Audio Source Separation Based on Residual Reprojection

Choongsang Cho, Je Woo Kim, and Sangkeun Lee

This paper describes an audio source separation that is based on nonnegative matrix factorization (NMF) and expectation maximization (EM). For stable and high-performance separation, an effective auxiliary source separation that extracts source residuals and reprojects them onto proper sources is proposed by taking into account an ambiguous region among sources and a source's refinement. Specifically, an additional NMF (model) is designed for the ambiguous region — whose elements are not easily represented by any existing or predefined NMFs of the sources. The residual signal can be extracted by inserting the aforementioned model into the NMF-EM-based audio separation. Then, it is refined by the weighted parameters of the separation and reprojected onto the separated sources. Experimental results demonstrate that the proposed scheme (outlined above) is more stable and outperforms existing algorithms by, on average, 4.4 dB in terms of the source distortion ratio.

Keywords: Audio mixing system, audio separation, expectation maximization, nonnegative matrix factorization, source residual projection.

Manuscript received Nov. 12, 2014; revised Apr. 28, 2015; accepted May 11, 2015.

This work was partially supported by the IT R&D program of MSIP/KEIT (10044569) by the NRF funded by the Ministry of Education, Science and Technology (No. NRF-2014S1A5B6037633 and No. NRF-2014R1A2A1A11049986) and by the Chung-Ang University Research Grant in 2014.

Choongsang Cho (ideafisher@keti.re.kr) and Sangkeun Lee (corresponding author, sangkny@cau.ac.kr) are with the Department of Imaging Engineering, the Graduate School of Advanced Imaging Science, Multimedia & Film at Chung-Ang University, Seoul, Rep. of Korea.

Je Woo Kim (jwkim@keti.re.kr) is with the Department of Multimedia IP Research Center, Korea Electronics Technology Institute, Seongnam, Rep. of Korea.

I. Introduction

Audio source separation from a multichannel mixture is a great research topic. At the same time, audio source separation is challenging since it is usually a mathematically ill-posed problem; one is required to have additional knowledge of the mixing process or source signals to obtain successful separation results [1]–[10]. Independent component analysis has been used for such a separation under the assumption of statistical independence of sources [5].

For the problem of *underdetermined* source separation, probability model–based approaches with generalized Gaussian priors or l_p-norm minimizations have been applied to this problem in an attempt to solve it [6].

Nonnegative matrix factorization (NMF) [11]–[13], which is useful in music transcription, was employed for audio source separation because it is suitable for use with polyphonic musical instruments [10]–[12]. In particular, NMF and expectation maximization (EM) were successfully incorporated to solve mathematically ill-posed source separation problems [8]. Additionally, a model parameter estimation procedure that uses an iterative generalized expectation maximization (GEM) algorithm was proposed with the incorporation of a priori knowledge [9], [14].

However, l_p-norm minimization approaches, NMF-EM-based approaches, and general flexible approaches are not suitable in the case where a mixed audio signal includes an ambiguous area that is not easily represented by any particular sources. Moreover, wrong source separation in an ambiguous area can cause quality degradation of the separated sources.

In this paper, a coarse-to-fine separation structure (namely, residual extraction) and reprojection schemes are proposed

© 2015 ETRI

for stable and efficient NMF-EM-based audio source separation. In the residual extraction step, in consideration of an ambiguous area of sources, an auxiliary NMF is attached to the NMF-EM-based system, and the residual signal is extracted using the inserted additional NMF model. Next, the residual signal is split into two categories — the remaining source components (which have similar characteristics to the original sources) and the rest of the components — using the weighted parameters and NMF-EM-based separation. Then, the remaining source components are reprojected onto the separated sources at the residual extraction step.

This paper is structured as follows. In Section II, the NMF-EM-based audio source separation is numerically explained. The proposed scheme of extracting and reprojecting the residuals is described in Section III. A comparison and analysis of the experimental results is given in Section IV. Finally, Section V concludes the paper with a summary of the proposed algorithm.

