BAYESIAN SINGING-VOICE SEPARATION

Po-Kai Yang, Chung-Chien Hsu and Jen-Tzung Chien

Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan
{niceallen.cm01g, chien.cm97g, jtchien}@nctu.edu.tw

ABSTRACT

This paper presents a Bayesian nonnegative matrix factorization (NMF) approach to extract singing voice from background music accompaniment. Using this approach, the likelihood function based on NMF is represented by a Poisson distribution and the NMF parameters, consisting of basis and weight matrices, are characterized by the exponential priors. A variational Bayesian expectation-maximization algorithm is developed to learn variational parameters and model parameters for monaural source separation. A clustering algorithm is performed to establish two groups of bases: one is for singing voice and the other is for background music. Model complexity is controlled by adaptively selecting the number of bases for different mixed signals according to the variational lower bound. Model regularization is tackled through the uncertainty modeling via variational inference based on marginal likelihood. The experimental results on MIR-1K database show that the proposed method performs better than various unsupervised separation algorithms in terms of the global normalized source to distortion ratio.

1. INTRODUCTION

Singing voice conveys important information of a song. This information is practical for many music-related applications including singer identification [11], music emotion annotation [21], melody extraction, lyric recognition and lyric synchronization [6]. However, singing voice is usually mixed with background accompaniment in a music signal. How to extract the singing voice from a single-channel mixed signal is known as a crucial issue for music information retrieval. Some approaches have been proposed to deal with single-channel singing-voice separation.

There are two categories of approaches to source separation: supervised learning [2] and unsupervised learning [8, 9, 13, 22]. Supervised approach conducts the single-channel source separation given by the labeled training data from different sources. In the application of singing-voice separation, the separate training data of singing voice and background music should be collected. But, it is more practical to conduct the unsupervised learning for blind source separation by using only the mixed test data. In [13], the repeating structure of the spectrogram of the mixed music signal was extracted and applied for separation of music and voice. The repeating components from accompaniment signal were separated from the non-repeating components from vocal signal. A binary time-frequency masking was applied to identify the repeating background accompaniment. In [9], a robust principal component analysis was proposed to decompose the spectrogram of mixed signal into a low-rank matrix for accompaniment signal and a sparse matrix for vocal signal. System performance was improved by imposing the harmonicity constraints [22]. A pitch extraction algorithm was inspired by the computational auditory scene analysis [3] and was applied to extract the harmonic components of singing voice.

In general, the issue of singing-voice separation is seen as a single-channel source separation problem which could be solved by using the learning approach based on the nonnegative matrix factorization (NMF) [10, 19]. Using NMF, a nonnegative matrix is factorized into a product of a basis matrix and a weight matrix which are nonnegative [10]. NMF can be directly applied in Fourier spectrogram domain for audio signal processing. In [7], the nonnegative sparse coding was proposed to conduct sparse learning for overcomplete representation based on NMF. Such sparse coding provides efficient and robust solution to NMF. However, how to determine the regularization parameter for sparse representation is a key issue for NMF. In addition, the time-varying envelopes of spectrogram convey important information. In [16], one dimensional convolutive NMF was proposed to extract the bases, which considered the dependencies across successive columns of input spectrogram, and was applied for supervised single-channel speech separation. In [14], two dimensional NMF was proposed to discover fundamental bases for blind musical instrument separation in presence of harmonic variations from piano and trumpet. Number of bases was empirically determined. Nevertheless, the selection of the number of bases is known as a model selection problem in signal processing and machine learning. How to tackle this regularization issue plays an important role to assure generalization for future data in ill-posed condition [1].

Basically, uncertainty modeling via probabilistic framework is helpful to improve model regularization for NMF.

The uncertainties in singing-voice separation may come from improper model assumption, incorrect model order and possible noise interference, nonstationary environment, reverberant distortion. Under probabilistic framework, nonnegative spectral signals are drawn from probability distributions. The nonnegative parameters are also represented by prior distributions. Bayesian learning is introduced to deal with uncertainty decoding and build a robust source separation by maximizing the marginal likelihood over the randomness of model parameters. In [15], Bayesian NMF (BNMF) was proposed for image feature extraction based on the assumption of Gaussian likelihood and exponential prior. In the BNMF [4], an approximate Bayesian inference based on variational Bayesian (VB) algorithm using Poisson likelihood for observation data and Gamma prior for model parameters was proposed for image reconstruction. Implementation cost was demanding due to the numerical calculation of shape parameter. Although NMF was presented for singing-voice separation in [19, 23], the regularization issue was ignored and the sensitivity of system performance due to uncertain model and ill-posed condition was serious.

