3-1-11

種々の実楽器信号に対する基底変形型教師ありNMFの分離精度評価[*]

☆北村大地, 猿渡洋, 鹿野清宏 (奈良先端大), 近藤多伸, 高橋祐 (ヤマハ株式会社)

1 Introduction

In recent years, source separation based on nonnegative matrix factorization (NMF) [1], 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 has been a convincing candidate used for source separation [2] in musical signal processing with a monaural format.

The methods of source separation based on NMF are roughly classified into unsupervised and supervised algorithms. The former method attempts the separation accompanied by various constraints, e.g., proposed in Refs. [2, 3]. However, these techniques have a problem in spectral pattern clustering owing to the blind approach. The latter method includes a priori training, requiring some sample sounds of a target instrument. In particular, a supervised approach that introduces a penalized condition in oder to prevent some absences of the target sound achieves a good performance [4]. However, such supervised techniques have a critical problem that a mismatch between the spectra trained in advance and the target actual sound reduces the accuracy of source separation.

In this paper, we propose a new advanced supervised NMF algorithm that includes a deformable term for the trained spectral bases and constraints for making the bases to fit into the real instrumental sound. The experimental results show that the proposed method outperforms the conventional method [4] even with treating a mixture of real instruments.

2 Conventional method

2.1 Overview of constrained supervised NMF

In the unsupervised approaches, it has difficultly in clustering the decomposed spectral patterns into a specific target instrumental sound. Furthermore, each basis may be threatened to include a multi-instrumental spectral pattern. To solve this problem, constrained supervised NMF (CSNMF) has been proposed as a supervised method [4]. CSNMF consists of two processes, a priori training and observed signal separation, as described below in detail.

2.2 Training process of supervision

In CSNMF, as the supervision, a priori spectral patterns (bases) should be trained in advance to achieve source separation. Hereafter, we assume that we can obtain the instrumental sounds, which is the target of the separation task. The trained bases are constructed by NMF as

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

where $\boldsymbol{Y}_{\mathrm{target}}(\in \mathbb{R}_{\geq 0}^{\Omega \times T_s})$ is an amplitude spectrogram of a specific signal for training, $\boldsymbol{F}(\in \mathbb{R}_{\geq 0}^{\Omega \times K})$ is a nonnegative matrix that involves bases of the target signal as column vectors, and $\boldsymbol{Q}(\in \mathbb{R}_{\geq 0}^{K \times T_s})$ is a nonnegative matrix that corresponds to the activation of each basis of $\boldsymbol{F}$. In addition, Ω is the number of frequency bins, T_s is the number of frames of the training signal, and K is the number of bases.

In the decomposition of NMF, the cost function can be constructed using some measures of the distance between $\boldsymbol{Y}$ and $\boldsymbol{FQ}$ as

$$\mathcal{J}_{\mathrm{NMF}} = \mathcal{D}\left(\boldsymbol{Y}|\boldsymbol{FQ}\right), \tag{2}$$

where $\mathcal{D}\left(\cdot|\cdot\right)$ is an arbitrary distance function, e.g., Itakura-Saito divergence (*IS divergence*), and generalized Kullback Leibler divergence (*KL divergence*). In this study, we propose to use KL divergence in the cost function. The multiplicative update rules for $\boldsymbol{F}$ and $\boldsymbol{Q}$ are given by

$$f_{\omega,k} \leftarrow \frac{f_{\omega,k}\sum_{t_s} y_{\omega,t_s} q_{k,t_s}\left(\sum_{k'} f_{\omega,k'} q_{k',t_s}\right)^{-1}}{\sum_{t_s} q_{k,t_s}}, \tag{3}$$

$$q_{k,t_s} \leftarrow \frac{q_{k,t_s}\sum_{\omega} y_{\omega,t_s} f_{\omega,k}\left(\sum_{k'} f_{\omega,k'} q_{k',t_s}\right)^{-1}}{\sum_{\omega} f_{\omega,k}}, \tag{4}$$

where y_{ω,t_s}, $f_{\omega,k}$, and q_{k,t_s} are the nonnegative entries of the matrices $\boldsymbol{Y}$, $\boldsymbol{F}$, and $\boldsymbol{Q}$, respectively.

