each source signal in the mixed signal has the same energy level. The single mode nature of the Gaussian distribution may limit the performance of the work in Wilson et al. (2008b) for realistic audio data which has a large variation. In addition, the used update rules for regularized NMF problem in Wilson et al. (2008b) do not handle the nonnegativity in a systematic fashion.

In our work, we handle energy differences between training and test data by using the logarithm of the normalized gains in the prior model and also consider a mixture model as a prior which represents the statistical nature of audio data in a much better way. Furthermore, our update rules for the solution of the regularized NMF problem handle the non-negativity constraints and the non-convex prior in an efficient manner. In addition, we show the importance of the regularization parameters which control the trade-off between the NMF divergence function and the prior log-likelihood. Finally, we show a comparison between conjugate prior and GMM prior for the gains matrix.

1.4. Organization of the paper

The remainder of this paper is organized as follows. In Section 2, a mathematical description of the single-channel source separation problem is given. In Section 3, we give a brief explanation about the different types of NMF cost functions and introduce the proposed regularized NMF. In Section 4, we show the training processes of the NMF bases models and the GMM prior gain models for the source signals. In Section 5, the separation process is described in detail. In the remaining sections, we present our observations and the results of our experiments.

2. Problem formulation

In single-channel source separation problems, the main aim is to find estimates of the source signals that had been mixed in a single observation $y(\tau)$. This problem is usually solved in the short time Fourier transform (STFT) domain because spectral power or magnitude is considered to be a more stable representation of audio signals than the time domain representation. Let $Y(t, f)$ be the STFT of $y(\tau)$, where t represents the frame index and f is the frequency-index. Due to the linearity of the STFT, we have:

$$Y(t, f) = \sum_{i=1}^{Z} S^{(i)}(t, f), \tag{1}$$

where $S^{(i)}(t,f)$ is the unknown STFT of source i in the mixed signal, and Z is the number of sources in the mixed signal. The phase angles of the STFT were usually ignored in source separation literature since the magnitude spectrum is convenient to work with and provides sufficient accuracy in separation performance (Helén and Virtanen, 2005; Virtanen, 2007). Hence, we follow this convention and consider spectral magnitude as the representation of any audio signal in this paper. We can approximate the magnitude spectrum of the measured signal as the sum of source signals' magnitude spectra as:

$$|Y(t,\ f)| \approx \sum_{i=1}^{Z} |S^{(i)}(t,\ f)|. \tag{2}$$

We can write the magnitude spectrograms $|Y(t,f)|$ and $|S^{(i)}(t,f)|$ in a matrix form, where the columns represent the time index t and the rows represent the frequency index f as follows:

$$\boldsymbol{Y} \approx \sum_{i=1}^{Z} \boldsymbol{S}^{(i)}, \tag{3}$$

where $\boldsymbol{S} = \{\boldsymbol{S}^{(1)}, \ldots, \boldsymbol{S}^{(i)}, \ldots, \boldsymbol{S}^{(Z)}\}$ are the unknown magnitude spectrograms of the source signals, and need to be estimated using the observed mixed signal and training data for each source. The columns of the magnitude spectrogram for the observed signal $\boldsymbol{Y}$ is obtained by taking the magnitude of the DFT of the windowed signal.

The main idea to solve for $\boldsymbol{S}$ is to use the magnitude spectra of the training data to train a set of nonnegative basis vectors $\boldsymbol{b}$ for each source. After observing the mixed signal, the estimate of each frame $\widetilde{\boldsymbol{s}}^{(i)}$ for each source i is found by decomposing its corresponding frame of the magnitude spectra of the observed mixed signal $\boldsymbol{y}$ as a nonnegative weighted linear combination of the trained set of nonnegative basis vectors $\boldsymbol{b}$ for all sources as follows:

$$\boldsymbol{y} \approx \underbrace{\sum_{q=1}^{Q^{(1)}} g_q^{(1)} \boldsymbol{b}_q^{(1)}}_{\widetilde{\boldsymbol{s}}^{(1)}} + \cdots + \underbrace{\sum_{q=1}^{Q^{(i)}} g_q^{(i)} \boldsymbol{b}_q^{(i)}}_{\widetilde{\boldsymbol{s}}^{(i)}} + \cdots + \underbrace{\sum_{q=1}^{Q^{(Z)}} g_q^{(Z)} \boldsymbol{b}_q^{(Z)}}_{\widetilde{\boldsymbol{s}}^{(Z)}}, \tag{4}$$

where $\boldsymbol{b}_q^{(i)}$ is the trained basis vector number q for source i, and $g_q^{(i)}$ is the gain that basis vector $\boldsymbol{b}_q^{(i)}$ gets in the mixed signal, and $Q^{(i)}$ is the number of trained basis vectors for source i. The nonnegativity constraint for the gains and bases ensures that the summation terms for each source will be nonnegative as they should be since they represent magnitude spectra. The nonnegativity constraints are considered to yield useful and meaningful representation of real world data as well (Lee and Seung, 1999; Cichockiy and Georgiev, 2003). In this work, the combination of the gain values $\boldsymbol{g}^{(i)} = [g_1^{(i)}, \ldots, g_q^{(i)}, \ldots, g_Q^{(i)}]^T$ are jointly encouraged to increase their log-likelihood with the trained gain prior GMM for each source i.

3. Nonnegative matrix factorization

Nonnegative matrix factorization is a matrix factorization algorithm with nonnegativity constraints. A nonnegative matrix $\boldsymbol{V}$ can be decomposed into a nonnegative basis vectors matrix $\boldsymbol{B}$ and a nonnegative gains matrix $\boldsymbol{G}$ as follows:

$$\boldsymbol{V} \approx \boldsymbol{BG}. \tag{5}$$

The columns of matrix $\boldsymbol{B}$ contain nonnegative basis vectors that are optimized to allow the data in $\boldsymbol{V}$ to be approximated as a nonnegative linear combination of its constituent vectors. To solve for matrix $\boldsymbol{B}$ and $\boldsymbol{G}$, a variety of cost functions can be used. The most used cost functions in source separation are the generalized Kullback–Leibler (KL–NMF) divergence cost function (Lee and Seung, 2001):

$$\min_{\boldsymbol{B},\boldsymbol{G}} D_{KL}(\boldsymbol{V}||\boldsymbol{BG}), \tag{6}$$

where

$$D_{KL}(\boldsymbol{V}||\boldsymbol{BG}) = \sum_{j,n}\left(\boldsymbol{V}_{j,n}\log\frac{\boldsymbol{V}_{j,n}}{(\boldsymbol{BG})_{j,n}} - \boldsymbol{V}_{j,n} + (\boldsymbol{BG})_{j,n}\right),$$

and the Itakura–Saito (IS–NMF) divergence cost function (Fevotte et al., 2009):

$$\min_{\boldsymbol{B},\boldsymbol{G}} D_{IS}(\boldsymbol{V}||\boldsymbol{BG}), \tag{7}$$

where

$$D_{IS}(\boldsymbol{V}||\boldsymbol{BG}) = \sum_{j,n}\left(\frac{\boldsymbol{V}_{j,n}}{(\boldsymbol{BG})_{j,n}} - \frac{\boldsymbol{V}_{j,n}}{(\boldsymbol{BG})_{j,n}} - 1\right).$$

These divergence cost functions were found to work better for audio source separation, and they are good measurements for the perceptual differences between different audio signals (Fevotte et al., 2009; Jaureguiberry et al., 2011; Wilson et al., 2008b).