This paper presents a new model-based singing-voice separation. The novelties of this paper are twofold. The first one is to develop Bayesian approach to unsupervised singing-voice separation. Model uncertainty is compensated to improve the performance of source separation of vocal signal and background accompaniment signal. Number of bases is adaptively determined from the mixed signal according to the variational lower bound of the logarithm of a marginal likelihood over NMF basis and weight matrices. The second one is the theoretical contribution in Bayesian NMF. We construct a new Bayesian NMF where the likelihood function in NMF is drawn from Poisson distribution and the model parameters are characterized by exponential distributions. A closed-form solution to hyperparameters using the VB expectation-maximization (EM) [5] algorithm is derived for ease of implementation and computation. This BNMF is connected to standard NMF with sparseness constraint. But, using the BNMF, the regularization parameters or hyperparameters are optimally estimated from training data without empirical selection from validation data. Beyond the approaches in [4, 15], the proposed BNMF completely considers the dependencies of the variational objective on hyperparameters and derives the analytical solution to singing-voice separation.

2. NONNEGATIVE MATRIX FACTORIZATION

Lee and Seung [10] proposed the standard NMF where no probabilistic distribution was assumed. Given a nonnegative data matrix $\mathbf{X} \in \mathcal{R}_+^{M\times N}$, NMF aims to decompose data matrix $\mathbf{X}$ into a product of two nonnegative matrices $\mathbf{B} \in \mathcal{R}_+^{M\times K}$ and $\mathbf{W} \in \mathcal{R}_+^{K\times N}$. The (m,n)-th entry of $\mathbf{X}$ is approximated by $X_{mn} \approx [\mathbf{BW}]_{mn} = \sum_k B_{mk}W_{kn}$. NMF parameters $\Theta = \{\mathbf{B},\mathbf{W}\}$ consist of basis matrix $\mathbf{B}$ and weight matrix $\mathbf{W}$. The approximation based on NMF is optimized by minimizing the Kullback-Leibler (KL) divergence $D_{\text{KL}}(\mathbf{X}\,\|\,\mathbf{BW})$ between the observed data $\mathbf{X}$ and the approximated data $\mathbf{BW}$

$$\sum_{m,n}\left(X_{mn}\log\frac{X_{mn}}{[\mathbf{BW}]_{mn}} + [\mathbf{BW}]_{mn} - X_{mn}\right) \tag{1}$$

2.1 Maximum Likelihood Factorization

NMF approximation is revisited by introducing the probabilistic framework based on maximum likelihood (ML) theory. The nonnegative latent variable Z_{mkn} is embedded in data entry X_{mn} by $X_{mn} = \sum_k Z_{mkn}$ and is represented by a Poisson distribution with mean $B_{mk}W_{kn}$, i.e. $Z_{mkn} \sim \text{Pois}(Z_{mkn}; B_{mk}W_{kn})$ [4]. Log likelihood function of data matrix $\mathbf{X}$ given parameters Θ is expressed by

$$\begin{aligned}\log p(\mathbf{X}|\mathbf{B},\mathbf{W}) &= \log\prod_{m,n}\text{Pois}(X_{mn};\sum_k B_{mk}W_{kn})\\ &= \sum_{m,n}(X_{mn}\log[\mathbf{BW}]_{mn} - [\mathbf{BW}]_{mn} - \log\Gamma(X_{mn}+1))\end{aligned} \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function. Maximizing the log likelihood function in Eq. (2) based on Poisson distribution is equivalent to minimizing the KL divergence between $\mathbf{X}$ and $\mathbf{BW}$ in Eq. (1). This ML problem with missing variables $\mathbf{Z} = \{Z_{mkn}\}$ can be solved according to EM algorithm. In E step, the expectation function of the log likelihood of data $\mathbf{X}$ and latent variable $\mathbf{Z}$ given new parameters $\mathbf{B}^{(\tau+1)}$ and $\mathbf{W}^{(\tau+1)}$ is calculated with respect to $\mathbf{Z}$ under current parameters $\mathbf{B}^{(\tau)}$ and $\mathbf{W}^{(\tau)}$. In M step, we maximize the resulting auxiliary function to obtain the updating of NMF parameters which is equivalent to that of standard NMF in [10].

