Bayesian Factorization and Selection for Speech and Music Separation

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

Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan

Abstract

This paper proposes a new Bayesian nonnegative matrix factorization (NMF) for speech and music separation. We introduce the Poisson likelihood for NMF approximation and the exponential prior distributions for the factorized basis matrix and weight matrix. A variational Bayesian (VB) EM algorithm is developed to implement an efficient solution to variational parameters and model parameters for Bayesian NMF. Importantly, the exponential prior parameter is used to control the sparseness in basis representation. The variational lower bound in VB-EM procedure is derived as an objective to conduct adaptive basis selection for different mixed signals. The experiments on single-channel speech/music separation show that the adaptive basis representation in Bayesian NMF via model selection performs better than the NMF with the fixed number of bases in terms of signal-to-distortion ratio.

Index Terms: nonnegative matrix factorization, model selection, Bayesian learning, source separation

10.21437/Interspeech.2014-262

1. Introduction

Speech-related applications using single microphone have been ubiquitous in modern world. Many practical systems have been developed to meet different requirements. Usually, target speech is contaminated with a variety of interferences such as ambient noise, competing speech and background music. It is important to conduct single-channel source separation to extract target speech from a mixture of speech and music signals and apply it for automatic speech recognition [4, 9]. The extension to multichannel source separation for instantaneous and convolutive mixtures is also crucial for the applications in real-world sound environments with multiple sources [17].

Nonnegative matrix factorization (NMF) has been successfully developed for source separation. Using NMF, the nonnegative data is factorized into a product of a nonnegative basis (or template) matrix and a nonnegative weight (or activation) matrix [13, 14]. For audio signals, NMF can be directly applied in Fourier spectrogram domain. In [11], the nonnegative sparse coding was proposed to learn sparse overcomplete representation based on NMF. Such sparse coding provides efficient and robust solution to NMF. However, it is a key issue to determine the regularization parameter in sparse representation. In addition, the time-varying envelopes of the spectrogram convey important information. In [22], the 1-dimensional convolutive NMF was proposed to extract the bases which considered the dependencies across successive columns of input spectrogram for supervised single-channel speech separation. In [18], the 2-dimensional NMF was proposed to discover fundamental bases or notes for blind musical instrument separation in presence of harmonic variations from piano and trumpet with shift-invariance along the log-frequency. Number of bases was empirically determined. Nevertheless, how to determine the number of bases or model order is also a concern.

More attractively, the nonnegative elements can be considered as being drawn from an underlying probability distribution. Extending from the latent topic model, the probabilistic latent semantic indexing [10] and the probabilistic latent component analysis [20, 21] decomposed the probability of nonnegative input data into a product of two conditional probabilities given latent variable via the expectation-maximization (EM) algorithm [8]. In [19], a Bayesian NMF (BNMF) was proposed for image feature extraction based on the assumption of Gaussian likelihood and exponential prior. The approximate Bayesian inference using Gibbs sampling was performed. In [7], the Bayesian group sparse learning for NMF was introduced by using the Laplacian priors for sparse coding and the groups of common bases and individual bases for music source separation. In addition to the Gibbs sampling, the full Bayesian inference based on variational Bayesian (VB) algorithm using Poisson likelihood and Gamma prior was proposed for image reconstruction [3]. Implementation cost was demanding due to the numerical calculation of shape parameter.

This paper presents a new BNMF with Poisson likelihood and exponential prior and applies it to a new task of single-channel speech and music separation. VB-EM algorithm is developed to estimate the hyperparameters for NMF by maximizing the lower bound of a marginal likelihood over NMF basis and weight matrices. The closed-form solution to BNMF hyperparameters is obtained for ease of implementation and computation. These hyperparameters serve as the sparsity-control parameters for basis representation. In particular, the number of bases for source signals is adaptively selected to meet different variations in contents and lengths of training samples according to the same variational lower bound. The rest of the paper is organized as follows. Section 2 reviews a series of NMF including standard NMF, maximum likelihood NMF and Bayesian NMF. Section 3 addresses the optimization criteria, VB-EM inference and implementation procedure for the proposed BNMF. Section 4 shows the experiments and Section 5 draws the conclusions.

