Log-Normal Matrix Factorization with Application to Speech-Music Separation

Takuya Yoshioka, Daichi Sakaue

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
yoshioka.takuya@lab.ntt.co.jp, dsakaue@kuis.kyoto-u.ac.jp

Abstract

This paper proposes a novel spectrogram factorization method, called log-normal matrix factorization (LogNMF). Conventional nonnegative matrix factorization (NMF) methods cannot efficiently capture random properties of actual spectra because these methods assume that speech and noise spectrograms can be precisely represented by combining a small number of temporally invariant spectral patterns, called basis vectors. This limitation results in unsatisfactory performance when NMF is used for speech enhancement. The proposed method overcomes this limitation by allowing each basis vector to change randomly at each time frame with a log-normal distribution. The use of the log-normal distribution is also desirable in that the degree of divergence between an observed spectrogram and a spectrogram model is measured based on squared errors of log power spectra, which are subjectively meaningful. Experimental results show that LogNMF is able to separate speech signals from background music signals more precisely than NMF.

Index Terms: matrix factorization, log-normal distribution, speech enhancement

1. Introduction

The acquisition of target speech signals with high accuracy using one microphone is paramount for various speech applications. Since a significant amount of background noise may be present in practically relevant environments, techniques for separating the speech from the background have long been studied. Conventional single-microphone speech enhancement methods can do this job effectively when the noise statistics are stationary or change slowly with time [1]. However, they are not able to cope with significantly non-stationary noise, typically caused by background music.

Spectrogram factorization approaches based on nonnegative matrix factorization (NMF) or similar techniques have been investigated extensively in recent years to overcome the above limitation of the conventional speech enhancement methods [2, 3]. At the heart of the NMF approaches lies the assumption that speech and noise spectrograms can be precisely represented by combining a small number of temporally invariant spectral patterns, called basis vectors. However, this assumption precludes these approaches from capturing random properties of actual spectra. To address this problem, we present an alternative method, called log-normal matrix factorization (LogNMF), which allows each basis vector to change randomly at each time frame with a log-normal distribution. The use of the log-normal distribution is also desirable in that the degree of divergence between an observed spectrogram and a spectrogram model is measured based on squared errors of log power spectra. The squared error measure on a logarithmic scale is known to be subjectively more meaningful than typical measures employed for NMF, including the Itakura-Saito divergence and the I divergence. We show that LogNMF is able to separate speech signals from background music signals with a higher separation precision than NMF.

The second author is currently with Kyoto University.

2. Nonnegative Matrix Factorization

NMF is a mathematical transformation for approximately factorizing a given nonnegative matrix $\mathsf{Y} \in \mathbb{R}^{\geq 0, F \times T}$ into two nonnegative matrices $\mathsf{X}^{\geq 0, F \times I}$ and $\mathsf{A}^{\geq 0, I \times T}$ as follows

$$\mathsf{Y} \approx \mathsf{XA}, \tag{1}$$

where $I \leq \mathrm{rank}(\mathsf{Y})$. To achieve this, conventional NMF algorithms minimize a predetermined divergence measure between Y and XA, such as the Itakura-Saito divergence [4] and the I divergence [5]. When $\boldsymbol{Y}_t$, $\boldsymbol{X}_i$, and $A_{i,t}$ denote the tth column of Y, the ith column of X, and the (i,t)th element of A, respectively, (1) can be equivalently rewritten as

$$\boldsymbol{Y}_t \approx \sum_{i=1}^{I} A_{i,t} \boldsymbol{X}_i, \ 1 \leq \forall t \leq T. \tag{2}$$

This means that each $\boldsymbol{Y}_t$ approximately resides in the space spanned by $\{\boldsymbol{X}_i\}_{1 \leq i \leq I}$. For this reason, each $\boldsymbol{X}_i$ is called a basis vector. Each $A_{i,t}$ is called an activation coefficient. A tutorial on NMF can be found in [6].