The KL–NMF solutions for Eq. (6) can be computed by alternating multiplicative updates of $\boldsymbol{B}$ and $\boldsymbol{G}$ as in Virtanen (2007):

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \otimes \frac{\frac{\boldsymbol{V}}{\boldsymbol{BG}}\boldsymbol{G}^T}{\mathbf{1}\boldsymbol{G}^T}, \tag{8}$$

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \otimes \frac{\boldsymbol{B}^T\frac{\boldsymbol{V}}{\boldsymbol{BG}}}{\boldsymbol{B}^T\mathbf{1}}, \tag{9}$$

where $\mathbf{1}$ is a matrix of ones which has the same size as $\boldsymbol{V}$, The IS–NMF solutions for Eq. (7) can be computed by alternating multiplicative updates of $\boldsymbol{B}$ and $\boldsymbol{G}$ as shown in Fevotte et al. (2009), and Jaureguiberry et al. (2011):

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \otimes \frac{\frac{\boldsymbol{V}}{(\boldsymbol{BG})^2}\boldsymbol{G}^T}{\frac{1}{\boldsymbol{BG}}\boldsymbol{G}^T}, \tag{10}$$

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \otimes \frac{\boldsymbol{B}^T\frac{\boldsymbol{V}}{(\boldsymbol{BG})^2}}{\boldsymbol{B}^T\frac{1}{\boldsymbol{BG}}}, \tag{11}$$

where the operation $\otimes$ is an element-wise multiplication, all divisions and $(\cdot)^2$ are element-wise operations.

In source separation applications, it is important to note that KL–NMF is used with matrices of magnitude spectrograms (Virtanen, 2007; Grais and Erdogan, 2011c), but IS–NMF is used with matrices of power spectral densities (Fevotte et al., 2009; Jaureguiberry et al., 2011). In case of using IS–NMF, the magnitude spectra in Eqs. (2)–(4) are replaced by power spectral densities (PSDs). We continue our explanation in general using KL–NMF and we will mention the differences in case of using IS–NMF later.

3.1. The proposed regularized nonnegative matrix factorization

The goal of regularized NMF is to incorporate prior information on the solutions of the matrices $\boldsymbol{B}$ and $\boldsymbol{G}$. In this work, we enforce a statistical prior on the solution of the gains matrix $\boldsymbol{G}$ only. We need the solution of $\boldsymbol{G}$ in Eq. (5) to minimize the KL-divergence cost function in Eq. (6), and the log-normalized columns of the gains matrix $\boldsymbol{G}$, namely $\log\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$, to maximize their log-likelihood under a trained GMM prior model. Hence, the solution of $\boldsymbol{G}$ can be found by minimizing the following regularized KL-divergence cost function:

$$C = D_{KL}(\boldsymbol{V}||\boldsymbol{BG}) - \alpha L(\boldsymbol{G}|\theta), \tag{12}$$

where $L(\boldsymbol{G}|\theta)$ is the log-likelihood of the log-normalized columns of the gains matrix $\boldsymbol{G}$ under the trained prior gain GMM with parameters θ, and α is a regularization parameter. The regularization parameter controls the trade-off

between the NMF cost function and the prior log-likelihood. The multivariate Gaussian mixture model (GMM) with parameters $\theta = \{w_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^{K}$ for a random variable $\boldsymbol{x}$ is defined as:

$$p(\boldsymbol{x}|\theta) = \sum_{k=1}^{K} \frac{w_k}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}, \tag{13}$$

where K is the number of Gaussian mixture components, w_k is the mixture weight, d is the vector dimension, $\boldsymbol{\mu}_k$ is the mean vector and Σ_k is the diagonal covariance matrix of the kth Gaussian model. In this section, we assume GMM parameters θ are given. We will mention the training of θ in Section 4. In this paper, the normalization is done using the ℓ^2 norm by modeling $\log \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$.

The reason for using the logarithm is because GMM is usually a better fit to the logarithm of the values between 0 and 1 due to wider support as observed in tandem speech recognition research (Wessel et al., 2000). The reason for normalization is to make the prior models insensitive to the change of the energy level of the signals, which makes the same prior models applicable for a wide range of energy levels and avoids the need to train a different prior model for different energy levels. The log-likelihood for the gains matrix $\boldsymbol{G}$ with N columns can be written as follows:

$$L(\boldsymbol{G}|\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} A_{k,n}(\theta), \tag{14}$$

where

$$A_{k,n}(\theta) = \frac{w_k}{(2\pi)^{(d/2)}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}\left(\log\frac{\boldsymbol{g}_n}{||\boldsymbol{g}_n||_2} - \boldsymbol{\mu}_k\right)^T \Sigma_k^{-1}\left(\log\frac{\boldsymbol{g}_n}{||\boldsymbol{g}_n||_2} - \boldsymbol{\mu}_k\right)\right\}, \tag{15}$$

and $\boldsymbol{g}_n$ is the column numbered n in the gains matrix $\boldsymbol{G}$. The multiplicative update rule for the basis matrix $\boldsymbol{B}$ for the cost function in Eq. (12) is the same as in Eq. (8). To find the multiplicative update rule solution for $\boldsymbol{G}$ in Eq. (12), we follow the same procedures as in Virtanen (2007), and Bertin et al. (2010). We express the gradient with respect to $\boldsymbol{G}$ of the cost function $\nabla_G C$ as the difference of two positive terms $\nabla_G^+ C$ and $\nabla_G^- C$ as:

$$\nabla_G C = \nabla_G^+ C - \nabla_G^- C. \tag{16}$$

The cost function is shown to be nonincreasing under the following update rule (Virtanen, 2007; Bertin et al., 2010):

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \otimes \frac{\nabla_G^- C}{\nabla_G^+ C}, \tag{17}$$

where the operations $\otimes$ and division are element-wise as in Eq. (9). We can write the gradients as:

$$\nabla_G C = \nabla_G D_{KL} - \alpha \nabla_G L(\boldsymbol{G}|\theta), \tag{18}$$

where $\nabla_G L(\boldsymbol{G}|\theta)$ is a matrix with the same size of $\boldsymbol{G}$. The gradient for the KL-cost function and the prior log-likelihood can also be divided as follows:

$$\nabla_G D_{KL} = \nabla_G^+ D_{KL} - \nabla_G^- D_{KL}, \tag{19}$$

$$\nabla_G L(\boldsymbol{G}|\theta) = \nabla_G^+ L(\boldsymbol{G}|\theta) - \nabla_G^- L(\boldsymbol{G}|\theta). \tag{20}$$

We can rewrite Eqs. (16) and (18) as:

$$\nabla_G C = (\nabla_G^+ D_{KL} + \alpha \nabla_G^- L(\boldsymbol{G}|\theta)) - (\nabla_G^- D_{KL} + \alpha \nabla_G^+ L(\boldsymbol{G}|\theta)). \tag{21}$$

The final update rule in Eq. (17) can be written as follows:

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \otimes \frac{\nabla_G^- D_{KL} + \alpha \nabla_G^+ L(\boldsymbol{G}|\theta)}{\nabla_G^+ D_{KL} + \alpha \nabla_G^- L(\boldsymbol{G}|\theta)}, \tag{22}$$

where

$$\nabla_G D_{KL} = \boldsymbol{B}^T \left(\boldsymbol{1} - \frac{\boldsymbol{V}}{(\boldsymbol{BG})}\right), \tag{23}$$

$$\nabla_G^- D_{KL} = \boldsymbol{B}^T \frac{\boldsymbol{V}}{(\boldsymbol{BG})}, \tag{24}$$

and

$$\nabla_G^+ D_{KL} = \boldsymbol{B}^T \mathbf{1}. \tag{25}$$

The row j and column n component of the gradient of the prior log-likelihood in Eq. (14) can be found as follows:

$$(\nabla_G L(\boldsymbol{G}|\theta))_{jn} = (\nabla_G^+ L(\boldsymbol{G}|\theta))_{jn} - (\nabla_G^- L(\boldsymbol{G}|\theta))_{jn}, \tag{26}$$

where

$$(\nabla_G^- \boldsymbol{L}(\boldsymbol{G}|\theta))_{jn} = \frac{\sum_{k=1}^K \left\{ -A_{k,n} (\sum_{k_{jj}})^{-1} \left(\frac{\boldsymbol{\mu}_{k_j}}{\boldsymbol{g}_{jn}} + \frac{\boldsymbol{g}_{jn}}{\|\boldsymbol{g}_n\|_2^2} \log \frac{\boldsymbol{g}_{jn}}{\|\boldsymbol{g}_n\|_2} \right) \right\}}{\sum_{k=1}^K A_{k,n}}, \tag{27}$$

$$(\nabla_G^+ \boldsymbol{L}(\boldsymbol{G}|\theta))_{jn} = \frac{\sum_{k=1}^K \left\{ -A_{k,n} (\sum_{k_{jj}})^{-1} \left(\frac{\boldsymbol{\mu}_{k_j} \boldsymbol{g}_{jn}}{\|\boldsymbol{g}_n\|_2^2} + \frac{1}{\boldsymbol{g}_{jn}} \log \frac{\boldsymbol{g}_{jn}}{\|\boldsymbol{g}_n\|_2} \right) \right\}}{\sum_{k=1}^K A_{k,n}}. \tag{28}$$

Since the GMMs are trained by log-normalized columns, we know that the values of the mean vectors $\boldsymbol{\mu}$ are always negative. The values of the vectors $\boldsymbol{g}$ are always positive, so the values from Eqs. (27) and (28) will be always positive. We can use Eqs. (24), (25), (27), and (28) to find the total gradients in Eq. (21) and then to derive the update rules for $\boldsymbol{G}$ in Eq. (22). The initialization of the matrix $\boldsymbol{G}$ is done by running one regular NMF iteration without any prior.

4. Training the source models

In the training stage, we aim to train a set of basis vectors for each source and a prior statistical GMM for the gain patterns that each set of basis vectors can receive for each source signal.

4.1. Sequential training

Given a set of training data for each source signal, the magnitude spectrogram $\boldsymbol{S}_{train}^{(i)}$ for each source i is calculated. The NMF is used to decompose $\boldsymbol{S}_{train}^{(i)}$ into basis matrix $\boldsymbol{B}^{(i)}$ and gains matrix $\boldsymbol{G}_{train}^{(i)}$. The gains matrix $\boldsymbol{G}_{train}^{(i)}$ is then used to train the prior GMM for each source. KL–NMF is used to decompose the magnitude spectrogram into bases and gains matrices as follows:

$$\boldsymbol{S}_{train}^{(i)} \approx \boldsymbol{B}^{(i)} \boldsymbol{G}_{train}^{(i)}, \tag{29}$$

$$\boldsymbol{B}^{(i)}, \boldsymbol{G}_{train}^{(i)} = arg\min_{\boldsymbol{B},\boldsymbol{G}} D_{KL}(\boldsymbol{S}_{train}^{(i)} || \boldsymbol{BG}).$$

After finding the basis and the gains matrices, the corresponding GMM parameters $\theta^{(i)}$ are then learned as follows:

$$\theta^{(i)} = arg\max_{\theta} L(\boldsymbol{G}^{(i)}|\theta). \tag{30}$$

We use multiplicative update rules in Eqs. (8) and (9) to find solutions for $\boldsymbol{B}^{(i)}$ and $\boldsymbol{G}^{(i)}$ in Eq. (29). All the matrices $\boldsymbol{B}$ and $\boldsymbol{G}_{train}$ are initialized by positive random noise with uniform distribution in the range $(\mathbf{0}, \mathbf{1}]$. In each iteration, we normalize the columns of $\boldsymbol{B}^{(i)}$ using the ℓ^2 norm and find $\boldsymbol{G}_{train}^{(i)}$ accordingly. After finding matrices $\boldsymbol{B}$ and $\boldsymbol{G}_{train}$ for all sources, all the basis matrices $\boldsymbol{B}$ are used in mixed signal decomposition as it is shown in Section 5. We use the gains matrices $\boldsymbol{G}_{train}$ to build statistical prior models. For each matrix $\boldsymbol{G}_{train}^{(i)}$, we normalize its columns and the logarithm is then calculated. These log-normalized columns are used to train a gain prior GMM for each source in Eq. (30) using the well-known expectation maximization (EM) algorithm (Dempster et al., 1977).

4.2. Joint training

In Section 4.1, the trained NMF basis and gains matrices for each source are found using Eqs. (8) and (9), and then the prior gain GMMs are trained using the logarithm of the normalized columns of the trained gains matrix. To match between the way the trained models are used during training with the way they are used during separation, we jointly train the basis vectors and the prior models simultaneously to minimize the regularized cost function:

$$(\boldsymbol{B}^{(i)}, \boldsymbol{G}^{(i)}_{train}, \theta^{(i)}) = arg \min_{\boldsymbol{B},\boldsymbol{G},\theta} D_{KL}(\boldsymbol{S}^{(i)}_{train} || \boldsymbol{BG}) - \alpha_{train} L(\boldsymbol{G}|\theta). \tag{31}$$

We use the trained NMF and GMM models from Section 4.1 as initializations for the source models, and then we update the model parameters by running alternating update (coordinate descent) iterations on $\boldsymbol{B}^{(i)}$, $\boldsymbol{G}^{(i)}_{train}$ and $\theta^{(i)}$ parameters. At each NMF iteration, we update the basis matrix $\boldsymbol{B}^{(i)}$ using update rule in (8) while keeping $\boldsymbol{G}^{(i)}$ fixed, and the gains matrix $\boldsymbol{G}^{(i)}_{train}$ is updated using update rule in (22) while keeping $\boldsymbol{B}^{(i)}$ and $\theta^{(i)}$ fixed. We use a fixed value for the regularization parameter α_{train} during training. The new gains matrix is then used to train a new GMM with its parameters $\theta^{(i)}$ using the EM algorithm initialized by the previous GMM parameters. By repeating this procedure at each NMF iteration during training, the basis matrix is learnt in a consistent way with the clustered structure of the gains matrix due to the usage of the GMM priors. Since the original NMF problem is non-convex and there may be many possible local minima, we conjecture that the prior term encourages an NMF solution which is more consistent with the GMM prior assumption of the gains matrix.

4.3. Determining the hyper-parameters

The hyper-parameters in our model are the number of basis vectors d, number of mixtures K, and the regularization parameter α_{train}. In addition, during testing, we may use different α parameters for each source depending on the energy ratios of source signals (speech-to-music or male-to-female energy ratios in our experiments) which yields better results than using fixed values as we explain in Sections 5 and 6.