2. Non-negative matrix factorization

2.1. Standard NMF

Lee and Seung [14] proposed the NMF which has been successfully developed for many applications such as source separation. Given a nonnegative data matrix $\mathbf{X} \in \mathcal{R}_+^{M\times N}$, NMF aims to decompose the data matrix into a product of two nonnegative matrices $\mathbf{B} \in \mathcal{R}_+^{M\times K}$ and $\mathbf{W} \in \mathcal{R}_+^{K\times N}$, as follows

$$X_{mn} \approx [\mathbf{BW}]_{mn} = \sum_k B_{mk} W_{kn} \tag{1}$$

where the NMF parameters $\Theta = \{\mathbf{B}, \mathbf{W}\}$ consists of basis matrix $\mathbf{B}$ and weight matrix $\mathbf{W}$. The approximation based on NMF is optimized by minimizing the Kullback-Leibler (KL) divergence between the observed data $\mathbf{X}$ and the approximated

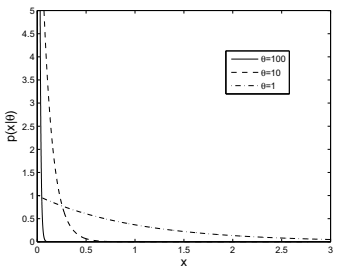

Figure 1: Exponential distribution with various θ.

data $\mathbf{BW}$

$$D_{\mathrm{KL}}(\mathbf{X}\,\|\,\mathbf{BW}) = \sum_{m,n}(X_{mn}\log\frac{X_{mn}}{[\mathbf{BW}]_{mn}} + [\mathbf{BW}]_{mn} - X_{mn}) \tag{2}$$

2.2. Maximum likelihood NMF

Standard NMF conducts the nonparametric approximation without considering parametric distributions for different variables. This approximation is revisited through the probabilistic framework based on the maximum likelihood (ML) theory by introducing the latent variable Z_{mkn} which is hidden in data component X_{mn} and is represented by a Poisson distribution with mean $B_{mk}W_{kn}$ [3]

$$X_{mn} = \sum_k Z_{mkn}, \;\; Z_{mkn} \sim \mathrm{Pois}(Z_{mkn}; B_{mk}W_{kn}) \tag{3}$$

where $\mathrm{Pois}(x;\theta) = \exp(x\log\theta - \theta - \log\Gamma(x+1))$ with gamma function $\Gamma(x+1) = x!$. The log likelihood function is expressed by

$$\begin{aligned}\log p(\mathbf{X}|\mathbf{B},\mathbf{W}) &= \log\sum_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z}|\mathbf{B},\mathbf{W}) \\ &= \log\prod_{m,n}\mathrm{Pois}(X_{mn};\sum_k B_{mk}W_{kn}) \\ &= \sum_{m,n}(X_{mn}\log[\mathbf{BW}]_{mn} - [\mathbf{BW}]_{mn} \\ &\qquad - \log\Gamma(X_{mn}+1)).\end{aligned} \tag{4}$$

We find that the maximization of log likelihood function based on Poisson distribution in Eq. (4) is equivalent to the minimization of information divergence [5] in Eq. (2). This ML problem with missing variables $\mathbf{Z} = \{Z_{mkn}\}$ could be solved according to the 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 solution to NMF parameters which is equivalent to that of standard NMF [14, 20]

$$\begin{aligned}B_{mk}^{(\tau+1)} &= B_{mk}^{(\tau)}\frac{\sum_n W_{kn}^{(\tau)}(X_{mn}/(\sum_j B_{mj}^{(\tau)}W_{jn}^{(\tau)}))}{\sum_n W_{kn}^{(\tau)}} \\ W_{kn}^{(\tau+1)} &= W_{kn}^{(\tau)}\frac{\sum_m B_{mk}^{(\tau)}(X_{mn}/(\sum_j B_{mj}^{(\tau)}W_{jn}^{(\tau)}))}{\sum_m B_{mk}^{(\tau)}}.\end{aligned} \tag{5}$$