When applying NMF to speech enhancement tasks, Y is assumed to contain short-time power spectra of an observed noisy speech signal in its columns [7]. As soon as we decrease the number I of basis vectors from $\mathrm{rank}(\mathsf{Y})$, each basis vector $\boldsymbol{X}_i$ begins to capture a spectral pattern that repeatedly manifests itself in Y. The activation coefficient $A_{i,t}$ represents the degree to which the ith basis vector, $\boldsymbol{X}_i$, contributes to the tth power spectrum, $\boldsymbol{Y}_t$. In other words, NMF automatically finds a small set of spectral patterns that are seen frequently in Y and evaluates the contribution of each basis vector. In the speech enhancement applications, we assume that each of the basis vectors used to represent the corrupted spectrogram Y shows a typical spectral pattern of either the target speech or background noise. Based on this assumption, the target speech spectrogram can be reconstructed by collecting only the spectrogram components corresponding to the 'speech' basis vectors.

Since the model given by (1) is excessively simple and cannot satisfactorily capture various aspects of actual noisy speech signals, many extensions have been proposed, including

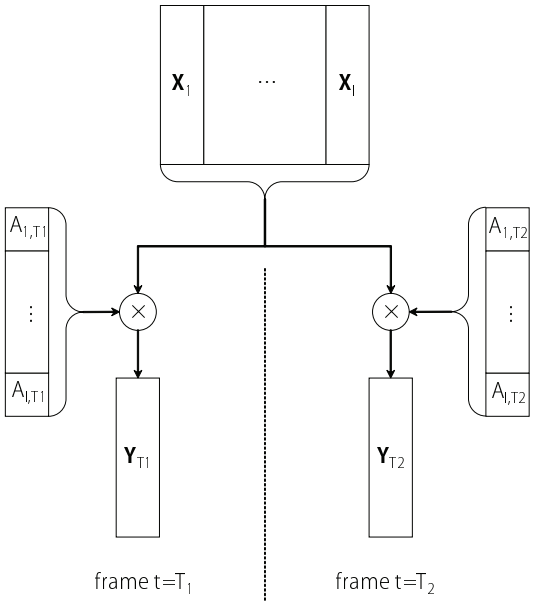

Figure 1: Spectrogram generation model of NMF.

convolutive NMF [8], the non-negative hidden Markov model (HMM) [2], and the infinite factorial infinite HMM [9], harmonic NMF [10, 11]. The methods presented in [2, 8, 9] are aimed at capturing the temporal dynamics of speech spectra while those described in [10, 11] attempt to account for harmonic structures characteristic of voiced sounds and many musical instruments.

In this paper, we address the following two issues with NMF, which have not been solved in the recent studies mentioned above.

1. *Temporal invariance assumption on basis vectors* — The basic NMF algorithm assumes that basis vectors do not change with time. This constraint hinders the accurate modeling of speech spectrograms because the power spectra of speech signals exhibit random behavior, which cannot be described by the temporally invariant basis vectors. This constraint would be harmful especially when NMF is employed for speech enhancement, where basis vectors used to represent target speech signals are trained in advance using a separate set of training data and fixed during the testing phase. To overcome this problem, each basis vector must be allowed to fluctuate randomly at each time frame.

2. *Subjectively improper divergence measure* — Divergence measures typically employed for NMF are not subjectively very meaningful although they are easy to manipulate. It is known that measures based on squared errors of log power spectra are more suitable for speech processing. Indeed, squared error measures on a logarithmic scale have been successfully employed for speech enhancement [12] and multitalker speech recognition [13].

Although the first problem may be mitigated by the method proposed in [9], which represents each basis vector with a small set of spectral atoms and randomly selects one atom at each time frame using an HMM, it would be desirable to jointly solve the above two problems. With this motivation, LogNMF is proposed as detailed in the following sections.