II. NMF-EM-Based Audio Source Separation

This section numerically describes the audio source separation using NMF [7]–[9], [11]. NMF is a popular data decomposition technique in the areas of machine learning; image and audio signal processing; and audio source separation [11], [13]. Audio source separation is performed by decomposing the power spectrogram of objects, which can be represented as two nonnegative matrices (matrix **W** for narrow spectral patterns and matrix **H** for the corresponding weights); thus an $F \times N$ audio power spectrogram, **V**, obtained by a short-time Fourier transform (STFT), can be expressed as follows:

$$\mathbf{V} \approx \mathbf{WH}, \tag{1}$$

where **W** and **H** are $F \times K$ and $K \times N$ nonnegative matrices, respectively. F and N indicate frequency bin and time-frequency index dimension, respectively. The decomposed matrices can represent the characteristics of audio objects [7]–[9], [11], and they are commonly used to analyze music characteristics [11]. The factorization is usually performed via cost function minimization as

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \Phi\left(\mathbf{V} \mid \mathbf{WH}\right), \tag{2}$$

where $\Phi(\cdot)$ is a cost function defined as

$$\Phi\left(\mathbf{V} \mid \mathbf{WH}\right) = \sum_{f=1}^{F}\sum_{n=1}^{N} \phi\left(\mathbf{V}_{fn} \mid \left[\mathbf{WH}\right]_{fn}\right), \tag{3}$$

where $\phi(x \mid y)$ is a scalar cost function composed of either a Euclidean distance, a Kullback–Leibler divergence, or an Itakura–Saito divergence. To solve the minimization problem, the EM or maximum likelihood algorithm can be applied using a statistical distribution that assumes a zero mean. In general, a mixed audio signal can be approximated in such a way that the audio sources are multiplied with an audio mixing system and additive noise is added. For example,

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_{fn}, \tag{4}$$

where $\mathbf{x}_{fn} = \left[x_{1,fn}, \ldots, x_{I,fn}\right]^{\mathrm{T}}$, $\mathbf{s}_{fn} = \left[s_{1,fn}, \ldots, s_{J,fn}\right]^{\mathrm{T}}$, $\mathbf{b}_{fn} = \left[b_{1,fn}, \ldots, b_{I,fn}\right]^{\mathrm{T}}$, and $\mathbf{A}_f = \left[a_{ij,f}\right]$ for $\langle i, j\rangle \in I \times J$.

Here, I and J indicate the number of input channels and sources to be separated, respectively. Furthermore, $\mathbf{x}_{fn}$, $\mathbf{A}_f$, and $\mathbf{s}_{fn}$ denote the mixed audio signal, the mixing system, and the mixed sources, respectively. Note that in the case of the instantaneous matrix $\mathbf{A}_{\text{inst}}$, the mixing system $\mathbf{A}_f$ is real-valued and shared between all of the frequency sub-bands; that is, $\mathbf{A}_f = \mathbf{A}_{\text{inst}}$ for all $\mathbf{A}_f, \mathbf{A}_{\text{inst}} \in R^{I\times J}$ [8]; $\mathbf{b}_{fn}$ represents the noisy data, which is assumed to have a Gaussian distribution and covariance $\sum_{\mathbf{b}}$, and is defined as

$$b_{i,fn} \sim N_c\left(0, \sigma_{i,f}^2\right) \text{ with } \Sigma_{\mathbf{b},f} = \mathrm{diag}\left(\left[\sigma_{i,f}^2\right]_i\right). \tag{5}$$

Here, diag(**u**) returns a square matrix with the elements of vector **u** on the main diagonal. In an audio source separation problem, each source (object) is represented as an NMF consisting of k_j-dimensional decomposition ($k_j \geq J$).

$$\begin{aligned} &\left|s_{j,fn}\right|^2 \approx W_j H_j,\ s_{j,fn} = \sum_{l\in k_j} c_{l,fn},\ c_{l,fn} = w_{fl} h_{nl}^{\mathrm{T}}, \\ &W_j = \left[w_{f1}, \ldots, w_{fk_j}\right],\ H_j = \left[h_{n1}, \ldots, h_{nk_j}\right]^{\mathrm{T}}. \end{aligned} \tag{6}$$

The total dimension of the objects is $k = \#\left\{k_1, \ldots, k_J\right\}$.