2.3. Bayesian NMF

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 [19], the approximation error of X_{mn} using $\sum_k B_{mk}W_{kn}$ is modeled by a zero-mean Gaussian distribution $X_{mn} \sim \mathcal{N}(X_{mn};\sum_k B_{mk}W_{kn},\sigma^2)$ and the priors of B_{mk} and W_{kn} are modeled by the exponential distributions with means $(\lambda_{mk}^b)^{-1}$ and $(\lambda_{kn}^w)^{-1}$

$$B_{mk} \sim \mathrm{Exp}(B_{mk};\lambda_{mk}^b),\; W_{kn} \sim \mathrm{Exp}(W_{kn};\lambda_{kn}^w) \tag{6}$$

where $\mathrm{Exp}(x;\theta) = \theta\exp(-\theta x)$ and σ^2 is distributed by an inverse gamma distribution. Figure 1 displays the exponential distribution with different parameter values. Typically, larger parameter θ produces a sparser exponential distribution. 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 [19], the hyperparameters $\{\lambda_{mk}^b, \lambda_{kn}^w\}$ were fixed and empirically determined so that system performance was limited. The other weakness in this BNMF 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 model parameters Θ was drawn by the corresponding conditional posterior probabilities.

Cemgil [3] proposed the BNMF for image reconstruction based on the Poisson likelihood function as given in Eq. (4) and using the gamma priors for basis matrix and weight matrix. The gamma distribution, represented by the shape parameter and the scale parameter, is known as conjugate prior to Poisson likelihood function. Variational Bayesian (VB) inference procedure was developed for NMF implementation. However, the closed-form solution to the shape parameter does not exist. The computation cost is relatively high.

3. Bayesian factorization and selection

This study aims to find an efficient solution to full Bayesian NMF and applies it for monaural speech and music separation with adaptive model order selection. VB-EM algorithm is derived for model construction. The sparsity control in basis representation is introduced.

3.1. Optimization criteria

In Bayesian framework, it is important to select probabilistic distributions for likelihood function and prior density so that a meaningful solution could be developed to meet the demands of the application [6]. Considering the spirit of standard NMF with Bayesian perspective, 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_{mk}^b and λ_{kn}^w, respectively. The 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 $\mathbf{BW}$

$$D_{\mathrm{KL}}(\mathbf{X}||\mathbf{BW}) + \sum_{m,k}\lambda_{mk}^b B_{mk} + \sum_{k,n}\lambda_{kn}^w W_{kn} \tag{7}$$

where the terms independent of B_{mk} and W_{kn} are ignored. Notably, the regularization terms in this objective are seen as

the ℓ_1 regularizers which are controlled by hyperparameters $\{\lambda^b_{mk}, \lambda^w_{kn}\}$. These regularizers impose sparseness in the estimated MAP parameters.

However, this paper presents full Bayesian framework for NMF through maximizing the marginal likelihood $p(\mathbf{X}|\Theta)$ over latent variables $\mathbf{Z}$ as well as NMF parameters $\{\mathbf{B}, \mathbf{W}\}$

$$\int \sum_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z}, \mathbf{B}, \mathbf{W}) p(\mathbf{Z}|\mathbf{B}, \mathbf{W}) p(\mathbf{B}, \mathbf{W}|\Theta) d\mathbf{B} d\mathbf{W} \quad (8)$$

and estimating the sparsity-controlled regularization parameters $\Theta = \{\lambda^b_{mk}, \lambda^w_{mk}\}$. The resulting evidence function could be also applied to judge which number of bases K should be selected. This number is adaptive to fit different experimental conditions with varying lengths and contents of experimental data. Model order selection is performed accordingly. But, the number of bases was empirically determined by using standard NMF. Model selection is performed to balance the tradeoff between data fitting and model complexity [1]. 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 [19] and the Poisson-Gamma BNMF in [3]. New benefit and application are introduced.