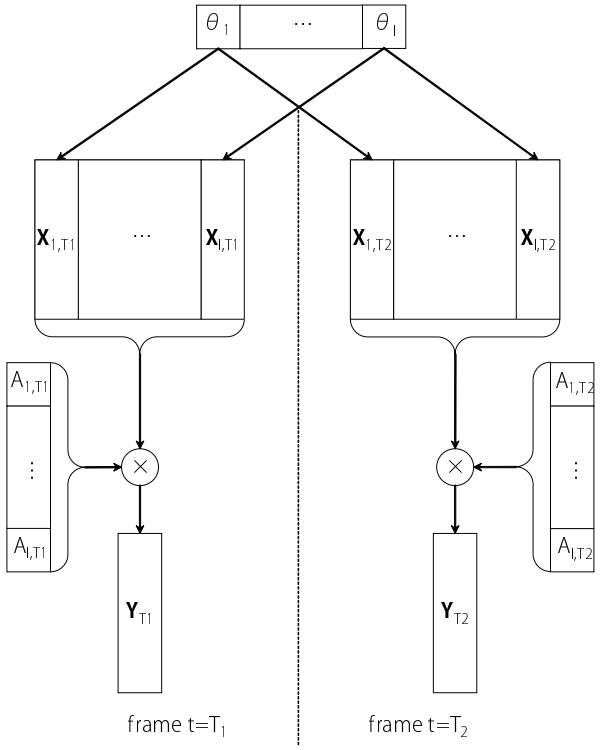

Figure 2: Spectrogram generation model of LogNMF.

3. Log-Normal Matrix Factorization

Let us begin by sketching the concept of the proposed LogNMF method. LogNMF makes the following assumptions to address the above two issues.

1. Instead of I time-invariant basis vectors, we have I log-normal distributions of basis vectors that are characterized by time-invariant parameters.
2. At each time frame, each of the I basis vectors is randomly sampled from the corresponding log-normal distribution.

These assumptions mean that the spectrogram model given by (2) is replaced by the following equations

$$\boldsymbol{Y}_t = \sum_{i=1}^{I} A_{i,t} \boldsymbol{X}_{i,t} \tag{3}$$

$$\boldsymbol{X}_{i,t} \sim \mathrm{LogNormal}(\theta_i), \tag{4}$$

where θ_i is a set of the parameters of the ith log-normal distribution. Note that, unlike in (2), both sides of (3) are assumed to be strictly equal because the approximation error in (2) can be accounted for by the stochastic property of the basis vectors. The goal of LogNMF is to obtain a set of appropriate values of $\{\theta_i\}_{1\le i\le I}$ and $\{A_{i,t}\}_{1\le i\le I, 1\le t\le T}$ that fits well with the observed spectrogram.

We can easily see that LogNMF handles the two issues with NMF described in the previous section. Equation (4) means that the logarithm of each basis vector is modeled by a normal distribution. The mean vector of the normal distribution can

be regarded as a 'canonical' logarithmic basis vector that does not change with time. On the other hand, the covariance matrix controls the degree to which the corresponding basis vector fluctuates at each time frame. By modeling the basis vector fluctuation in this way, LogNMF copes with the first issue. Furthermore, since the log-normal distribution is based on squared errors on a logarithmic scale, the second issue is also dealt with appropriately. Figures 1 and 2 contrast the spectrogram generation models of NMF and LogNMF. We can see that LogNMF models the basis vector fluctuation by sharing basis vector distributions among different time frames instead of using a common basis vector set.

Before proceeding further, let us define some additional notations. To avoid the difficulty resulting from directly manipulating the log-normal distribution, we represent all relevant variables in the log-spectral domain. Thus, we employ the following notations

$$\boldsymbol{y}_t = \log \boldsymbol{Y}_t \tag{5}$$

$$a_{i,t} = \log A_{i,t} \tag{6}$$

$$\boldsymbol{x}_{i,t} = \log \boldsymbol{X}_{i,t}. \tag{7}$$

Using these notations, (3) and (4) are rewritten as follows by using a normal distribution:

$$\boldsymbol{y}_t = \log\left(\sum_{i=1}^{I} \exp(a_{i,t} + \boldsymbol{x}_{i,t})\right) \tag{8}$$

$$\boldsymbol{x}_{i,t} \sim \mathrm{Normal}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{9}$$

where the covariance matrix $\boldsymbol{\Sigma}_i$ is assumed to be diagonal so that we have $\boldsymbol{\Sigma}_i = \mathrm{diag}(\tau_{1,i}^{-1}, \cdots, \tau_{F,i}^{-1})$. Hereafter, we call $\boldsymbol{x}_{i,t}$ and $a_{i,t}$ a logarithmic basis vector and shift coefficient, respectively, in light of the fact that the elements of the logarithmic basis vector $\boldsymbol{x}_{i,t}$ are shifted by $a_{i,t}$ as shown in (8). The goal of LogNMF is also redefined as estimating a set of appropriate values of $\{\boldsymbol{\mu}_i\}_{1\le i\le I}$, $\{\boldsymbol{\tau}_i\}_{1\le i\le I}$ and $\{a_{i,t}\}_{1\le i\le I, 1\le t\le T}$, given an observed log spectrogram $\{\boldsymbol{y}_t\}_{1\le t\le T}$, where $\boldsymbol{\tau}_i$ is the vector consisting of the precisions $\tau_{1,i}, \cdots, \tau_{F,i}$.