These hyper-parameters, especially α value(s), may be learned using a fully Bayesian treatment by putting priors on them and using the evidence framework or the integrate-out method (MacKay, 1999). For Bayesian learning of number of mixtures in the GMM and the number of basis vectors, one needs to use nonparametric Bayesian methods of Dirichlet process mixtures (Ferguson, 1973) and Bayesian nonparametric NMF (Blei et al., 2010) which enable variable number of mixtures and NMF basis components respectively. This overall Bayesian treatment is possible since the divergence cost functions D_{KL} and D_{IS} can be seen as negative log-likelihood functions that depend on parameters of the NMF decomposition under the probabilistic interpretations of NMF (Cemgil, 2008; Fevotte et al., 2009). However, Bayesian solutions involve highly complicated computations due to sampling techniques and are pretty cumbersome to implement. We consider these approaches as out of scope for this paper and leave them as future work. Thus, we take the conventional approach of determining these parameters using grid search on validation data. Basically, we perform different experiments with a range of reasonable values for each of these hyper-parameters and choose the values that provide the best results on validation data.

5. Signal separation

After observing the mixed signal $y(\tau)$, the magnitude spectrogram $\boldsymbol{Y}$ of the mixed signal is computed using STFT. To find the contribution of every source in the mixed signal magnitude spectra, we use KL–NMF to decompose the magnitude spectra $\boldsymbol{Y}$ with the trained bases matrices $\boldsymbol{B}$ that were found from solving Eq. (29) as follows:

$$\boldsymbol{Y} \approx [\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(i)}, \ldots, \boldsymbol{B}^{(Z)}]\boldsymbol{G}. \tag{32}$$

Let $\boldsymbol{B} = [\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(i)}, \ldots, \boldsymbol{B}^{(Z)}]$. The only unknown here is the gains matrix $\boldsymbol{G}$ since the matrix $\boldsymbol{B}$ and the trained GMM parameters $\boldsymbol{\Theta} = \{\theta^{(1)}, \ldots, \theta^{(i)}, \ldots, \theta^{(Z)}\}$ were found during the training stage and they are fixed in the separation

stage. The matrix $\boldsymbol{G}$ is a combination of submatrices, and every column n of $\boldsymbol{G}$ is a concatenation of subcolumns as follows:

$$\begin{bmatrix} \boldsymbol{G}^{(1)} \\ . \\ . \\ \boldsymbol{G}^{(i)} \\ . \\ . \\ \boldsymbol{G}^{(Z)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{g}_1^{(1)} & . & . & \boldsymbol{g}_n^{(1)} & . & . & \boldsymbol{g}_N^{(1)} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ \boldsymbol{g}_1^{(i)} & . & . & \boldsymbol{g}_n^{(i)} & . & . & \boldsymbol{g}_N^{(i)} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ \boldsymbol{g}_1^{(Z)} & . & . & \boldsymbol{g}_n^{(Z)} & . & . & \boldsymbol{g}_N^{(Z)} \end{bmatrix}, \tag{33}$$

where N is the maximum number of columns in matrix $\boldsymbol{G}$, and $\boldsymbol{g}_n^{(i)}$ is the column number n in the gain submatrix $\boldsymbol{G}^{(i)}$ for source signal i. Each submatrix represents the gain combinations that their corresponding basis vectors in the bases matrix have in the mixed signal. For the log-normalized columns of the submatrix $\boldsymbol{G}^{(i)}$ there is a corresponding trained gain prior GMM. We need the solution of $\boldsymbol{G}$ in Eq. (32) to minimize the KL-divergence cost function in Eq. (6), and the log-normalized columns of each submatrix $\boldsymbol{G}^{(i)}$ in $\boldsymbol{G}$ to maximize the log-likelihood with its corresponding trained gain prior GMM. Combining these two objectives, the solution of $\boldsymbol{G}$ can be found by minimizing the following regularized KL-divergence cost function as in Eq. (12):

$$C = D_{KL}(\boldsymbol{Y}||\boldsymbol{BG}) - R(\boldsymbol{G}|\boldsymbol{\Theta}), \tag{34}$$

where $R(\boldsymbol{G})$ is the weighted sum of the log-likelihoods of the log-normalized columns of the gain submatrices in matrix $\boldsymbol{G}$. For each log-likelihood of the gain submatrix $\boldsymbol{G}^{(i)}$ there is a corresponding regularization parameter $\alpha^{(i)}$ and GMM parameters $\theta^{(i)}$. $R(\boldsymbol{G})$ can be written as follows:

$$R(\boldsymbol{G}|\boldsymbol{\Theta}) = \sum_{i=1}^{Z} \alpha^{(i)} L(\boldsymbol{G}^{(i)}|\theta^{(i)}), \tag{35}$$

where $L(\boldsymbol{G}^{(i)}|\theta^{(i)})$ is the log-likelihood for the submatrix $\boldsymbol{G}^{(i)}$ for source i as in Eq. (14). The regularization parameters play an important role in the separation performance as we show later. Each source subcolumns $[\boldsymbol{g}_1^{(i)}, \ldots, \boldsymbol{g}_n^{(i)}, \ldots, \boldsymbol{g}_N^{(i)}]$ in matrix $\boldsymbol{G}$ in Eq. (33) are normalized and treated separately than other subcolumns sets, and each set of subcolumns is associated with its corresponding trained gain prior GMM.

The multiplicative update rule solution for $\boldsymbol{G}$ can be found using Eqs. (22), (24), and (25) as follows:

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \otimes \frac{\nabla_G^- D_{KL} + \nabla_G^+ R(\boldsymbol{G}|\boldsymbol{\Theta})}{\nabla_G^+ D_{KL} + \nabla_G^- R(\boldsymbol{G}|\boldsymbol{\Theta})}, \tag{36}$$

where

$$\nabla_G R(\boldsymbol{G}|\boldsymbol{\Theta}) = \nabla_G^+ R(\boldsymbol{G}|\boldsymbol{\Theta}) - \nabla_G^- R(\boldsymbol{G}|\boldsymbol{\Theta}), \tag{37}$$

$\nabla_G R(\boldsymbol{G}|\boldsymbol{\Theta})$ is a matrix with the same size of $\boldsymbol{G}$ and it is a combination of submatrices as follows:

$$\nabla_G R(\boldsymbol{G}|\boldsymbol{\Theta}) = \begin{bmatrix} \alpha^{(1)} \nabla_G L(\boldsymbol{G}^{(1)}|\theta^{(1)}) \\ . \\ . \\ \alpha^{(i)} \nabla_G L(\boldsymbol{G}^{(i)}|\theta^{(i)}) \\ . \\ . \\ \alpha^{(Z)} \nabla_G L(\boldsymbol{G}^{(Z)}|\theta^{(Z)}) \end{bmatrix}, \tag{38}$$

and $\nabla_G L(\boldsymbol{G}^{(i)}|\theta^{(i)})$ can be found for each source i using Eqs. (26)–(28).

