Audio Signal Separation Using Supervised NMF with Time-Variant All-Pole-Model-Based Basis Deformation

Hiroaki Nakajima*, Daichi Kitamura†, Norihiro Takamune*, Shoichi Koyama*,
Hiroshi Saruwatari*, 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
§Research & Development Division, Yamaha Corporation, 203 Matsunokijima, Iwata, Shizuoka, 438–0192, Japan

***Abstract*—We address a novel nonnegative matrix factorization (NMF) with a new basis deformation method to handle various music sounds. Conventional supervised NMF has a critical problem that a mismatch between bases trained in advance and an actual target sound reduces the accuracy of separation. To solve this problem, we proposed an advanced supervised NMF that applies a single time-invariant filter to the bases for making them fit into the target sound. However, this method suffers from limitations on basis deformation ability, especially for transient instrumental sounds. In this paper, we propose a new time-variant all-pole-model-based basis deformation method. Our proposed deformation method consists of two types of filter that individually deforms attack and sustain parts in one note. Each of the all-pole models can be automatically selected and adapted to the open data via a statistical signal sampling approach. Experimental results show that the proposed method outperforms conventional methods in many types of instrumental sound.**

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], [4], [5]. The latter method is called *supervised NMF* (SNMF), which includes an a priori training process and requires some sound samples of a target signal [6], [7]. 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.

To reduce the mismatch problem, we proposed a new SNMF that applies a single time-invariant filter to the bases for making them fit into the target sound [8]. However, this method suffers from limitations on basis deformation ability, especially for transient instrumental sounds, e.g., a piano sound that is generated via different types of vibration mechanism in each time duration. In this paper, we propose a new time-variant all-pole-model-based basis deformation method. Our proposed deformation method consists of two types of filter that individually deforms attack and sustain parts in one note. Each of the all-pole models can be automatically selected and adapted to the open data via a statistical signal sampling approach. Experimental results show that the proposed method outperforms conventional methods in many types of instrumental sound.

II. Conventional Methods

A. SNMF

SNMF [6] 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, $\boldsymbol{FG}$ 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.

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 [7], [9], [10]. For example, in penalized SNMF (PSNMF) [7], the cost function is described as

$$\begin{aligned}&\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} \mathcal{D}_{\mathrm{P\_KL}}(\boldsymbol{Y}_{\mathrm{mix}}|\boldsymbol{FG}+\boldsymbol{HU})\\ &\quad=\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} \mathcal{D}_{\mathrm{KL}}(\boldsymbol{Y}_{\mathrm{mix}}|\boldsymbol{FG}+\boldsymbol{HU})+\mu||\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}||^2_{\mathrm{Fr}},\end{aligned} \tag{4}$$

where μ is a weight parameter and $||\cdot||^2_{\mathrm{Fr}}$ is Frobenius-norm.

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 [11]. In this method, the following equation represents the decomposition in SNMF-ABD with trained supervision $\boldsymbol{F}$:

$$\boldsymbol{Y}_{\mathrm{mix}} \simeq (\boldsymbol{F}+\boldsymbol{D})\boldsymbol{G}+\boldsymbol{HU}, \tag{5}$$

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

$$\begin{aligned}\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} &\mathcal{D}_{\mathrm{KL}}(\boldsymbol{Y}_{\mathrm{mix}}|(\boldsymbol{F}+\boldsymbol{D})\boldsymbol{G}+\boldsymbol{HU})+\mu_1||\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}||^2_{\mathrm{Fr}}\\ &+\mu_2||\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}||^2_{\mathrm{Fr}}+\mu_3||\boldsymbol{D}^{\mathrm{T}}\boldsymbol{H}||^2_{\mathrm{Fr}},\end{aligned} \tag{6}$$

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.

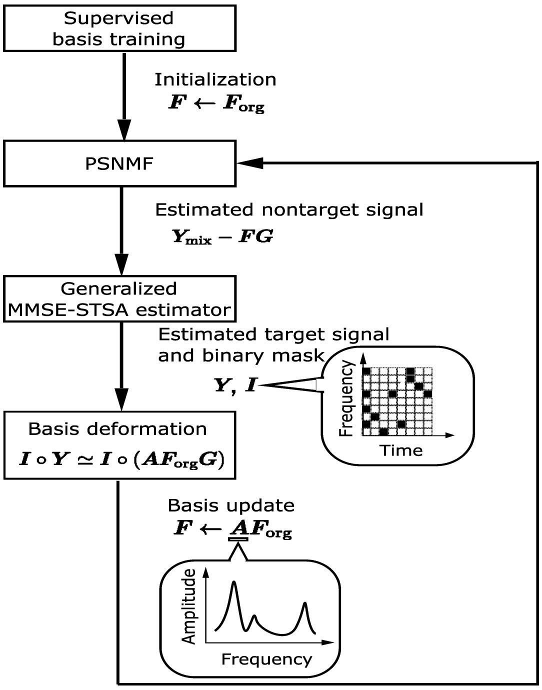
Supervised basis training
Initialization
$\boldsymbol{F} \leftarrow \boldsymbol{F}_{\mathrm{org}}$
PSNMF
Estimated nontarget signal
$\boldsymbol{Y}_{\mathrm{mix}}-\boldsymbol{FG}$
Generalized MMSE-STSA estimator
Estimated target signal and binary mask
$\boldsymbol{Y}, \boldsymbol{I}$
Frequency
Time
Basis deformation
$\boldsymbol{I}\circ\boldsymbol{Y}\simeq\boldsymbol{I}\circ(\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G})$
Basis update
$\boldsymbol{F} \leftarrow \underline{\boldsymbol{A}}\boldsymbol{F}_{\mathrm{org}}$
Amplitude
Frequency