3.1. Likelihood function

To perform LogNMF, i.e., to obtain appropriate estimates of $\{\boldsymbol{\mu}_i\}_{1\le i\le I}$, $\{\boldsymbol{\tau}_i\}_{1\le i\le I}$ and $\{a_{i,t}\}_{1\le i\le I, 1\le t\le T}$, we need to define the marginal probability density function (pdf) of the observed log spectra $\{\boldsymbol{y}_t\}_{1\le t\le T}$, which gives the likelihood function of these parameters. In the following, we represent the fth frequency components of $\boldsymbol{y}_t$, $\boldsymbol{x}_{i,t}$, and $\boldsymbol{\mu}_i$ with $y_{f,t}$, and $x_{f,i,t}$, and $\mu_{f,i}$ respectively. In addition, we define $u_{f,i,t}$ as follows

$$u_{f,i,t} = a_{i,t} + x_{f,i,t}, \tag{10}$$

which allows us to rewrite (8) as follows:

$$y_{f,t} = \log\left(\sum_{i=1}^{I} \exp(u_{f,i,t})\right). \tag{11}$$

First, due to (9), the pdf of each frequency component of a logarithmic basis vector is written as

$$\begin{aligned} p(x_{f,i,t}|\mu_{f,i}, \tau_{f,i}) &= f_{\mathrm{Normal}}(x_{f,i,t}; \mu_{f,i}, \tau_{f,i}^{-1}) \\ &= \sqrt{\frac{\tau_{f,i}}{2\pi}} \exp\left(-\frac{\tau_{f,i}}{2}(x_{f,i,t} - \mu_{f,i})^2\right). \end{aligned} \tag{12}$$

Because the distribution of each $u_{f,i,t}$ is obtained by shifting that of $x_{f,i,t}$ by $a_{i,t}$, we obtain the following pdf of $u_{f,i,t}$

$$\begin{aligned} p(u_{f,i,t}|a_{i,t}, \mu_{f,i}, \tau_{f,i}) &= f_{\mathrm{Normal}}(u_{f,i,t}; a_{i,t} + \mu_{f,i}, \tau_{f,i}^{-1}) \\ &= \sqrt{\frac{\tau_{f,i}}{2\pi}} \exp\left(-\frac{\tau_{f,i}}{2}(x_{f,i,t} - a_{i,t} - \mu_{f,i})^2\right). \end{aligned} \tag{13}$$

Next, we can see from (11) that the conditional pdf of each observed spectral component $y_{f,t}$, given the latent components $\{u_{f,i,t}\}_{1\le i\le I}$ at the corresponding time-frequency slot, is written as

$$p(y_{f,t}|\{u_{f,i,t}\}_{1\le i\le T}) = \delta\left(y_{f,t} - \log\left(\sum_{i=1}^{I} \exp(u_{f,i,t})\right)\right), \tag{14}$$

where δ is the Dirac delta function. Unfortunately, the log-sum-exp form in (14) is difficult to manipulate. To avoid this difficulty, we employ the max approximation method [13], which approximates (14) by the following equation

$$p(y_{f,t}|\{u_{f,i,t}\}_{1\le i\le I}) = \delta\left(y_{f,t} - \max_{1\le i\le I} u_{f,i,t}\right). \tag{15}$$

The max approximation method has been successfully used for multitalker speech recognition [13] and single-channel speech separation [14]. A discussion of the accuracy of the max approximation can be found in [13].