Normalizing vectors in the prior models slightly increases the derivation complexity and the computational requirements of the multiplicative update rule of the gains matrix, but it is beneficial in situations where the source signals occur with varying energy levels. Normalizing the training and testing gain matrices gives the prior models the chance to be applicable for any energy level that the source signals can take in the mixed signal regardless of the energy levels of the training signals. It is important to note that, normalization during the separation process is done only for maximizing the prior log-likelihood. The general solution for the cost function in Eq. (34) is not normalized. After finding the suitable solution for the matrix $\boldsymbol{G}$, the initial magnitude spectral estimate of each source i is found as follows:

$$\widetilde{\boldsymbol{S}}^{(i)} = \boldsymbol{B}^{(i)}\boldsymbol{G}^{(i)}. \tag{39}$$

5.1. Reconstruction of source signals and spectral masks

We can directly use the initial estimated spectrograms of the source signals in Eq. (39) as the final spectrogram estimate of every source, but the estimated spectra $\widetilde{\boldsymbol{S}}$ may not sum up to the mixture $\boldsymbol{Y}$. We usually get nonzero decomposion error.

$$\boldsymbol{Y} \approx \sum_{i=1}^{Z} \widetilde{\boldsymbol{S}}^{(i)}.$$

Assuming noise is negligible in our mixed signal, the spectrogram of the source signals' sum should be directly equal to the spectrogram of the mixed signal. To achieve this, we use the initial estimates $\widetilde{\boldsymbol{S}}$ to build a spectral mask (Grais and Erdogan, 2011b,c; Grais et al., 2012) as follows:

$$\boldsymbol{H}^{(i)} = \frac{(\boldsymbol{B}^{(i)}\boldsymbol{G}^{(i)})^p}{\sum_{j=1}^{Z}(\boldsymbol{B}^{(j)}\boldsymbol{G}^{(j)})^p}, \tag{40}$$

where $p>0$ is a parameter and $(.)^p$, and the division are element-wise operations. Note that elements of $\boldsymbol{H} \in [0, 1]$. Using different p values leads to different kinds of masks. The value of p controls the saturation level of the ratio in (40). When $p>1$, the larger component will dominate more in the mixture. At $p=\infty$, we achieve a binary mask (hard mask) which will choose the larger source component as the only component. In KL–NMF, $p=1$ is usually used (Raj et al., 2010, 2011; Virtanen and Cemgil, 2009; Nakano et al., 2011). These masks will scale every time-frequency component in the observed mixed signal spectrogram in Eq. (1) with a ratio that determines how much each source contributes in the mixed signal such that

$$\hat{S}^{(i)}(t, f) = H^{(i)}(t, f)Y(t, f), \tag{41}$$

where $\hat{S}^{(i)}(t, f)$ is the final estimated STFT for $S^{(i)}(t, f)$ in Eq. (1) for source i, and $H^{(i)}(t, f)$ is the column t and row f entry of the spectral mask $\boldsymbol{H}^{(i)}$ in Eq. (40). When $p=2$ in Eq. (40) the mask $H(t, f)$ can be considered as a Wiener filter and $\hat{S}^{(i)}(t, f)$ can be seen as the minimum mean square error (MMSE) estimate for the STFT of source i (Fevotte et al., 2009). As we can see, $\hat{S}^{(i)}(t, f)$ has the same phase angles as $Y(t, f)$ since $\boldsymbol{H}$ is a real filter. After finding the contribution of each source signal in the mixed signal, the estimated source signal $\hat{s}^{(i)}(\tau)$ can be found by using inverse STFT of $\hat{S}^{(i)}(t, f)$.

5.2. Signal separation using IS–NMF

In case of using IS–NMF rather than using KL–NMF, we only need to replace the gradients in Eqs. (23)–(25), respectively with

$$\nabla_G D_{IS} = \boldsymbol{B}^T \frac{\boldsymbol{1}}{\boldsymbol{BG}} - \boldsymbol{B}^T \frac{\boldsymbol{V}}{(\boldsymbol{BG})^2}, \tag{42}$$

$$\nabla_G^- D_{IS} = \boldsymbol{B}^T \frac{\boldsymbol{V}}{(\boldsymbol{BG})^2}, \tag{43}$$

and

$$\nabla_G^+ D_{IS} = \boldsymbol{B}^T \frac{\mathbf{1}}{\boldsymbol{BG}}. \tag{44}$$

These gradients are used to find the update rules in Eqs. (22) and (36). It is also important to note that the gradients in Eqs. (27), (28), and (38) will be the same in the IS–NMF framework. Training the bases in Section 4 is done by using the IS–NMF update rules. The IS–NMF is used in training and separation stages with power spectral density (PSD) matrices rather than using magnitude spectra as in the case of KL–NMF. In practice, we just use the squared magnitude spectra as PSD estimates. By using IS–NMF, the value $\widetilde{\boldsymbol{S}}^{(i)} = \boldsymbol{B}^{(i)}\boldsymbol{G}^{(i)}$ in Eqs. (39) and (40) is the PSD of the source i. The spectral mask that is usually used in IS–NMF is the Wiener filter (Fevotte et al., 2009), which means $p = 1$ in Eq. (40) since the values of the product $\boldsymbol{B}^{(i)}\boldsymbol{G}^{(i)}$ in IS–NMF represent PSD estimates for the sources.

6. Experiments and discussion

We applied the proposed algorithm to two different problems: the first problem is speech–music separation, and the second one is speech–speech separation. In each case, we tested our separation algorithm using both KL–NMF and IS–NMF. This procedure results in four different sets of experiments. The spectrograms for the training and testing signals were calculated by using the STFT: a Hamming window with 480 points length and 60% overlap was used and the FFT was taken at 512 points, the first 257 FFT points only were used since the conjugate of the remaining 255 points are involved in the first FFT points. In case of using KL–NMF we chose the value of the spectral mask parameter $p = 1$ in Eq. (40). In case of using IS–NMF we chose the Wiener filter to be the spectral mask in Eq. (40) as in Fevotte et al. (2009).

Performance measurement of the separation algorithms was done using the signal to noise ratio (SNR) as in Virtanen (2007), Virtanen and Cemgil (2009), and Radfar et al. (2010).

6.1. Speech–music separation

In this experiment, we used the proposed algorithm to separate a speech signal from a background piano music signal. Our main goal was to get the clean speech signal from the mixture of speech and piano signals. We simulated our algorithm on a collection of speech and piano data at 16 kHz sampling rate. For speech data, we used a male Turkish speech data for a single speaker[1]. The data was recorded using a headset microphone in clear office environment. The data contains 560 short utterances with approximate duration 4 s each. For training speech data, we used 540 short utterances, we used other 20 utterances for validation and testing with 10 utterances each. For music data, we downloaded piano music data from piano society web site (URL, 2009). We used 12 pieces with approximate 50 min total duration from different composers but from a single artist for training and left out one piece for testing. We trained 128 basis vectors for each source, which makes the size of each matrix $\boldsymbol{B}^{(speech)}$ and $\boldsymbol{B}^{(music)}$ to be 257×128. The simulated mixed data was formed by adding random portions of the test music file to the 20 speech utterance test and validation files at a different speech-to-music ratio (SMR) values in dB. The audio power levels of each file were found using the "audio voltmeter" program from the G.191 ITU-T STL software suite (URL, 2009). For each SMR value, we obtained 20 mixed utterances this way. The first 10 mixed files for each SMR were used as a validation set to choose the suitable values for regularization parameters. The other 10 mixed files were used for testing. The proposed algorithm was run first on the validation set by using different values for the regularization parameters. We started with very small value 0.0001 for the regularization parameters, and we gradually increased their values by a multiple of ten as long as the SNR results had been improved, until the SNR started to decrease, then we searched close to the tried values for the regularization parameters that gave the highest SNR. The suitable values of the regularization parameters that were found using the validation set were then used on the test set. The shown results for all experiments are the average SNR of the 10 mixed test utterances.

The suitable number of mixture components K was chosen by trying different values as we can see from Fig. 1. The figure shows the SNR in dB of the estimated speech signal at $SMR = -5$ dB, with joint training of the source models

[1] The speech data is available at http://students.sabanciuniv.edu/grais/DataSets/TurkishSpeech.