The EM algorithm is widely adopted to solve the mathematically ill-posed problem and optimize the cost function. We let a complete data set and a parameter set be $Z = (\mathbf{X}, \mathbf{S})$ and $\theta = \left\{\mathbf{A}_f, \mathbf{W}, \mathbf{H}, \Sigma_{\mathbf{b}}\right\}$, respectively. Here, $\mathbf{X} = (\mathbf{x}_{fn})_I$ and $\mathbf{S} = (\mathbf{s}_{fn})_J$. Then, to solve the equation, we use the log-likelihood function [15], which is defined by

$$Q(\Theta, \Theta^{i-1}) = \mathrm{E}\left[\log p(\mathbf{X}, \mathbf{S} \mid \Theta) \mid \mathbf{X}, \Theta^{i-1}\right]. \tag{7}$$

The resulting criterion can be expressed as

$$\Theta^* = \arg\max_{\Theta} Q(\Theta, \Theta^{i-1}), \tag{8}$$

where Θ^{i-1} represents the current parameters used to evaluate expectation $\mathrm{E}[\cdot]$, and Θ represents the new parameters that optimize Q to maximize the expectation. The log-likelihood function [8]–[9] for audio source separation can be expressed as

$$Q(\Theta,\Theta^{i-1})$$
$$= \sum_{fn}\left[\log\left|\Sigma_{\mathbf{b},fn}\right| + (\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{S}_{fn})^H \Sigma_{\mathbf{b}}^{-1}(\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{S}_{fn})\right]$$
$$+ \sum_k\sum_{fn}\log\left(w_{fk}h_{kn}\right) + \frac{\left|x_{fn}\right|^2}{w_{kf}h_{kn}}. \quad (9)$$

The parameters [8]–[9] are simply computed by using the partial derivatives of the log-likelihood with respect to A, W, H, and $\Sigma_{\mathbf{b}}$ as

$$\mathbf{A}_f = \mathbf{R}_{\mathbf{xs},f}\mathbf{R}_{\mathbf{ss},f}^{-1},$$
$$\Sigma_{\mathbf{b},f} = \mathrm{diag}\left[\mathbf{R}_{\mathbf{xx},f} - \mathbf{A}_f\mathbf{R}_{\mathbf{xx},f} - \mathbf{R}_{\mathbf{xs},f}\mathbf{A}_f - \mathbf{A}_f\mathbf{R}_{\mathbf{xx},f}\mathbf{A}_f^H\right], \quad (10)$$
$$w_{fk_j} = \frac{1}{N}\sum_n \frac{u_{k_j,fn}}{h_{k_jn}}, h_{kn} = \frac{1}{F}\sum_f \frac{u_{k_j,fn}}{w_{k_jn}},$$

with

$$\mathbf{R}_{\mathbf{xx},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{x}_{fn}^H, \quad \mathbf{R}_{\mathbf{xs},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{s}_{fn}^H,$$
$$\mathbf{R}_{\mathbf{ss},f} = \frac{1}{N}\sum_n \mathbf{s}_{fn}\mathbf{s}_{fn}^H, \quad u_{k_j,fn} = \left|c_{k_j,fn}\right|^2. \quad (11)$$

III. Two-Step Audio Source Separation: Residual Extraction and Reprojection Scheme

In audio source separation, a mixed audio signal, $\mathbf{x}$, can be described by two concepts according to the type of source combinations, as shown in Fig. 1. When the audio sources (original objects) are disjoint (that is, $\cap_{j=1}^{J}\mathbf{s}_j = \varnothing$), the mixed signal is represented as shown in Fig. 1(a). On the other hand, the signal is represented as shown in Fig. 1(b) when an intersection among the audio sources exists; that is, $\cap_{j=1}^{J}\mathbf{s}_j \neq \varnothing$.

To successfully separate a mixed audio signal, it should be split into each object source. In the case of null intersection among sources, a point in the mixed signal region is exactly matched to a point in a source. However, a point in the mixed signal is mapped into at least two sources when an ambiguous region among sources exists. Thus, a wrong source-separation in the mixed signal region can cause quality degradation in the original sources because a source in the mixed signal can be incorrectly mapped to a different source.

To improve the separation performance in the case where an ambiguous region exists, a coarse-to-fine structure-based audio separation (residual extraction and reprojection steps) is proposed as shown in Fig. 2. To extract a residual signal in an ambiguous region, an auxiliary NMF, $\mathbf{W}_{J+1}\mathbf{H}_{J+1}$, for the residual signal, which has difficulty in belonging to a particular

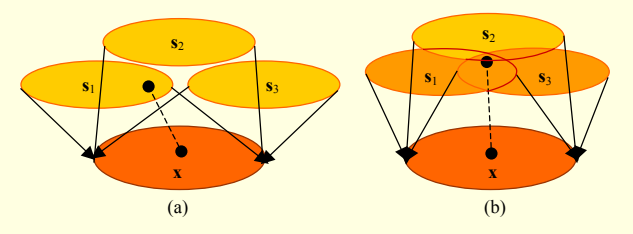


Fig. 1. Concepts to represent audio source mixing: (a) case 1 with null intersection and (b) case 2 with intersection.