Finally, multiplying (13) and (15) and integrating out the latent components $\{u_{f,i,t}\}_{1\le i\le I}$, we can obtain the marginal pdf of each observed spectral component as follows (see [13] for the marginalization of $u_{f,i,t}$)

$$\begin{aligned} p(y_{f,t}|\Theta) = \sum_{i=1}^{I}\Bigg(& f_{\mathrm{Normal}}(y_{f,t}; a_{i,t} + \mu_{f,i}, \tau_{f,i}^{-1}) \\ & \times \prod_{j\ne i} F_{\mathrm{Normal}}(y_{f,t}; a_{j,t} + \mu_{f,j}, \tau_{f,j}^{-1})\Bigg), \end{aligned} \tag{16}$$

where Θ is the set of all the parameters, i.e.,

$$\Theta = \{\{a_{i,t}\}_{1\le i\le I, 1\le t\le T}, \{\mu_{f,i}\}_{1\le f\le F, 1\le i\le I}, \{\tau_{f,i}\}_{1\le f\le F, 1\le i\le I}\}, \tag{17}$$

and $F_{\mathrm{Normal}}(x; \mu, \sigma)$ denotes the cumulative distribution function (cdf) of the normal distribution with mean μ and variance σ. The normal distribution's cdf is given by

$$F_{\mathrm{Normal}}(x; \mu, \sigma) = \Phi\left(\frac{x - \mu}{\sqrt{\sigma}}\right), \tag{18}$$

where Φ is the cdf of the standard normal distribution, which can be expressed using the error function (erf) as follows

$$\Phi(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right). \tag{19}$$

Assuming that each observed spectral component is independently created when the parameter values are specified, we can obtain the following marginal pdf of the observed log spectrogram

$$p(\{\boldsymbol{y}_t\}_{1\le t\le T}|\Theta) = \prod_{f=1}^{F}\prod_{t=1}^{T} p(y_{f,t}|\Theta). \tag{20}$$

(20) defines the likelihood of Θ.

3.2. Prior distribution

It is possible to place a prior distribution over the set, Θ, of the parameters to be estimated. In this paper, we use a normal prior distribution for each shift coefficient and a normal-gamma distribution to model the joint prior distribution of $\mu_{f,i}$ and $\tau_{f,i}$ so that we have

$$p(a_{i,t}) = f_{\text{Normal}}(a_{i,t}; \theta_0, \sigma_0) \tag{21}$$

$$p(\mu_{f,i}, \tau_{f,i}) = f_{\text{Normal}}\left(\mu_{f,i}; \nu_0, \frac{1}{\xi_0 \tau_{f,i}}\right) f_{\text{Gamma}}\left(\tau_{f,i}; \frac{\rho_0}{2}, \frac{\lambda_0}{2}\right). \tag{22}$$

In these equations, θ_0, σ_0, ξ_0, ρ_0, and λ_0 are predetermined hyperparameters. With these definitions, the prior distribution of Θ is given by

$$p(\Theta) = \left(\prod_{i=1}^{I}\prod_{t=1}^{T} p(a_{i,t})\right)\left(\prod_{f=1}^{F}\prod_{i=1}^{I} p(\mu_{f,i}, \tau_{f,i})\right). \tag{23}$$

Using the likelihood function and the prior distribution, given by (20) and (23), respectively, we obtain estimates of the parameters with the maximum a posteriori (MAP) estimation method. An algorithm for performing the MAP estimation is shown in Appendix A.

3.3. Reconstruction of target speech signals

When using LogNMF for speech enhancement, we first perform LogNMF using an observed log power spectrogram to obtain estimates of the model parameters. Then, we estimate the logarithmic power spectral components of a target speech signal with the minimum mean square error (MMSE) estimation method as described below.

We assume that the basis vectors obtained from the observed spectrogram can be classified into speech and noise categories. When $\mathbb{I}^{\text{S}}$ and $s_{n,l}$ denote a set of the basis vector indices corresponding to the speech category and a logarithmic power spectral component of the target speech signal, respectively, $s_{n,l}$ is given by

$$s_{n,l} = \log\left(\sum_{i \in \mathbb{I}^{\text{S}}} \exp(u_{f,i,t})\right). \tag{24}$$

With the MMSE estimation method, $s_{n,l}$ is estimated by

$$\hat{s}_{n,l} = \int s_{n,l} p(s_{n,l}|y_{n,l}, \Theta) ds_{n,l}, \tag{25}$$

where $p(s_{n,l}|y_{n,l}, \Theta)$ is the posterior pdf of $s_{n,l}$. To obtain $\hat{s}_{n,l}$ using (25), we use the vector Taylor series approximation approach [15], which approximates the log-sum-exp form in (24) with a linear function.