Fig. 1. Block diagram of TID.

C. SNMF with time-invariant basis deformation (TID)

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. Therefore, we proposed a new SNMF introducing that deform the basis carried out with a linear time-invariant filter, namely an all-pole model, that consists of fewer parameters [8].

A block diagram of the method is shown in Fig. 1. First, we perform PSNMF with a current supervised basis $\boldsymbol{F}$. Second, using a generalized minimum meansquare error short-time spectral amplitude (MMSE-STSA) estimator [12] (one of the Bayesian estimators) 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 $\boldsymbol{F}_{\mathrm{org}}$ 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. The above-mentioned concepts are described as

$$\boldsymbol{I}\circ\boldsymbol{Y}\simeq\boldsymbol{I}\circ(\boldsymbol{A}\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G}), \tag{7}$$

where $\boldsymbol{I}$ is an $\Omega \times T$ binary mask matrix with entries $i_{\omega,t}$, which was obtained from 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

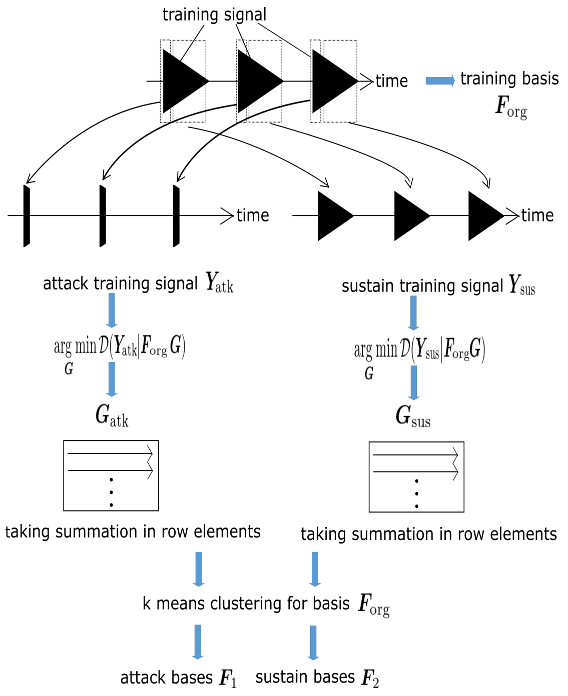
training signal
time
training basis F_{org}
time
time
attack training signal Y_{atk}
sustain training signal Y_{sus}
$\arg\min_G \mathcal{D}(Y_{\mathrm{atk}}|F_{\mathrm{org}}G)$
$\arg\min_G \mathcal{D}(Y_{\mathrm{sus}}|F_{\mathrm{org}}G)$
G_{atk}
G_{sus}
taking summation in row elements
taking summation in row elements
k means clustering for basis F_{org}
attack bases F_1
sustain bases F_2

Fig. 2. Block diagram of basis division procedure.

of $\boldsymbol{A}$ are described as

$$A_{\omega,\omega} = \frac{1}{|1 - \sum_{k=1}^{p} \alpha_k \exp(-\pi j k \frac{\omega}{\Omega})|}, \tag{8}$$

where p is the order and α_k are the coefficients of the all-pole model.

III. Proposed Method

A. Overview of proposed method

In the method described in the previous section, we deform the trained basis using the linear time-invariant filter so as to be suitable for source separation. In other words, supposing that musical instruments are linear time-invariant systems, it is possible for this method to eliminate the mismatch between the trained basis and the spectrogram of the actual target sound in open data. However, common systems of musical instruments are not time-invariant but time-varying systems, especially in the front of one note (called *attack*) and the rear end of one note (called *sustain*) [13]. Therefore, the trained basis can be deformed more suitably for source separation if the trained basis is deformed independently according to attack and sustain. In other words, the basis should be deformed using time-variant filters to improve the quality of source separation.

Meanwhile, raising freedom of basis deformation is equivalent to raising a risk to deform excessively, resulting in wrong basis transformation into non-target signals because the generalized MMSE-STSA estimator is a Bayesian estimator that allows the statistical error to some extent. Therefore, it is necessary to restrict freedom of deformation rationally.