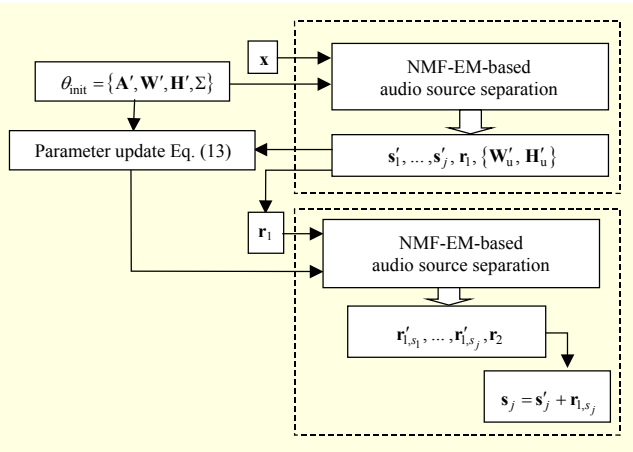


Fig. 2. Residual extraction and reprojection-based separation structure.

single NMF of a source, is considered. By adding a model for the ambiguous area of a mixed signal, the number of objects, which are exchangeable with the sources or channels, for the NMF-EM-based separation is increased and a component of the mixing system is then added by inserting the auxiliary NMF. Then, the initial NMF parameters for the EM-based audio source separation are represented as

$$\mathbf{W}'\mathbf{H}' = \left\{W_1H_1, \ldots, W_JH_J, W_{J+1}H_{J+1}\right\}, \quad (12)$$
$$\mathbf{A}' = \left\{\mathbf{A}, a_{i(J+1)}\right\} = \left[a_{ij}\right] \quad \text{for } \langle i,j\rangle \in I\times(J+1).$$

The residual signal, which is not represented by any particular NMF of the sources, is assumed to be a random signal because an ambiguous region can belong to more than two sources.

In this paper, NMF components in the ambiguous region, $\mathbf{W}_{J+1}$ and $\mathbf{H}_{J+1}$, are generated by a normally distributed pseudorandom method [16]. Specifically, an NMF for an ambiguous region is modeled as a random signal with zero mean and certain variance, and its power spectrogram, $|\mathbf{W}_{J+1}\mathbf{H}_{J+1}|$, is adopted and illustrated (see Fig. 3).

In the first step, which is the residual extraction stage, a mixed audio signal is separated into source signals, $\{\mathbf{s}'_1, \ldots, \mathbf{s}'_J\}$, and a residual signal, $\mathbf{r}_1$, using the NMF-EM-based separation. The separated signals contain the dominant

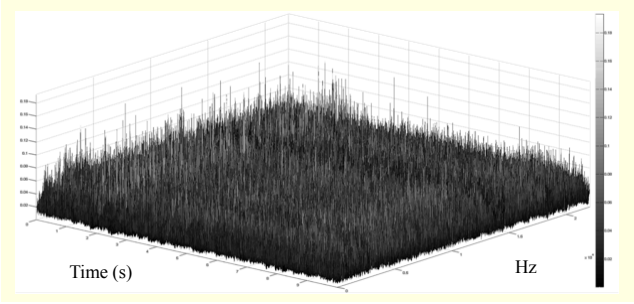


Fig. 3. Power spectrogram of inserted NMF.

components of each original source before mixing, and residual signal $\mathbf{r}_1$ contains a common set of sources that is difficult to split into the exact sources. To separate the residual signal $\mathbf{r}_1$ into exact residuals of the objects, a second separation is performed by the NMF-EM algorithm using the weighted initial parameters, which is called the "source residual projection stage." To consider the estimated characteristics of objects from the mixed audio, the weighted initial parameters, $\mathbf{W}'_{\mathrm{n}}$ and $\mathbf{H}'_{\mathrm{n}}$, are obtained from the weighted sum of the initial parameters, $\mathbf{W}'$ and $\mathbf{H}'$, and their updated parameters, $\mathbf{W}'_{\mathrm{u}}$ and $\mathbf{H}'_{\mathrm{u}}$, in the previous step, as