2.3 Signal separation process

The following equation represents the decomposition process of CSNMF using the trained supervision $\boldsymbol{F}$:

$$\boldsymbol{Y} \simeq \boldsymbol{FG} + \boldsymbol{HU}, \tag{5}$$

where $\boldsymbol{Y}(\in \mathbb{R}_{\geq 0}^{\Omega \times T})$ is an observed spectrogram, $\boldsymbol{G}(\in \mathbb{R}_{\geq 0}^{K \times T})$ is an activation matrix that corresponds to $\boldsymbol{F}$, $\boldsymbol{H}(\in \mathbb{R}_{\geq 0}^{\Omega \times L})$ is the residual spectral patterns that cannot be expressed by $\boldsymbol{FG}$, and $\boldsymbol{U}(\in \mathbb{R}_{\geq 0}^{L \times T})$ is an activation matrix that corresponds to $\boldsymbol{H}$. Hence, $\boldsymbol{FG}$ represents the target instrumental components, and $\boldsymbol{HU}$ represents other different components from the target sounds ideally. Moreover, L is the number of frames of the observed spectrogram.

2.4 Cost function with constrained condition

After the decomposition by Eq. (5), $\boldsymbol{FG}$ is expected to represent the spectrogram that corresponds to the signal known in advance, and $\boldsymbol{HU}$ expresses the spectrogram of other signals. However, if some of the spectral patterns in $\boldsymbol{F}$ and $\boldsymbol{H}$ are the same, the corresponding activations separately appear in $\boldsymbol{G}$ and $\boldsymbol{U}$, degrading the separation performance. To cope with this problem, a constraint is imposed in the cost functions of CSNMF as

$$\mathcal{J}_{\mathrm{CSNMF}} = \mathcal{D}\left(\boldsymbol{Y}|\boldsymbol{FG}+\boldsymbol{HU}\right) + \mu\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{F}}, \tag{6}$$

[*] "Evaluation of Separation Accuracy for Various Real Instruments Based on Supervised NMF with Basis Deformation," by Daichi Kitamura, Hiroshi Saruwatari, Kiyohiro Shikano (Nara Institute of Science and Technology), Kazunobu Kondo, Yu Takahashi (Yamaha Corporation).

where μ is weighting parameter, and $\|\cdot\|_{\mathrm{F}}$ represents Frobenius norm.

2.5 Multiplicative update rules of CSNMF

The update rules for Eq. (6) are given by

$$g_{k,t} \leftarrow \frac{g_{k,t}\sum_{\omega} f_{\omega,k} y_{\omega,t} \left(\sum_{k'} f_{\omega,k'} g_{k',t} + \sum_{l'} h_{\omega,l'} u_{l',t}\right)^{-1}}{\sum_{\omega} f_{\omega,k}}, \tag{7}$$

$$h_{\omega,l} \leftarrow \frac{h_{\omega,l}\sum_{t} y_{\omega,t} u_{l,t} \left(\sum_{k'} f_{\omega,k'} g_{k',t} + \sum_{l'} h_{\omega,l'} u_{l',t}\right)^{-1}}{\sum_{t} u_{l,t} + 2\mu \sum_{k'} f_{\omega,k'} \sum_{\omega'} f_{\omega',k'} h_{\omega',l}}, \tag{8}$$

$$u_{l,t} \leftarrow \frac{u_{l,t}\sum_{\omega} h_{\omega,l} y_{\omega,t} \left(\sum_{k'} f_{\omega,k'} g_{k',t} + \sum_{l'} h_{\omega,l'} u_{l',t}\right)^{-1}}{\sum_{\omega} h_{\omega,l}}, \tag{9}$$

where $g_{k,t}$, $h_{\omega,l}$, and $u_{l,t}$ are the nonnegative entries of the matrices $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$, respectively.

2.6 Problem of supervised method

The supervised techniques such as CSNMF involve an inherent problem that a mismatch between the bases trained in advance and the target actual sound reduces the accuracy of separation. Even if the trained bases are constructed using the same type of instrument as the target sound, the spectra are different according to, e.g., an individual style of playing and the timbre individuality for each instrument. Therefore, it is impossible to provide perfect supervision and to predict more realistic supervision in practice. To solve this problem, in the next section, we propose a new advanced CSNMF that includes a deformable term for the trained bases and the additional constraint for making the bases to fit into the target sound.

