Unsupervised Singing Voice Separation Using Gammatone Auditory Filterbank and Constraint Robust Principal Component Analysis

Feng Li and Masato Akagi
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
E-mail: {lifeng, akagi}@jaist.ac.jp

***Abstract*—This paper presents an unsupervised singing voice separation algorithm which using an extension of robust principal component analysis (RPCA) with rank-1 constraint (CRPCA) based on gammatone auditory filterbank on cochleagram. Unlike the conventional algorithms that focus on spectrogram analysis or its variants, we develop an extension of RPCA on cochleagram using an alternative time-frequency representation based on gammatone auditory filterbank. We also apply time-frequency masking to improve the results of separated low-rank and sparse matrices by using CRPCA method. Evaluation results demonstrate that the proposed algorithm can achieve better separation performance on MIR-1K dataset.**

I. Introduction

Over the past few years, singing voice separation has attracted considerable attention and interest in many real-world applications. The goal of singing voice separation approaches are to separate singing voice from the mixture music signal, which is a significant technology for chord recognition [1], music information retrieval (MIR) [2], leading instrument detection [3], and karaoke applications [4]. However, the current separation results of the state of art methods are still far behind human hearing capability. The existing problems of singing voice separation are faced with great challenges [5]. Such as the large variety of sound type, the abundant mixing conditions, and the unclear mechanism to distinguish sources, especially for the similar sounds.

Many algorithms have been proposed in literature with the goal of overcoming the difficulty in separation task. Although the approaches based on deep neural network (DNN) [6]-[9], have recently proved to be powerful tools for singing voice separation task, they need a large number of training data to be available in advanced. Unsupervised algorithms therefore still remain the attraction for singing voice separation particularly where only a limited amount of singing voice data is available or without using any additional information. The classical algorithm Non-negative Matrix Factorization (NMF) [10] for music separation decomposes an input the given spectrogram into a sum of a spectral basic matrix and its activation matrix. Rafii et al [11] proposed a repeatable accompaniment idea about background music and used Repeating Pattern Extraction Technique (REPET) approach for separating the repeating music part from the non-repeating singing voice in a mixture signal. The main method was to identify the periodically repeating.

Recently, Huang et al. [12] proposed a robust principal component analysis (RPCA) method for singing voice separation, which decomposed an input matrix into a sparse matrix plus a low-rank matrix. Yang [13] proposed new sparse and low-rank matrices that were based on the incorporation of harmonicity priors and a back-end drum removal procedure. He [14] also proposed the multiple low-rank representations (MLRR) to decompose a magnitude spectrogram into two low-rank matrices. RPCA as an effective method to separate singing voice from the mixture signal, which decomposes a given amplitude spectrogram (matrix) of a mixture signal into the sum of a low-rank matrix (music accompaniment) and a sparse matrix (singing voice). Since music instruments can reproduce the same sounds each time in the same music, so its magnitude spectrogram can be considered as a low-rank structure part. Singing voice, on the contrary, varies significantly and has a sparse distribution in the spectrogram domain owning to its harmonic structure part, resulting in a spectrogram with a sparse structure part.

Inspired by a sparse and low-rank model, in our previous work, we proposed an effective extension of RPCA with rank-1 constraint (CRPCA) [15]. Although it can get better separation results than RPCA in singing voice separation task, there is still exists a lot of room for improvement. Recently a study was published hinting that cochleagram, as an alternative time-frequency analysis based on gammatone filterbank, is more suitable than spectrogram for source separation [16] [17]. This is because, cochleagram is derived from non-uniform time-frequency transform whereas time-frequency units in low-frequency regions have higher resolutions than in the high-frequency regions, which closely resembles the functions of the human ear. Similarly, singing voice performances are quite different from music accompaniment on cochleagram. The spectral energy centralizes in a few time-frequency units for singing voice and thus can be assumed to be sparse. On the other hand, music accompaniment on the cochleagram has similar spectral patterns and structures that can be captured by a few basis vectors, so it can be hypothesized as a low-rank subspace. Therefore, it is promising to separate singing voice via sparse and low-rank decomposition on cochleagram instead of the spectrogram.

To improve the separation performance, we combine gammatone auditory filterbank with cochleagram by using CRPCA algorithm. In addition, we further apply time-frequency masking estimation [18] to enforce the constraints between an input mixture music signal and the output results.

