Residual Drum Sound Estimation for RPCA Singing Voice Extraction

Shiori Mikami, Arata Kawamura and Youji Iiguni
Osaka University, Osaka, Japan
E-mail: mikami@sip.sys.es.osaka-u.ac.jp Tel: +81-06-6850-6580

***Abstract*—In this paper, we propose an efficient singing voice extraction method from a music sound consisting of a singing voice and a drum sound. The proposed method is based on robust principal component analysis (RPCA) which is a technique to separate a given matrix into a sparse matrix and a low rank matrix. Dealing with a spectrogram of the music sound as the given matrix, RPCA gives an extracted singing voice spectrogram as a sparse spectrogram. We improve the capability of RPCA method by removing a residual drum sound from the sparse spectrogram. The residual drum sound is estimated as spectral components which repeatedly arise in the sparse spectrogram. Removing the estimated residual drum sound spectrogram from the sparse spectrogram gives an improved singing voice spectrogram. Simulation results show that the proposed method can improve 10dB of GNSDR as compared with the conventional RPCA singing voice extraction method.**

I. Introduction

Singing voice extraction from a music sound including a singing voice and accompaniment sounds is an interesting topic due to its various applications such as lyric recognition [1],[2], singer identification [3], music touch-up [4],[5], active music listening [6], and so on. Many conventional singing voice extraction methods use spectrogram of a music signal which represents a spectral power of the music signal in time-frequency domain.

Some methods focus on temporal/spectral continuities of an accompaniment sound and sparsity of a singing voice [7],[8]. In this method, the continuities with time or frequency of music spectrogram are evaluated in each time-frequency bin. Tachibana et al. proposed singing voice extraction using harmonic/percussive sound separation (HPSS) based on isotropic nature of harmonic/percussive sound [7]. Fitzgerald et al. apply different median filters to the spectrogram along the time and frequency direction for evaluating the continuity [8]. Both methods are assumed that the harmonic sounds are stable in several analysis frames and the percussive sounds are widely spread in the frequency domain.

Another approach is exploiting the repetition of accompaniment sound where this approach inherently does not rely on the spectral continuities. Nonnegative matrix factorization (NMF) has been used for separating an observed spectrogram into multiple singing voice components and accompaniment components by representing the observed spectrogram as a basis matrix and an activation matrix [9]. Using NMF, it is difficult to determine an appropriate number of basis and select bases belonging to the singing voice. On the other hand, robust principle component analysis (RPCA) is also used for singing voice extraction, where RPCA can directly separate the spectrogram into two spectrograms, i.e., a singing voice spectrogram and an accompaniment spectrogram. RPCA is a technique to separate the sum of a low rank matrix and a sparse matrix into each other [10]. Under the assumption that an accompaniment spectrum repeatedly arises and a singing voice spectrum is sparse, RPCA can extract the singing voice spectrogram as a sparse matrix from an observed spectrogram which consists of the singing voice and accompaniment spectrograms [11]. Here, the repeated accompaniment spectrum is separated as the low rank matrix. Ikemiya et al. proposed a combined method of RPCA with F0 estimation to extract the singing voice and its pitch simultaneously [12]. Unfortunately, an accompaniment spectrum often becomes sparse while it repeatedly arises. For example, attenuation parts of a drum sound has such characteristics, and they are separated as a singing voice spectrum.

We focus on a RPCA-based method which do not rely on the continuity of accompaniment sound or the number of basis. The aim of this study is to remove the residual drum sound from the estimated singing voice spectrogram obtained by RPCA. We propose a residual drum sound estimation (RDSE) which is also based on the repetition of the drum spectrum. We gather the repeated residual drum sound spectrum, and estimate a representative spectrum which is obtained by taking the median value of the gathered drum sound spectra. Putting the representative spectrum on each drum sound position, we have a residual drum sound spectrogram. Finally, subtracting the residual drum sound spectrogram from the RPCA sparse spectrogram gives an improved singing voice spectrogram. Simulation results show that the proposed method can improve the performance of singing voice extraction as compared with the conventional RPCA singing voice extraction method.