3 Proposed method

3.1 Supervised NMF with basis deformation

The proposed method also uses the pre-recorded sound that is similar to the target instrument and available in advance for a priori training, and composes the trained bases $\boldsymbol{F}$. For instance, when the target signal is the real specific instrumental sound, it is allowed to use MIDI sound of the same type of instrument because we can easily generate the training sound via MIDI. The decomposition model is represented as

$$\boldsymbol{Y} \simeq (\boldsymbol{F} + \boldsymbol{D})\,\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}, \tag{10}$$

where $\boldsymbol{D}(\in \mathbb{R}_{\geq 0}^{\Omega\times K})$ is a basis matrix which shares the activation matrix $\boldsymbol{G}$ with $\boldsymbol{F}$. In this decomposition, to adapt the bases into the target sound that cannot be represented by $\boldsymbol{F}$, another basis matrix $\boldsymbol{D}$ is imposed as a deformation term for $\boldsymbol{F}$. Furthermore, $\boldsymbol{D}$ is constructed under the following constraints,

$$\eta f_{\omega,k} + d_{\omega,k} \geq 0, \tag{11}$$

$$0 \leq \eta \leq 1, \tag{12}$$

$$g_{k,t} \geq 0, \tag{13}$$

$$h_{\omega,l} \geq 0, \tag{14}$$

$$u_{l,t} \geq 0, \tag{15}$$

where $d_{\omega,k}$ is the entry of the matrix $\boldsymbol{D}$, which probably has *positive and negative* values, and η is the parameter that represents an allowable range of negative deformation of $\boldsymbol{F}$.

3.2 Cost function with constraints

Since the deformation term $\boldsymbol{D}$ shares the activation $\boldsymbol{G}$ with $\boldsymbol{F}$, $\boldsymbol{D}$ can be threatened to include other instrumental components that simultaneously sound with the target such as *unison*. This yields a nonnegligible leakage of the undesired instrumental components into the resultant output $(\boldsymbol{F}+\boldsymbol{D})\,\boldsymbol{G}$. To avoid this phenomenon, the penalized terms for orthogonalization are imposed in the cost function as

$$\begin{aligned}\mathcal{J} =& \mathcal{D}(\boldsymbol{Y}|(\boldsymbol{F}+\boldsymbol{D})\boldsymbol{G}+\boldsymbol{H}\boldsymbol{U}) + \mu_1\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}\|_{\mathrm{F}}^2 \\ &+\mu_2\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{F}}^2 + \mu_3\|\boldsymbol{D}^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{F}}^2 + \mu_4\|(\boldsymbol{F}+\boldsymbol{D})^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{F}}^2,\end{aligned} \tag{16}$$

where μ_1, μ_2, μ_3, and μ_4 are the wighting parameters for each penalized term. These penalized terms indicate that $\boldsymbol{F}$, $\boldsymbol{D}$, and $\boldsymbol{H}$ are forced to become uncorrelated each other, and the target component bases $(\boldsymbol{F}+\boldsymbol{D})$ and other component bases $\boldsymbol{H}$ also become uncorrelated each other.

3.3 Multiplicative update rules

In this section, we derive the update rules based on Eq. (16). Equation (16) can be rewritten as

$$\begin{aligned}\mathcal{J} =& \sum_{\omega,t}\left(-y_{\omega,t}\log r_{\omega,t} + r_{\omega,t} + C_{\mathcal{J}}\right) \\ &+\mu_1\sum_{k,k'}\Big(\sum_{\omega} f_{\omega,k'} d_{\omega,k}\Big)^2 + \mu_2\sum_{k,l}\Big(\sum_{\omega} f_{\omega,k} h_{\omega,l}\Big)^2 \\ &+\mu_3\sum_{k,l}\Big(\sum_{\omega} d_{\omega,k} h_{\omega,l}\Big)^2 + \mu_4\sum_{k,l}\Big(\sum_{\omega} b_{\omega,k} h_{\omega,l}\Big)^2,\end{aligned} \tag{17}$$

where $r_{\omega,t}$, $b_{\omega,k}$, and $C_{\mathcal{J}}$ are given by

$$r_{\omega,t} = \sum_{k} b_{\omega,k} g_{k,t} + \sum_{l} h_{\omega,l} u_{l,t}, \tag{18}$$

$$b_{\omega,k} = f_{\omega,k} + d_{\omega,k}, \tag{19}$$

$$C_{\mathcal{J}} = y_{\omega,t}\log y_{\omega,t} - y_{\omega,t}. \tag{20}$$

Since it is difficult to analytically derive the optimal $\boldsymbol{D}$, $\boldsymbol{G}$, and $\boldsymbol{H}$ that minimize Eq. (17), we define an auxiliary function, witch represents the upper bound of $\mathcal{J}$ as described bellow.