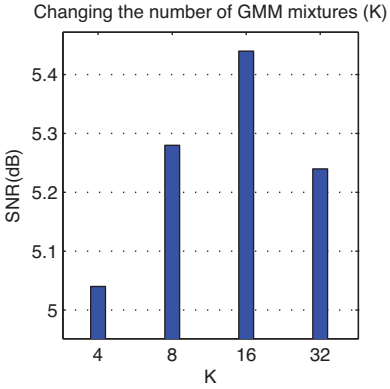

Fig. 1. The effect of changing the number of GMM mixture K for speech–music separation using KL–NMF at $SMR = -5$ dB, $\alpha^{(speech)} = \alpha^{(music)} = 0.005$, $\alpha_{train} = 0.0001$.

as shown in Section 4.2, with $\alpha_{train} = 0.0001$ for both sources, and $\alpha^{(speech)} = \alpha^{(music)} = 0.005$. We tried $K \in \{4, 8, 16, 32\}$. We got slightly better results for $K = 16$. We fixed the value of $K = 16$ for all other experiments.

To show the performance difference between using sequential training in Section 4.1 and using joint training in Section 4.2, we used KL–NMF with two different training cases. Table 1 shows the SNR of the separated speech signal using KL–NMF and sequential training for the source models. In this case, the regularization parameters $\alpha_{train} = 0$ for both sources. The second column shows the separation results of using NMF without using the GMM gain prior models in training and separation, which means the regularization parameters for separation $\alpha^{(speech)} = \alpha^{(music)} = 0$. In the third column, we show the case where the same values for the regularization parameters improve the separation results for all SMR cases compared to using NMF without any prior information. If we know some information about SMR of the mixed signal or estimate it online, we can choose different values for the regularization parameters for each SMR case, that can lead to better results as we can see in the last three columns of the same table. In Tables 1–5, the bold numbers are the best found SNR values and their corresponding regularization parameters are shown in the last two columns.

Table 2 shows the results with the same data as in Table 1 but with joint training for the source models. The second column in Table 2 shows the separation results of using NMF without using the GMM gain prior models in training and separation, which means $\alpha_{train} = 0$, $\alpha^{(speech)} = \alpha^{(music)} = 0$ for both sources. In the third column, we show the case where the same values for the regularization parameters improve the separation results for all SMR cases. In the last three columns of the table, better results based on better choices of the regularization parameters are shown assuming the SMR is known. The values of the regularization parameters during training stage are $\alpha_{train} = 0.0001$ for both sources in the third and fourth columns in Table 2. We can see that the results of jointly training the models in Table 2 are better than their corresponding results in Table 1 for the case of training the models separately.

Fig. 2 shows the signal to interference ratio (SIR) of the estimated speech signal for different cases. SIR is defined as the ratio of the target energy to the interference error due to the music signal only (Vincent et al., 2006). The figure

Table 1
SNR in dB for the speech signal for speech–music separation using regularized KL–NMF with $\alpha_{train} = 0$ and different values of the regularization parameters in testing $\alpha^{(speech)}$ and $\alpha^{(music)}$.

SMR	$\alpha^{(speech)} = 0$	$\alpha^{(speech)} = 0.01$	Best found values		
dB	$\alpha^{(music)} = 0$	$\alpha^{(music)} = 0.01$		$\alpha^{(speech)}$	$\alpha^{(music)}$
−5	4.33	4.53	**4.71**	0.1	0.05
0	7.96	8.14	**8.14**	0.01	0.01
5	9.71	9.86	**9.86**	0.01	0.01
10	11.75	11.86	**11.86**	0.01	0.01
15	13.72	13.88	**15.61**	0.01	0.05
20	14.09	14.25	**18.09**	0.001	1

Table 2
SNR in dB for the speech signal for speech–music separation using regularized KL–NMF with different values of the regularization parameters $\alpha^{(speech)}$, $\alpha^{(music)}$ and $\alpha_{train} = 0.0001$ for the third and fourth columns.

SMR	$\alpha^{(speech)} = 0$	$\alpha^{(speech)} = 0.005$	Best found values		
dB	$\alpha^{(music)} = 0$	$\alpha^{(music)} = 0.005$		$\alpha^{(speech)}$	$\alpha^{(music)}$
−5	4.33	5.44	**5.55**	0.01	0.01
0	7.96	8.70	**8.70**	0.005	0.005
5	9.71	10.25	**10.33**	0.001	0.005
10	11.75	12.03	**12.26**	0.001	0.005
15	13.72	14.20	**18.00**	0.001	0.1
20	14.09	14.44	**18.69**	0.001	0.1

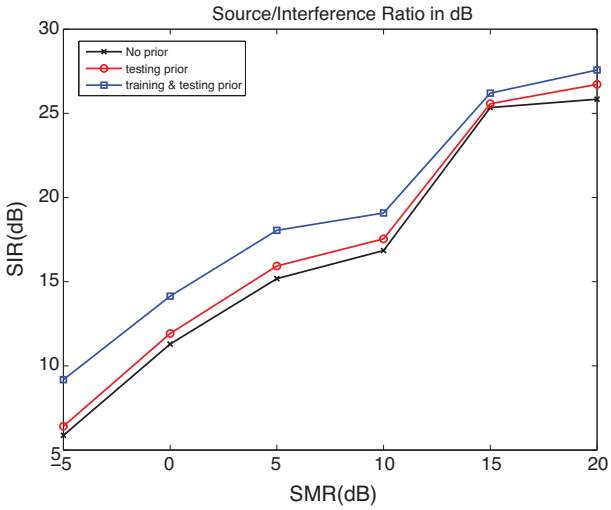

Fig. 2. The SIR for the case of using no priors during training and separation stages, the case of using prior only during testing, and the case of using prior during training and separation stages.

shows in black color the SIR corresponding to the case of using no prior in the second column in Tables 1 or 2. The SIR corresponding to the third column in Table 1 is shown in this figure in red color; in this case the priors were used during separation without performing joint training. The blue line in this figure shows the SIR corresponding to the third column in Table 2 where the joint training was applied with $\alpha_{train} = 0.0001$ for both sources and $\alpha^{(speech)} = \alpha^{(music)} = 0.005$. We can see from Fig. 2 and Tables 1 and 2 that using joint training improves the performance of the separation process. The shown values of the regularization parameters were selected based on the validation set. Since the joint training of the source models gives better results than the sequential training, we used joint training for our other experiments.

Table 3
SNR in dB for the speech signal for speech–music separation using regularized IS–NMF with different values of the regularization parameters $\alpha^{(speech)}$ and $\alpha^{(music)}$.

SMR	$\alpha^{(speech)} = 0$	$\alpha^{(speech)} = 0.5$	Best found values		
dB	$\alpha^{(music)} = 0$	$\alpha^{(music)} = 0.5$		$\alpha^{(speech)}$	$\alpha^{(music)}$
−5	3.66	4.19	**5.09**	0.5	0.1
0	8.02	8.51	**8.81**	0.5	0.1
5	10.54	10.62	**10.62**	0.5	0.5
10	12.64	12.84	**12.84**	0.5	0.5
15	16.45	17.02	**18.55**	0.1	1
20	16.80	17.28	**19.13**	0.1	1

Table 4
SNR in dB for the male speech signal for speech–speech separation using regularized KL–NMF with different values of the regularization parameters $\alpha^{(male)}$ and $\alpha^{(female)}$.