2.2 Bayesian Factorization

ML estimation is prone to find an over-trained model [1]. To improve model regularization, Bayesian approach is introduced to establish NMF for single-source separation. ML NMF was improved by considering the priors of basis matrix $\mathbf{B}$ and weight matrix $\mathbf{W}$ for Bayesian NMF (BNMF). Different specifications of likelihood function and prior distribution result in different solutions with different inference procedures. In [15], the approximation error of X_{mn} using $\sum_k B_{mk}W_{kn}$ is modeled by a zero-mean Gaussian distribution

$$X_{mn} \sim \mathcal{N}(X_{mn};\sum_k B_{mk}W_{kn},\sigma^2) \tag{3}$$

with the variance parameter σ^2 which is distributed by an inverse gamma prior. The priors of nonnegative B_{mk} and W_{kn} are modeled by the exponential distributions

$$B_{mk} \sim \text{Exp}(B_{mk};\lambda_{mk}^b),\;\; W_{kn} \sim \text{Exp}(W_{kn};\lambda_{kn}^w) \tag{4}$$

where $\text{Exp}(x;\theta) = \theta\exp(-\theta x)$, with means $(\lambda_{mk}^b)^{-1}$ and $(\lambda_{kn}^w)^{-1}$, respectively. Typically, the larger the exponential hyperparameter θ is involved, the sparser the exponential distribution is shaped. The sparsity of basis parameter B_{mk} and weight parameter W_{kn} is controlled by hyperparameters λ_{mk}^b and λ_{kn}^w, respectively. In [15], the hyperparameters $\{\lambda_{mk}^b, \lambda_{kn}^w\}$ were fixed and empirically determined. The Gaussian likelihood does not adhere to the assumption of nonnegative data matrix $\mathbf{X}$. The other weakness in the BNMF [15] is that the exponential distribution is not conjugate prior to the Gaussian likelihood function for NMF. There was no closed-form solution. The parameters $\Theta = \{\mathbf{B}, \mathbf{W}, \sigma^2\}$ were accordingly estimated by Gibbs sampling procedure where a sequence of posterior samples of Θ was drawn by the corresponding conditional posterior probabilities.

Cemgil [4] proposed the BNMF for image reconstruction based on the Poisson likelihood function as given in Eq. (2) and the gamma priors for basis and weight matrices. The gamma distribution, represented by a shape parameter and a scale parameter, is known as the conjugate prior to Poisson likelihood function. Variational Bayesian (VB) inference procedure was developed for NMF implementation. However, the shape parameter was implemented by the numerical solution. The computation cost was relatively high. Some dependencies of variational lower bound on model parameters were ignored in [4]. The resulting parameters did not reach true optimum of variational objective.

3. NEW BAYESIAN FACTORIZATION

This study aims to find an analytical solution to full Bayesian NMF by considering all dependencies of variational lower bound on regularization parameters. Regularization parameters are optimally estimated.

3.1 Bayesian Objectives

In accordance with the Bayesian perspective and the spirit of standard NMF, we adopt the Poisson distribution as likelihood function and the exponential distribution as *conjugate prior* for NMF parameters B_{mk} and W_{kn} with hyperparameters λ^b_{mk} and λ^w_{kn}, respectively. Maximum *a posteriori* (MAP) estimates of parameters $\Theta = \{\mathbf{B}, \mathbf{W}\}$ are obtained by maximizing the posterior distribution or minimizing $-\log p(\mathbf{B}, \mathbf{W}|\mathbf{X})$ which is arranged as a regularized KL divergence between $\mathbf{X}$ and $\mathbf{BW}$

$$D_{\text{KL}}(\mathbf{X}||\mathbf{BW}) + \sum_{m,k} \lambda^b_{mk} B_{mk} + \sum_{k,n} \lambda^w_{kn} W_{kn} \tag{5}$$

where the terms independent of B_{mk} and W_{kn} are treated as constants. Notably, the regularization terms (2nd and 3rd terms) in this objective are nonnegative and seen as the ℓ_1 regularizers [18] which are controlled by hyperparameters $\{\lambda^b_{mk}, \lambda^w_{kn}\}$. These regularizers impose sparseness in the estimated MAP parameters.