4. Experimental Results

We conducted speech-music separation experiments to evaluate the efficacy of our proposed LogNMF method. Specifically, LogNMF was employed to enhance a speech signal corrupted by a music signal when a single-channel mixture of the speech and music was observed. As in [2, 3], we adopted a speaker-dependent semi-supervised approach, where we trained basis vector parameters corresponding to the target speech signal in advance using a separate set of training data. In the testing phase, the speech basis vector parameters were fixed and only the remaining unspecified parameters were estimated using the corrupted signal. The estimated parameter values were used to reconstruct the target speech signal. The experimental procedure and results are described below.

We selected four speakers (two male and two female) from the TIMIT database and took four utterances for each speaker. We mixed each utterance with four different music excerpts (two jazz and two classical) and one white noise. Therefore, we performed a total of 80 experiments for a given signal-to-noise ratio (SNR).

The basis vector parameters, i.e., the means and precisions of logarithmic basis vectors, for each speaker were trained as follows. First, we created the training data for a given speaker by concatenating ten utterances of the same speaker that were different from the utterances used for evaluation. Using the spectrogram extracted from these training data, we performed NMF using the Itakura-Saito divergence [4] to obtain I basis vectors for the given speaker, where I was set at eight. Then, we computed the logarithms of these basis vectors and those of the activation coefficients, which were also obtained by the NMF, in order to initialize the logarithmic basis vector means and shift coefficients, respectively. We also initialized the logarithmic basis vector precisions using the reciprocals of the frequency-dependent variances of the training spectrogram. Starting with these initial values, we iteratively updated the parameters using the optimization algorithm shown in Appendix A. The hyperparameters were adjusted to make the prior non-informative, i.e., θ_0 and ν_0 were set at 0, σ_0 was set at a sufficiently large value, and ξ_0, ρ_0, and λ_0 were set at sufficiently small values. The logarithmic basis vector means and precisions obtained after convergence were employed for the target speech in the testing phase.

The result of each experiment was evaluated in terms of separation index (SI) and fidelity index (FI). The SI is a power ratio between speech- and music-derived components of an estimated speech signal. When the power spectral elements of the estimated speech signal and the corresponding clean and corrupted signals are denoted by $\hat{S}_{f,t}$, $S_{f,t}$, and $Y_{f,t}$, respectively, the speech-derived component, $\hat{S}^{\text{S}}_{f,t}$, of $\hat{S}_{f,t}$ was computed by

$$\hat{S}^{\text{S}}_{f,t} = \frac{\hat{S}_{f,t}}{Y_{f,t}} S_{f,t} \tag{26}$$

while the music-derived component was calculated as $\hat{S}^{\text{M}}_{f,t} = \hat{S}_{f,t} - \hat{S}^{\text{S}}_{f,t}$. The SI measures the degree to which the background music was reduced, ignoring the spectral change imposed on the target speech signal. On the other hand, the FI measures the degree to which the spectral shape of the target speech signal was retained. It is defined as a power ratio between the original speech signal and the signal representing the spectral change imposed on the target speech signal, where the spectral change was computed by

$$\hat{S}^{\text{C}}_{f,t} = \left(\frac{Y_{f,t} - \hat{S}_{f,t}}{Y_{f,t}}\right) S_{f,t}. \tag{27}$$

A larger FI indicates a smaller amount of spectral change.

Figures 3 and 4 show the average separation and fidelity indices, respectively, obtained by using all the experimental results. We can see that LogNMF significantly outperformed NMF in the separation index while LogNMF decreased the fidelity index. These results show that LogNMF separated the speech signals from the background music more precisely than the conventional NMF method by allowing a larger amount of spectral change. The superiority of LogNMF was pronounced especially when the input SNR was 0 dB, where a significant

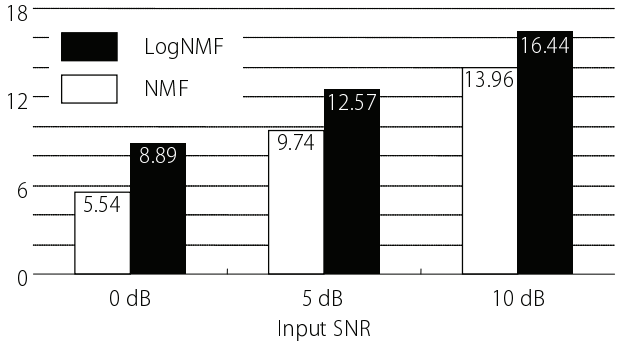

Figure 3: Separation index in dB scale for three different input SNRs.

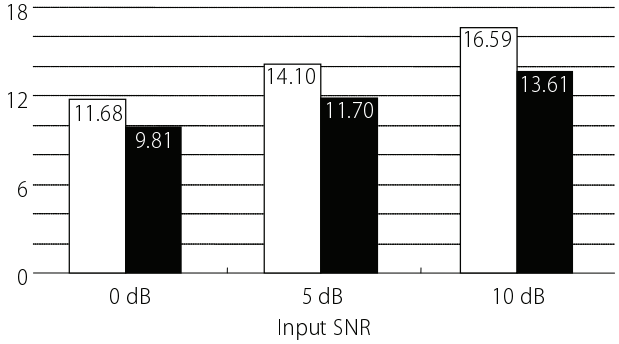

Figure 4: Fidelity index in dB scale for three different input SNRs.

improvement in the SI was achieved at the expense of a small decrease in the FI. By employing informal listening tests, we found that LogNMF reduced a much larger part of the background music than NMF and that the speech signal estimates obtained by LogNMF tended to sound like high-pass filtered versions of the original speech signals.

5. Conclusion

We have presented a novel spectrogram factorization method, called LogNMF. The proposed LogNMF method differs from the conventional NMF methods in that the former allows each basis vector to change randomly at each time frame with a log-normal distribution. This property enables the proposed method to capture the randomness seen in actual spectra and to use squared errors of log power spectra when measuring the divergence between an observed spectrogram and a spectrogram model. The experimental results show that LogNMF provided better separation results than NMF. To show the full potential of LogNMF, modifications of the method are desirable. For example, the separation performance would be further improved by modifying the method to account for other aspects of speech generation and perception processes, including the temporal dynamics of speech spectra and harmonic structures of voiced sounds.

6. References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[2] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 17–20.

[3] A. Ozerov, C. Fevotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. Workshop Appl. Signal Process. Audio, Acoust.*, 2009, pp. 121–124.

[4] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis," *Neural Comp.*, vol. 21, no. 3, pp. 763–830, 2009.

[5] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances Neural, Inf. Process. Syst.*, T. K. Leen, T. G. Dietterch, and V. Tresp, Eds. MIT Press, 2001, pp. 556–562.

[6] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley, 2009.

[7] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. Workshop Appl. Signal Process. Audio, Acoust.*, 2003, pp. 177–180.

[8] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 1–12, 2007.

[9] M. Nakano, J. Le Roux, H. Kameoka, T. Nakamura, N. Ono, and S. Sagayama, "Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model," in *Proc. Workshop Appl. Signal Process. Audio, Acoust.*, 2011, pp. 325–328.

[10] S. A. Raczynski, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. Int. Conf. Music Inform. Retrieval*, 2007.

[11] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 109–112.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[13] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 66–80, 2010.

[14] R. J. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Speech Commun.*, vol. 24, no. 1, pp. 16–29, 2010.

[15] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environmental-independent speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 1996, pp. 733–736.

A. Optimization Algorithm

The MAP estimation of the model parameters is performed by maximizing the objective function obtained by multiplying (20) and (23) and taking the logarithm of the resultant function.

Specifically, the objective function is given by

$$
\begin{aligned}
g(\Theta) = \sum_{t=1}^{T}\sum_{f=1}^{F} \log\Bigg\{ \sum_{i=1}^{I}\Bigg(& f_{\text{Normal}}(y_{f,t}; a_{i,t} + \mu_{f,i}, \tau_{f,i}^{-1}) \\
& \times \prod_{j\neq i} F_{\text{Normal}}(y_{f,t}; a_{j,t} + \mu_{f,j}, \tau_{f,j}^{-1}) \Bigg) \Bigg\} \\
& + \sum_{i=1}^{I}\sum_{t=1}^{T} \log f_{\text{Normal}}(a_{i,t}; \theta_0, \sigma_0) \\
& + \sum_{f=1}^{F}\sum_{i=1}^{I} \Bigg\{ \log f_{\text{Normal}}\Bigg(\mu_{f,i}; \nu_0, \frac{1}{\xi_0 \tau_{f,i}}\Bigg) \\
& \qquad + \log f_{\text{Gamma}}\Bigg(\tau_{f,i}; \frac{\rho_0}{2}, \frac{\lambda_0}{2}\Bigg) \Bigg\}. \quad (28)
\end{aligned}
$$