3.2. Variational Bayesian inference

As we know, the exact Bayesian solution to optimization problem in Eq. (8) is not analytically tractable. It is because the posterior probability of three latent variables $\{\mathbf{Z}, \mathbf{B}, \mathbf{W}\}$ given the observed mixtures $\mathbf{X}$ could not be factorized for Bayesian inference. To deal with this issue, the variational Bayesian expectation-maximization (VB-EM) algorithm is developed to implement the Poisson-Exponential BNMF. VB-EM algorithm applies the Jensen's inequality and maximizes the lower bound of logarithm of marginal likelihood

$$\log p(\mathbf{X}|\Theta) \geq \int \sum_{\mathbf{Z}} q(\mathbf{Z}, \mathbf{B}, \mathbf{W}) \log \frac{p(\mathbf{X}, \mathbf{Z}, \mathbf{B}, \mathbf{W}|\Theta)}{q(\mathbf{Z}, \mathbf{B}, \mathbf{W})} \times d\mathbf{B} d\mathbf{W} = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z}, \mathbf{B}, \mathbf{W}|\Theta)] + H[q(\mathbf{Z}, \mathbf{B}, \mathbf{W})] \quad (9)$$

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 $p(\mathbf{Z}, \mathbf{B}, \mathbf{W}|\mathbf{X}, \boldsymbol{\Theta})$. 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]

$$\log \hat{q}_j \propto \mathbb{E}_{q_{(i \neq j)}}[\log p(\mathbf{X}, \mathbf{Z}, \mathbf{B}, \mathbf{W}|\Theta)]. \quad (10)$$

Given the variational distributions defined by

$$\begin{aligned} q(B_{mk}) &\propto \text{Gam}(B_{mk}; \alpha^b_{mk}, \beta^b_{mk}) \\ q(W_{kn}) &\propto \text{Gam}(W_{kn}; \alpha^w_{kn}, \beta^w_{kn}) \\ q(Z_{mkn}) &\propto \text{Mult}(Z_{mkn}; P_{mkn}) \end{aligned} \quad (11)$$

the variational parameters in three distributions are estimated by

$$\begin{aligned} \alpha^b_{mk} &= 1 + \sum_n \langle Z_{mkn} \rangle, \;\; \beta^b_{mk} = \left(\sum_n \langle W_{kn} \rangle + \lambda^b_{mk} \right)^{-1} \\ \alpha^w_{kn} &= 1 + \sum_m \langle Z_{mkn} \rangle, \;\; \beta^w_{kn} = \left(\sum_k \langle B_{mk} \rangle + \lambda^w_{kn} \right)^{-1} \\ P_{mkn} &= \frac{\exp(\langle \log B_{mk} \rangle + \langle \log W_{kn} \rangle)}{\sum_j \exp(\langle \log B_{mj} \rangle + \langle \log W_{jn} \rangle)} \end{aligned} \quad (12)$$

where the expectation function $\mathbb{E}_q[\cdot]$ is replaced by $\langle \cdot \rangle$ for simplicity. By substituting the variational distribution into Eq. (9), the variational lower bound is obtained by

$$\begin{aligned} \mathcal{B}_L = &- \sum_{m,n,k} \langle B_{mk} \rangle \langle W_{kn} \rangle \\ &+ \sum_{m,n} (-\log \Gamma(X_{mn} + 1) - \sum_k \langle Z_{mkn} \rangle \log P_{mkn}) \\ &+ \sum_{m,k} \langle \log B_{mk} \rangle \sum_n \langle Z_{mkn} \rangle + \sum_{kn} \langle \log W_{kn} \rangle \sum_m \langle Z_{mkn} \rangle \\ &+ \sum_{m,k} (\log \lambda^b_{mk} - \lambda^b_{mk} \langle B_{mk} \rangle) + \sum_{k,n} (\log \lambda^w_{kn} - \lambda^w_{kn} \langle W_{kn} \rangle) \\ &+ \sum_{m,k} (-(\alpha^b_{mk} - 1)\Psi(\alpha^b_{mk}) + \log \beta^b_{mk} + \alpha^b_{mk} + \log \Gamma(\alpha^b_{mk})) \\ &+ \sum_{k,n} (-(\alpha^w_{kn} - 1)\Psi(\alpha^w_{kn}) + \log \beta^w_{kn} + \alpha^w_{kn} + \log \Gamma(\alpha^w_{kn})) \end{aligned} \quad (13)$$