In this section, we propose time-variant discriminative basis deformation method introducing the following schemes. (a) We focus on attack and sustain in one note (this is a time-variant system) and deform the trained basis respectively. (b) The deformation is carried out in a discriminative manner to avoid the excess deformation and balance the deformation and separation. The following subsections describe the detail algorithms in each scheme.

B. Basis deformation with time-variant all-pole model using generalized MMSE-STSA estimator

In this section, we propose a deformation method that individually deforms attack and sustain parts in one note using the generalized MMSE-STSA estimator. In the following states, as shown in Fig. 2, we divide $\boldsymbol{F}_{\mathrm{org}}$ into two sub-matrices, $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ that respectively represent attack and sustain part in one note. First, we separate the training signal into two parts corresponding to attack parts and sustain parts. Then we convert them into spectrograms $\boldsymbol{Y}_{\mathrm{atk}}$ and $\boldsymbol{Y}_{\mathrm{sus}}$ and optimize (9) and (10) as

$$\boldsymbol{G}_{\mathrm{atk}} = \arg\min_{\boldsymbol{G}} \mathcal{D}(\boldsymbol{Y}_{\mathrm{atk}}|\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G}), \tag{9}$$

$$\boldsymbol{G}_{\mathrm{sus}} = \arg\min_{\boldsymbol{G}} \mathcal{D}(\boldsymbol{Y}_{\mathrm{sus}}|\boldsymbol{F}_{\mathrm{org}}\boldsymbol{G}). \tag{10}$$

Taking summation of $\boldsymbol{G}_{\mathrm{atk}}$ and $\boldsymbol{G}_{\mathrm{sus}}$ in row elements, we obtain excitation degree of each basis in the physical states of musical sound, namely attack and sustain. Next, using this excitation degree of each basis, we separate $\boldsymbol{F}_{\mathrm{org}}$ into $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ using k-means method. Note that the relation, $[\boldsymbol{F}_1|\boldsymbol{F}_2] = \boldsymbol{F}_{\mathrm{org}}$, holds regardless of the row-wise permutation in $\boldsymbol{F}_{\mathrm{org}}$.

Second, we deform the clustered bases $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ according to

$$\boldsymbol{I} \circ \boldsymbol{Y} \simeq \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{G}_2), \tag{11}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal matrices in which the diagonal elements are described using the all-pole model.

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

$$\mathcal{J} = \sum_{\omega,t} i_{\omega,t} \Big\{ -y_{\omega,t} + \frac{\sum_k f_{\omega,k,1} g_{k,t,1}}{|A_\omega|} + \frac{\sum_l f_{\omega,l,2} g_{l,t,2}}{|B_\omega|} + y_{\omega,t} \log \frac{y_{\omega,t}}{\sum_k f_{\omega,k,1} g_{k,t,1}/|A_\omega| + \sum_l f_{\omega,l,2} g_{l,t,2}/|B_\omega|} \Big\}, \tag{12}$$

where $y_{\omega,t}$, $f_{\omega,k,1}$, $f_{\omega,l,2}$, $g_{k,t,1}$ and $g_{l,t,2}$ are the nonnegative elements of matrices $\boldsymbol{Y}$, $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$, respectively. In addition, A_ω represents $1 - \sum_{k=1}^{p} \alpha_k \exp(-\pi j k \frac{\omega}{\Omega})$, B_ω represents $1 - \sum_{k=1}^{p} \beta_k \exp(-\pi j k \frac{\omega}{\Omega})$ and β_k is the coefficient of the all-pole model of $\boldsymbol{B}$. This cost function can be optimized using auxiliary function technique (see Appendix) and we obtain the update rule of $g_{k,t,1}$, $g_{l,t,2}$, α_k and β_k as

$$g_{k,t,1} \leftarrow g_{k,t,1} \Big(\sum_\omega \frac{i_{\omega,t} y_{\omega,t} f_{\omega,k,1}}{\sum_k f_{\omega,k,1} g_{k,t,1} + |A_\omega| \sum_l f_{\omega,l,2} g_{k,l,2}/|B_\omega|} \Big) \Big/ \Big(\sum_\omega \frac{i_{\omega,t} f_{\omega,k,1}}{|A_\omega|} \Big), \tag{13}$$

$$\begin{aligned}
g_{l,t,2} \leftarrow & g_{l,t,2} \\
& \Big(\sum_{\omega} \frac{i_{\omega,t} y_{\omega,t} f_{\omega,l,2}}{\sum_l f_{\omega,l,2} g_{l,t,2} + |B_\omega| \sum_k f_{\omega,k,1} g_{k,t,1}/|A_\omega|}\Big) \\
& /\Big(\sum_{\omega} \frac{i_{\omega,t} f_{\omega,l,2}}{|B_\omega|}\Big), \qquad (14) \\
\boldsymbol{\alpha} =& \boldsymbol{R}^{-1}\boldsymbol{r}, \qquad (15) \\
\boldsymbol{\beta} =& \boldsymbol{W}^{-1}\boldsymbol{w}, \qquad (16)
\end{aligned}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the vectors of coefficients in the all-pole model weight matrix $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. In addition, $\boldsymbol{R}$, $\boldsymbol{r}$, $\boldsymbol{W}$ and $\boldsymbol{w}$ are given by