The rest of this paper is organized as follows. In Section 2, we review the conventional RPCA and CRPCA methods for singing voice separation. In Section 3, we describe the proposed framework of unsupervised singing voice separation. In Section 4, we evaluate the proposed method on MIR-1K dataset. Finally, we draw conclusions and describe future work in Section 5.

II. Background

In this section, we introduce the principles of RPCA and CRPCA methods. And they are applied to singing voice separation.

A. Principle of RPCA

RPCA can decompose an input matrix $M \in \mathbb{R}_{m\times n}$ into the sum of a low-rank matrix $L \in \mathbb{R}_{m\times n}$ and a sparse matrix $S \in \mathbb{R}_{m\times n}$. The convex model can be defined as follows:

$$\begin{aligned} &\text{minimize } |L|_* + \lambda |S|_1, \\ &\text{subject to } M = L + S. \end{aligned} \tag{1}$$

where $|\cdot|_*$ denotes the nuclear norm (sum of singular values), $|\cdot|_1$ is the L_1-norm (sum of absolute values of matrix entries), and λ is a positive constant parameter between the low-rank matrix L and the sparsity matrix S. Candés et al. suggested $\lambda = 1/\sqrt{max(m,n)}$ [19] can obtain better results. Furthermore, this convex program can be solved by accelerated proximal gradient (APG) or augmented Lagrange multipliers (ALM) [20] (we used an inexact version of ALM in a baseline experiment).

B. RPCA for singing voice separation

Huang et al. assumed that RPCA method can be applied to the task of separating singing voice and music accompaniment from the mixture music signal [12]. On account of the music accompaniment part, music instruments can reproduce the same sounds each time in the same music, so its magnitude spectrogram can be considered as a low-rank matrix structure. Singing voice part, in contrast, varies significantly and has a sparse distribution in the spectrogram domain due to its harmonic structure part, resulting in a spectrogram with a sparse matrix structure. Therefore, we can use the RPCA method to decompose an input matrix into a sparse matrix (singing voice) and a low-rank matrix (music accompaniment). However, it makes some strong assumptions. For instance, drums may lie in the sparse subspace instead of being low-rank, which decreases the separation performance in the mixture music signal.

Algorithm 1 CRPCA for singing voice separation [15]

Input: Mixture signal $M \in \mathbb{R}_{m\times n}$.
1: **Initialize**: $\rho > 1, \mu_0 > 0, k = 0, L_0 = S_0 = 0.$
2: While not convergence,
3: **do** :
4: $L_{k+1} = P_{1,\mu_k^{-1}}(M - S_k + \mu_k^{-1} J_k).$
5: $S_{k+1} = Q_{\lambda\mu_k^{-1}}(M - L_{k+1} + \mu_k^{-1} J_k).$
6: $J_{k+1} = J_k + \mu_k (M - L_{k+1} - S_{k+1}).$
7: $\mu_{k+1} = \rho * \mu_k.$
8: $k = k + 1.$
9: **end while**.
Output: $L_{m\times n},\ S_{m\times n}.$

C. Principle of CRPCA

CRPCA is a novel extension of RPCA, which exploiting rank-1 constraint for singing voice separation. We define the model as follows:

$$\begin{aligned} &\text{minimize } \sum_{i=2}^{\min(m,n)} \delta_i(L) + \lambda |S|_1, \\ &\text{subject to } M = L + S. \end{aligned} \tag{2}$$

where L is the value of low-rank matrix, S is the value of sparse matrix. $M \in \mathbb{R}_{m\times n}$ is the value of an input matrix, which consists of $L \in \mathbb{R}_{m\times n}$ and $S \in \mathbb{R}_{m\times n}$, and $\lambda > 0$ is a positive constant parameter between the sparse matrix S and the low-rank matrix L. And $\delta_i(L)$ is the i-th singular value of L. We used $\lambda = 1/\sqrt{\max(m,n)}$ as suggested in [19]. We also used an efficient iALM [20] method to solve this convex model in this work. The augmented Lagrangian function can be defined as follows:

$$\begin{aligned} J(M,L,S,\mu) = \ &\min \sum_{i=2}^{\min(m,n)} \delta_i(L) + \lambda |S|_1 \\ &+ \langle J, M - L - S\rangle + \frac{\mu}{2} |M - L - S|_F^2. \end{aligned} \tag{3}$$

where J is the Lagrange multiplier and μ is a positive scaler. The process of separating singing voice from the mixture music signal can be seen in **Algorithm 1** CRPCA for singing voice separation. The value of M is a mixture music signal from the observed audio data. After separated by using CRPCA, we can get a low-rank matrix L (music accompaniment) and a sparse matrix S (singing voice).