where $\Psi(x) \triangleq \frac{d}{dx} \log \Gamma(x)$ is a digamma function. In VB-M step, the optimal regularization parameters $\Theta = \{\lambda^b_{mk}, \lambda^w_{kn}\}$ are derived by maximizing Eq. (13). A closed-form solution to NMF hyperparameters Θ is found by

$$\lambda^b_{mk} = \frac{1}{\langle B_{mk} \rangle}, \quad \lambda^w_{kn} = \frac{1}{\langle W_{kn} \rangle} \quad (14)$$

where $\langle B_{mk} \rangle = \alpha^b_{mk} \beta^b_{mk}$, $\langle W_{kn} \rangle = \alpha^w_{kn} \beta^w_{kn}$. VB-E step and VB-M step are alternatively and iteratively performed to attain BNMF estimates Θ with convergence.

3.3. Implementation for speech and music separation

This study presents the BNMF approach to single-channel speech and music separation. We would like to separate a mixture of speech and background music into two source signals. The separation problem is tackled in the short-time Fourier transform (STFT) domain. The observed magnitude spectrogram $\mathbf{X}$ is viewed as an addition of the speech spectrogram $\mathbf{X}^s$ and the music spectrogram $\mathbf{X}^m$. NMF is performed in supervised fashion with a training procedure given by training data of speech and music signals. The 1024-point STFT is calculated to obtain the Fourier magnitude spectrograms with frame duration of 40 ms and frame shift of 10ms. The proposed method is applied to factorize the magnitude spectrograms of training data to find speech bases $\mathbf{B}^s$ and music bases $\mathbf{B}^m$ via $\mathbf{X}^s \approx \mathbf{B}^s \mathbf{W}^s$ and $\mathbf{X}^m \approx \mathbf{B}^m \mathbf{W}^m$, respectively. Importantly, we determine the number of speech bases $\mathbf{B}^s$ and music bases $\mathbf{B}^m$ based on the same training criterion in Eq. (9). The trained speech and music bases $\mathbf{B}^s$ and $\mathbf{B}^m$ are then applied during test phase where the mixed magnitude spectrogram of a test sample $\mathbf{X}$ is represented by using the trained bases

$$\mathbf{X} \approx [\mathbf{B}^s \; \mathbf{B}^m] \, \mathbf{W}. \quad (15)$$

In the implementation, ML-NMF is applied for initialization of BNMF similar to [15]. We run 20 iterations to find posterior means of basis and weight parameters. The estimated spectrograms of speech and music are found by multiplying basis matrix with the corresponding weight matrix $\hat{\mathbf{W}}$ estimated from test data $\hat{\mathbf{X}}^{\mathbf{s}} = \mathbf{B}^s \hat{\mathbf{W}}^s$ and $\hat{\mathbf{X}}^{\mathbf{m}} = \mathbf{B}^m \hat{\mathbf{W}}^m$. In addition, the soft mask based on Wiener gain is applied to improve the spectrograms for speech source $\tilde{\mathbf{X}}^s$ and music source $\tilde{\mathbf{X}}^m$. Finally, the separated speech and music signals are obtained by the overlap-and-add method using the original phase.

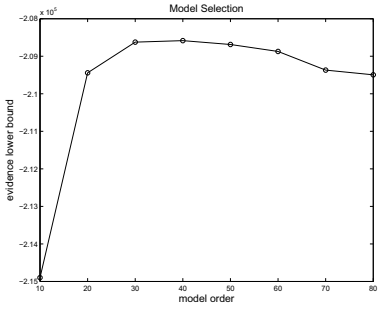

Figure 2: An example of Bayesian selection for model order.