$$\begin{aligned}
R_{m,q} =& \sum_{\omega,t} \Bigg[i_{\omega,t} \Big(\sum_k \frac{f_{\omega,k,1} g_{k,t,1}}{|A_\omega|^3} + \frac{y_{\omega,t}}{2|A_\omega|^2} \\
& \frac{\sum_k f_{\omega,k,1} g_{k,t,1}}{\sum_k f_{\omega,k,1} g_{k,t,1} + |A_\omega| \sum_l f_{\omega,l,2} g_{k,l,2}/|B_\omega|} \Big) \\
& \Big(\exp\big(-\pi j \frac{\omega}{\Omega}(m-q)\big) + \exp\big(\pi j \frac{\omega}{\Omega}(m-q)\big) \Big) \Bigg],
\end{aligned} \qquad (17)$$

$$\begin{aligned}
r_q =& \sum_{\omega,t} i_{\omega,t} \Bigg[\big(\sum_k f_{\omega,k,1} g_{k,t,1} \frac{1}{|A_\omega|^3} + \frac{y_{\omega,t}}{2|A_\omega|^2} \\
& \frac{\sum_k f_{\omega,k,1} g_{k,t,1}}{\sum_k f_{\omega,k,1} g_{k,t,1} + |A_\omega| \sum_l f_{\omega,l,2} g_{k,l,2}/|B_\omega|} \big) \\
& \big(\exp(-\pi j \frac{\omega}{\Omega} q) + \exp(\pi j \frac{\omega}{\Omega} q) \big) \\
& - 3 \sum_k f_{\omega,k,1} g_{k,t,1} \mathcal{R}e \Big[\frac{(A_\omega)^*}{|A_\omega|^3} \exp\big(-\pi j \frac{\omega}{\Omega} q)\big) \Big] \Bigg],
\end{aligned} \qquad (18)$$

$$\begin{aligned}
W_{m,q} =& \sum_{\omega,t} \Bigg[i_{\omega,t} \Big(\sum_l f_{\omega,l,2} g_{l,t,2} \frac{1}{|B_\omega|^3} + \frac{y_{\omega,t}}{2|B_\omega|^2} \\
& \frac{\sum_l f_{\omega,l,2} g_{l,t,2}}{\sum_l f_{\omega,l,2} g_{l,t,2} + |B_\omega| \sum_k f_{\omega,k,1} g_{k,t,1}/|A_\omega|} \Big) \\
& \Big(\exp\big(-\pi j \frac{\omega}{\Omega}(m-q)\big) + \exp\big(\pi j \frac{\omega}{\Omega}(m-q)\big) \Big) \Bigg],
\end{aligned} \qquad (19)$$

$$\begin{aligned}
w_q =& \sum_{\omega,t} i_{\omega,t} \Bigg[\big(\sum_l f_{\omega,l,2} g_{l,t,2} \frac{1}{|B_\omega|^3} + \frac{y_{\omega,t}}{2|B_\omega|^2} \\
& \frac{\sum_l f_{\omega,l,2} g_{l,t,2}}{\sum_l f_{\omega,l,2} g_{l,t,2} + |B_\omega| \sum_k f_{\omega,k,1} g_{k,t,1}/|A_\omega|} \big) \\
& \big(\exp(-\pi j \frac{\omega}{\Omega} q) + \exp(\pi j \frac{\omega}{\Omega} q) \big) \\
& - 3 \sum_l f_{\omega,l,2} g_{l,t,2} \mathcal{R}e \Big[\frac{(B_\omega)^*}{|B_\omega|^3} \exp\big(-\pi j \frac{\omega}{\Omega} q)\big) \Big] \Bigg].
\end{aligned} \qquad (20)$$

C. Discriminative basis deformation

In our method, we use the generalized MMSE-STSA estimator as a sampler to deform the bases $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$. 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 [14], which learns supervised bases while paying attention to the separability of signals, to our proposed basis deformation. Note that the method in [14] 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

$$\begin{aligned}
& \boldsymbol{A}, \boldsymbol{B} \\
& \quad = \mathop{\arg\min}_{\boldsymbol{A},\boldsymbol{B}} \mathcal{D}\Big(\boldsymbol{I} \circ \boldsymbol{Y} | \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_1\boldsymbol{G}_{1\mathrm{s}} + \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{G}_{2\mathrm{s}}) \Big) \\
& \text{s.t. } \boldsymbol{G}_{1\mathrm{s}}, \boldsymbol{G}_{2\mathrm{s}} \\
& \quad = \mathop{\arg\min}_{\boldsymbol{G}_1,\boldsymbol{G}_2,\boldsymbol{H},\boldsymbol{U}} \mathcal{D}\Big(\boldsymbol{I} \circ \boldsymbol{Y}_{\mathrm{mix}} | \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{G}_2 + \boldsymbol{H}\boldsymbol{U}) \Big).
\end{aligned} \qquad (21)$$