However, MAP estimates are seen as point estimates. The randomness of parameters is not considered in model construction. To conduct full Bayesian treatment, BNMF is developed by maximizing the marginal likelihood $p(\mathbf{X}|\Theta)$ over latent variables $\mathbf{Z}$ as well as NMF parameters $\{\mathbf{B}, \mathbf{W}\}$

$$\int \sum_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z}, \mathbf{B}, \mathbf{W}) p(\mathbf{Z}|\mathbf{B}, \mathbf{W}) p(\mathbf{B}, \mathbf{W}|\Theta) d\mathbf{B} d\mathbf{W} \tag{6}$$

and estimating the sparsity-controlled hyperparameters or regularization parameters $\Theta = \{\lambda^b_{mk}, \lambda^w_{mk}\}$. The resulting evidence function is meaningful to act as an objective for model selection which balances the tradeoff between data fitness and model complexity [1]. In the singing-voice separation based on NMF, this objective is used to judge which number of bases K should be selected. The selected number is adaptive to fit different experimental conditions with varying lengths and the variations from different singers, genders, songs, genres, instruments and music accompaniments. Model regularization is tackled accordingly. But, using NMF without Bayesian treatment, the number of bases was fixed and empirically determined.

3.2 Variational Bayesian Inference

The exact Bayesian solution to optimization problem in Eq. (6) does not exist because the posterior probability of three latent variables $\{\mathbf{Z}, \mathbf{B}, \mathbf{W}\}$ given the observed mixtures $\mathbf{X}$ could not be factorized. To deal with this issue, the variational Bayesian expectation-maximization (VB-EM) algorithm is developed to implement Poisson-Exponential BNMF. VB-EM algorithm applies the Jensen's inequality and maximizes the lower bound of the logarithm of marginal likelihood

$$\begin{aligned} \log p(\mathbf{X}|\Theta) \geq \int \sum_{\mathbf{Z}} q(\mathbf{Z}, \mathbf{B}, \mathbf{W}) \log \frac{p(\mathbf{X}, \mathbf{Z}, \mathbf{B}, \mathbf{W}|\Theta)}{q(\mathbf{Z}, \mathbf{B}, \mathbf{W})} \\ \times d\mathbf{B} d\mathbf{W} = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z}, \mathbf{B}, \mathbf{W}|\Theta)] + H[q(\mathbf{Z}, \mathbf{B}, \mathbf{W})] \end{aligned} \tag{7}$$

where $H[\cdot]$ is an entropy function. The factorized variational distribution $q(\mathbf{Z}, \mathbf{B}, \mathbf{W}) = q(\mathbf{Z})q(\mathbf{B})q(\mathbf{W})$ is assumed to approximate the true posterior distribution $p(\mathbf{Z}, \mathbf{B}, \mathbf{W}|\mathbf{X}, \boldsymbol{\Theta})$.

3.2.1 VB-E Step

In VB-E step, a general solution to variational distribution q_j of an individual latent variable $j \in \{\mathbf{Z}, \mathbf{B}, \mathbf{W}\}$ is obtained by [1]

$$\log \hat{q}_j \propto \mathbb{E}_{q_{(i \neq j)}}[\log p(\mathbf{X}, \mathbf{Z}, \mathbf{B}, \mathbf{W}|\Theta)]. \tag{8}$$

Given the variational distributions defined by

$$\begin{aligned} q(B_{mk}) &= \text{Gam}(B_{mk}; \alpha^b_{mk}, \beta^b_{mk}) \\ q(W_{kn}) &= \text{Gam}(W_{kn}; \alpha^w_{kn}, \beta^w_{kn}) \\ q(Z_{mkn}) &= \text{Mult}(Z_{mkn}; P_{mkn}) \end{aligned} \tag{9}$$

the variational parameters $\{\alpha^b_{mk}, \beta^b_{mk}, \alpha^w_{kn}, \beta^w_{kn}, P_{mkn}\}$ in three distributions are estimated by

$$\begin{aligned} \hat{\alpha}^b_{mk} &= 1 + \sum_n \langle Z_{mkn} \rangle, \ \hat{\beta}^b_{mk} = \left(\sum_n \langle W_{kn} \rangle + \lambda^b_{mk} \right)^{-1} \\ \hat{\alpha}^w_{kn} &= 1 + \sum_m \langle Z_{mkn} \rangle, \ \hat{\beta}^w_{kn} = \left(\sum_k \langle B_{mk} \rangle + \lambda^w_{kn} \right)^{-1} \\ \hat{P}_{mkn} &= \frac{\exp(\langle \log B_{mk} \rangle + \langle \log W_{kn} \rangle)}{\sum_j \exp(\langle \log B_{mj} \rangle + \langle \log W_{jn} \rangle)} \end{aligned} \tag{10}$$

where the expectation function $\mathbb{E}_q[\cdot]$ is replaced by $\langle \cdot \rangle$ for simplicity. By substituting the variational distribution into