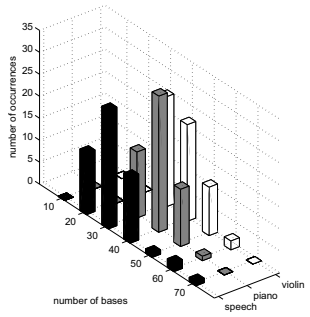

Figure 3: Histogram of model order for source signals.

3.4. Bayesian selection

In addition to deal with regularization issue in NMF, Bayesian approach is beneficial to solve the problem of model selection, or equivalently determine the number of bases K for source separation. The evidence function in Eq. (8) plays an important role for model selection [1]. Using VB-EM algorithm, the variational lower bound in Eq. (13) is applied. The same criterion is optimized to choose the model order K as well as to estimate the NMF hyperparameters Θ. Figure 2 illustrates the variational lower bound versus different number of speech bases. The optimal model order $K = 30$ with the highest lower bound $\mathcal{B}_L$ is chosen in this example. Figure 3 displays the histogram of the selected model order K for three source signals . We can see that the optimal model order is changed sample by sample. Most of model orders are selected in the range between 20 and 40. This adaptive basis selection is helpful for BNMF-based speech and music source separation.

4. Experiments

We evaluated the proposed method on single-channel supervised speech separation from background music. The speech samples were extracted from the TIMIT corpus [24]. We randomly selected 60 sentences with 3 males and 3 females from TIMIT corpus. Each sentence had a length of 2-3 seconds. A set of high quality music recordings were sampled from Saarland Music Data (SMD) [16]. The SMD dataset consists of two music collections. The first collection contains MIDI-Audio pairs of piano music and the second collection contains various Western classical music repertoire. We selected one piano and one violin pieces composed by Bach from the second collections and down-sampled the signals to 16 KHz sampling frequency. The test signals were generated by corrupting with a randomly selected music segments at 0 dB speech-to-music ration (SMR).

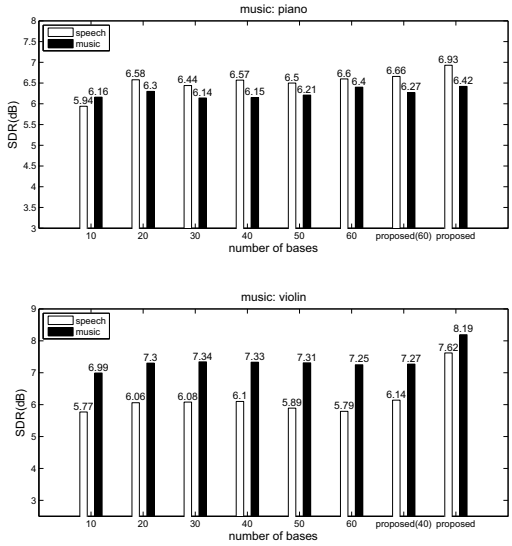

Figure 4: Comparison of SDR using NMF with the fixed number of bases (first 1-6 pairs of bars), BNMF with the best fixed number of bases (7^{th} pair of bars) and BNMF with adaptive number of bases (8^{th} pair of bars).

The desired SMR levels were ensured during speech segments. The 10-fold cross validation for each speaker was performed in our evaluation. For each test speech, the other sentences from the remaining data were concatenated to obtain its corresponding spectrogram so as to learn speech bases $\mathbf{B}^s$. The music bases $\mathbf{B}^m$ were computed from a disjoint music piece of the same track which was used for generating the test sample.

System performance is measured in terms of signal-to-distortion ratio (SDR) [23] of the separated speech and music. SDR is a measure commonly used in evaluation of signal separation. Figure 4 shows the averaged SDR scores where each SDR is averaged over 60 test sentences. The upper diagram is the result for mixture of speech and piano signals while the below diagram is the result for mixture of speech and violin signals. We fixed the same number of bases on speech and music signals when performing standard NMF. It is obvious that the proposed BNMF with adaptive number of bases outperforms various NMFs with the fixed number of bases.