First, for the logarithmic term (hereinafter, referred to as $\mathcal{J}_{\log}$) in the first term of the right-hand side in Eq. (17), the upper bound function $Q_{\log}$ is defined using auxiliary variables $\alpha_{k,\omega,t} \geq 0$, $\beta_{l,\omega,t} \geq 0$, $\gamma_1 \geq 0$, and $\gamma_2 \geq 0$ that satisfy $\sum_k \alpha_{k,\omega,t} = 1$, $\sum_l \beta_{l,\omega,t} = 1$, and $\gamma_1 + \gamma_2 = 1$. Applying Jensen's inequality to this, we have

$$\begin{aligned}\mathcal{J}_{\log} =& -y_{\omega,t}\log r_{\omega,t} \\ \leq& -y_{\omega,t}\sum_{k,l}\alpha_{k,\omega,t}\beta_{l,\omega,t}\log(A+B) + C_{\mathcal{J}_{\log}} \\ \leq& -y_{\omega,t}\sum_{k,l}\alpha_{k,\omega,t}\beta_{l,\omega,t}(\log A^{\gamma_1} + \log B^{\gamma_2}) + C'_{\mathcal{J}_{\log}} \\ \equiv& \, Q_{\log},\end{aligned} \tag{21}$$

where A, B, $C_{\mathcal{J}_{\log}}$, and $C'_{\mathcal{J}_{\log}}$ are given by

$$A = \beta_{l,\omega,t} b_{\omega,k} g_{k,t}, \tag{22}$$

$$B = \alpha_{k,\omega,t} h_{\omega,l} u_{l,t}, \tag{23}$$

$$C_{\mathcal{J}_{\log}} = -\textstyle\sum_{k,l} \alpha_{k,\omega,t}\beta_{l,\omega,t} \log \alpha_{k,\omega,t}\beta_{l,\omega,t}, \tag{24}$$

$$C'_{\mathcal{J}_{\log}} = C_{\mathcal{J}_{\log}} - \textstyle\sum_{k,l} \alpha_{k,\omega,t}\beta_{l,\omega,t}(\gamma_1 \log \gamma_1 + \gamma_2 \log \gamma_2). \tag{25}$$

The equality in Eq. (21) holds if and only if the auxiliary variables are set to as follows:

$$\alpha_{k,\omega,t} = \frac{b_{\omega,k} g_{k,t}}{\sum_{k'} b_{\omega,k'} g_{k',t}}, \tag{26}$$

$$\beta_{l,\omega,t} = \frac{h_{\omega,l} u_{l,t}}{\sum_{l'} h_{\omega,l'} u_{l',t}}, \tag{27}$$

$$\gamma_1 = \frac{A}{A+B}, \tag{28}$$

$$\gamma_2 = \frac{B}{A+B}. \tag{29}$$

Second, for the penalized terms (hereinafter, referred to as $\mathcal{J}_{\mathrm{p}}$) in Eq. (17), the upper bound function Q_{p} is defined using the auxiliary variables $\delta_{\omega,k',k} \geq 0$, $\epsilon_{\omega,k,l} \geq 0$, $\theta_{\omega,k,l} \geq 0$, and $\lambda_{\omega,k,l} \geq 0$ that satisfy $\sum_k \delta_{\omega,k',k}{=}1$, $\sum_l \epsilon_{\omega,k,l}{=}1$, $\sum_l \theta_{\omega,k,l}{=}1$, and $\sum_l \lambda_{\omega,k,l}{=}1$. Similarly to Eq. (21), we obtain