II. RPCA singing voice extraction

Let $x(n)$ be a segmented and windowed music signal at time n $(0 \leq n < N)$ given as

$$x(n) = v(n) + h(n) + d(n), \tag{1}$$

where $v(n)$, $h(n)$, and $d(n)$ denote singing voice, harmonic accompaniment sound, and drum sound, respectively. Taking the short-time Fourier transform (STFT) of (1), we have

$$X_l(k) = V_l(k) + H_l(k) + D_l(k), \tag{2}$$

where l and k denote the frame index $(0 \leq l < L)$ and the frequency index $(0 \leq k < N)$, respectively. We define the $(N/2+1) \times L$ matrix $\mathbf{X}$ given as

$$\mathbf{X} = \begin{bmatrix} |X_0(0)|^2 & \cdots & |X_{L-1}(0)|^2 \\ \vdots & \ddots & \vdots \\ |X_0(N/2+1)|^2 & \cdots & |X_{L-1}(N/2+1)|^2 \end{bmatrix}, \quad (3)$$

where $|\cdot|$ denotes the absolute value. When representing the elements of $\mathbf{X}$ as a brightness, $\mathbf{X}$ is often called as the spectrogram. We additionally define $(N/2+1) \times L$ matrices $\mathbf{V}$, $\mathbf{H}$, $\mathbf{D}$ so that each (l,k) elements are expressed as $|V_l(k)|^2$, $|H_l(k)|^2$, $|D_l(k)|^2$, respectively. Also, $\angle\mathbf{X}$ is defined as a matrix of the phase spectrogram whose each element is $\exp(j\angle X_l(k))$, where $j = \sqrt{-1}$ and $\angle\{\cdot\}$ denotes the phase spectrum. We assume that $\mathbf{X} = \mathbf{V} + \mathbf{H} + \mathbf{D}$. When $\mathbf{V}$ is a sparse matrix and both $\mathbf{H}$ and $\mathbf{D}$ are low rank matrices, RPCA is useful to extract $\mathbf{V}$ from $\mathbf{X}$ [11].

RPCA works for satisfying the following condition [10].

$$\text{minimize} \quad ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1 \quad (4)$$
$$\text{subject to} \quad \mathbf{L} + \mathbf{S} = \mathbf{X}, \quad (5)$$

where $\mathbf{L}$ and $\mathbf{S}$ represent the low rank matrix and the sparse matrix, respectively. Also, $||\cdot||_*$ represents nuclear norm and $||\cdot||_1$ represents L_1 norm. The parameter λ is a nonnegative value to control the sparsity of $\mathbf{S}$. In the literature [10], $\lambda = 1/\sqrt{\max(I,J)}$ is recommended when $\mathbf{X} \in \mathbb{R}^{I \times J}$. To solve the optimization problem (4), an inexact version of the augmented Lagrange multiplier (ALM) algorithm has been proposed in the literature [13]. The matrix $\mathbf{S}$ obtained by RPCA is corresponding to the estimated singing voice spectrogram.

III. RESIDUAL DRUM SOUND ESTIMATION

Actually, it is difficult to perfectly extract $\mathbf{V}$ as $\mathbf{S}$ by using RPCA, and $\mathbf{S}$ often includes a residual background sound e.g. drum, guitar, piano, etc. Especially, a residual drum sound is uncomfortable due to repeatedly arising. Hence, we focus on removal of the residual drum sound.

We applied STFT with $N = 1024$ to the drum sound of 4 seconds to generate $\mathbf{D}$, where the drum sound is a snare drum sound sampled at 16kHz taken from RWC Music Database [15]. Then we performed RPCA to $\mathbf{D}$ with $\lambda = 1/\sqrt{\max(N/2+1, L)}$. The result is shown in Fig. 1, where $\mathbf{D_L}$ and $\mathbf{D_S}$ denote the low rank spectrogram and the sparse spectrogram, respectively. Since $\mathbf{D}$ has repeated structure, it should be separated as a low rank. Indeed, $\mathbf{D_L}$ has power spreading over the frequency. However, $\mathbf{D_S}$ has large power in the narrow band that has both characteristics of sparse and low rank. Although the separation ratio can be adjusted by λ, singing voice extraction performance may be impaired when putting λ so that $\mathbf{D_S}$ to $\mathbf{0}$. Hence, we remove $\mathbf{D_S}$ after the RPCA processing.