From the augmented Lagrangian function, we solve the

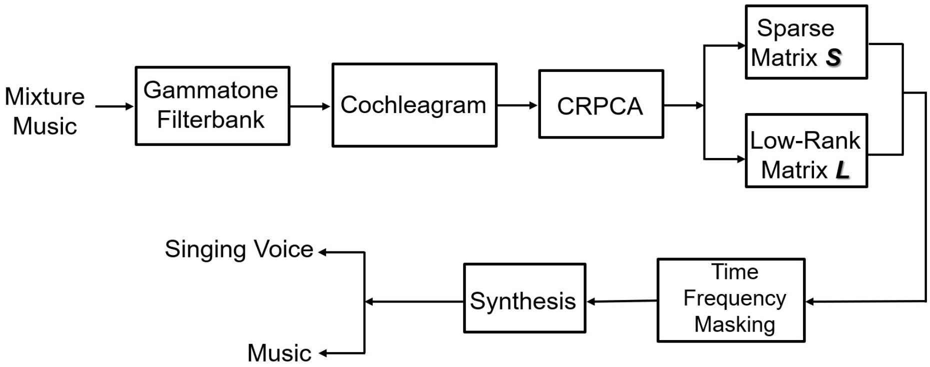

Fig. 1. Block diagram of unsupervised singing voice separation system.

following two sub-problems about L and S:

$$
\begin{aligned}
L_{k+1} = & \min_{L} \sum_{i=2}^{\min(m,n)} \delta_i(L) + \langle J_k, M - L - S_k \rangle \\
& + \frac{\mu_k}{2} |M - L - S_k|_F^2. \qquad (4) \\
S_{k+1} = & \min_{S} \lambda |S|_1 + \langle J_k, M - L_k - S \rangle \\
& + \frac{\mu_k}{2} |M - L_k - S|_F^2. \qquad (5)
\end{aligned}
$$

As suggested by Oh et al. [21], the update rules of L and S are solved as the following two equations (6) and (7):

$$L_{k+1} = P_{1,\mu_k^{-1}}(M - S_k + \mu_k^{-1} J_k) \qquad (6)$$

$$S_{k+1} = Q_{\lambda\mu_k^{-1}}(M - L_{k+1} + \mu_k^{-1} J_k) \qquad (7)$$

$P_{1,\mu_k^{-1}}(\cdot)$ can be defined as follows:

$$P_{1,\mu_k^{-1}}(Y) = U_Y(D_{Y_1} + Q_{\mu_k^{-1}}(D_{Y_2}))V_Y^T \qquad (8)$$

where $Y = Y_1$+Y_2 ($Y \in \mathbb{R}_{m\times n}$), $D_{Y_1} = diag(\delta_1, 0, ..., 0)$, $Q_{\mu_k^{-1}}(D_{Y_2}) = sign(D_{Y_2}) \cdot max(|D_{Y_2}| - \mu_k^{-1}, 0)$ is the soft-thresholding operator [22], $D_{Y_2} = diag(0, \delta_2, ..., \delta_{\min(m,n)})$, δ_1 and δ_2 are the first and second singular values.

III. Proposed Method

In this section, we explain the proposed framework for unsupervised singing voice separation.

A. Gammatone filterbank and cochleagram

The Gammatone filterbank [23] is a cochlear filtering representation which decomposes an input signal into the time-frequency domain using a lot of gammatone filters. The impulse response of a gammatone filter centered at frequency w is obtained as follow:

$$g(w,t) = \begin{cases} t^{h-1}e^{-2\pi vt}cos(2\pi wt), \ t > 0 \\ 0, \ others \end{cases} \qquad (9)$$

where h represents the order of filter, v stands for the rectangular bandwidth which increases as the center frequency w increases. The filter output response $r(c,t)$ can be expressed as follow:

$$r(c,t) = x(t) * g(w_c, t) \qquad (10)$$

where '*' indicates the convolution in time domain, c is a particular filter channel and the center frequency is w_c. So this function can be shifted backwards by using (h-1)/($2\pi v$) to compensate for the filter delay. The output of each filter channel is cut into time-frequency with half of overlap between the consecutive frames. And finally, the time-frequency spectra of all the filter outputs are constructed to form the cochleagram.