MFR	$\alpha^{(male)}=0$	$\alpha^{(male)}=0.01$	Best found values		
dB	$\alpha^{(female)}=0$	$\alpha^{(female)}=0.01$		$\alpha^{(male)}$	$\alpha^{(female)}$
−5	1.23	1.38	**1.61**	0.1	0.01
0	4.05	4.44	**4.44**	0.05	0.05
5	6.04	6.48	**6.64**	0.01	0.1
10	7.23	7.61	**10.03**	0.01	10
15	7.88	8.17	**14.51**	0.01	10
20	8.14	8.48	**19.12**	0.01	10

Table 3 shows the results with the same data in Table 2 with the same values of α_{train} but using IS–NMF with Wiener filter as a spectral mask.

We can see from the last three columns in both Tables 2 and 3 that, at low SMR we get better results when the values of $\alpha^{(speech)}$ is slightly higher than their values at high SMR. This means, when the speech signal has less energy in the mixed signal, we rely more on the prior model for the speech signal. As the energy level of the speech signal increases, the values of the speech prior parameter decreases and the value of the music prior parameter increases since the energy level of the music signal is decreased.

6.2. Speech–speech separation

In this experiment, we used the proposed regularized NMF algorithm to separate a male speech signal from a background female speech signal. Our main goal was to get a clean male speech signal from a mixture of male and female speech signals. We simulated our algorithm on a collection of male and female speech signals using the TIMIT database. For the training speech data, we used around 550 utterances from multiple male and female speakers from the training data of the TIMIT database. The validation and test data were formed using the TIMIT test data by adding 20 different female speech files to the 20 different male speech files at a different male-to-female ratio (MFR) values in dB. For each MFR value, we obtained 10 utterances for each test and validation set. We trained 32 basis vectors for each source, which makes the size of each matrix $\boldsymbol{B}^{(male)}$ and $\boldsymbol{B}^{(female)}$ to be 257×32. The number of the GMM components K is also 16 in this experiment.

Table 4 shows the signal to noise ratio of the separated male speech signal using KL–NMF. In the second column where no prior is used, the regularization parameters in training and testing are all equal to zero. For the third and fourth columns, the training regularization parameters $\alpha_{train}=0.001$ for both sources, and indicated values for the regularization parameters are used in testing.

Table 5 shows the results of using IS–NMF with different values of the regularization parameters $\alpha^{(male)}$, $\alpha^{(female)}$, and $\alpha_{train}=0.001$ for third and fourth columns.

Table 5
SNR in dB for the male speech signal for speech–speech separation using regularized IS–NMF with different values of the regularization parameters $\alpha^{(male)}$ and $\alpha^{(female)}$.

MFR	$\alpha^{(male)}=0$	$\alpha^{(male)}=1.5$	Best found values		
dB	$\alpha^{(female)}=0$	$\alpha^{(female)}=1.5$		$\alpha^{(male)}$	$\alpha^{(female)}$
−5	1.59	1.63	**1.66**	1.5	1
0	3.23	3.29	**3.45**	1.5	5
5	4.22	4.36	**5.64**	0.1	10
10	4.89	5.13	**9.59**	0.001	10^4
15	5.15	5.74	**14.29**	0.001	10^4
20	5.34	5.84	**18.04**	0.001	10^4

We can see from all tables that, comparing with no prior case, incorporating statistical prior information with NMF improves the performance of the separation algorithm. We also observe that, our proposed algorithm improves the performance of NMF regardless of the application and the used NMF cost function. In addition we found that, the same trained GMM prior model works for a wide range of energy levels avoiding the need to train different GMM model for each different energy level.

6.3. Comparison with conjugate prior

In this section we are trying to compare our proposed method of using GMM as a prior on the solution of NMF with the conjugate prior models for the case of KL–NMF. Instead of using GMM as a prior for the solution of the gains matrix during the separation process, the conjugate prior model is used as a prior for the gains matrix in this section. The probabilistic conjugate prior model for the solution of the gains matrix $\boldsymbol{G}$ for KL–NMF is the Gamma distribution as shown in Canny (2004). The probability distribution function (pdf) of the Gamma distribution with parameters a and b of a random variable x is defined as follows:

$$p(x) = \frac{x^{a-1}e^{-(x/b)}}{b^a\Gamma(a)}, \tag{45}$$

where $\Gamma(a)$ is the gamma function. The parameter a is known as the shape parameter and b is the scale parameter. These parameters can be selected individually for each gains matrix entry. Here, we fix the values for the parameters a and b for all entries of the gains matrix for each source. The update rule of the solution of the gains matrix in the separation stage that solve the cost function in Eq. (34) with Gamma prior is defined as follows (Canny, 2004):

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \otimes \frac{\boldsymbol{B}^T \frac{\boldsymbol{Y}}{\boldsymbol{BG}} + \frac{a.\hat{\mathbf{1}}-\hat{\mathbf{1}}}{\boldsymbol{G}}}{\boldsymbol{B}^T\mathbf{1} + \frac{\hat{\mathbf{1}}}{b.\hat{\mathbf{1}}}}, \tag{46}$$

where $\hat{\mathbf{1}}$ is a matrix of ones with the same size of $\boldsymbol{G}$, the operation $a.\hat{\mathbf{1}}$ means multiplying each entry of the matrix $\hat{\mathbf{1}}$ with a, and $\mathbf{1}$ is a matrix of ones with the same size of $\boldsymbol{Y}$. When the parameter $a = 1$ the prior distribution is an exponential distribution, and solving for $\boldsymbol{G}$ in the separation stage is equivalent to solving the following sparse KL–NMF problem (Virtanen and Cemgil, 2009)

$$C(\boldsymbol{G}) = D_{KL}(\boldsymbol{Y}||\boldsymbol{BG}) + \lambda\sum_{j,n}\boldsymbol{G}_{j,n}, \tag{47}$$

where the regularization parameter $\lambda = (1/b)$. In this case the update rule of $\boldsymbol{G}$ in (46) can be simplified as follows (Virtanen and Cemgil, 2009):

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \otimes \frac{\boldsymbol{B}^T \frac{\boldsymbol{Y}}{\boldsymbol{BG}}}{\boldsymbol{B}^T\mathbf{1} + \lambda.\hat{\mathbf{1}}}. \tag{48}$$

We repeated the speech-music separation experiment using KL–NMF in Section 6.1 with the same number of bases and $p = 1$ but using conjugate prior update rule in Eq. (46). We chose different values of the scale parameter for each source, b^s for speech and b^m for music. We used the same value of the shape parameter a for both sources. We tried different values of the parameters on the validation data and the parameter values that gave the best results were then used on the test data. Table 6 shows the signal to noise ratio of the separated speech signal using conjugate prior models in the case of KL–NMF with different values of the shape and scale parameters of the conjugate Gamma prior model for each source.

Comparing the results in Table 2 with the results in Table 6 we can see that, the third column results in Table 2 are better than their corresponding results in the third column in Table 6. Comparing the best found results in the last columns of both tables, we can see that using the GMM prior models give better results than using conjugate prior models at most SMR cases, and conjugate prior gives better results at $SMR = 20$ dB. As we can see in both cases there are many parameter values to be chosen and exact comparison cannot be achieved since we cannot test all possible combinations of the parameters. From running many experiments, we observed that, the performance in the case of using conjugate prior is very sensitive to small changes in the combination choices of the prior parameter values especially the shape parameter a. For each NMF divergence cost function there is a corresponding conjugate prior

Table 6
SNR in dB for the speech signal for speech–music separation using conjugate prior KL–NMF with different values of the prior parameters.