5. Conclusions

In this paper, we have presented a new Bayesian factorization and selection for NMF-based single-channel source separation. Instead of empirically determining the number of bases and the parameters for sparsity control, the proposed method tackled these two issues automatically in training stage. The objective for solving Bayesian NMF was illustrated. A kind of sparse learning was performed for NMF approximation. Compared with various manual setups in NMF and Bayesian NMF, the proposed method obtained higher signal-to-distortion ratio in the separated speech and music signals. In future works, we are extending the proposed method for unsupervised and semi-supervised source separation. The optimization procedure considering complete variable dependency shall be explored. The sparsity control in Bayesian NMF [2, 12] shall be investigated.

6. References

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

[2] M. Carlin, N. Malyska and T. Quatieri, "Speech enhancement using sparse convolutive non-negative matrix factorization with basis adaptation", *Proceedings of Annual Conference of International Speech Communication Association*, pp. 583-586, 2012.

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

[4] J.-T. Chien and B.-C. Chen, "A new independent component analysis for speech recognition and separation", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1245-1254, 2006.

[5] J.-T. Chien and H.-L. Hsieh, "Convex divergence ICA for blind source separation", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 290-301, 2012.

[6] J.-T. Chien and H.-L. Hsieh, "Nonstationary source separation using sequential and variational Bayesian learning, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 681-694, 2013.

[7] J.-T. Chien and H.-L. Hsieh, "Bayesian group sparse learning for music source separation", *EURASIP Journal on Audio, Speech, and Music Processing*, 2013:18, 2013. (doi: 10.1186/1687-4722-2013-18)

[8] 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)*, vol. 39, no. 1, pp. 1-38, 1977.

[9] J. R. Hershey, S. J. Rennie, P. A. Olsen and T. T. Kristjansson, "Super-human multi-talker speech recognition: a graphical model approach", *Computer Speech and Language*, vol. 24, no. 1, pp. 45-66, 2010.

[10] T. Hofmann, "Probabilistic latent semantic indexing", *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50-57, 1999.

[11] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints", *Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, 2004.

[12] C. Joder, F. Weninger, D. Virette and Bjrn Schuller, "A comparative study on sparsity penalties for NMF-based speech separation: beyond LP-norms", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 858-862, 2013.

[13] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization", *Advances in Neural Information Processing Systems*, vol. 13, pp. 556-562, 2000.

[15] N. Mohammadiha, J. Taghia and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4561-4564, 2012.

[16] M. Muller, V. Konz, W. Bogler and V. Arifi-Muller, "Saarland Music Data (SMD)", *Proceeding of Annual Conference of International Society for Music Information Retrieval*, 2011.

[17] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550-563, 2010.

[18] M. N. Schmidt and M. Morup, "Non-negative matrix factor 2-D deconvolution for blind single channel source separation", *Proceeding of International Conference on Independent Component Analysis and Blind Signal Separation*, pp. 700-707, 2006.

[19] M. N. Schmidt, O. Winther and L. K. Hansen, "Bayesian non-negative matrix factorization", *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 540-547, 2009.

[20] M. Shashanka, B. Raj and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations", *Computational Intelligence and Neuroscience*, Article ID 947438, 2008.

[21] P. Smaragdis, B. Raj and M. Shashanka, "A probabilistic latent variable model for acoustic modeling", *Advances in Models for Acoustic Processing Workshop (NIPS)*, 2006.

[22] P. Smaragdis, "Convolutive speech bases and their application to speech separation", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1-12, 2007.

[23] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation", *IEEE Transaction on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462-1469, 2006.

[24] V. Zue, S. Seneff and J. Glass, "Speech database development at MIT: TIMIT and beyond", *Speech Communication*, vol. 9, no. 4, pp. 351-356, 1990.