Music Signal Separation Using Supervised NMF with All-Pole-Model-Based Discriminative Basis Deformation

Hiroaki Nakajima*, Daichi Kitamura†, Norihiro Takamune*, Shoichi Koyama*, Hiroshi Saruwatari*,
Nobutaka Ono‡, Yu Takahashi§ and Kazunobu Kondo§
*The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–8656, Japan
†SOKENDAI (The Graduate University for Advanced Studies), 2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan
‡National Institute of Information, 2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan
§Research & Development Division, Yamaha Corporation, 203 Matsunokijima, Iwata, Shizuoka, 438–0192, Japan

***Abstract*—In this paper, we address the music signal separation problem and propose a new supervised nonnegative matrix factorization (SNMF) algorithm employing the deformation of a spectral supervision basis trained in advance. Conventional SNMF has a problem that the separation accuracy is degraded by a mismatch between the trained basis and the spectrogram of the actual target sound in open data. To reduce the mismatch problem, we propose a new method with two features. First, we introduce a deformation with an all-pole model that is optimized to make the trained basis fit the spectrogram of the target signal, even if the true target component is hidden in the observed mixture. Next, to avoid an excess deformation, we limit the degree of freedom in the deformation by performing discriminative training. Our experimental evaluation reveals that the proposed method outperforms conventional SNMFs.**

I. Introduction

In recent years, source separation based on nonnegative matrix factorization (NMF), which is a type of sparse representation algorithm, has been a very active area of signal processing research. NMF for acoustical signals decomposes an input spectrogram into a product of a spectral basis matrix and its activation matrix. In particular, NMF is a promising candidate for source separation in music signal processing with a monaural format [1].

The methods of source separation based on NMF are roughly classified into unsupervised and supervised algorithms. The former method attempts the separation without using any training sequences [2], [3]. The latter method is called *supervised NMF* (SNMF), which includes an a priori training process and requires some sound samples of a target signal [4], [5]. However, such supervised techniques have the critical problem that the separation accuracy is markedly degraded by a mismatch between the trained basis and the spectrogram of the actual target sound in open data.

In this paper, to reduce the mismatch problem, we propose a new SNMF method with two novel features. First, we introduce a basis deformation algorithm with an all-pole model that can be controlled by a statistical postfilter. The all-pole model is optimized to make the trained basis fit the spectrogram of the target signal, even if the true target component is hidden in the observed mixture, owing to the "target signal sampling" ability of the postfilter. Next, to avoid a potential *excess deformation* that can represent nontarget components such as a residual of interference, we propose to limit the degree of freedom in the deformation by performing *discriminative training*. Our experimental evaluation reveals that the proposed method outperforms conventional SNMFs.

II. Conventional Methods

A. SNMF

SNMF [4] consists of two processes, namely, a priori training and observed signal separation, as described below in detail. A priori basis training is carried out via NMF, expressed as

$$\boldsymbol{Y}_{\mathrm{target}} \simeq \boldsymbol{F}\boldsymbol{G}_{\mathrm{t}}, \tag{1}$$

where $\boldsymbol{Y}_{\mathrm{target}}$ is an $\Omega \times T_s$ nonnegative matrix that represents an amplitude spectrogram of the specific signal used for training, $\boldsymbol{F}$ is an $\Omega \times K$ nonnegative matrix that comprises the basis vectors of the target signal as column vectors, and $\boldsymbol{G}_{\mathrm{t}}$ is a $K \times T_s$ nonnegative matrix that corresponds to the activation of each basis vector of $\boldsymbol{F}$. In addition, Ω is the number of frequency bins, K is the number of supervised basis vectors, and T_s is the number of frames of the training signal. Therefore, the basis matrix $\boldsymbol{F}$ is constructed by the supervision of the target instrumental signal.

The following equation represents the decomposition of SNMF with the trained supervision $\boldsymbol{F}$:

$$\boldsymbol{Y}_{\mathrm{mix}} \simeq \boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}, \tag{2}$$

where $\boldsymbol{Y}_{\mathrm{mix}}$ is an $\Omega \times T$ observed spectrogram, $\boldsymbol{G}$ is a $K \times T$ activation matrix that corresponds to $\boldsymbol{F}$, $\boldsymbol{H}$ is an $\Omega \times L$ matrix comprising the residual spectral patterns that cannot be expressed by $\boldsymbol{F}\boldsymbol{G}$, and $\boldsymbol{U}$ is an $L \times T$ activation matrix that corresponds to $\boldsymbol{H}$. Moreover, T is the number of frames of the observed signal and L is the number of basis vectors of $\boldsymbol{H}$. In SNMF, the matrices $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ are optimized under the condition that $\boldsymbol{F}$ is known in advance. Hence, $\boldsymbol{F}\boldsymbol{G}$

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

ideally represents the target instrumental components and $\boldsymbol{HU}$ represents the other components after the decomposition.

The cost function for (2) is defined as

$$\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} \mathcal{D}_{\mathrm{KL}}(\boldsymbol{Y}_{\mathrm{mix}}|\boldsymbol{FG}+\boldsymbol{HU}), \tag{3}$$

where $\mathcal{D}_{\mathrm{KL}}(\cdot|\cdot)$ is a generalized KL divergence defined as

$$\mathcal{D}_{\mathrm{KL}}(\boldsymbol{P}|\boldsymbol{Q}) = \sum_{m,n} p_{m,n}\log\frac{p_{m,n}}{q_{m,n}} + q_{m,n} - p_{m,n}, \tag{4}$$

where $\boldsymbol{P} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{Q} \in \mathbb{R}^{M\times N}$ are matrices whose entries are $p_{m,n}$ and $q_{m,n}$, respectively.

There are methods that involve imposing an orthogonal (or probabilistic) restriction on the relationship between the target signal and the nontarget signal in (3) to improve the separation [5], [6], [7]. For example, in penalized SNMF (PSNMF) [5], the cost function is described as

$$\begin{aligned}\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} &\mathcal{D}_{\mathrm{P\_KL}}(\boldsymbol{Y}_{\mathrm{mix}}|\boldsymbol{FG}+\boldsymbol{HU}) \\ &= \min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} \mathcal{D}_{\mathrm{KL}}(\boldsymbol{Y}_{\mathrm{mix}}|\boldsymbol{FG}+\boldsymbol{HU}) + \mu||\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}||^2_{\mathrm{Fr}},\end{aligned} \tag{5}$$

where μ is a weight parameter.

B. SNMF with additive basis deformation (SNMF-ABD)

Conventional SNMF has the critical problem that a mismatch between the trained bases and the target signal spectrogram reduces the accuracy of separation. To solve this problem, SNMF-ABD has been proposed [8]. In this method, the following equation represents the decomposition in SNMF-ABD with trained supervision $\boldsymbol{F}$:

$$\boldsymbol{Y}_{\mathrm{mix}} \simeq (\boldsymbol{F}+\boldsymbol{D})\boldsymbol{G} + \boldsymbol{HU}, \tag{6}$$

where $\boldsymbol{D}$ is an $\Omega \times M$ additive basis matrix describing the deformation and shares the activation matrix $\boldsymbol{G}$ with $\boldsymbol{F}$. In addition, M is the number of basis vectors of $\boldsymbol{D}$. In this decomposition, to adapt the supervised bases to the target sound that cannot be represented by $\boldsymbol{F}$, another basis matrix $\boldsymbol{D}$ is imposed as a deformation term for $\boldsymbol{F}$. Although $\boldsymbol{D}$ is not exclusively nonnegative, some restrictions are imposed on $\boldsymbol{D}$ so that $\boldsymbol{F}+\boldsymbol{D}$ is nonnegative. The cost function for (6) is given by

$$\begin{aligned}\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} &\mathcal{D}_{\mathrm{KL}}(\boldsymbol{Y}_{\mathrm{mix}}|(\boldsymbol{F}+\boldsymbol{D})\boldsymbol{G}+\boldsymbol{HU}) + \mu_1||\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}||^2_{\mathrm{Fr}} \\ &+ \mu_2||\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}||^2_{\mathrm{Fr}} + \mu_3||\boldsymbol{D}^{\mathrm{T}}\boldsymbol{H}||^2_{\mathrm{Fr}},\end{aligned} \tag{7}$$

where μ_1, μ_2, and μ_3 are weight parameters.

However, this method has three problems. First, it is difficult to adjust the three weight parameters. Second, this model strongly depends on the initial values because of the complexity of the cost function. These two problems are caused by the difficulty of simultaneously optimizing deformation and separation. Finally, this deformation is nonlinear. Therefore, there is a risk that $\boldsymbol{D}$ will excessively deform the basis and make it possible for an unwanted basis to describe the nontarget signal.

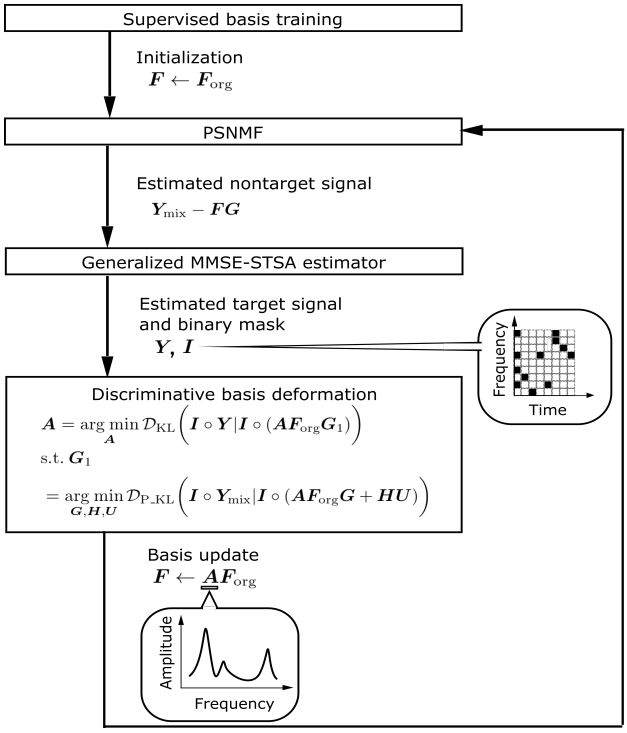

Fig. 1. Block diagram of proposed method.

III. Proposed Method

A. Overview of proposed method

As described above, it is necessary to adapt the supervised basis to the target signal spectrogram to deal with real music sounds. However, it is difficult for SNMF-ABD to perform optimal basis deformation because it is a nonlinear deformation and it optimizes the deformation and separation simultaneously. In this paper, we propose a new SNMF introducing the following schemes. (a) Apart from the source separation process, the deformation process is separately carried out with a linear time-invariant filter, namely an all-pole model, that consists of fewer parameters. (b) The parameters of the all-pole model can be optimized by utilizing "sampled convincing target components" obtained by a generalized minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [9]. (c) We introduce a discriminative deformation to avoid the excess deformation and balance the deformation and separation.

A block diagram of the proposed method is shown in Fig. 1. First, we perform PSNMF with a current supervised basis $\boldsymbol{F}$. Second, using the generalized MMSE-STSA estimator with an estimated nontarget signal $\boldsymbol{Y}_{\mathrm{mix}} - \boldsymbol{FG}$, we obtain an estimated target signal $\boldsymbol{Y}$ and a binary mask $\boldsymbol{I}$ that extracts seldom overlapping components with the nontarget signal from the estimated target signal $\boldsymbol{Y}$. Finally, we deform the original supervised basis $\boldsymbol{F}_{\mathrm{org}}$ discriminatively and update $\boldsymbol{F}$ as a deformed basis. After some iterations of the procedures, we conduct PSNMF using the deformed basis and obtain the improved separation.

B. Convincing component sampler using statistical spectral amplitude estimator

The generalized MMSE-STSA estimator calculates the spectrum gain $\boldsymbol{J}$ that minimizes the average squared error between the target signal and an estimated signal given the a priori probability distribution of the target signal. This process is expressed as follows:

$$\boldsymbol{Y} = \boldsymbol{J} \circ \boldsymbol{Y}_{\mathrm{mix}}, \tag{8}$$

$$J_{\omega,t} = \frac{\sqrt{v_{\omega,t}}}{\tilde{\gamma}_{\omega,t}} \cdot \Big(\frac{\Gamma(\rho + 0.5)}{\Gamma(\rho)} \cdot \frac{\Phi(0.5 - \rho, 1, -v_{\omega,t})}{\Phi(1 - \rho, 1, -v_{\omega,t})} \Big)^{1/\beta}, \tag{9}$$

where $\boldsymbol{Y}$ is the target signal estimated by the generalized MMSE-STSA estimator, $\circ$ is a Hadamard product, $J_{\omega,t}$ is an element of $\boldsymbol{J}$, $\Gamma(\cdot)$ is the gamma function, $\Phi(a, b; k) = F_1(a, b; k)$ is the confluent hypergeometric function, β is the amplitude compression parameter, and ρ is the shape parameter of the chi-squared distribution used as the prior distribution of the target signal. In addition, $v_{\omega,t}$ is defined using an a priori SNR $\tilde{\epsilon}_{\omega,t}$ and a posteriori SNR $\tilde{\gamma}_{\omega,t}$ as

$$v_{\omega,t} = \tilde{\gamma}_{\omega,t} \tilde{\epsilon}_{\omega,t} \Big(1 + \tilde{\epsilon}_{\omega,t} \Big)^{-1}. \tag{10}$$