SMR	No prior	$a=1$	Best found values			
dB		$b^s=b^m=10^4$		a	b^s	b^m
−5	4.33	4.33	**4.33**	1	10^4	10^4
0	7.96	7.96	**8.02**	1	10^3	10^3
5	9.71	9.72	**9.80**	1	10^3	10^3
10	11.75	11.79	**11.86**	1	10^3	10^3
15	13.72	13.73	**16.71**	1.2	10^3	1
20	14.09	14.11	**19.45**	1.1	10^4	1

The best found SNR values are indicated in bold numbers.

distribution that must be chosen. In case of KL–NMF the conjugate prior distribution is the Gamma distribution, in IS–NMF case the conjugate prior distribution is the inverse-Gamma pdf (Fevotte et al., 2009). The GMM prior models can be applied regardless of the type of the NMF cost function.

7. Conclusion

In this work, we introduced a new regularized NMF algorithm for single channel source separation. The energy independent prior GMM was used to force the NMF solution to satisfy the statistical nature of the estimated source signals. The gains found in NMF solution were encouraged to increase their likelihood with the prior gain models of the source signals. Gaussian mixture models were used to model the log-normalized gain prior to improve the separation results. Our experiments indicate that the proposed approach is a promising method in single channel speech–music and speech–speech separation using various target-to-background energy ratios and different NMF divergence functions. As future work, we may consider the extension of our model using the Bayesian framework discussed in Section 4.3. In addition, we plan to model the gain prior using Hidden Markov Models (HMMs), which consider the statistical and dynamic nature of the source signals.

Acknowledgements

This research is partially supported by Turk Telekom under Grant Number 3014-06, project entitled “Single Channel Source Separation”, support year 2012. We would also like to thank and acknowledge the anonymous reviewers for many helpful comments and suggestions.

References

Bertin, N., Badeau, R., Vincent, E., 2009. Fast Bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

Bertin, N., Badeau, R., Vincent, E., 2010. Enforcing harmonicity and smoothness in bayesian nonnegative matrix factorization applied to polyphonic music transcription. IEEE Transactions on Audio, Speech, and Language Processing 18, 538–549.

Blei, D., Hoffman, M., Cook, P., 2010. Bayesian nonparametric matrix factorization for recorded music. In: International Conference on Machine Learning.

Canny, J., 2004. GaP: a factor model for discrete data. In: Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Cemgil, A., 2008. Bayesian inference in non-negative matrix factorisation models. Technical Report. CUED/F-INFENG/TR.609, University of Cambridge.

Cemgil, A.T., Dikmen, O., 2007. Conjugate Gamma Markov random fields for modelling nonstationary sources. In: International Conference on Independent Component Analysis and Signal Separation.

Chen, Z., Cichocki, A., Rutkowski, T.M., 2006. Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer’s disease. In: IEEE International Conference on Acoustics, Speech, and Signal Processing.

Cichocki, A., Zdunek, R., Amari, S., 2006. New algorithms for nonnegative matrix factorization in applications to blind source separation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Cichockiy, A., Georgiev, P., 2003. Blind source separation algorithms with matrix constraints. In: IEICE Transaction on Fundamentals of Electronics, Communications and Computer Sciences.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society.
Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1.
Fevotte, C., Bertin, N., Durrieu, J.L., 2009. Nonnegative matrix factorization with the itakura-saito divergence, With application to music analysis. Neural Computation 21, 793–830.
Grais, E.M., Erdogan, H., 2011a. Adaptation of speaker-specific bases in non-negative matrix factorization for single channel speech–music separation. In: Annual Conference of the International Speech Communication Association (INTERSPEECH).
Grais, E.M., Erdogan, H., 2011b. Single channel speech-music separation using matching pursuit and spectral masks. In: IEEE Conference on Signal Processing and Communications Applications (SIU).
Grais, E.M., Erdogan, H., 2011c. Single channel speech music separation using nonnegative matrix factorization and spectral masks. In: International Conference on Digital Signal Processing.
Grais, E.M., Erdogan, H., 2011d. Single channel speech music separation using nonnegative matrix factorization with sliding window and spectral masks. In: Annual Conference of the International Speech Communication Association (INTERSPEECH).
Grais, E.M., Topkaya, I.S., Erdogan, H., 2012. Audio–visual speech recognition with background music using single-channel source separation. In: IEEE Conference on Signal Processing and Communications Applications (SIU).
Helén, M., Virtanen, T., 2005. Separation of drums from polyphonic music using nonnegative matrix factorization and support vector machine. In: European Signal Processing Conference.
Jaureguiberry, X., Leveau, P., Maller, S., Burred, J.J., 2011. Adaptation of source-specific dictionaries in non-negative matrix factorization for source separation. In: IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP).
Lee, D.D., Seung, H.S., 1999. Learning of the parts of objects by non-negative matrix factorization. Nature 401, 788–791.
Lee, D.D., Seung, H.S., 2001. Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems 13, 556–562.
MacKay, D.J.C., 1999. Comparison of approximate methods for handling hyperparameters. Neural Computation 11, 1035–1068.
Nakano, S., Yamamoto, K., Nakagawa, S., 2011. Speech recognition in mixed sound of speech and music based on vector quantization and non-negative matrix factorization. In: Annual Conference of the International Speech Communication Association (INTERSPEECH).
Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ.
Radfar, M.H., Wong, W., Dansereau, R.M., Chan, W.Y., 2010. Scaled factorial Hidden Markov Models: a new technique for compensating gain differences in model-based single channel speech separation. In: IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP).
Raj, B., Singh, R., Virtanen, T., 2011. Phoneme-dependent NMF for speech enhancement in monaural mixtures. In: Annual Conference of the International Speech Communication Association (INTERSPEECH).
Raj, B., Virtanen, T., Chaudhure, S., Singh, R., 2010. Non-negative matrix factorization based compensation of music for automatic speech recognition. In: Annual Conference of the International Speech Communication Association (INTERSPEECH).
Schmidt, M.N., Olsson, R.K., 2006. Single-channel speech separation using sparse non-negative matrix factorization. In: International Conference on Spoken Language Processing (INTERSPEECH).
URL, 2009. http://pianosociety.com
URL, 2009. http://www.itu.int/rec/T-REC-G.191/en
Vincent, E., Gribonval, R., Fevotte, C., 2006. Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 14, 1462–1469.
Virtanen, T., 2007. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Transactions on Audio Speech, and Language Processing 15, 1066–1074.
Virtanen, T., Cemgil, A.T., 2009. Mixtures of gamma priors for non-negative matrix factorization based speech separation. In: International Conference on Independent Component Analysis and Blind Signal Separation.
Virtanen, T., Cemgil, A.T., Godsill, S., 2008. Bayesian extensions to non-negative matrix factorization for audio signal modeling. In: IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP).
Wang, B., Plumbley, M.D., 2006. Investigating single-channel audio source separation methods based on non-negative matrix factorization. In: International Workshop of the ICA Research Network.
Wessel, F., Schluter, R., Ney, H., 2000. Using posterior word probabilities for improved speech recognition. In: IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP).
Wilson, K.W., Raj, B., Smaragdis, P., 2008a. Regularized non-negative matrix factorization with temporal dependencies for speech denoising. In: Annual Conference of the International Speech Communication Association (INTERSPEECH).
Wilson, K.W., Raj, B., Smaragdis, P., Divakaran, A., 2008b. Speech denoising using nonnegative matrix factorization with priors. In: IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP).