INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2021.15.5 Volume 15, 2021

Monaural Singing Voice Separation Using Robust Principal Component Analysis with Weighted Values

Feng Li*, Hao Chang
Department of Computer Science and Technology,
Anhui University of Finance and Economics,
Bengbu, 233030, China

Received: June 8, 2020. Revised: December 15, 2020. Accepted: January 19, 2021. Published: January 28, 2021.

Abstract- This paper proposes an extension of robust principal component analysis (RPCA) with weighted values for monaural singing voice separation.Although the conventional RPCA is an effective method to separate singing voice and music accompaniment from the mixted audio signal, it fails when one singular value is much larger than all others. For example, drums may lie in the sparse subspace instead of being lowrank, which lead that the separation performance is decreased in many real world applications, especially for drums existing in the mixture music signal. Therefore, in order to solve this problem, we utilize different weighted values between sparse (singing voice) and low-rank matrices (music accompaniment). Evaluation results on ccMixter and DSD100 datasets show that the proposed method achieves better separation performance than the conventional RPCA.

Keywords- Singing voice separation, Robust principal component analysis (RPCA), Low-rank and sparse matrices, Weighted values

I. Introduction

RECENTLY, monaural singing voice separation has attracted more considerable interests and attentions in many real world applications. It attempts to separate the singing voice and music accompaniment parts of a music recording, which is very significant technology for lyric recognition [1] and alignment [2], music information retrieval (MIR) [3], singer identification [4] and chord recognition [5]. However, current state-of-the-art results are still far behind human hearing capability. The existing problems of singing voice separation are still more challenginging [6, 7, 8, 9].

Many previous separation algorithms have been proposed with the goal of overcoming the difficulty in separation tasks. Most of them have attempted to use the distinctive characteristic of each source. Rafii *et al.* [10] proposed a repeat idea about music accompaniment and used REpeating Pattern Extraction Technique (REPET) approach for separating the repeating music accompaniment from the non-repeating vocals in a mixture music signal. The basic idea was to identify the periodically repeating segments in the audio, compared them to a repeating segment model derived from them, and extracted the repeating patterns via time-frequency masking. Huang *et al.* [11] proposed a robust principal component analysis (RPCA) for singing voice separation, which decomposed an input matrix into a low-rank matrix plus a sparse matrix. Inspired by low-rank and sparse model, Yang [12] proposed a new low-rank and sparse matrix based on the incorporation of harmonicity priors and a back-end drum removel procedure. Moreover, he [13] also proposed a multiple low-rank representation (MLRR) to decompose a magnitude spectrogram into two low-rank matrices. Some relevant studies can be found in [14] and [15].

As mentioned above, RPCA is an effective algorithm for separating singing voice from the mixed music signals. It decomposes the given amplitude spectrogram of the music signal into the sum of a low-rank matrix and a sparse matrix. Since music accompaniment tends to have a similar phrases, resulting in a spectrogram with the low-rank structure part. While singing voice varies significantly and continuously over time, resulting that a spectrogram has a sparse structure part. Although RPCA has been successfully applied to singing voice separation, it has a strong assumption. For example, drums may lie in the sparse subspace instead of being low-rank, which lead that the separation performance is decreased in many real world applications, especially for the drums existing in music signal. Therefore, to overcome this problem, in this paper, we propose a weighted method to make sure different scale values to describe sparse and low-rank matrices called Weighted Robust Principal Component Analysis (WRPCA), which is choose different weighted values between low-rank and sparse matrices. Figure 1 describes the process of monaural singing voice separation.

The rest of this paper is structured as follows. In section II, we brief introduce the related work focusing on principal of RPCA. The proposed method is described in section III. Then, the results and analysis of the proposed method on the two databases are provided in section IV. Fianlly, we conclude with a brief summary.

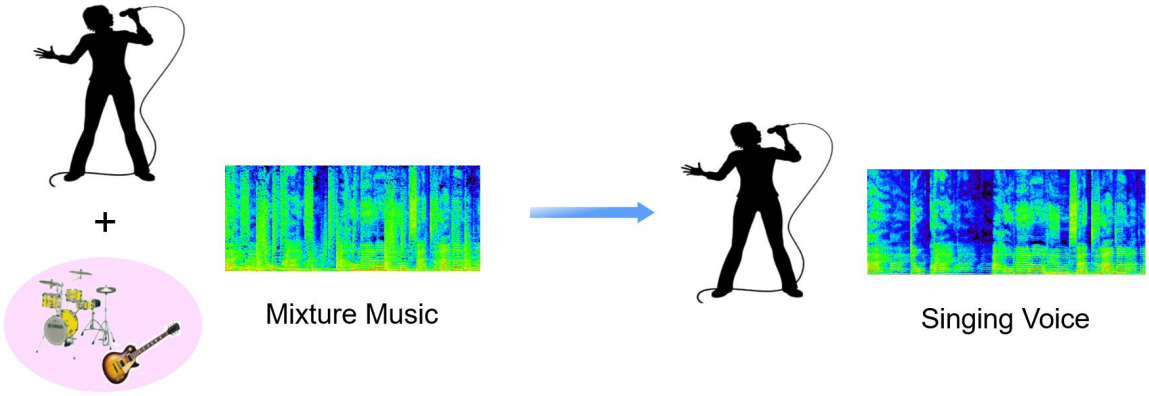


Fig. 1: Illustration of monaural singing voice separation.

II. Robust Principal Component Analysis

Candés *et al*. [16] presented a convex program RPCA, which decomposed an input matrix $M \in \mathbb{R}_{m\times n}$ into as the sum of a low-rank matrix $L \in \mathbb{R}_{m\times n}$ plus a sparse matrix $S \in \mathbb{R}_{m\times n}$. The problem can be formulated as follows:

$$\begin{aligned} &min\ |L|_* + \lambda|S|_1, \\ &s.t.\ M = L + S. \end{aligned} \tag{1}$$

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 $\lambda = 1/\sqrt{max(m,n)}$ [16]. Furthermore, this convex program can be solved by accelerated proximal gradient (APG) or augmented Lagrange multipliers (ALM) [17] (we use inexact version of ALM as the baseline experiment).

Huang *et al*. supposed that RPCA method can be applied to the task of separating singing voice and music accompaniment from the mixture music signal [11]. On account of music accompaniment part, instruments can reproduce the same sounds each time in the same music, so we can think the magnitude spectrogram of music as a low-rank matrix. Singing voice part, on the contrary, is sparse distribution owning to its harmonic structure part in the spectrogram domain.

Thus, we can use RPCA method to decompose an input matrix into a sparse matrix (singing voice) and a low-rank matrix (music accompaniment). However, it has a strong assumption. For instance, drums may lie in the sparse subspace instead of being low-rank, which lead that the separation performance is decreased in the mixture music signal, especially for drums existing.

III. Proposed Method

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

A. Principal of WRPCA

WRPCA is an extension of RPCA, which has a different scale values between sparse and low-rank matrices.

Algorithm 1 WRPCA for Singing Voice Separation

Input: Mixture signal $M \in \mathbb{R}_{m\times n}$, weight w.
Initialization: $\rho, \mu_0, L_0 = M, J_0 = 0, k = 0$.
While not convergence **do**
repeat
$S_{k+1} = \arg\min_S |S|_1 + \frac{\mu_k}{2}|M + \mu_k^{-1}J_k - L_k - S|_F^2$.
$L_{k+1} = \arg\min_L |L|_{w,*} + \frac{\mu_k}{2}|M + \mu_k^{-1}J_k - S_{k+1} - L|_F^2$.
$J_{k+1} = J_k + \mu_k(M - L_{k+1} - S_{k+1})$.
$\mu_{k+1} = \rho * \mu_k$.
$k \leftarrow k + 1$.
end while.
Output: $S_{m\times n},\ L_{m\times n}$.