$$\begin{aligned}\mathcal{J}_{\mathrm{p}} &= \mu_1 \textstyle\sum_{k',k}\left(\sum_\omega f_{\omega,k'} d_{\omega,k}\right)^2 + \mu_2 \sum_{k,l}\left(\sum_\omega f_{\omega,k} h_{\omega,l}\right)^2 \\ &\quad + \mu_3 \textstyle\sum_{k,l}\left(\sum_\omega d_{\omega,k} h_{\omega,l}\right)^2 + \mu_4 \sum_{k,l}\left(\sum_\omega b_{\omega,k} h_{\omega,l}\right)^2 \\ &\leq \mu_1 \textstyle\sum_{k',k,\omega} \dfrac{f^2_{\omega,k'} d^2_{\omega,k}}{\delta_{\omega,k',k}} + \mu_2 \sum_{k,l,\omega} \dfrac{f^2_{\omega,k} h^2_{\omega,l}}{\epsilon_{\omega,k,l}} \\ &\quad + \mu_3 \textstyle\sum_{k,l,\omega} \dfrac{d^2_{\omega,k} h^2_{\omega,l}}{\theta_{\omega,k,l}} + \mu_4 \sum_{k,l,\omega} \dfrac{b^2_{\omega,k} h^2_{\omega,l}}{\lambda_{\omega,k,l}} \\ &\equiv Q_{\mathrm{p}},\end{aligned} \tag{30}$$

where the equality in Eq. (30) holds if and only if the auxiliary variables are set to as follows:

$$\delta_{\omega,k',k} = \frac{f_{\omega,k'} d_{\omega,k}}{\sum_{\omega'} f_{\omega',k'} d_{\omega',k}}, \tag{31}$$

$$\epsilon_{\omega,k,l} = \frac{f_{\omega,k} h_{\omega,l}}{\sum_{\omega'} f_{\omega',k} h_{\omega',l}}, \tag{32}$$

$$\theta_{\omega,k,l} = \frac{d_{\omega,k} h_{\omega,l}}{\sum_{\omega'} d_{\omega',k} h_{\omega',l}}, \tag{33}$$

$$\lambda_{\omega,k,l} = \frac{b_{\omega,k} h_{\omega,l}}{\sum_{\omega'} b_{\omega',k} h_{\omega',l}}. \tag{34}$$

Finally, by using Eq. (21) and Eq. (30), we can define the upper bound function $\mathcal{J}^+$ for $\mathcal{J}$ as

$$\mathcal{J} \leq \mathcal{J}^+ = \textstyle\sum_{\omega,t}(Q_{\log} + r_{\omega,t} + C_{\mathcal{J}}) + Q_{\mathrm{p}}. \tag{35}$$

The update rules for $\mathcal{J}^+$ with respect to each variable are determined by setting the gradient to zero. From $\partial \mathcal{J}^+ / \partial d_{\omega,k} = 0$, we obtain

$$\begin{aligned}&\textstyle\sum_t \left(g_{k,t} - \dfrac{y_{\omega,t}\alpha_{k,\omega,t} \sum_l \beta_{l,\omega,t}\gamma_1}{b_{\omega,k}} \right) \\ &+ 2\mu_1 \textstyle\sum_{k'} \dfrac{f^2_{\omega,k'} d_{\omega,k}}{\delta_{\omega,k',k}} + 2\mu_3 \sum_l \dfrac{d_{\omega,k} h^2_{\omega,l}}{\theta_{\omega,k,l}} \\ &\qquad + 2\mu_4 \textstyle\sum_l \dfrac{b_{\omega,k} h^2_{\omega,l}}{\lambda_{\omega,k,l}} = 0.\end{aligned} \tag{36}$$

By substituting Eqs. (26)-(29) and Eqs. (31)-(34) into Eq. (36), we can rewrite Eq. (36) as

$$\textstyle\sum_t g_{k,t} + v_{\omega,k} + 2\mu_4 \sum_l h_{\omega,l} \sum_{\omega'} b_{\omega',k} h_{\omega',l} = \sum_t \dfrac{y_{\omega,t} g_{k,t}}{r_{\omega,t}}, \tag{37}$$

where $v_{\omega,k}$ is given by

$$\begin{aligned} v_{\omega,k} &= 2\mu_1 \textstyle\sum_{k'} f_{\omega,k'} \sum_{\omega'} f_{\omega',k'} d_{\omega',k'} \\ &\quad + 2\mu_3 \textstyle\sum_l h_{\omega,l} \sum_{\omega'} d_{\omega',k} h_{\omega',l}. \end{aligned} \tag{38}$$