$$\mathbf{W}'_{\mathrm{n}}\mathbf{H}'_{\mathrm{n}} = \omega_2 \times \left[\omega_1 \left\{\mathbf{W}'\mathbf{H}'\right\} + \left(1-\omega_1\right)\left\{\mathbf{W}'_{\mathrm{u}}\mathbf{H}'_{\mathrm{u}}\right\}\right], \quad (13)$$

where ω_1 denotes the weighting value to combine the initial and updated parameters. In (13), ω_2, which is meant for consideration of the variation of the input signal, is defined by the ratio of the absolute power spectrograms of the mixed and residual signals, since the input for the second step is residual, as

$$\omega_2 = \sqrt{\frac{\frac{1}{F\times N}\sum_{f,n}\left|X_{f,n}\right|}{\frac{1}{F\times N}\sum_{f,n}\left|R_{1_{f,n}}\right|}}, \quad (14)$$

where X_{fn} and $R_{1_{f,n}}$ denote the power spectrograms of the mixed and residual signals, respectively.

Finally, the mixed residual $\mathbf{r}_1$ is refined using the NMF-EM-based separation into the remaining signal of the audio sources, $\{\mathbf{r}_{1,s_1}, \ldots, \mathbf{r}_{1,s_J}\}$ (it can be interpreted as real source residue), and (mixed source region) residual signal ($\mathbf{r}_2$). In other words, the remaining signals are sifted signals of the source residues by the second separation with parameters $\mathbf{W}'_{\mathrm{n}}\mathbf{H}'_{\mathrm{n}}$.

"Remaining signal" means a signal that is highly likely to belong to a source, in terms of statistical probability. To improve the separation performance by considering the refined residual signal, the sifted signals, $\mathbf{r}_{1,s_j}$, are projected onto the audio sources, $\mathbf{s}'_j$, separated during the residual extraction step as

$$\mathbf{s}_j = \mathbf{s}'_j + \mathbf{r}_{1,s_j}. \quad (15)$$

IV. Experiment

For the performance evaluation, we compare the proposed scheme with previous approaches — l_p-norm minimization approach (l_pNM) [6], NMF-EM-based approach (NMF-EM) [8], [17], and a general flexible framework–based algorithm (GFF) [9], [18]. For the test set, five 10 s clips and one 30 s clip from a variety of Korean pop (K-pop) songs, which were commercially recorded at 44.1 kHz with 16-bit resolution and stereo for object audio services in South Korea, are applied in the separation. The original objects of the given music files are independently provided, and the mixed signals from three objects — vocal, drum, and keyboard — of each K-pop song are generated by an instantaneous mixing system using a 2×3 matrix whose elements are in the range [0, 1]. Specifically, the fourth and fifth files are different sections of the same content. The fourth file has weak artificial effects from a vocoder and virtual sound technology [19], which are widely used for commercial music, and the fifth one has strong artificial effects for approximately 30% of the file length.

For the quantitative evaluation, the source distortion ratio (SDR) and source-to-interference ratio (SIR) [20]–[21] of the separated result are measured against the original files. To generate the initial parameters for the given system, the original objects of the contents are applied to the NMF generation on the basis of the EM approach with 1,000 iterations; k_j for the objects is set to {6; 4; 4}. All of the parameters used in this test are summarized in Table 1. To be more specific, the elements of the initial mixing system are set equal to the same value ($1/J$), and the mixed signal is separated by the proposed approach with 20 iterations and 10 iterations for the two stages. Additionally, weight ω_1 is set to 0.9 for the second stage. Note that the weight for the updated parameters is lower than that for the initial parameters because the updated parameters can have less accurate characteristics for objects that are estimated from

Table 1. Parameter settings for experiment.

Parameters	Set value
Vocal NMF order	6
Drum NMF order	4
Keyboard NMF order	4
Common set NMF order	4
STFT window size	1,024
Iteration # of EM for initial model	1,000

Table 2. SDR scores of proposed method and existing algorithms (dB).