In the generalized MMSE-STSA estimator, it is necessary to obtain the power spectrum of the nontarget signal to calculate $\tilde{\gamma}_{\omega,t}$. In this study, we use $\boldsymbol{Y}_{\mathrm{mix}} - \boldsymbol{FG}$ for this purpose. In addition, we use the method proposed in [10] to estimate ρ.

C. Basis deformation with all-pole model using generalized MMSE-STSA estimator

In this section, we propose basis deformation with an all-pole model controlled by the generalized MMSE-STSA estimator. Note that several studies on NMF have introduced this all-pole model to describe a spectral envelope in a music signal [11]. However, to the best of our knowledge, our method is the first approach to apply the model to the basis deformation problem.

In our method, we calculate the trained supervision and deform the basis $\boldsymbol{F}_{\mathrm{org}}$ with reference to the estimated target signal $\boldsymbol{Y}$. Since the estimated target signal $\boldsymbol{Y}$ still has low accuracy, it is necessary to extract only a sufficient number of reliable components to deform the basis correctly. Otherwise, the basis deforms excessively and cannot accomplish the separation. Therefore, to avoid this, the thresholding of the spectrum gain $\boldsymbol{J}$ used to extract seldom overlapping components with the nontarget signal is introduced. In addition, although the few components are sampled by the thresholding that yields many blanks in the spectrogram, they are still sufficient to decide the all-pole model because the model has the time-invariant and frequency-interpolation properties. The above-mentioned concepts are described as

$$\boldsymbol{I} \circ \boldsymbol{Y} \simeq \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G}), \tag{11}$$

where $\boldsymbol{I}$ is an $\Omega \times T$ binary mask matrix with entries $i_{\omega,t}$, which was obtained from the spectrum gain matrix $\boldsymbol{J}$ of the generalized MMSE-STSA estimator, the entries of which were subjected to thresholding (e.g., if $J_{\omega,t} > 0.8$, then $i_{\omega,t} = 1$; otherwise $i_{\omega,t} = 0$). In addition, $\boldsymbol{A}$ is a diagonal matrix in which the diagonal elements are described using the all-pole model. The elements of $\boldsymbol{A}$ are described as

$$A_{\omega,\omega} = \frac{1}{|1 - \sum_{k=1}^{p} \alpha_k \exp(-\pi j k \frac{\omega}{\Omega})|}, \tag{12}$$

where p is the order and α_k are the coefficients of the all-pole model. In addition, we define $A_\omega = 1 - \sum_{k=1}^{p} \alpha_k \exp(-\pi j k \frac{\omega}{\Omega})$ to simplify the calculations.

D. Cost function and update rule

The cost function for (11) based on the generalized KL divergence is given by

$$\mathcal{J} = \sum_{\omega,t} i_{\omega,t} \Big\{ -y_{\omega,t} + \frac{\sum_k f_{\omega,k} g_{k,t}}{|A_\omega|} + y_{\omega,t} \log \frac{y_{\omega,t}}{\sum_k f_{\omega,k} g_{k,t} / |A_\omega|} \Big\}, \tag{13}$$

where $y_{\omega,t}$, $f_{\omega,k}$, and $g_{k,t}$ are the nonnegative elements of matrices $\boldsymbol{Y}$, $\boldsymbol{F}_{\mathrm{org}}$, and $\boldsymbol{G}$, respectively. Since it is difficult to analytically derive the optimal $\boldsymbol{A}$ and $\boldsymbol{G}$, we define an auxiliary function that represents the upper bound of $\mathcal{J}$, as described below. First, applying Jensen's inequality to $\log \sum_k f_{\omega,k} g_{k,t}$ and the tangent inequality to $\log |A_\omega| = 1/2 \log |A_\omega|^2$, we have

$$\mathcal{J} \leq \sum_{\omega,t} i_{\omega,t} \Big\{ \frac{\sum_k f_{\omega,k} g_{k,t}}{|A_\omega|} + y_{\omega,t} \big(\frac{1}{2\rho_\omega} |A_\omega|^2 - \sum_k \zeta_{\omega,t,k} \log \frac{f_{\omega,k} g_{k,t}}{\zeta_{\omega,t,k}} \big) + C_{\omega,t} \Big\}, \tag{14}$$

where $C_{\omega,t}$ are unnecessary constants when calculating the update rules of the activation matrix $\boldsymbol{G}$ and the all-pole-model weight matrix $\boldsymbol{A}$, and ρ_ω and $\zeta_{\omega,t,k}$ are auxiliary variables. The equality in (14) holds if and only if the auxiliary variables are set to $\rho_\omega = |A_\omega|^2$ and $\zeta_{\omega,t,k} = f_{\omega,k} g_{k,t} / \sum_k f_{\omega,k} g_{k,t}$. Second, to make the auxiliary function a quadratic form of $|A_\omega|$, we conduct a Taylor expansion around τ_ω,

$$\mathcal{J} \leq \sum_{\omega,t} i_{\omega,t} \Big\{ \sum_k f_{\omega,k} g_{k,t} \big(\frac{1}{\tau_\omega^3} |A_\omega|^2 - 3 \frac{1}{\tau_\omega^2} |A_\omega| + \frac{3}{\tau_\omega} \big) + y_{\omega,t} \big(\frac{1}{2\rho_\omega} |A_\omega|^2 - \sum_k \zeta_{\omega,t,k} \log \frac{f_{\omega,k} g_{k,t}}{\zeta_{\omega,t,k}} \big) + C_{\omega,t} \Big\}. \tag{15}$$

The equality of (15) holds if and only if $\tau_\omega = |A_\omega|$. This approximation does not meet the condition of an auxiliary function, but if τ_ω is updated as $|A_\omega|$, this approximation is equivalent to Newton's method. Finally, using the inequality $\mathcal{R}e[\theta_\omega^* A_\omega] \leq |A_\omega|$, we can define the upper bound function $\mathcal{J}^+$ for $\mathcal{J}$ as

$$\mathcal{J} \leq \sum_{\omega,t} i_{\omega,t} \Big\{ \sum_k f_{\omega,k} g_{k,t} \big(\frac{1}{\tau_\omega^3} |A_\omega|^2 - 3 \frac{1}{\tau_\omega^2} \mathcal{R}e[\theta_\omega^* A_\omega] + \frac{3}{\tau_\omega} \big) + y_{\omega,t} \big(\frac{1}{2\rho_\omega} |A_\omega|^2 - \sum_k \zeta_{\omega,t,k} \log \frac{f_{\omega,k} g_{k,t}}{\zeta_{\omega,t,k}} \big) + C_{\omega,t} \Big\}, \tag{16}$$

where $\mathcal{R}e[\cdot]$ is a real part of $\cdot$ and $|\theta_\omega| = 1$. The equality of (16) holds if and only if $\theta_\omega = A_\omega/|A_\omega|$.

1) Multiplicative update rule for activation matrix $\boldsymbol{G}$: The update rule for $\mathcal{J}^+$ with respect to the activation matrix $\boldsymbol{G}$ is determined by setting the gradient to zero. From $\partial\mathcal{J}^+/\partial g_{k,t} = 0$, we obtain

$$\sum_\omega i_{\omega,t}\Big\{f_{\omega,k}(\frac{1}{\tau_\omega^3}|A_\omega|^2 - 3\frac{1}{\tau_\omega^2}\mathcal{R}e[\theta_\omega^* A_\omega] + \frac{3}{\tau_\omega}) + y_{\omega,t}(-\zeta_{\omega,t,k}g_{k,t}^{-1})\Big\} = 0. \tag{17}$$

By substituting the auxiliary variables into (17) and simplifying it, we obtain the multiplicative update rule of $g_{k,t}$ as

$$g_{k,t} \leftarrow g_{k,t}\frac{\sum_\omega i_{\omega,t}y_{\omega,t}f_{\omega,k}/(\sum_\kappa f_{\omega,\kappa}g_{\kappa,t})}{\sum_\omega i_{\omega,t}f_{\omega,k}/|A_\omega|}. \tag{18}$$

2) Multiplicative update rule for all-pole-model weight matrix $\boldsymbol{A}$: First, by differentiating $\mathcal{J}^+$ partially with respect to α_q and setting it to zero, we obtain

$$\begin{aligned}
&\sum_{k=1}^{p}\alpha_k\sum_{\omega,t}\Bigg[i_{\omega,t}(\sum_k f_{\omega,k}g_{k,t}\frac{1}{\tau_\omega^3} + y_{\omega,t}\frac{1}{2\rho_\omega})\\
&\Big(\exp\big(-\pi j\frac{\omega}{\Omega}(k-q)\big) + \exp\big(\pi j\frac{\omega}{\Omega}(k-q)\big)\Big)\Bigg]\\
&-\sum_{\omega,t}i_{\omega,t}\Bigg[(\sum_k f_{\omega,k}g_{k,t}\frac{1}{\tau_\omega^3} + y_{\omega,t}\frac{1}{2\rho_\omega})\big(\exp(-\pi j\frac{\omega}{\Omega}q)\\
&+\exp(\pi j\frac{\omega}{\Omega}q)\big) - \frac{3}{\tau_\omega^2}\sum_k f_{\omega,k}g_{k,t}\mathcal{R}e[\theta_\omega^*\exp(-\pi j\frac{\omega}{\Omega}q))]\Bigg]\\
&= 0,
\end{aligned} \tag{19}$$

where $1 \le q \le p$. Second, we define $\boldsymbol{R}$ and $\boldsymbol{r}$ as

$$\begin{aligned}
R_{k,q} = &\sum_{\omega,t}\Bigg[i_{\omega,t}(\sum_k f_{\omega,k}g_{k,t}\frac{1}{\tau_\omega^3} + y_{\omega,t}\frac{1}{2\rho_\omega})\\
&\Big(\exp\big(-\pi j\frac{\omega}{\Omega}(k-q)\big) + \exp\big(\pi j\frac{\omega}{\Omega}(k-q)\big)\Big)\Bigg],
\end{aligned} \tag{20}$$

$$\begin{aligned}
r_q = &\sum_{\omega,t}i_{\omega,t}\Bigg[(\sum_k f_{\omega,k}g_{k,t}\frac{1}{\tau_\omega^3} + y_{\omega,t}\frac{1}{2\rho_\omega})\\
&\big(\exp(-\pi j\frac{\omega}{\Omega}q) + \exp(\pi j\frac{\omega}{\Omega}q)\big)\\
&-\frac{3}{\tau_\omega^2}\sum_k f_{\omega,k}g_{k,t}\mathcal{R}e[\theta_\omega^*\exp(-2\pi j\frac{\omega}{\Omega}q))]\Bigg].
\end{aligned} \tag{21}$$

By substituting (20) and (21) into (19), we obtain

$$\boldsymbol{R}\boldsymbol{\alpha} = \boldsymbol{r}, \tag{22}$$

where $\boldsymbol{\alpha}$ is the vector of coefficients in the all-pole model. Since $\boldsymbol{R}$ is a Toeplitz matrix, we can derive $\boldsymbol{\alpha}$ using the Levinson–Durbin algorithm with a computationally efficient form.

E. Discriminative basis deformation

In our method, we use the generalized MMSE-STSA estimator as a sampler to deform the basis $\boldsymbol{F}_{\mathrm{org}}$. However, its signal enhancement ability is not perfect. Since the output of the estimator is still contaminated with residual nontarget signals, there is a risk that the basis will be deformed to be suitable for partially representing the nontarget signals if we optimize only (11). In addition, a basis suitable for representing the target signal is not necessarily suitable for separation. Therefore, we apply the idea of discriminative NMF [12], which learns supervised bases while paying attention to the separability of signals, to our proposed basis deformation. Note that the method in [12] requires full supervision (i.e., all training samples of all the instruments are needed in advance), but our method only requires semi-supervision (only the target sample). First, we formulate this problem as bilevel optimization as

$$\begin{aligned}
&\boldsymbol{A} = \arg\min_{\boldsymbol{A}}\mathcal{D}_{\mathrm{KL}}\Big(\boldsymbol{I}\circ\boldsymbol{Y}\,|\,\boldsymbol{I}\circ(\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G}_1)\Big)\\
&\text{s.t. } \boldsymbol{G}_1\\
&= \arg\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}}\mathcal{D}_{\mathrm{P_KL}}\Big(\boldsymbol{I}\circ\boldsymbol{Y}_{\mathrm{mix}}|\boldsymbol{I}\circ(\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U})\Big).
\end{aligned} \tag{23}$$

This bilevel optimization searches for the optimal basis deformation matrix $\boldsymbol{A}$ under the constraint of minimizing $\mathcal{D}(\boldsymbol{I}\circ\boldsymbol{Y}_{\mathrm{mix}}|\boldsymbol{I}\circ(\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}))$ with respect to $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$. To minimize (23), it is reasonable for $\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G}$ and $\boldsymbol{H}\boldsymbol{U}$ to be independent. This means that the basis deformation is prevented from representing the nontarget signal and is thus able to represent the estimated target signal well. Since it is difficult to solve the bilevel optimization problem, we propose the following iterative algorithm that can derive an approximate solution to the optimization.

Step 1 : Initialization

$$\boldsymbol{A}_{\mathrm{s}} = \arg\min_{\boldsymbol{A},\ \boldsymbol{G}}\mathcal{D}_{\mathrm{KL}}\Big(\boldsymbol{I}\circ\boldsymbol{Y}\,|\,\boldsymbol{I}\circ(\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G})\Big). \tag{24}$$

Step 2 : Modeling of Mixture $\boldsymbol{Y}_{\mathrm{mix}}$

$$\boldsymbol{G}_{\mathrm{s}} = \arg\min_{\boldsymbol{G},\ \boldsymbol{H},\ \boldsymbol{U}}\mathcal{D}_{\mathrm{P\_KL}}\Big(\boldsymbol{I}\circ\boldsymbol{Y}_{\mathrm{mix}}|\boldsymbol{I}\circ(\boldsymbol{A}_{\mathrm{s}}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U})\Big). \tag{25}$$

Step 3 : Modeling of Target $\boldsymbol{Y}$

$$\boldsymbol{A}_{\mathrm{s}} = \arg\min_{\boldsymbol{A}}\mathcal{D}_{\mathrm{KL}}\Big(\boldsymbol{I}\circ\boldsymbol{Y}\,|\,\boldsymbol{I}\circ(\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G}_{\mathrm{s}})\Big). \tag{26}$$

Return to Step 2

This algorithm searches for the basis deformation matrix $\boldsymbol{A}$ that minimizes $\mathcal{D}(\boldsymbol{I}\circ\boldsymbol{Y}_{\mathrm{mix}}|\boldsymbol{I}\circ(\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G}+\boldsymbol{H}\boldsymbol{U}))$ in the vicinity of the minimal $\mathcal{D}(\boldsymbol{I}\circ\boldsymbol{Y}|\boldsymbol{I}\circ(\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G}))$.

IV. Experiment

A. Experimental conditions

To evaluate the proposed algorithm, we compared the conventional methods (SNMF, PSNMF, and SNMF-ABD) and

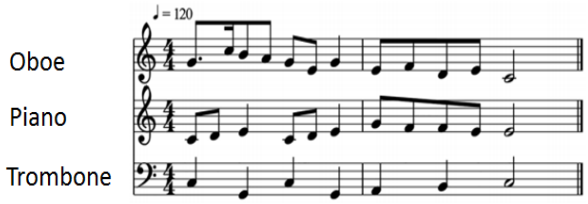

Fig. 2. Scores of each instrument.

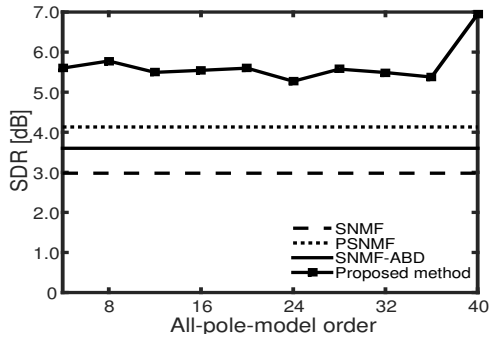

Fig. 3. Example of SDR for separating Pf. from the mixture of Pf. and Ob.

the proposed method by applying them to the separation of two monaural instrumental sources. In this experiment, we used three instruments, namely, a piano (Pf.), oboe (Ob.), and trombone (Tb.). We separately generated three melodies depicted in Fig. 2 using Microsoft GS Wavetable SW Synth software (as artificial MIDI sounds), and two of the three sources were selected and mixed with an input SNR of 0 dB. Training sounds were generated by Garritan Personal Orchestra software (as different MIDI sound from the mixed sound generator). Training sounds contain two octave notes that cover all the notes of the target signal in the mixed signal. The sampling frequency of all the signals was 44.1 kHz. The spectrograms were computed using a 92 ms rectangular window with a 76 ms overlap shift. The number of iterations used in the training and the separation was 1000. Moreover, the number of supervised bases $\boldsymbol{F}$ was 100 and that of bases for matrix $\boldsymbol{H}$ was 30. We used the signal-to-distortion ratio (SDR) as the evaluation score [13]. The SDR indicates the total quality of the separated target sound, evaluating the degree of separation between the target sound and other sounds and the absence of artificial distortion. In proposed method, the all-pole-model order is varied from 1 to 40. In addition, the number of iterations of the whole processing in Fig. 1 is 8.

B. Experimental results

Figure 3 shows a typical example of the SDR for SNMF, PSNMF, SNMF-ABD, and the proposed method for the task of separating Pf. from the mixture of Pf. and Ob. It can be seen that the proposed method outperforms the conventional methods.

Table 1 shows SDRs of SNMF, PSNMF, SNMF-ABD, and the proposed method for extracting the target instrument sound (the first of the two sounds) from each combination of the instruments. All the parameters of each method were manually optimized. From these results, it can be confirmed that the proposed method increases the separation performance compared with the conventional methods in all cases, except for the pair of Ob. and Pf.

TABLE I
MAXIMUM VALUE OF SDR IN EACH MIXTURE [dB]

	SNMF	PSNMF	SNMF-ABD	Proposed method
Ob. & Pf.	7.6	6.7	**8.1**	7.1
Ob. & Tb.	1.5	2.4	2.6	**3.0**
Pf. & Ob.	3.0	4.1	3.6	**7.0**
Pf. & Tb.	1.9	3.1	3.2	**5.0**
Tb. & Ob.	-0.6	0.7	0.2	**2.6**
Tb. & Pf.	1.8	2.9	2.6	**4.5**
Average	2.5	3.3	3.4	**4.9**

V. CONCLUSION

In this paper, we propose a new advanced SNMF that includes discriminative deformation of the trained basis to make it fit the target sound. From the experimental results, it was confirmed that the proposed method outperforms the conventional methods in many cases.

REFERENCES

[1] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. ASLP*, vol. 15, no. 3, pp. 1066–1074, 2007.

[2] A. Ozerov, C. Fevotte and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," *Proc. WASPAA*, pp. 121–124, 2009.

[3] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino and S. Sagayama, "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints," *Proc. ICASSP*, pp. 5365–5368, 2012.

[4] P. Smaragdis, B. Raj, M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. LVA/ICA, LNCS 6365*, pp. 140—148, 2010.

[5] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans. Fundamentals*, vol. E97-A, no. 5, pp. 1113–1118, 2014.

[6] E.M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation," *Proc. Interspeech*, 2013.

[7] N. Mohammadiha, P. Smaragdis, A. Leijon, "Low-artifact source separation usig probablistic latent component analysis," *Proc. WASPAA*, 2013.

[8] D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, Y. Takahashi, "Music signal separation by supervised nonnegative matrix factorization with basis deformation," *Proc. IEEE 18th International Conference on Digital Signal Processing (DSP2013)*, no. T3P(C)-1, 2013.

[9] C. Breihaupt, M. Krawczyk, R. Martin "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," *Proc. ICASSP*, pp. 4037–4040, 2008.

[10] Y. Murota, D. Kitamura, S. Nakai, H. Saruwatari, S. Nakamura, K. Shikano, Y. Takahashi, K, Kondo,"Music signal separation based on bayesian spectral amplitude estimator with automatic target prior adaptation," *Proc. ICASSP*, pp. 7490–7494, 2014.

[11] N. Yasuraoka, H. Okuno, "Musical audio signal modeling for joint estimation of harmonic, inharmonic, and timbral structure and its application to source separation," *IPSJ-MUS*, vol. 27, pp. 1–8, 2012 (in Japanese).

[12] F. Weninger, J. Le Roux, J. R. Hershey, S. Watanabe, "Discriminative NMF and its application to single-channel source separation," *Proc. ISCA Interspeech 2014*, 2014.

[13] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.