Thanks to the constraints Eqs. (11)-(15), all the terms in Eq. (37) except $v_{\omega,k}$ are nonnegative. In addition, $v_{\omega,k}$ can become both positive and negative values because $d_{\omega,k}$ has a possibility to be both positive and negative. For keeping the nonnegativity of Eq. (37), we divide the update rules of $d_{\omega,k}$ in cases that $v_{\omega,k} \geq 0$ and $v_{\omega,k} < 0$. If $v_{\omega,k} \geq 0$, both sides of Eq. (37) are nonnegative. Then we can obtain the update rule by multiplying both sides of Eq. (37) by $\eta f_{\omega,k} + d_{\omega,k}$, as

$$d_{\omega,k} \leftarrow \frac{(\eta f_{\omega,k} + d_{\omega,k}) \sum_t y_{\omega,t} g_{k,t} r^{-1}_{\omega,t}}{\sum_t g_{k,t} + v_{\omega,k} + 2\mu_4 \sum_l h_{\omega,l} \sum_{\omega'} b_{\omega',k} h_{\omega',l}} - \eta f_{\omega,k} \quad (v_{\omega,k} \geq 0). \tag{39}$$

If $v_{\omega,k} < 0$, we also have the update rule by transposing $v_{\omega,k}$ to the right-hand side and multiplying both sides of Eq. (37) by $\eta f_{\omega,k} + d_{\omega,k}$:

$$d_{\omega,k} \leftarrow \frac{(\eta f_{\omega,k} + d_{\omega,k})\left(\sum_t y_{\omega,t} g_{k,t} r^{-1}_{\omega,t} - v_{\omega,k}\right)}{\sum_t g_{k,t} + 2\mu_4 \sum_l h_{\omega,l} \sum_{\omega'} b_{\omega',k} h_{\omega',l}} - \eta f_{\omega,k} \quad (v_{\omega,k} < 0). \tag{40}$$

Similarly to Eqs. (39) and (40), the update rules of the other variables are obtained as follows:

$$h_{\omega,l} \leftarrow \frac{h_{\omega,l} \sum_t y_{\omega,t} u_{l,t} r^{-1}_{\omega,t}}{\sum_t u_{l,t} + 2\mu_2 \sum_k f_{\omega,k} \sum_{\omega'} f_{\omega',k} h_{\omega',l} + w_{\omega,l} + s_{\omega,l}} \quad (w_{\omega,k} \geq 0), \tag{41}$$

$$h_{\omega,l} \leftarrow \frac{h_{\omega,l} \left(\sum_t y_{\omega,t} u_{l,t} r^{-1}_{\omega,t} - w_{\omega,l}\right)}{\sum_t u_{l,t} + 2\mu_2 \sum_k f_{\omega,k} \sum_{\omega'} f_{\omega',k} h_{\omega',l} + s_{\omega,l}} \quad (w_{\omega,k} < 0), \tag{42}$$

$$g_{k,t} \leftarrow \frac{g_{k,t} \sum_\omega b_{\omega,k} y_{\omega,t} r^{-1}_{\omega,t}}{\sum_\omega b_{\omega,k}}, \tag{43}$$

$$u_{l,t} \leftarrow \frac{u_{l,t} \sum_\omega h_{\omega,l} y_{\omega,t} r^{-1}_{\omega,t}}{\sum_\omega h_{\omega,l}}, \tag{44}$$

where $w_{\omega,l}$ and $s_{\omega,l}$ are given by

$$w_{\omega,l} = 2\mu_3 \textstyle\sum_k d_{\omega,k} \sum_{\omega'} d_{\omega',k} h_{\omega',l}, \tag{45}$$

$$s_{\omega,l} = 2\mu_4 \textstyle\sum_k b_{\omega,k} \sum_{\omega'} b_{\omega',k} h_{\omega',l}. \tag{46}$$

4 Evaluation experiment

4.1 Experimental conditions

To confirm the effectiveness of the proposed algorithm, we compared the conventional method (CSNMF) and new NMF with Eqs. (39)-(44), applying them to the separation task for monaural multiple instrumental sources. In this experiment,