B. CRPCA using time-frequency masking

After separated by using CRPCA, in order to improve the separation performance, we apply binary time-frequency masking estimation to further improve the separation results. We define b_m as follows:

$$b_m = \begin{cases} 1 & S_{ij} \geq L_{ij} \\ 0 & S_{ij} < L_{ij} \end{cases} \qquad (11)$$

where S_{ij} and L_{ij} are the values of sparse and low-rank matrices.

A block diagram of our proposed unsupervised singing voice separation system can be illustrated in Fig. 1. For each mixture music audio in the test dataset, we calculate the cochleagram of the mixture music audio under the condition of gammatone filterbank, after that decompose the matrix into low-rank matrix L (music accompaniment) and sparse matrix S (singing voice) by using CRPCA method, and then, we deal with the separated sparse and low-rank matrices by using time-frequency masking. Finally, the separated matrices can be synthesized as described in [24].

IV. Experimental Evaluation

In this section, we show how evaluated the proposed unsupervised singing voice separation method by using MIR-1K dataset [1] [25], and how we compared it with the conventional RPCA method.

[1]https://sites.google.com/site/unvoicedsoundseparation/mir-1k/

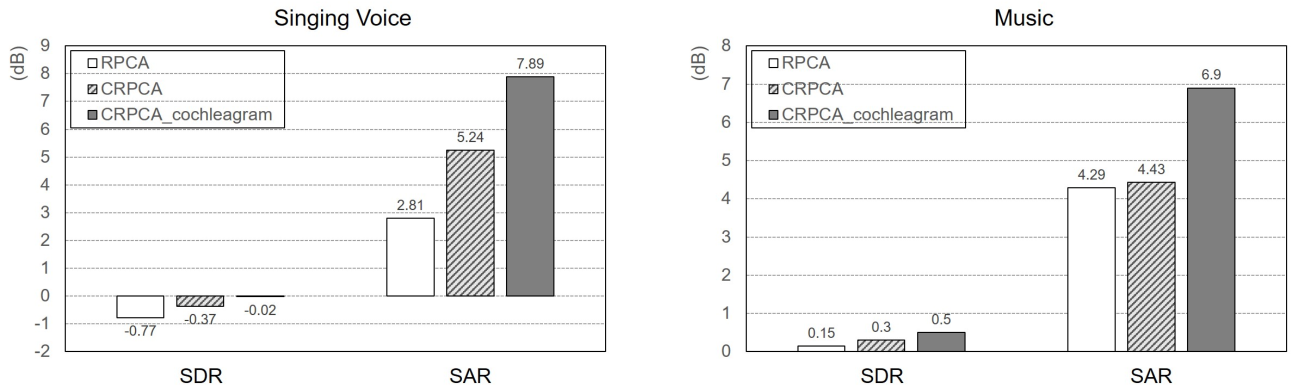

Fig. 2. Comparison of unsupervised singing voice separation results on MIR-1K dataset among of the conventional RPCA, CRPCA and CRPCA on cochleagram, respectively.

A. Dataset and condition

In our experiments, we evaluated the proposed method on MIR-1K dataset, which contains 1000 song clips with durations ranging from 4 to 13 seconds. The data were extracted from 110 Chinese karaoke pop songs. The dataset includes annotations of the pitch contours, lyrics, indices and types for unvoiced frames, and indices of the vocal and non-vocal frames. We mainly focused on monaural singing voice separation in our experiments. This is even more difficult than multichannel source separation since only a single channel is available. All experiment data were sampled at 16 kHz. We set parameters for cochleagram analysis: 128 channels, 40$\sim$8000 Hz frequency range, and 256 frequency length. To compare the results with those obtained with CRPCA, we calculated the input feature by using short-time Fourier transform (STFT) and inverse STFT (ISTFT), which is a part of baseline experiments that have been performed on spectrogram for conventional RPCA method. We used a window size of 1024 samples, a hop size of 256 samples for the STFT and an FFT size of 1024.

To confirm the effectiveness of the proposed method, we assessed its quality of separation in terms of the source-to-distortion ratio (SDR) and the source-to-artifact ratio (SAR) by using the BSS-EVAL 3.0 metrics[2] [26] is defined as

$$\hat{S}(t) = S_{target}(t) + S_{interf}(t) + S_{artif}(t). \tag{12}$$

where $S_{target}(t)$ is the allowable deformation of the target sound, $S_{interf}(t)$ is the allowable deformation of the sources that account for the interferences of the undesired sources, and $S_{artif}(t)$ is an artifact term that may correspond to the artifact of the separation method. The formulas for the SDR and SAR are defined as

$$SDR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t \{e_{interf}(t) + e_{artif}(t)\}^2}. \tag{13}$$

$$SAR = 10\log_{10}\frac{\sum_t \{S_{target}(t) + e_{interf}(t)\}^2}{\sum_t e_{artif}(t)^2}. \tag{14}$$

[2]http://bass-db.gforge.inria.fr/bss_eval/

The higher values of the SDR and SAR represent that the method exhibits better separation performance in source separation task. The SDR represents the quality of the separated target sound signals. The SAR represents the absence of artificial distortion. All the evaluation metrics are expressed in dB.

B. Results

To examine the proposed method, we evaluated it on MIR-1K dataset. Fig. 2 shows the comparison results of conventional RPCA, CRPCA and CRPCA on cochleagram, respectively. All methods were run by using binary time-frequency masking estimation. From the experiment results, we can see that the proposed method can improve the separation performance between singing voice and music. In terms of singing voice, the separation performance is worse than the part of music in SDR. On the contrary, the SAR of the proposed method has the highest value among them. In addition, the SAR obtains a significant improvement on cochleagram between the parts of singing voice and music.

V. Conclusion

In this paper, a novel unsupervised method to address the singing voice separation task has been proposed. From the experiment results on MIR-1K dataset, we can see clearly that the proposed method outperforms the conventional RPCA method. In future work, since melody extraction is significant for separating singing voice from the mixture music signal, we therefore will combine with it to improve the separation results.

Acknowledgments

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan Scholarship and the China Scholarship Council (CSC) of China Scholarship.

REFERENCES

[1] T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music," in *Proc. ICMC*, pp. 464-467, 1999.

[2] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: current directions and future challenges," *Proc. IEEE*, vol. 96, no. 4, pp. 668-696, 2008.

[3] J. L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180-1191, 2011.

[4] A. J. R. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: extracting vocals from musical mixtures using a convolutional deep neural network," in *Proc. LVA/ICA*, pp. 429-436, 2015.

[5] A. Liutkus, F. R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. LVA/ICA*, pp. 323-332, 2017.

[6] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. ISMIR*, pp. 477-482, 2014.

[7] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. ICASSP*, pp. 2135-2139, 2015.

[8] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, pp. 31-35, 2016.

[9] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: stronger together," in *Proc. ICASSP*, pp. 61-65, 2017.

[10] V. Tuomas, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. ALSP*, vol. 15, no. 3, pp. 1066-1074, 2007.

[11] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): a simple method for music/voice separation," *IEEE Trans. ALSP*, vol. 21, no. 1, pp. 73-84, 2013.

[12] P. S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. ICASSP*, pp. 57-60, 2012.

[13] Y. H. Yang, "On sparse and low-rank matrix decomposition for singing voice separation," in *Proc. ACM Multimedia*, pp. 757-760, 2012.

[14] Y. H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *Proc. ISMIR*, pp. 427-432, 2013.

[15] F. Li and M. Akagi, "Unsupervised singing voice separation based on robust principal component analysis exploiting rank-1 constraint," in *Proc. EUSIPCO*, 2018, pp. 1929-1933.

[16] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised single-channel separation of nonstationary signals using gammatone filterbank and itakura–saito nonnegative matrix two-dimensional factorizations," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 3, pp. 662-675, 2013.

[17] F. Li and M. Akagi, "Weighted robust principal component analysis with gammatone auditory filterbank for singing voice separation," in *Proc. ICONIP*, vol. 6, 2017, pp. 849-858.

[18] Y. P. Li and D. L. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, vol 51, no. 3, pp. 230-239, 2009.

[19] E. J. Candés, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, 2011.

[20] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[21] T. H. Oh, Y. W. Tai, J. C. Bazin, H. Kim, and I. S. Kweon, "Partial sum minimization of singular values in robust PCA: algorithm and applications," *IEEE Trans. PAMI*, vol. 38, no. 4, pp. 744-758, 2016.

[22] E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for ℓ_1-minimization: methodology and convergence," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1107-1130, 2008.

[23] G. N. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on neural networks*, vol. 15, no. 5, pp. 1135-1150, 2004.

[24] D. L. Wang and G. J. Brown, "Computational auditory scene analysis: principles, algorithms, and applications," *Wiley-IEEE Press, Hoboken*, 2006.

[25] C. L. Hsu, and J. S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Trans. ALSP*, vol. 18, no. 2, pp. 310-319, 2010.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ALSP*, vol. 14, no. 4, pp. 1462-1469, 2006.