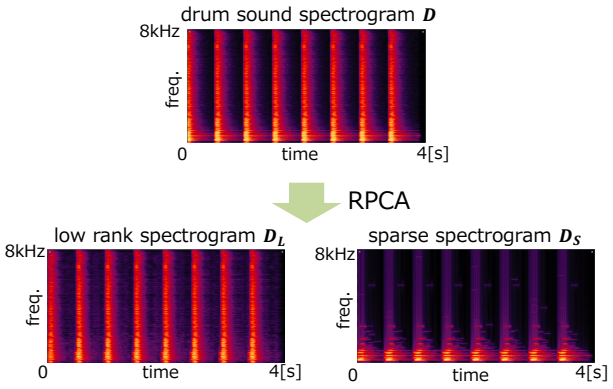


Fig. 1: RPCA result for drum sound.

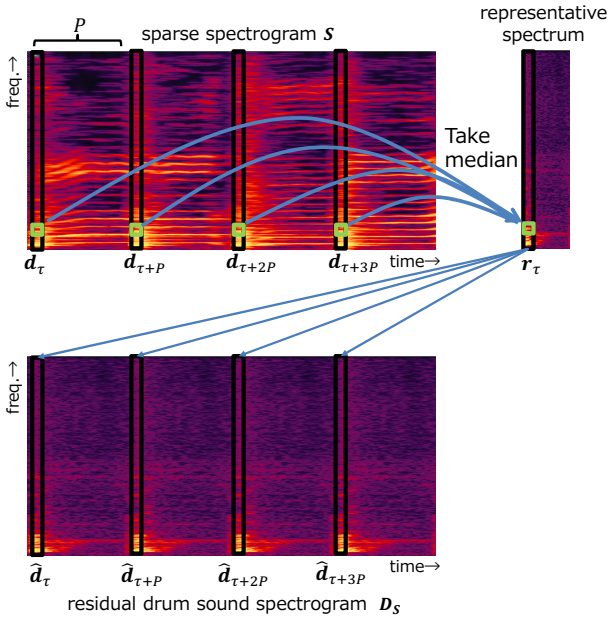


Fig. 2: Procedure of RDSE. Taking median value of $\hat{\mathbf{d}}_{\tau+mP}$ to each element for all m and generating representative spectrum $\mathbf{r}_\tau$. Then, creating $\hat{\mathbf{D}}_\mathbf{S}$ by arranging $\mathbf{r}_\tau$ in $(\tau + Pm)$-th frame.

We assume

$$\mathbf{S} = \mathbf{V} + \mathbf{D_S}, \quad (6)$$
$$\mathbf{L} = \mathbf{H} + \mathbf{D_L}, \quad (7)$$

where

$$\mathbf{D} = \mathbf{D_S} + \mathbf{D_L}. \quad (8)$$

We call $\mathbf{D_S}$ as the residual drum sound spectrogram.

We represent $\mathbf{S}$ and $\mathbf{D_S}$ with column vectors as

$$\mathbf{S} = [\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_{L-1}], \quad (9)$$
$$\mathbf{D_S} = [\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{L-1}], \quad (10)$$

where $\mathbf{s}_l$ and $\mathbf{d}_l$ $(0 \leq l < L)$ are $(N/2+1) \times 1$ vectors, respectively. We assume that the drum sound repeatedly occurs

at the same interval in the processing area that is L frames. It implies that

$$\mathbf{d}_\tau = \mathbf{d}_{\tau+mP}, \tag{11}$$

where P denotes the number of frames corresponding to the single interval, τ $(I \leq \tau < I+J)$ denotes a frame index supporting the first drum sound in the processing area. Here, I denotes the first frame index of the first drum sound, and J denotes the number of frames corresponding to the duration of the single drum sound. The nonnegative integer m must satisfy

$$m < \frac{L-I-J+1}{P}. \tag{12}$$

Since the columns of $\mathbf{V}$ is sparse and varied, we estimate the representative drum sound spectrum as a median spectrum given as

$$\mathbf{r}_\tau = \mathrm{Med}_m\{S_{\tau+mP}\} \tag{13}$$

where $\mathrm{Med}\{\cdot\}$ is an operator to take an element-wise median value of vectors.