Eq. (7), the variational lower bound is obtained by

$$\begin{aligned}
\mathcal{B}_L = &- \sum_{m,n,k} \langle B_{mk}\rangle \langle W_{kn}\rangle \\
&+ \sum_{m,n} (-\log\Gamma(X_{mn}+1) - \sum_{k} \langle Z_{mkn}\rangle \log \hat{P}_{mkn}) \\
&+ \sum_{m,k} \langle \log B_{mk}\rangle \sum_{n} \langle Z_{mkn}\rangle + \sum_{k,n} \langle \log W_{kn}\rangle \sum_{m} \langle Z_{mkn}\rangle \\
&+ \sum_{m,k} (\log \lambda^b_{mk} - \lambda^b_{mk} \langle B_{mk}\rangle) + \sum_{k,n} (\log \lambda^w_{kn} - \lambda^w_{kn} \langle W_{kn}\rangle) \\
&+ \sum_{m,k} (-(\hat{\alpha}^b_{mk} - 1)\Psi(\hat{\alpha}^b_{mk}) + \log \hat{\beta}^b_{mk} + \hat{\alpha}^b_{mk} + \log\Gamma(\hat{\alpha}^b_{mk})) \\
&+ \sum_{k,n} (-(\hat{\alpha}^w_{kn} - 1)\Psi(\hat{\alpha}^w_{kn}) + \log \hat{\beta}^w_{kn} + \hat{\alpha}^w_{kn} + \log\Gamma(\hat{\alpha}^w_{kn}))
\end{aligned} \tag{11}$$

where $\Psi(\cdot)$ is the derivative of the log gamma function, and is known as a digamma function.

3.2.2 VB-M Step

In VB-M step, the optimal regularization parameters $\Theta = \{\lambda^b_{mk}, \lambda^w_{kn}\}$ are derived by maximizing Eq. (11) with respect to Θ and yielding

$$\begin{aligned}
\frac{\partial \mathcal{B}_L}{\partial \lambda^b_{mk}} &= \frac{1}{\lambda^b_{mk}} - \langle B_{mk}\rangle + \frac{\partial \log \beta^b_{mk}}{\partial \lambda^b_{mk}} = 0 \\
\frac{\partial \mathcal{B}_L}{\partial \lambda^w_{kn}} &= \frac{1}{\lambda^w_{kn}} - \langle W_{kn}\rangle + \frac{\partial \log \beta^w_{kn}}{\partial \lambda^w_{kn}} = 0.
\end{aligned} \tag{12}$$

Accordingly, the solution to BNMF hyperparameters is derived by solving a quadratic equation where nonnegative constraint is considered to find positive values of hyperparameters by

$$\begin{aligned}
\hat{\lambda}^b_{mk} &= \frac{1}{2}\left(-\sum_{n} \langle W_{kn}\rangle + \sqrt{(\sum_{n} \langle W_{kn}\rangle)^2 + 4 \frac{\sum_n \langle W_{kn}\rangle}{\langle B_{mk}\rangle}} \right) \\
\hat{\lambda}^w_{kn} &= \frac{1}{2}\left(-\sum_{m} \langle B_{mk}\rangle + \sqrt{(\sum_{m} \langle B_{mk}\rangle)^2 + 4 \frac{\sum_m \langle B_{mk}\rangle}{\langle W_{kn}\rangle}} \right)
\end{aligned} \tag{13}$$

where $\langle B_{mk}\rangle = \alpha^b_{mk}\beta^b_{mk}$ and $\langle W_{kn}\rangle = \alpha^w_{kn}\beta^w_{kn}$ are obtained as the means of gamma distributions. VB-E step and VB-M step are alternatively and iteratively performed to estimate BNMF parameters Θ with convergence. It is meaningful to select the best number of bases (K) with the largest lower bound of the log marginal likelihood which integrates out the parameters of weight and basis matrices.