The model can be defined as follows:

$$min\ |L|_{w,*} + \lambda|S|_1, \quad s.t.\ M = L + S. \tag{2}$$

where $|L|_{w,*}$ is the low-rank matrix with different weighted values, while S is the sparse matrix. $M \in \mathbb{R}_{m\times n}$ is 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 trade-off constant parameter between the sparse matrix S and the low-rank matrix L. We use $\lambda = 1/\sqrt{max(m,n)}$ as suggested in [16]. Moreover, adopt an efficient inexact version of the augmented Lagrange multiplier (ALM) [17] to solve this convex model. The corresponding augmented Lagrange function is defined as follows:

$$\begin{aligned} J(M,L,S,\mu) = &\ |L|_w, * + \lambda|S|_1 + < J, M - \\ &L - S > + \frac{\mu}{2}|M - L - S|_F^2. \end{aligned} \tag{3}$$

where J is the Lagrange multiplier and μ is a positive scaler. The corresponding to the separation process of mixture music signal can be seen in **Algorithm 1** WRPCA for monaural singing voice separation. The value of M is a mixture music signal from the observed data, after the separation by using WRPCA, and finally, we can obtain a sparse matrix S (singing voice) and a low-rank matrix L (music accompaniment).

B. Weighted values

In this paper, we adopt different weighted values to trim low-rank matrix during separation process.

Lemma 1. Set $M = U\sum V^T$ is the singular value decomposition (SVD) of $M \in \mathbb{R}_{m\times n}$, where

$$\sum = \begin{pmatrix} diag(\delta_1(M), \delta_2(M), ..., \delta_n(M)) \\ 0 \end{pmatrix}, \quad (4)$$

and $\delta_i(M)$ denotes the i-th singular value of M. If the positive regularization parameter C exists and the positive value $\varepsilon < min(\sqrt{C}, \frac{C}{\delta_1(M)})$ holds, by using the reweighting formula $W_i^l = \frac{C}{\delta_i(L_l)+\varepsilon}$ [18] with initial estimation $L_0 = M$, the reweighted problem has the closed-from solution:
$L^* = U\sum' V^T$, where

$$\sum' = \begin{pmatrix} diag(\delta_1(L^*), \delta_2(L^*), ..., \delta_n(L^*)) \\ 0 \end{pmatrix}, \quad (5)$$

and

$$\delta_i(L^*) = \begin{cases} 0 \\ \dfrac{c_1 + \sqrt{c_2}}{2} \end{cases} \quad (6)$$

where $c_1 = \delta_i(M) - \varepsilon$ and $c_2 = (\delta_i(M) + \varepsilon)^2 - 4C$. The more specific proof of the *Lemma* 1 can be found in [19]. In our experiments, the regularization parameter C is empirically set as the maximum size of matrix, the separation performance can be obtained the best results, e.g., $C = max(m, n)$.

As mentioned above, we use different values to adjust the weighted scales to optimize the conventional RPCA, the corresponding separation results of spectrograms of example are excerpted from 'AlexBeroza_-_To_Be_Sensitive_(with_mind-mapthat)' in the set of ccMixter in Figure 2. The left three spectrograms are singing voice; on the contrary, the right ones are music accompaniment. And the above two spectrograms are original signal (singing voice and music accompaniment), while the middle two spectrograms and the below two spectrograms are separated by RPCA and WRPCA, respectively.

IV. Experimental Evaluation

In this part, we evaluate the proposed WRPCA method using two different datasets.

A. Experimental Datasets

One is ccMixter dataset[1], which contains 50 full stereo songs with durations ranging from 1'17" to 7'36". Each audio contains three parts: singing voice, music accompaniment and their mixture, respectively. To reduce computations, we use only 30-second fragments (from 0'30" to 1'00") at the same time of each song, which is the maximum period of all songs containing singing voice, but there are still exist 2 songs with no singing voice during this period, we adopt to another period (from 1'30" to 2'00") in this 2 songs.

The other is DSD100 dataset[2]. It contains 100 full stereo songs with durations ranging from 2'21" to 7'15" as also used in the 2016 Signal Separation Evaluation Campaign (SiSEC) [7], which is divided into 50 development songs (***dev*** data) and 50 test songs (***test*** data). To reduce computations, we also use only 30-second fragments (from 1'45" to 2'15"), which is the only period where all 100 full stereo songs contain singing voice. Because there are 4 sources (e.g., bass, drums, vocals and others) for each track, we consider the sum of bass, drums and others as music accompaniment part.

B. Experiment Conditions

In this study, we mainly focusing on single-channel source separation. It is even more difficult than separation the multi-channel audio signal since only one single channel was available from the mixed data. The two-channel stereo mixtures were downmixted into a single mono channel and obtained an average value of each channel. All experiment data are sampled at 44100Hz. The input feature is calculated using short time Fourier transform (STFT) and inverse STFT (inverse short time Fourier transform). FFT size is 1024., and A window size of 1024 samples and a hop size of 256 samples for the STFT.

To evaluate the effectiveness of the proposed method, the quality of separation is assessed in terms of source-to-interference ratio (SIR) and source-to-distortion ratio (SDR) by using the BSS-EVAL 3.0 metrics[3] [20] and the normalized of SDR (NSDR). The estimated signal $\hat{S}$(t) is defined as

$$\hat{S}(t) = S_{target}(t) + S_{interf}(t) + S_{artif}(t). \quad (7)$$

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 SDR, SIR and NSDR are defined as

$$SIR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t S_{interf}(t)^2}, \quad (8)$$

$$SDR = 10log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t \{e_{interf}(t) + e_{artif}(t)\}^2}, \quad (9)$$

$$NSDR(\hat{v}, v, x) = SDR(\hat{v}, v) - SDR(x, v). \quad (10)$$

where $\hat{v}$ is the separated voice part, v is the original clean signal, and x is the original mixture. The NSDR is used to estimate the overall improvement in the SDR between x and $\hat{v}$.

The higher values of SDR, SIR, and NSDR represent the method that exhibits better separation performance.

[1] https://members.loria.fr/ALiutkus/kam/
[2] http://liutkus.net/DSD100.zip
[3] http://bass-db.gforge.inria.fr/bss_eval/