This bilevel optimization searches for the optimal basis deformation matrix $\boldsymbol{A}$ and $\boldsymbol{B}$ under the constraint of minimizing $\mathcal{D}(\boldsymbol{I} \circ \boldsymbol{Y}_{\mathrm{mix}} | \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{G}_2 + \boldsymbol{H}\boldsymbol{U}))$ with respect to $\boldsymbol{G}_1$, $\boldsymbol{G}_2$, $\boldsymbol{H}$, and $\boldsymbol{U}$. To minimize (21), it is reasonable for $\boldsymbol{A}\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{G}_2$ 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

$$\boldsymbol{A}, \boldsymbol{B} = \mathop{\arg\min}_{\boldsymbol{A},\boldsymbol{G}_1,\boldsymbol{B},\boldsymbol{G}_2} \mathcal{D}\Big(\boldsymbol{I} \circ \boldsymbol{Y} | \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{G}_2) \Big). \qquad (22)$$

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

$$\begin{aligned}
\boldsymbol{G}_1, \boldsymbol{G}_2 = & \mathop{\arg\min}_{\boldsymbol{G}_1,\boldsymbol{G}_2,\boldsymbol{H},\boldsymbol{U}} \\
& \mathcal{D}\Big(\boldsymbol{I} \circ \boldsymbol{Y}_{\mathrm{mix}} | \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{G}_2 + \boldsymbol{H}\boldsymbol{U}) \Big).
\end{aligned} \qquad (23)$$

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

$$\boldsymbol{A}, \boldsymbol{B} = \mathop{\arg\min}_{\boldsymbol{A},\boldsymbol{B}} \mathcal{D}\Big(\boldsymbol{I} \circ \boldsymbol{Y} | \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{G}_2) \Big). \qquad (24)$$

Return to Step 2

This algorithm searches for the basis deformation matrix $\boldsymbol{A}$ and $\boldsymbol{B}$ that minimizes $\mathcal{D}(\boldsymbol{I} \circ \boldsymbol{Y}_{\mathrm{mix}} | \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{G}_2 + \boldsymbol{H}\boldsymbol{U}))$ in the vicinity of the minimal $\mathcal{D}(\boldsymbol{I} \circ \boldsymbol{Y} | \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{B}\boldsymbol{F}_2\boldsymbol{G}_2))$.

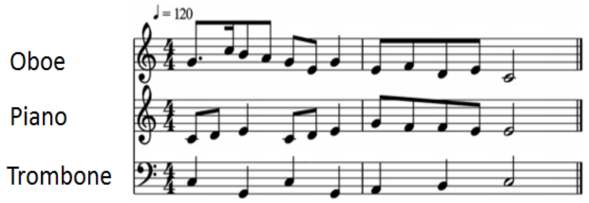

Fig. 3. Scores of each instrument.

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

	SNMF	PNMF	SNMF-ABD	TID	Proposed
Ob. & Pf.	7.6	6.7	**8.1**	6.7	7.0
Ob. & Tb.	1.5	2.4	2.6	2.8	**2.9**
Pf. & Ob.	3.0	4.1	3.6	5.2	**6.1**
Pf. & Tb.	1.9	3.1	3.2	**4.5**	**4.5**
Tb. & Ob.	-0.6	0.7	0.2	2.4	**2.8**
Tb. & Pf.	1.8	2.9	3.4	3.9	**4.4**
Average	2.5	3.3	3.4	4.3	**4.9**

IV. EXPERIMENT

A. *Experimental conditions*

To evaluate the proposed algorithm, we compared the conventional methods (SNMF, PSNMF, SNMF-ABD and TID) and 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. 3 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 [15]. 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 TID and the 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 4 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, TID 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.

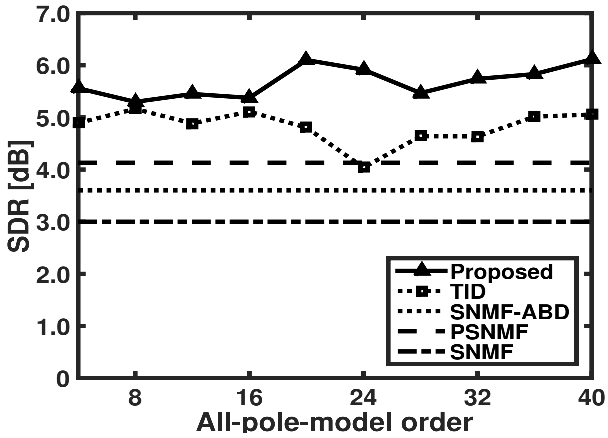

Fig. 4. Example of SDR for separating Pf. from mixture of Pf. and Ob.

V. CONCLUSIONS

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

APPENDIX
DERIVATION OF UPDATE RULES FOR TIME-VARIANT BASIS DEFORMATION

This section expresses how to derive the update rules for time-variant basis deformation, namely, $\boldsymbol{G}_1$, $\boldsymbol{G}_2$, $\boldsymbol{A}$ and $\boldsymbol{B}$ in (11). The cost function for (11) based on the generalized KL divergence is given by

$$\begin{aligned}\mathcal{J} = \sum_{\omega,t} i_{\omega,t}\Big\{&\frac{\sum_k f_{\omega,k,1} g_{k,t,1}}{|A_\omega|} + \frac{\sum_l f_{\omega,l,2} g_{l,t,2}}{|B_\omega|}\\ &- y_{\omega,t}\log(\frac{\sum_k f_{\omega,k,1} g_{k,t,1}}{|A_\omega|} + \frac{\sum_l f_{\omega,l,2} g_{l,t,2}}{|B_\omega|}) + C_{\omega,t}\Big\},\end{aligned} \tag{25}$$

where $C_{\omega,t}$ are unnecessary constants when calculating the update rules. Since it is difficult to analytically derive the optimal $\boldsymbol{G}_1$, $\boldsymbol{G}_2$, $\boldsymbol{A}$ and $\boldsymbol{B}$, we define an auxiliary function that represents the upper bound of $\mathcal{J}$, as described below. First, applying Jensen's inequality to $\log(\sum_k f_{\omega,k,1} g_{k,t,1}/|A_\omega| +$

$\sum_l f_{\omega,l,2}g_{l,t,2}/|B_\omega|)$, we have

$$\begin{aligned}\mathcal{J} \le \sum_{\omega,t} i_{\omega,t}\Big\{&\frac{\sum_k f_{\omega,k,1}g_{k,t,1}}{|A_\omega|} + \frac{\sum_l f_{\omega,l,2}g_{l,t,2}}{|B_\omega|}\\ &- y_{\omega,t}\eta_{\omega,t,1}\log(\frac{\sum_k f_{\omega,k,1}g_{k,t,1}}{|A_\omega|\eta_{\omega,t,1}})\\ &- y_{\omega,t}\eta_{\omega,t,2}\log(\frac{\sum_l f_{\omega,l,2}g_{l,t,2}}{|B_\omega|\eta_{\omega,t,2}}) + C_{\omega,t}\Big\},\end{aligned} \tag{26}$$

where $\eta_{\omega,t,1}$ and $\eta_{\omega,t,2}$ are auxiliary variables. The equality in (26) holds if and only if the auxiliary variables are set to

$$\eta_{\omega,t,1} = \frac{\sum_k f_{\omega,k,1}g_{k,t,1}}{\sum_k f_{\omega,k,1}g_{k,t,1} + |A_\omega|\sum_l f_{\omega,l,2}g_{k,l,2}/|B_\omega|}, \tag{27}$$

$$\eta_{\omega,t,2} = \frac{\sum_l f_{\omega,l,2}g_{l,t,2}}{|B_\omega|\sum_k f_{\omega,k,1}g_{k,t,1}/|A_\omega| + \sum_l f_{\omega,l,2}g_{k,l,2}}. \tag{28}$$

Second, applying Jensen's inequality and the tangent inequality, we have

$$\begin{aligned}\mathcal{J} \le \sum_{\omega,t} i_{\omega,t}\Big\{&\frac{\sum_k f_{\omega,k,1}g_{k,t,1}}{|A_\omega|} + \frac{\sum_l f_{\omega,l,2}g_{l,t,2}}{|B_\omega|}\\ &- y_{\omega,t}\eta_{\omega,t,1}\sum_k \zeta_{\omega,t,k}\log\frac{f_{\omega,k,1}g_{k,t,1}}{\zeta_{\omega,t,k}}\\ &+ y_{\omega,t}\eta_{\omega,t,1}(\frac{1}{2\rho_\omega}|A_\omega|^2 + \frac{1}{2}\log\rho_\omega - \frac{1}{2})\\ &- y_{\omega,t}\eta_{\omega,t,2}\sum_l \epsilon_{\omega,t,l}\log\frac{f_{\omega,l,2}g_{l,t,2}}{\epsilon_{\omega,t,l}}\\ &+ y_{\omega,t}\eta_{\omega,t,2}(\frac{1}{2\sigma_\omega}|B_\omega|^2 + \frac{1}{2}\log\sigma_\omega - \frac{1}{2}) + C_{\omega,t}\Big\},\end{aligned} \tag{29}$$