3.3 Poisson-Exponential Bayesian NMF

To the best of our knowledge, this is the first study where a Bayesian approach is developed for singing-voice separation. The uncertainties in singing-voice separation due to a variety of singers, songs and instruments could be compensated. Model selection problem is tackled as well. In this study, total number of basis vectors K is adaptively selected for individual mixed signal according to the variational lower bound in Eq. (11) with the converged variational parameters $\{\hat{\alpha}^b_{mk}, \hat{\beta}^b_{mk}, \hat{\alpha}^w_{kn}, \hat{\beta}^w_{kn}, \hat{P}_{mkn}\}$ and model parameters $\{\hat{\lambda}^b_{mk}, \hat{\lambda}^w_{kn}\}$.

Considering the pairs of likelihood function and prior distribution in NMF, the proposed method is also called the Poisson-Exponential BNMF which is different from the Gaussian-Exponential BNMF in [15] and the Poisson-Gamma BNMF in [4]. The superiorities of the proposed method to the BNMFs in [15, 4] are twofold. First, assuming the exponential priors provides a BNMF approach with tractable solution as given in Eq. (13). Gibbs sampling in [15] and Newton's solution in [4] are computationally expensive. Second, the dependencies of three terms of the variational lower bound in Eq. (11) on hyperparameters λ^b_{mk} or λ^w_{kn} are all considered in finding the true optimum while some dependencies were ignored in the solution to Poisson-Gamma BNMF [4]. Also, the observations in Gaussian-Exponential BNMF [15] were not constrained to be nonnegative.

4. EXPERIMENTS

4.1 Experimental Setup

We used the MIR-1Kdataset [8] to evaluate the proposed method for unsupervised singing-voice separation from background music accompaniment. The dataset consisted of 1000 song clips extracted from 110 Chinese karaoke pop songs performed by 8 female and 11 male amateurs. Each clip recorded at 16 KHz sampling frequency with the duration ranging from 4 to 13 seconds. Since the music accompaniment and the singing voice were recorded at left and right channels, we followed [8, 9, 13] and simulated three different sets of monaural mixtures at signal-to-music-ratios (SMRs) of 5, 0, and -5 dB where the singing-voice was treated as signal and the accompaniment was treated as music. The separation problem was tackled in the short-time Fourier transform (STFT) domain. The 1024-point STFT was calculated to obtain the Fourier magnitude spectrograms with frame duration of 40 ms and frame shift of 10 ms. In the implementation of BNMF, ML-NMF was adopted as the initialization and 50 iterations were run to find the posterior means of basis and weight parameters. To evaluate the performance of singing-voice separation, we measure the signal-to-distortion ratio (SDR) [20] and then calculate the normalized SDR (NSDR) and the global NSDR (GNSDR) as

$$\begin{aligned}
\text{NSDR}(\hat{\mathbf{V}}, \mathbf{V}, \mathbf{X}) &= \text{SDR}(\hat{\mathbf{V}}, \mathbf{V}) - \text{SDR}(\mathbf{X}, \mathbf{V}) \\
\text{GNSDR}(\hat{\mathbf{V}}, \mathbf{V}, \mathbf{X}) &= \frac{\sum_{n=1}^{\tilde{N}} l_n \text{NSDR}(\hat{\mathbf{V}}_n, \mathbf{V}_n, \mathbf{X}_n)}{\sum_{n=1}^{\tilde{N}} l_n}
\end{aligned} \tag{14}$$

where $\hat{\mathbf{V}}, \mathbf{V}, \mathbf{X}$ denote the estimated singing voice, the original clean singing voice, and the mixture signal, respectively, $\tilde{N}$ is the total number of the clips and l_n is the length of the nth clip. NSDR is used to measure the improvement of SDR between the estimated singing voice $\hat{\mathbf{V}}$ and the mixture signal $\mathbf{X}$. GNSDR is used to calculate the overall separation performance by taking the weighted mean of the NSDRs.

4.2 Unsupervised Singing-Voice Separation

We implemented the unsupervised singing-voice separation where total number of bases (K) and the grouping of these bases into vocal source and music source were both

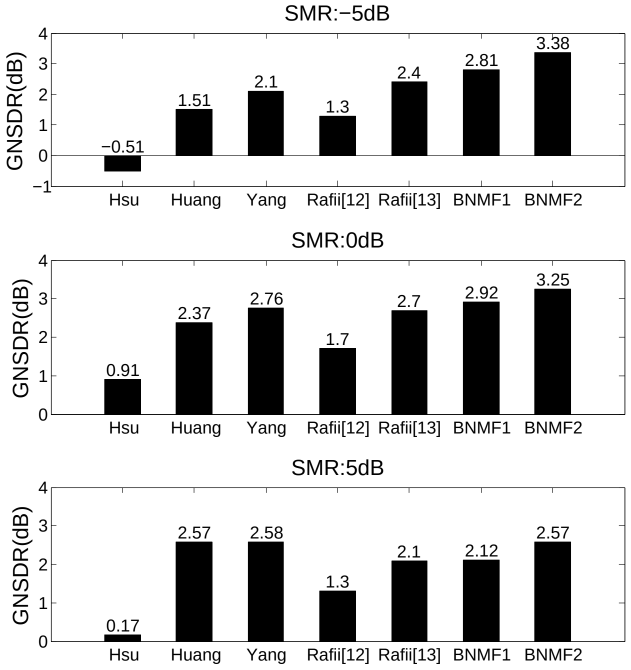

Figure 1. Performance comparison using BNMF1 (K-means clustering) and BNMF2 (NMF-clustering) and five competitive methods (Hsu [8], Huang [9], Yang [22], Rafii [12], Rafii [13]) in terms of GNSDR under various SMRs.

learned from test data in an unsupervised way. No training data were required. Model complexity based on K was determined in accordance with the variational lower bound of log marginal likelihood in Eq. (11) while the grouping of bases for two sources was simply performed via the clustering algorithms using the estimated basis vectors in $\mathbf{B}$ or equivalently from the estimated variational parameters $\{\alpha^b_{mk}, \beta^b_{mk}\}$. Following [17], we conducted the K-means clustering algorithm based on the basis vectors $\mathbf{B}$ in Mel-frequency cepstral coefficient (MFCC) domain. Each basis vector was first transformed to the Mel-scaled spectrum by applying 20 overlapping triangle filters spaced on the Mel scale. Then, we took the logarithm and applied the discrete cosine transform to obtain nine MFCCs. Finally, we normalized each coefficient to zero mean and unit variance. The K-means clustering algorithm was applied to partition the feature set into two clusters through an iterative procedure until convergence. However, it is more meaningful to conduct NMF-based clustering for the proposed BNMF method. To do so, we transformed the basis vectors $\mathbf{B}$ into Mel-scaled spectrum to form the Mel-scaled basis matrix. ML-NMF was applied to factorize this Mel-scaled basis matrix into two matrices $\tilde{\mathbf{B}}$ of size N-by-2 and $\tilde{\mathbf{W}}$ of size 2-by-K. The soft mask scheme based on Wiener gain was applied to smooth the separation of $\mathbf{B}$ into basis vectors for vocal signal and music signal. This same soft mask was performed for the separation of mixed signal X into vocal signal and music signal based on the K-means clustering and NMF clustering. Finally, the separated singing voice and music accompaniment signals were obtained by the overlap-and-add method using the original phase.

	NMF (30)	NMF (40)	NMF (50)	BNMF (adaptive)
K-means clustering	2.69	2.58	2.47	2.92
NMF clustering	3.15	3.13	2.97	3.25

Table 1. Comparison of GNSDR at SMR = 0 dB using NMF with fixed number of bases $\{30, 40, 50\}$ and BNMF with adaptive number of bases.

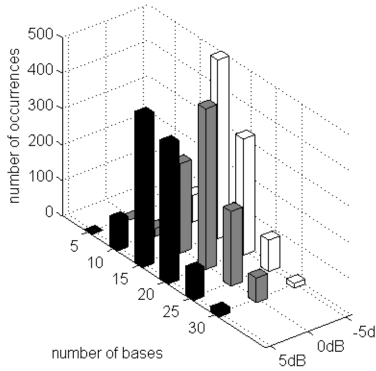

Figure 2. Histogram of the selected number of bases using BNMF under various SMRs.

4.3 Experimental Results

The unsupervised single-channel separation using BNMFs (BNMF1 using K-means clustering and BNMF2 using NMF clustering) and the other five competitive systems (Hsu [8], Huang [9], Yang [22], Rafii [12], Rafii [13]) is compared in terms of GNSDR as depicted in Figure 1. Using K-means clustering in MFCC domain, the resulting BNMF1 outperforms the other five methods under SMRs of 0 dB and -5 dB while the results using Huang [9] and Yang [22] perform better than BNMF1 under 5 dB condition. This is because the methods in [9, 22] used additional pre- and/or post-processing techniques as provided in [13, 22] which were not applied in BNMF1 and BNMF2. Nevertheless, using BNMF factorization with NMF clustering (BNMF2), the overall evaluation consistently achieves around 0.33$\sim$0.57 dB relative improvement in GNSDR compared with BNMF1 including the SMR condition at 5dB. In addition, we evaluate the effect on the adaptive basis selection using BNMF. Table 1 reports the comparison of BNMF1 and BNMF2 with adaptive basis selection and ML-NMF with fixed number of bases under SMR of 0 dB. Two clustering methods were also carried out for NMF with different K. BNMF factorization combined with NMF clustering achieves the best performance in this comparison. Figure 2 shows the histogram of the selected number of bases K using BNMF. It is obvious that this adaptive basis selection plays an important role to find suitable amount of bases to fit different experimental conditions.

5. CONCLUSIONS

We proposed a new unsupervised Bayesian nonnegative matrix factorization approach to extract the singing voice from background music accompaniment and illustrated the novelty on an analytical and true optimum solution to the Poisson-Exponential BNMF. Through the VB-EM inference procedure, the proposed method automatically selected different number of bases to fit various experimental conditions. We conducted two clustering algorithms to find the grouping of bases into vocal and music sources. Experimental results showed the consistent improvement of using BNMF factorization with NMF clustering over the other singing-voice separation methods in terms of GNSDR. In future works, the proposed BNMF shall be extended to multi-layer source separation and applied to detect unknown number of sources.

6. REFERENCES

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science, 2006.

[2] N. Boulanger-Lewandowski, G. J. Mysore, and M. Hoffman. Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation. In *Proc. of ICASSP*, pages 337–344, 2014.

[3] A. S. Bregman. *Auditory Scene Analysis: the Perceptual Organization of Sound*. MIT Press, 1990.

[4] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, (Article ID 785152), 2009.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society (B)*, 39(1):1–38, 1977.

[6] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno. Lyricsynchronizer: automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1252–1261, 2011.

[7] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

[8] C.-L. Hsu and J.-S. R. Jang. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Transactions on Audio, Speech, Language Processing*, 18(2):310–319, 2010.

[9] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *Proc. of ICASSP*, pages 57–60, 2012.

[10] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, pages 556–562, 2000.

[11] A. Mesaros, T. Virtanen, and A. Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *Proc. of Annual Conference of International Society for Music Information Retrieval*, pages 375–378, 2007.

[12] Z. Rafii and B. Pardo. A simple music/voice separation method based on the extraction of the repeating musical structure. In *Proc. of ICASSP*, pages 221–224, 2011.

[13] Z. Rafii and B. Pardo. Repeating pattern extraction technique (REPET): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech, Language Processing*, 21(1):73–84, Jan. 2013.

[14] M. N. Schmidt and M. Morup. Non-negative matrix factor 2-D deconvolution for blind single channel source separation. In *Proc. of ICA*, pages 700–707, 2006.

[15] M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In *Proc. of ICA*, pages 540–547, 2009.

[16] P. Smaragdis. Convolutive speech bases and their application to speech separation. *IEEE Transactions on Audio, Speech, Language Processing*, 15(1):1–12, 2007.

[17] M. Spiertz and V. Gnann. Source-Filter based clustering for monaural blind source separation. In *Proc. of International Conference on Digital Audio Effects*, pages 1–4, 2009.

[18] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[19] S. Vembu and S. Baumann. Separation of vocals from polyphonic audio recordings. In *Proc. of ISMIR*, pages 375–378, 2005.

[20] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transaction on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

[21] D. Yang and W. Lee. Disambiguating music emotion using software agents. In *Proc. of ISMIR*, pages 52–57, 2004.

[22] Y.-H. Yang. On sparse and low-rank matrix decomposition for singing voice separation. In *Proc. of ACM International Conference on Multimedia*, pages 757–760, 2012.

[23] B. Zhu, W. Li, R. Li, and X. Xue. Multi-stage non-negative matrix factorization for monaural singing voice separation. *IEEE Transactions on Audio, Speech, Language Processing*, 21(10):2096–2107, 2013.