Object	Term	1	2	3	4	5	6	Avg.
Vocal	l_pNM [5]	0.2	3.8	0.7	0.1	1.7	0.2	1.1
	NMF-EM [7]	7.3	11.3	10.2	5.4	–4.3	6.6	6.1
	GFF [8]	2.2	3.8	3.2	3.8	2.2	3.2	3.1
	Proposed	9.7	10.0	11.1	9.4	6.8	9.8	9.5
		10.2	11.8	11.2	10.6	9.3	9.9	10.5
Drum	l_pNM [5]	–1.8	0.2	–0.6	–3.7	1.2	–1.8	–1.1
	NMF-EM [7]	7.0	5.8	6.8	4.4	–5.6	1.0	3.2
	GFF [8]	4.7	4.3	5.6	6.8	1.2	2.6	4.2
	Proposed	10.0	9.5	8.6	6.7	2.8	5.9	7.3
		11.1	9.8	8.7	9.8	5.4	7.4	8.7
Keyboard	l_pNM [5]	–0.4	–3.2	–1.6	–8.8	–2.8	–0.4	–2.9
	NMF-EM [7]	0.7	1.7	1.6	–0.3	–3.2	0.5	0.2
	GFF [8]	–3.6	–3.9	–6.3	–6.2	–4.6	–4.9	–4.9
	Proposed	0.9	0.9	1.5	0.5	1.9	0.1	1.0
		2.9	3.8	2.4	2.0	2.4	1.3	2.5

Table 3. SIR scores of proposed method and existing algorithms (dB).

Object	Term	1	2	3	4	5	6	Avg.
Vocal	l_pNM [5]	–4.6	6.2	–0.7	0.9	2.0	–4.6	–0.1
	NMF-EM [7]	9.4	15.5	13.0	11.4	–4.7	10.3	9.2
	GFF [8]	1.0	6.6	4.6	5.9	2.9	3.5	4.1
	Proposed	15.9	16.8	16.3	17.1	14.1	12.4	15.4
		15.2	17.6	16.3	17.0	16.5	12.4	15.8
Drum	l_pNM [5]	–1.8	1.4	–5.7	–2.2	5.2	–1.8	–0.8
	NMF-EM [7]	14.0	8.3	17.7	17.0	9.8	13.4	13.4
	GFF [8]	8.7	8.1	8.5	14.1	4.8	7.5	8.6
	Proposed	23.9	23.5	18.1	20.7	12.0	17.0	19.2
		22.8	23.0	17.8	20.6	11.0	16.6	18.6
Keyboard	l_pNM [5]	5.9	2.3	2.7	0.6	–8.0	5.9	1.6
	NMF-EM [7]	7.1	3.4	4.1	1.0	–4.0	6.5	3.0
	GFF [8]	–0.9	–3.2	–3.8	–5.5	–8.8	–3.1	–4.2
	Proposed	9.4	5.0	4.4	6.9	8.5	7.0	6.9
		8.7	6.1	4.4	6.8	9.4	6.2	6.9

the mixed audio at the first stage.

To compare SDR and SIR scores, the separation performances of the proposed algorithm for the three objects are shown in Tables 2 and 3, respectively. An SDR score is a measure that is used to evaluate the spatial distortion, interference, and artifacts at the initial equal weights, and an SIR score evaluates relative amounts of interference errors between original and recovered signals [20]–[21]. Higher scores indicate better separation performance. As shown in the SDR score comparison of the existing and proposed schemes, the vocal and drum objects show better separation performance than the keyboard object. In the case of the vocal object, the NMF-EM approach has a higher separation performance than the l_pNM and GFF approaches for the five files except the fifth file. The average SDR of the NMF-EM approach is approximately 8.2 dB except for the fifth file. The average SDR score including the entire contents is approximately 6.1 dB. The difference comes from the lower separation performance in the fifth file, as it contains strong artificial effects. In contrast, the average SDR of the proposed method is approximately 10.7 dB and 10.5 dB without and with the fifth content, respectively. The proposed method shows higher scores than the existing schemes, and it exhibits a stable separation performance even in the presence of strong artificial effects, regardless of content. Similarly, the proposed scheme outperforms the existing schemes with a stable performance in the drum and keyboard objects. The proposed residual reprojection scheme shows an average SDR increase in the vocal, drum, and keyboard objects by approximately 1.0 dB, 1.4 dB, and 1.5 dB, respectively. In the SIR score of vocal, the NMF-EM approach has the higher performance than the l_pNM and GFF approaches in the five files except the fifth file. Moreover, in the drum and keyboard objects, the NMF-EM outperforms the l_pNM and GFF schemes in all the test files. The average SIR scores of the NMF-EM approach give the best performance among the existing algorithms. The proposed method produces higher SIR scores for the three objects than the existing methods. When it is compared to the NMF-EM approach, the scores for the three objects are increased by approximately 6.6 dB, 5.2 dB, and 3.9 dB, respectively. It is noted that the residual extraction in the proposed algorithm is performed to refine the residues. Therefore, to analyze and evaluate the characteristics of the extracted residues, the cross correlation between sources is calculated by

$$c_s = \frac{\mathrm{corr}\left(|s_1|,|s_2|\right)+\mathrm{corr}\left(|s_1|,|s_3|\right)+\mathrm{corr}\left(|s_2|,|s_3|\right)}{3}, \quad (16)$$

where s_1, s_2, and s_3 indicate the power spectrograms of the sources and $\mathrm{corr}(\cdot)$ is the cross-correlation operator. Similarly, the cross-correlation between the residue and the source is measured as

$$c_{r,i} = \frac{\mathrm{corr}\left(|r_i|,|s_1|\right)+\mathrm{corr}\left(|r_i|,|s_2|\right)+\mathrm{corr}\left(|r_i|,|s_3|\right)}{3} \text{ for } i=1,2, \quad (17)$$

http://dx.doi.org/10.4218/etrij.15.0114.1311

Table 4. Cross-correlation between sources and residuals.

File #	1	2	3	4	5	6	Avg.
c_s	0.16	0.13	0.20	0.11	0.10	0.18	0.15
$c_{r,1}$	0.48	0.36	0.49	0.39	0.35	0.26	0.39
$c_{r,2}$	0.32	0.24	0.35	0.25	0.22	0.12	0.25

where r_i indicates the residual power spectrogram during the residual extraction ($i = 1$) and projection ($i = 2$) steps. Table 4 shows that cross-correlation $c_{r,1}$ is much higher than cross-correlation c_s. In other words, the sources have a non-zero correlation relative to each other in the case when the proposed concept is assumed for audio source mixing, and a residual signal from a high-correlation region between sources is extracted during the residual extraction step. Cross-correlation $c_{r,2}$ decreases during the residual projection stage compared with $c_{r,1}$, which indicates that the level of uncertainty in the ambiguous region decreases because the inherent signal is reprojected from the mixed residue onto each source.

It is noted that a source code is available at http://mmc.cau.ac.kr/publications/publications.php.

V. Conclusion

This paper has presented an NMF-EM-based audio source separation scheme using residual extraction and reprojection. For this, a coarse-to-fine separation structure is used to consider the ambiguous area in sources. In particular, a residual signal was extracted by inserting an auxiliary NMF to represent a mixed signal area that is difficult to be represented by any particular NMF of the sources. Then, the residual signal was reprojected onto the separated sources. The experimental results for real commercial contents showed that the proposed audio source separation scheme could provide much higher performance than the state-of-the-art approaches. Additionally, the proposed method produced much more stable results even in a content generated with artificial sound effects. However, the proposed coarse-to-fine source separation structure may increase the computational complexity compared to the existing NMF-based approach when original sources have more correlations. Nevertheless, we believe that the proposed scheme can be a useful tool for audio source separation using the NMF-EM algorithm.

References

[1] H. Attias, "New EM Algorithm for Source Separation and Deconvolution with a Microphone Array," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Hong Kong, China, Apr. 6–10, 2003, pp. 297–300.

[2] C.J. Chun and H.K. Kim, "Sound Source Separation Using Interaural Intensity Difference in Real Environments," *Proc. AES Convention*, Oct. 2013.

[3] N.J. Bryan and G.J. Mysore, "Interactive Refinement of Supervised and Semi-supervised Sound Source Separation Estimates," *IEEE Inter. Conf. Acoustics, Speech, Signal Process.*, Vancouver, Canada, May 26–31, 2013, pp. 883–887.

[4] C. Fevotte and C. Doncarli, "Two Contributions to Blind Source Separation Using Time-Frequency Distributions," *IEEE Signal Process. Lett.*, vol. 11, no. 3, Mar. 2004, pp. 386–389.

[5] G-S. Fu et al., "Blind Source Separation by Entropy Rate Minimization," *IEEE Trans. Signal Process.*, vol. 62, no. 16, June 2014, pp. 4245–4255.

[6] E. Vincent, "Complex Nonconvex l_p Norm Minimization for Underdetermined Source Separation," *Int. Conf. Ind. Compon. Anal.*, London, UK, Sept. 9–12, 2007, pp. 430–437.

[7] P. Smragdis et al., "Static and Dynamic Source Separation Using Nonnegative Factorizations: A Unified View," *IEEE Signal Process. Mag.*, vol. 31, no. 3, May 2014, pp. 66–75.

[8] A. Ozerov and C. Fevotte, "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, Mar. 2010, pp. 550–563.

[9] A. Ozerov, E. Vincent, and F. Bimbot, "A General Flexible Framework for the Handling of Prior Information in Audio Source Separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, May 2012, pp. 1118–1133.

[10] P. Smaragdis, "Convolutive Speech Bases and Their Application to Supervised Speech Separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, Jan. 2007, pp. 1–12.

[11] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Comput.*, vol. 21, no. 3, Mar. 2009, pp. 793–830.

[12] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, Mar. 2007, pp. 1066–1074.

[13] D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Nonnegative Matrix Factorization," *Nature*, vol. 401, Oct. 1999, pp. 788–791.

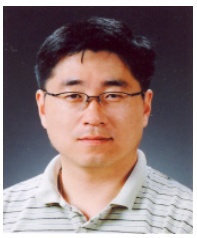
[14] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via EM Algorithm," *J. Royal Statistic Soc. Series B (Methodological)*, vol. 39, no. 1, 1977, pp. 1–38.

[15] T.K. Moon, "*Mathematical Methods and Algorithms for Signal Processing*," NJ, USA: Prentice Hall, 2009.

[16] C. Moler, "*Numerical Computing with MATLAB*," Electronic

Edition, The Mathworks: Natick, MA, USA, 2004.
[17] Example Web Page. Accessed Nov. 11, 2014. http://www.irisa.fr/metiss/ozerov/demos.html#ieeetaslp09
[18] A. Ozerov, E. Vincent, and F. Bimbot, *Flexible Audio Source Separation Toolbox (FASST) Version 1.0 User Guide*. Accessed Nov. 11, 2014. http://bass-db.gforge.inria.fr/fasst/FASST_UserGuide_v1.pdf
[19] R. Bianchini and A. Cipriani, "*Virtual Sound: Sound Synthesis and Signal Processing-Theory and Practice with Csound*," Rome, Italy: ComTempo, 2000.
[20] E. Vincent, R. Gribonval, and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, July 2006, pp. 1462–1469.
[21] E. Vincent et al., "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," *Int. Conf. Independent Compon. Anal. Signal Separation*, London, UK, Sept. 9–12, 2007, pp. 552–559.

Choongsang Cho received his BS degree in electronic engineering from Suwon University, Rep. of Korea, in 2006 and his MS degree in information and communications from Gwangju Institute of Science and Technology, Rep. of Korea. Since 2008, he has been working as a researcher at Multimedia IP Research Center, Korea Electronics Technology Institute, Seongnam, Rep. of Korea. Currently, he is pursuing his PhD degree in imaging engineering at the Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University, Seoul, Rep. of Korea. His research interests include numerical signal processing, audio separation, digital holograms, and image segmentation.



Je Woo Kim received his BS and MS degrees in control & instrumentation engineering from the University of Seoul, Rep. of Korea, in 1997 and 1999, respectively. In 1999, he joined the Korea Electronics Technology Institute, Seongnam, Rep. of Korea, where he was involved in the development of video codecs, video transcoders, and multi-view video systems. He is currently a managerial researcher at the Multimedia IP Research Center, Korea Electronics Technology Institute. His research interests include audio-visual codecs and their applications, in particular UHD systems.

Sangkeun Lee received his BS and MS degrees in electronic engineering from Chung-Ang University, Seoul, Rep. of Korea, in 1996 and 1999, respectively. He received his PhD degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, USA, in 2003. He is an associate professor at the Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University. From 2003 to 2008, he was a staff research engineer with the Digital Media Solutions Lab, Samsung Information Systems America, Irvine, CA, USA, where he was involved in the development of video processing and enhancement algorithms (DNIe) for Samsung's HDTV. His current research and development interests include computer vision, digital video, and image processing; especially augmented reality, video analysis/synthesis, denoising, compression for HDTV and multimedia applications, and CMOS image sensors. He is a senior member of the IEEE.