We use $\mathbf{r}_\tau$ to obtain the estimated residual drum sound spectrogram $\hat{\mathbf{D}}_\mathbf{S}$ given as

$$\hat{\mathbf{D}}_\mathbf{S} = [\hat{\mathbf{d}}_0,\ \hat{\mathbf{d}}_1,\ \ldots,\ \hat{\mathbf{d}}_{L-1}], \tag{14}$$

$$\hat{\mathbf{d}}_l = \begin{cases} \mathbf{r}_\tau, & l = \tau + mP \\ 0, & \text{otherwise} \end{cases}. \tag{15}$$

We call the procedure from (13) to (15) as RDSE. The estimated singing voice spectrogram $\hat{\mathbf{V}}$ is obtained as

$$\hat{\mathbf{V}} = \mathbf{S} - \hat{\mathbf{D}}_\mathbf{S}. \tag{16}$$

The block diagram of the proposed method is shown in Fig. 3, where x denotes a whole music signal to be processed. The signal x is segmented into L frames by a window function, where the segmented signal of the single frame provides a single column of $\mathbf{X}$. RPCA is applied to $\mathbf{X}$, we have $\mathbf{S}$ as a pre-estimated singing voice spectrogram. Applying RDSE to $\mathbf{S}$ gives a residual drum sound spectrogram as $\hat{\mathbf{D}}_\mathbf{S}$. We have an improved singing voice spectrogram $\hat{\mathbf{V}}$ by subtracting $\hat{\mathbf{D}}_\mathbf{S}$ from $\mathbf{S}$. Finally, inverse-STFT is applied to each column of the matrix obtained as the Hadamard product of $\hat{\mathbf{V}}$ and $\angle\mathbf{X}$, and we have the estimated singing voice $\hat{v}$ in time domain.

IV. Simulation

We carried out singing voice extraction simulations to confirm the capability of the proposed method.

A. Dataset and condition

We used 50 female and 50 male singing voices of Chinese pop song taken from MIR-1K dataset [14]. Here, a duration of the singing voices are 4 to 13 seconds and the sampling rate is 16kHz. Most of singers are amateurs who do not have professional music training. Also, we used 5 kinds of drum sounds as snare drum, bass drum, high tom, closed hi-hat cymbal, and crash cymbal taken from the instrument database

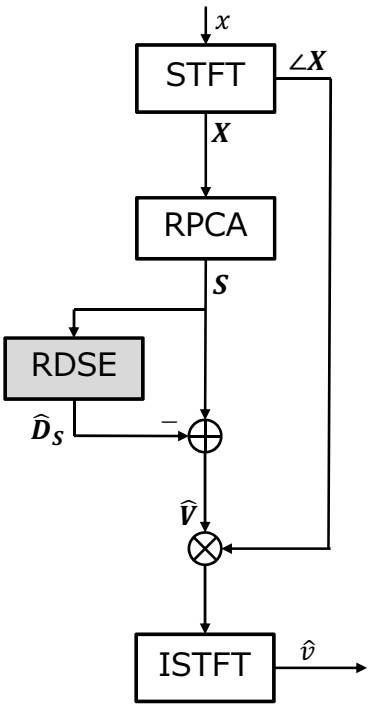


Fig. 3: Block diagram of proposed method

of the RWC Music Database [15]. We added the drum sound to the singing voice for every 8192 samples which is once per 16 frames in $N = 1024$, where the observed signal is generated as 6dB of SNR. We have prepared the observed signals of the 500 clips.

To calculate $\mathbf{X}$, we used Hanning window of $N = 1024$, and the frame shift is $N/2$. As suggested in literatures [11], [10], we put $\lambda = 1/\sqrt{\max(N/2+1, L)}$. We assumed that the position of the drum sound is known in the following simulations.

B. Evaluation criteria

As the evaluation criteria, we used signal to noise ratio (SNR) and spectral distance (SD) given as

$$\mathrm{SNR} = 10\log\frac{\sum_{q=0}^{Q-1} v^2(q)}{\sum_{q=0}^{Q-1}(v(q)-\hat{v}(q))^2}, \tag{17}$$

$$\mathrm{SD} = \sum_{l=0}^{L-1}\frac{\sum_{k=1}^{N-1}(V_l(k)-\hat{V}_l(k))^2}{N}, \tag{18}$$

where $v(q)$ represents the clean singing voice at time q, $\hat{V}_l(k)$ is the STFT of $\hat{v}(q)$, and Q denotes the length of $v(q)$. Additionally, we used a global normalized source to distortion ratio (GNSDR) given as [12],[11]

$$\mathrm{GNSDR} = \frac{\sum_{m=1}^{M} l_m \mathrm{NSDR}(\hat{v}_k, v_k, x_k)}{\sum_{m=1}^{M} l_m}, \tag{19}$$

$$\mathrm{NSDR}(\hat{v}, v, x) = \mathrm{SDR}(\hat{v}, v) - \mathrm{SDR}(x, v), \tag{20}$$

TABLE I: Evaluation result of SNR, SD, and GNSDR. The averaged value (av.) and standard deviation (s.d.) for SNR and SD are shown. Welch's t-test was conducted on the average, and the items with a significant difference at the 5% significance level are shown in bold.

		SNR		SD		GNSDR	
		RPCA	prop.	RPCA	prop.	RPCA	prop.
Bass Drum	av.	**6.59**	**15.55**	**0.56**	**0.09**	0.58	11.33
	s.d.	0.64	3.91	0.23	0.11		
Snare Drum	av.	**7.20**	**14.66**	**0.42**	**0.10**	1.16	10.12
	s.d.	0.48	3.34	0.17	0.12		
High Tom	av.	**6.40**	**15.10**	**0.51**	**0.10**	0.33	10.93
	s.d.	0.29	3.80	0.21	0.12		
Closed Hi-Hat	av.	**14.83**	**16.11**	0.08	0.08	9.40	11.86
	s.d.	1.68	3.65	0.04	0.11		
Crash Cymbal	av.	**14.92**	**16.73**	0.08	0.07	9.46	12.95
	s.d.	1.32	3.97	0.04	0.11		

where m is the clip number $(1 \leq m \leq M)$ and l_m is the length of the m-th clip. SDR is calculated as given

$$\mathrm{SDR}(\hat{v}, v) = 10 \log_{10} \frac{||v||^2}{||e_{\mathrm{interf}} + e_{\mathrm{artif}}||^2}, \qquad (21)$$

where $\hat{v}(q) = v(q) + e_{\mathrm{interf}}(q) + e_{\mathrm{noise}}(q) + e_{\mathrm{artif}}(q)$, $e_{\mathrm{interf}}(q)$, $e_{\mathrm{noise}}(q)$ and $e_{\mathrm{artif}}(q)$ denotes the interference of the unwanted sources, perturbing noise, and artifacts in the separated signals.

SNR and GNSDR show better separation performance with larger values, and SD shows better with smaller values.

C. Simulation result and discussion

Simulation results are shown in Table I, where av. and s.d. denote the averaged value and standard deviation of the results, respectively. We applied Welch's t-test on the averaged values of SNR and SD, and a significant difference at the significance level of 5% is observed for all drum sounds at SNR and bass drum, snare drum, and high tom at SD. In GNSDR, the increase of the value by the proposed method was 10.95dB, 8.96dB, and 10.60dB for bass drum, snare drum and high tom, respectively in comparison to the conventional RPCA method. For closed hi-hat cymbal and crash cymbal, these values are 2.46dB and 3.49dB respectively, that the large increase like the former three percussions was not observed.

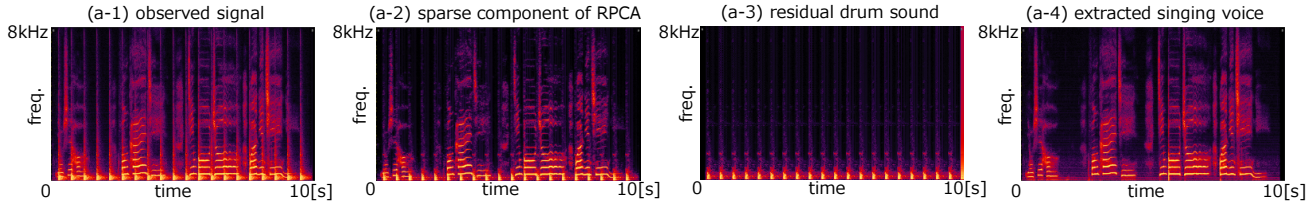
Fig. 4 (a) shows spectrograms of the simulation results when the proposed method is superior to the conventional RPCA method in NSDR. In this example, bass drum sounds were added to clip "amy_1_01". NSDR of extracted singing voice by the conventional RPCA is 0.90dB, and the proposed method is 18.34dB. In (a-2), the low frequency component of the bass drum can not be removed by RPCA. The bass drum is one kind of membranophone which produces sound by vibrating a stretched membrane and resonating with the body. The resonance frequency is determined by the shape or size of the drum body, and the power concentrates in the resonance frequency band. It means that the spectrogram of attenuated sound of membranophone is sparse and it is separated as the singing voice spectrum by RPCA. The proposed method can appropriately estimate the residual drum sounds as shown in Fig. 4 (a-3). Subtracting (a-3) from (a-2), only the singing voice is extracted as shown in (a-4). Snare drum and high tom are also membranophone, and their evaluation values also improved.

Fig. 4 (b) shows the spectrograms when the proposed and conventional methods gave almost the same capability in NSDR. In this example, closed hi-hat cymbal sounds were added to the clip "abjones_1_01". The RPCA singing voice extraction result is shown in (b-2), and result of the proposed method is shown in (b-4). Both the NSDRs are around 12.08dB. Since the power of the hi-hat cymbal spectrum is evenly distributed from the high to the low frequency band, most of hi-hat sound are separated as a low rank in the RPCA. As shown in (b-2), the RPCA sparse matrix contains only a little residual hi-hat cymbal. When the drum sound is not sparse, the RDSE method has less effect.

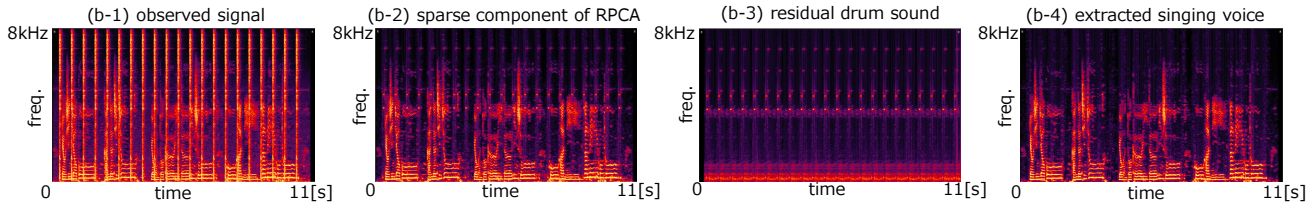
Fig. 4 (c) shows spectrograms of an extraction example where the proposed method is inferior to the conventional RPCA method in NSDR. The observed signal is obtained by adding the closed hi-hat cymbal to the clip "bobon_1_01". The conventional RPCA singing voice extraction result shown in (c-2) gave NSDR of 7.16dB, and the proposed method shown in (c-4) gave 3.36dB. Clip "bobon_1_01" has a singing voice whose pitch constantly exists during the clip. When choosing a representative spectrum from such a signal, the representative spectrum includes the singing voice component. As a result, singing voice components are included in the residual drum spectrum as shown in the circle in Fig. 4 (c-3). It leads to a degradation of speech quality, while the residual drum sound is effectively removed.

V. CONCLUSION

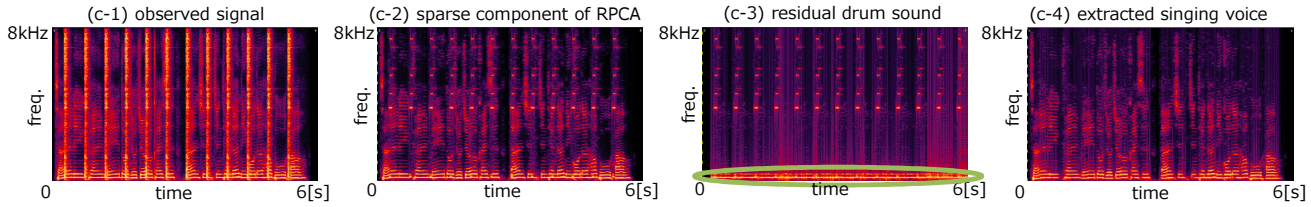
This paper proposed a singing voice extraction method based on RPCA. We developed the RDSE method to remove a residual drum sound from a singing voice spectrum estimated by RPCA. Under the assumption that the residual drum sound spectrum repeatedly occurs, the representative residual drum spectrum is obtained by taking the median value of the residual drum spectral candidates. Subtracting the representative



(a) Proposed method is superior to conventional RPCA method in NSDR.



(b) Proposed and conventional RPCA methods are almost the same in NSDR.



(c) Proposed method is inferior to conventional RPCA method in NSDR.

Fig. 4: Spectrograms of simulation resutls.

spectrum from the RPCA singing voice spectrum, we have an improved singing voice signal. Simulation result showed that the proposed method improved 10dB of GNSDR in comparison to the conventional RPCA method. Note that the position of the drum sound is assumed to be known in this paper. In a practical situation, we have to detect the position of the drum sound. Also, we intend to compare this proposed method with other methods based on NMF or deep neural network.

REFERENCES

[1] M. Suzuki, T. Hosoya, A. Ito, and S. Makino, "Music information retrieval from a singing voice using lyrics and melody information," in EURASIP J. Adv. Signal Process., Article ID: 038727, 2007.

[2] A Mesaris and T. Virtanen, "Automatic recognition of lyrics in singing," in EURASIP J. Audio, Speech, Music Process., Article ID: 546047, 2010

[3] H.Fujihara, T. Kitahara, M. Goto, K. Komatani, and T. Ogata, "Singer identification based on accompaniment sound reduction and reliable frame selection," in Proc. ISMIR, pp.329-336, 2005.

[4] K. Yoshii, M. Goto, H. G. Okuno, "INTER:D: A drum sound equalizer for controlling volume and timbre of drums," in Proc. Eur. Workshop Integration of Knowledge Semantics Digital Media Technol., pp.205-212, 2005.

[5] Y. Ikemiya, K. Itoyama, H. G. Okuno, "Transcribing vocal expression from polyphonic music," in Proc. ICASSP, pp.3127-3131, 2014.

[6] M. Goto, "Active music listening interfaces based on signal processing," in Proc. Int. Conf. Acoust., Speech, and Signal Process., pp.1441-1444, 2007.

[7] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," in IEEE Trans. Audio, Speech, Lang. Process, vol.22, no.1, pp.228-237, 2014.

[8] D. Fitzgerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," in ISAST Trans. on Electronic and Signal Processing, vol.4, no.1, pp.62-73, 2010.

[9] A. Chanrungutai and C.A. Ratanamahan, "Singing voice separation in mono-channel music using non-negative matrix factorization," in Proc. Int. Conf. Adv. Technol. Commun., pp.243-246, 2008.

[10] Emmanuel J. Cand'es, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," in J. ACM, vol. 58, no.11, pp.1-37, Jun., 2011.

[11] P. S. Huang, S. D. Chen, P. Smaragdis, and M. H. Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in Proc. Int. Conf. Acoust., Speech, and Signal Process., pp. 57-60, 2012

[12] Y. Ikemiya, "Singing voice separation and vocal F0 estimation based on mutual combination of robust principal component analysis and subharmonic summation," in IEEE/ACM Trans. on Audio, Speech, and Language. Process., vol.24, no.11, pp.2084-2095, 2016.

[13] Y. M. Z. Lin, M. Chen, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," in Mathematical Programming, 2009.

[14] MIR-1K Dataset https://sites.google.com/site/unvoicedsoundseparation/mir-1k

[15] M. Goto, "Development of the RWC music database," in Proc. Int. Congr. Acoustics, pp. I-553-I-556, 2004.