This objective function is inconvenient in terms of deriving an efficient parameter estimation algorithm because it involves the log-sum form in the first term, resulting from the sum in (16). A widely used technique for avoiding this inconvenience is to view the sum in (16) as a mixture of distributions. Specifically, we introduce an additional random variable $d_{f,t}$ and define the joint pdf of $d_{f,t}$ and $y_{f,t}$ as follows

$$
\begin{aligned}
p(d_{f,t} = i, y_{f,t}|\Theta) = & f_{\text{Normal}}(y_{f,t}; a_{i,t} + \mu_{f,i}, \tau_{f,i}^{-1}) \\
& \times \prod_{j\neq i} F_{\text{Normal}}(y_{f,t}; a_{j,t} + \mu_{f,j}, \tau_{f,j}^{-1}). \quad (29)
\end{aligned}
$$

Then, the marginal pdf of each observed spectral coefficient, given by (16), is obtained by marginalizing $d_{f,t}$, i.e.,

$$
p(y_{f,t}|\Theta) = \sum_{i=1}^{I} p(d_{f,t} = i, y_{f,t}|\Theta). \quad (30)
$$

Equation (30) enables us to represent the marginal pdf using the following factorized form

$$
p(y_{f,t}|\Theta) = \sum_{i=1}^{I} p(y_{f,t}|d_{f,t} = i, \Theta) p(d_{f,t} = i|\Theta). \quad (31)
$$

With (31), we can regard $d_{f,t}$ as a latent variable, which allows us to employ the expectation-maximization (EM) algorithm to perform the MAP estimation.

With the EM algorithm, we start with a set of initial parameter estimates and iteratively update the set by performing the E-step and M-step described below. Hereafter, we call each $d_{f,t}$ a dominant basis indicator because $d_{f,t}$ indicates which of $\{u_{f,i,t}\}_{1\leq i\leq I}$ dominates the corresponding time-frequency slot.

A.1. E-step

In the E-step, we calculate the posterior probability, $\gamma_{f,i,t}$, of each dominant basis indicator being i as follows

$$
\begin{aligned}
\gamma_{f,i,t} = & p(d_{f,t} = i|y_{f,t}, \hat{\Theta}) \\
= & \frac{p(d_{f,t} = i, y_{f,t}|\hat{\Theta})}{\sum_{i=1}^{I} p(d_{f,t} = i, y_{f,t}|\hat{\Theta})}, \quad (32)
\end{aligned}
$$

where $\hat{\Theta}$ denotes a tentative estimate of Θ. This posterior probability measures the dominance of the ith basis vector at the corresponding time-frequency slot.

A.2. M-step

In the M-step, we update the parameter estimates by maximizing the following auxiliary function

$$
Q(\Theta) = \sum_{f=1}^{F}\sum_{i=1}^{I}\sum_{t=1}^{T} \log p(d_{f,t} = i, y_{f,t}|\Theta) + \log p(\Theta), \quad (33)
$$

where the first term is given by (29). Since it is difficult to jointly optimize all the parameters to maximize this auxiliary function, we sequentially compute the updated estimates of the logarithmic basis vector means, those of the logarithmic basis vector precisions, and those of the shift coefficients. The updates of the estimates of the logarithmic basis vector means and those of the shift coefficients can be performed with Newton's method while the update of the estimates of the logarithmic basis vector precisions can be achieved by using the bisection method. The first and second derivatives of the logarithm of the standard normal distribution's cdf $\Phi(x)$, needed by Newton's method, are given by

$$
(\log \Phi(x))' = \sqrt{\frac{2}{\pi}}\, \text{erfcx}\Bigg(-\frac{x}{\sqrt{2}}\Bigg)^{-1} \quad (34)
$$

$$
(\log \Phi(x))'' = -(\log \Phi(x))'(x + (\log \Phi(x))'), \quad (35)
$$

respectively, where erfcx is the scaled complementary error function[1].

[1]The scaled complementary error function can be computed by `erfcx` function when using MATLAB.