Table 1 Evaluation scores of conventional and proposed methods

Target sound	Other sound	Conventional method			Proposed method		
		SDR	SIR	SAR	SDR	SIR	SAR
Piano	Clarinet	2.4	8.4	4.3	**8.4**	**14.8**	**9.6**
Piano	Trombone	3.1	15.8	3.5	**11.0**	**24.4**	**11.2**
Clarinet	Flute	0.1	1.8	**7.3**	**0.7**	**2.6**	7.2
Clarinet	Trombone	3.2	14.6	3.7	**9.6**	**23.9**	**9.8**
Flute	Piano	5.8	12.8	7.0	**7.0**	**14.9**	**7.9**
Trombone	Clarinet	2.1	12.4	2.8	**4.7**	**19.3**	**4.9**

we four types of real instruments, namely, piano, clarinet, flute, and trombone, as the target sound. These sound sources were separately recorded, and the observed signals $\boldsymbol{Y}$ were produced by mixing two sources selected from four sources with the input SNR of 0 dB. In addition, we used artificial MIDI sounds of the target instruments as supervision for a priori training. The training sounds contain two octave notes that cover all notes of the target signal in the observed signal. The sampling frequency of all signals was 44.1 kHz. The spectrograms were computed using a 92-ms long rectangle window with a 46-ms overlap shift. The number of iterations for the training was 500 and for separation was 400. Moreover, the number of a priori bases was 100, and the number of bases for the matrix $\boldsymbol{H}$ was 30. In this experiment, the wighting parameters were empirically determined.

We used the signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and sources-to-artifacts ratio (SAR) defined in Ref. [5] as the evaluation scores. SDR indicates the quality of separated target sound. SIR indicates the degree of separation between target and other sounds. SAR indicates absence of artificial distortion such as musical noise.

4.2 Experimental results

Table 1 shows the evaluated scores that are the maximum of SDR within 10 trials. Also, Fig. 1 shows an example of the separation result obtained by the conventional and proposed algorithms. As noted in Table 1, the separation performance is increased in the proposed method. The resultant scores with respect to the mixture of clarinet and flute is slightly lower than those of other mixtures because both of the spectral patterns are relatively similar. Figure 1 shows that the target piano sound of the proposed method can be more enhanced than that of conventional CSNMF because the trained bases are fitted by the deformable ability in the proposed method.

5 Conclusions

In this paper, we propose a new advanced supervised NMF that includes the deformable term for the trained spectral bases and constraints for making the bases to fit into the target sound. From the experimental results, it can be confirmed that the proposed method increases the separation performance for the real instruments compared with the conventional method.

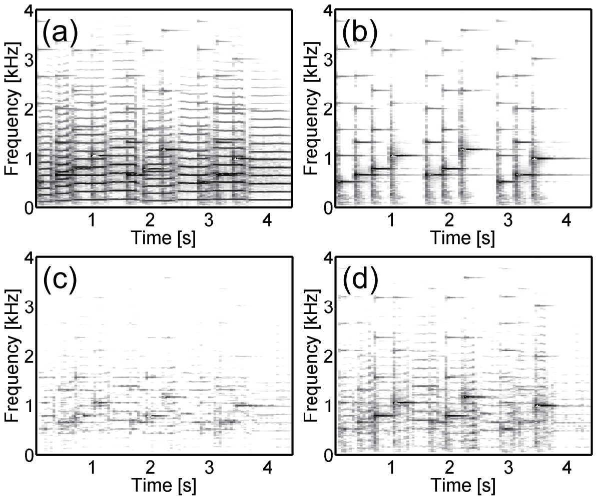

Fig. 1 Spectrograms of (a) observed signal consisting of piano and trombone, (b) oracle signal of the target piano signal, (c) extracted piano signal by conventional method, and (d) extracted piano signal by proposed method.

References

[1] D. D. Lee, H. S. Seung, “Algorithms for non-negative matrix factorization,” *Neural Info. Process. Syst.*, vol.13, pp.556–562, 2001.

[2] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Trans. Audio, Speech and Language Processing*, vol.15, pp.1066–1074, 2007.

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

[4] K. Yagi, Y. Takahashi, H. Saruwatari, K. Shikano, K. Kondo, “Music signal separation by orthogonality and maximum-distance constrained nonnegative matrix factorization with target signal information,” *Proc. Audio Engineering Society 45th International Conference*, 2012.

[5] E. Vincent, R. Gribonval, C. Fevotte. “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech and Language Processing*, vol.14, no.4, pp.1462–1469, 2006.