where $\zeta_{\omega,t,k}$, $\epsilon_{\omega,t,l}$, ρ_ω and σ_ω are auxiliary variables. The equality in (29) holds if and only if the auxiliary variables are set to $\zeta_{\omega,t,k} = f_{\omega,k,1}g_{k,t,1}/(\sum_k f_{\omega,k,1}g_{k,t,1})$, $\epsilon_{\omega,t,l} = f_{\omega,l,2}g_{l,t,2}/(\sum_l f_{\omega,l,2}g_{l,t,2})$ and $\rho_\omega = |A_\omega|^2$, $\sigma_\omega = |B_\omega|^2$. Third, to make the auxiliary function a quadratic form of $|A_\omega|$ and $|B_\omega|$, we conduct a Taylor expansion around τ_ω and υ_ω respectively,

$$\begin{aligned}\mathcal{J} \le \sum_{\omega,t} i_{\omega,t}\Big\{&\sum_k f_{\omega,k,1}g_{k,t,1}(\frac{1}{\tau_\omega^3}|A_\omega|^2 - 3\frac{1}{\tau_\omega^2}|A_\omega| + \frac{3}{\tau_\omega})\\ &+ \sum_l f_{\omega,l,2}g_{l,t,2}(\frac{1}{\upsilon_\omega^3}|B_\omega|^2 - 3\frac{1}{\upsilon_\omega^2}|B_\omega| + \frac{3}{\upsilon_\omega})\\ &- y_{\omega,t}\eta_{\omega,t,1}\sum_k \zeta_{\omega,t,k}\log\frac{f_{\omega,k,1}g_{k,t,1}}{\zeta_{\omega,t,k}} + \frac{y_{\omega,t}\eta_{\omega,t,1}}{2\rho_\omega}|A_\omega|^2\\ &- y_{\omega,t}\eta_{\omega,t,2}\sum_l \epsilon_{\omega,t,l}\log\frac{f_{\omega,l,2}g_{l,t,2}}{\epsilon_{\omega,t,l}}\\ &+ \frac{y_{\omega,t}\eta_{\omega,t,2}}{2\sigma_\omega}|B_\omega|^2 + C_{\omega,t}\Big\}.\end{aligned} \tag{30}$$

The equality of (30) holds if and only if $\tau_\omega = |A_\omega|$ and $\upsilon_\omega = |B_\omega|$. This approximation does not meet the condition of an auxiliary function, but if τ_ω is updated as $|A_\omega|$ and υ_ω is B_ω, this approximation is equivalent to Newton's method. Finally, using the inequality $\mathcal{R}e[\kappa_\omega^* A_\omega] \le |A_\omega|$ and $\mathcal{R}e[\theta_\omega^* B_\omega] \le |B_\omega|$, we can define the upper bound function $\mathcal{J}^+$ for $\mathcal{J}$ as

$$\begin{aligned}\mathcal{J} \le \sum_{\omega,t} i_{\omega,t}\Big\{&\sum_k f_{\omega,k,1}g_{k,t,1}(\frac{1}{\tau_\omega^3}|A_\omega|^2 - 3\frac{1}{\tau_\omega^2}\mathcal{R}e[\kappa_\omega^* A_\omega])\\ &+ \sum_l f_{\omega,l,2}g_{l,t,2}(\frac{1}{\upsilon_\omega^3}|B_\omega|^2 - 3\frac{1}{\upsilon_\omega^2}\mathcal{R}e[\theta_\omega^* B_\omega])\\ &- y_{\omega,t}\eta_{\omega,t,1}\sum_k \zeta_{\omega,t,k}\log\frac{f_{\omega,k,1}g_{k,t,1}}{\zeta_{\omega,t,k}} + \frac{y_{\omega,t}\eta_{\omega,t,1}}{2\rho_\omega}|A_\omega|^2\\ &- y_{\omega,t}\eta_{\omega,t,2}\sum_l \epsilon_{\omega,t,l}\log\frac{f_{\omega,l,2}g_{l,t,2}}{\epsilon_{\omega,t,l}}\\ &+ \frac{y_{\omega,t}\eta_{\omega,t,2}}{2\sigma_\omega}|B_\omega|^2 + C_{\omega,t}\Big\}.\end{aligned} \tag{31}$$

where $\mathcal{R}e[\cdot]$ is a real part of $\cdot$, $|\kappa_\omega| = 1$ and $|\theta_\omega| = 1$. The equality of (31) holds if and only if $\kappa_\omega = A_\omega/|A_\omega|$ and $\theta_\omega = B_\omega/|B_\omega|$.

A. Multiplicative update rules for activation matrices $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$

The update rule for $\mathcal{J}^+$ with respect to the activation matrix $\boldsymbol{G}_1$ is determined by setting the gradient to zero. From $\partial\mathcal{J}^+/\partial g_{k,t,1} = 0$, we obtain

$$\begin{aligned}&\sum_\omega i_{\omega,t}\Big\{f_{\omega,k,1}(\frac{1}{\tau_\omega^3}|A_\omega|^2 - 3\frac{1}{\tau_\omega^2}\mathcal{R}e[\kappa_\omega^* A_\omega] + \frac{3}{\tau_\omega})\\ &- \frac{y_{\omega,t}\eta_{\omega,t,1}\zeta_{\omega,t,k}}{g_{k,t,1}})\Big\}\\ &= 0.\end{aligned} \tag{32}$$

By substituting the auxiliary variables into and simplifying it, we obtain the multiplicative update rule of $g_{k,t,1}$ as (13). In addition, we similarly obtain the multiplicative update rule of $g_{l,t,2}$ as (14).

B. Update rule for all-pole-model weight matrices $\boldsymbol{A}$ and $\boldsymbol{B}$

First, by differentiating $\mathcal{J}^+$ partially with respect to α_q, which is an coefficient of the all-pole model weight matrix $\boldsymbol{A}$, and setting it to zero, we obtain

$$\begin{aligned}&\sum_{m=1}^{p}\alpha_m\sum_{\omega,t}\Bigg[i_{\omega,t}(\sum_k f_{\omega,k,1}g_{k,t,1}\frac{1}{\tau_\omega^3} + y_{\omega,t}\eta_{\omega,t,1}\frac{1}{2\rho_\omega})\\ &\Big(\exp(-\pi j\frac{\omega}{\Omega}(m-q)) + \exp\big(\pi j\frac{\omega}{\Omega}(m-q)\big)\Big)\Bigg]\\ &- \sum_{\omega,t} i_{\omega,t}\Bigg[(\sum_k f_{\omega,k,1}g_{k,t,1}\frac{1}{\tau_\omega^3} + y_{\omega,t}\eta_{\omega,t,1}\frac{1}{2\rho_\omega})\\ &\big(\exp(-\pi j\frac{\omega}{\Omega}q) + \exp(\pi j\frac{\omega}{\Omega}q)\big)\\ &- \frac{3}{\tau_\omega^2}\sum_k f_{\omega,k,1}g_{k,t,1}\mathcal{R}e[\kappa_\omega^*\exp(-\pi j\frac{\omega}{\Omega}q))]\Bigg]\\ &= 0,\end{aligned} \tag{33}$$

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

$$R_{m,q} = \Biggl[i_{\omega,t} (\sum_k f_{\omega,k,1} g_{k,t,1} \frac{1}{\tau_\omega^3} + y_{\omega,t} \eta_{\omega,t,1} \frac{1}{2\rho_\omega}) \Bigl(\exp\bigl(-\pi j \frac{\omega}{\Omega}(m-q)\bigr) + \exp\bigl(\pi j \frac{\omega}{\Omega}(m-q)\bigr) \Bigr) \Biggr], \tag{34}$$

$$r_q = \sum_{\omega,t} i_{\omega,t} \Biggl[(\sum_k f_{\omega,k,1} g_{k,t,1} \frac{1}{\tau_\omega^3} + y_{\omega,t} \eta_{\omega,t,1} \frac{1}{2\rho_\omega}) \bigl(\exp(-\pi j \frac{\omega}{\Omega} q) + \exp(\pi j \frac{\omega}{\Omega} q) \bigr) - \frac{3}{\tau_\omega^2} \sum_k f_{\omega,k,1} g_{k,t,1} \mathcal{R}e[\kappa_\omega^* \exp(-\pi j \frac{\omega}{\Omega} q))] \Biggr]. \tag{35}$$

By substituting (34) and (35) into (33), we obtain

$$\boldsymbol{R}\boldsymbol{\alpha} = \boldsymbol{r}, \tag{36}$$

where $\boldsymbol{\alpha}$ is the vector of coefficients in the all-pole model weight matrix $\boldsymbol{A}$. Since $\boldsymbol{R}$ is a Toeplitz matrix, we can derive $\boldsymbol{\alpha}$ using the Levinson–Durbin algorithm with a computationally efficient form. In addition, we similarly obtain the update rule of $\boldsymbol{\beta}$, which is the vector of coefficients for the all-pole model weight matrix $\boldsymbol{B}$, as (16).

Acknowledgment

This work was partly supported by SECOM Science and Technology Foundation.

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] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Trans. ASLP*, vol. 24, no. 9, 16pages, 2016. (in press)

[5] Y. Mitsufuji, S. Koyama and H. Saruwatari, "Multichannel blind source separation based on non-negative tensor factorization in wavenumber domain," *Proc. ICASSP*, pp. 56–60, 2016.

[6] 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.

[7] 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.

[8] H. Nakajima, D. Kitamura, N. Takamune, S. Koyama, H. Saruwatari, N. Ono, Y. Takahashi, K. Kondo, "Music signal separation using supervised NMF with all-pole-model-based basis deformation," *Proc. ASJ Autumn Meeting*, pp. 573–576, 2015. (in Japanese)

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

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

[11] 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.

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

[13] N. H. Fletcher, T. D. Rossing, "The Physics of Musical Instruments," *Springer-Verlag*, 1991.

[14] 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.

[15] 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.