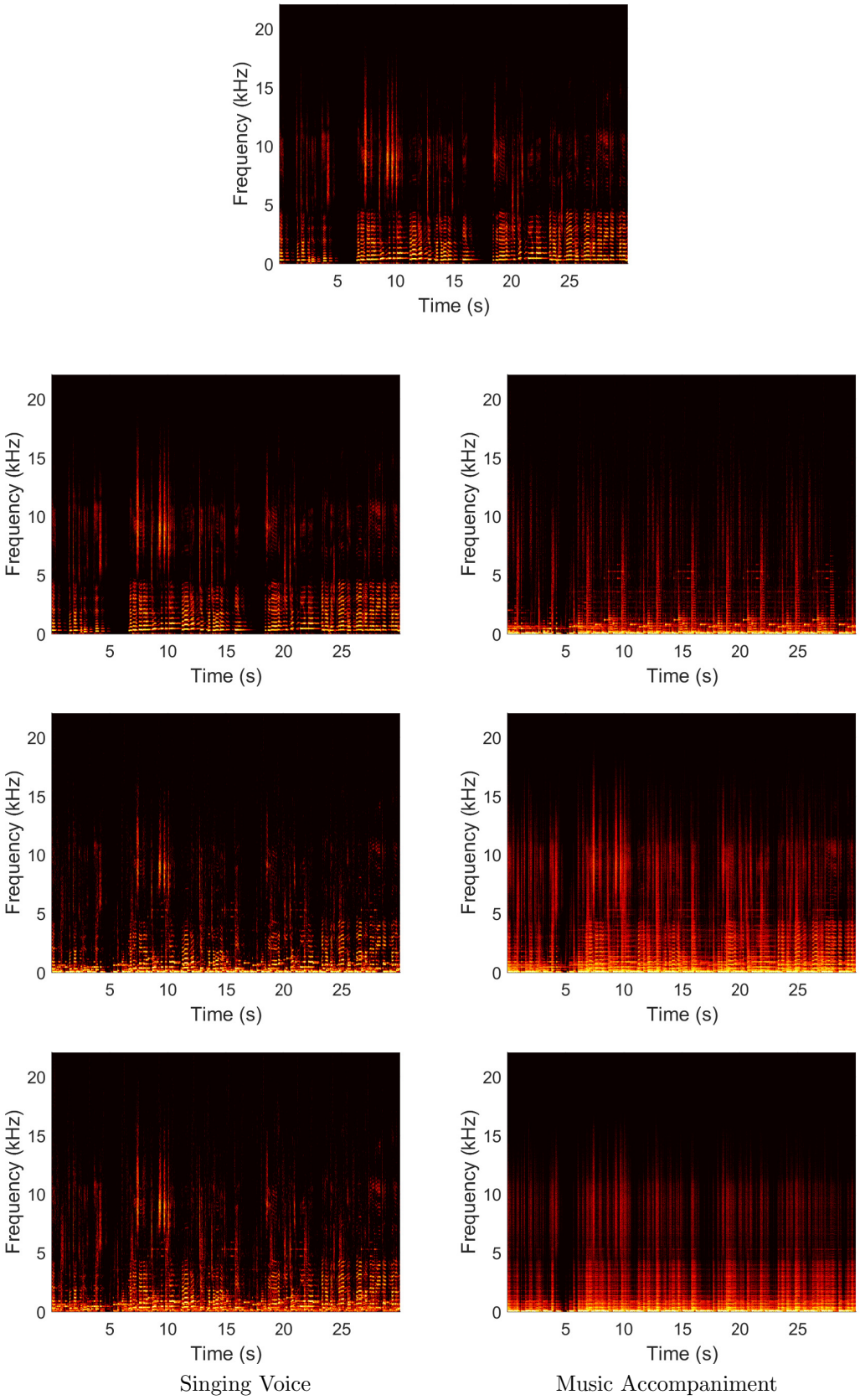
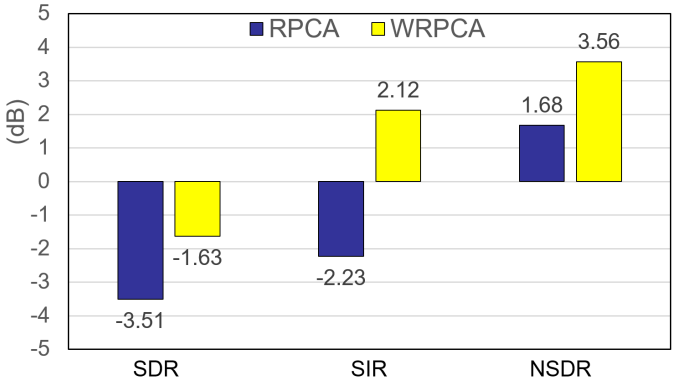
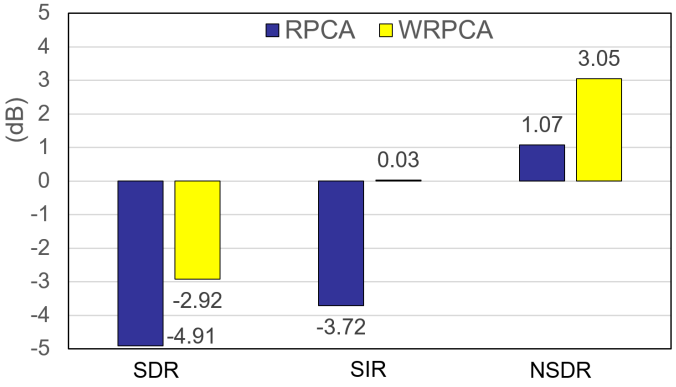


Fig. 2: Spectrograms of example are excerpted from music in the set of ccMixter. The left three spectrograms are singing voice and the right ones are the corresponding of music accompaniment from the mixture signal. The above two spectrograms are **original** signal, the middle spectrograms are separated by **RPCA** and the below spectrograms are separated by **WRPCA**.





(a)

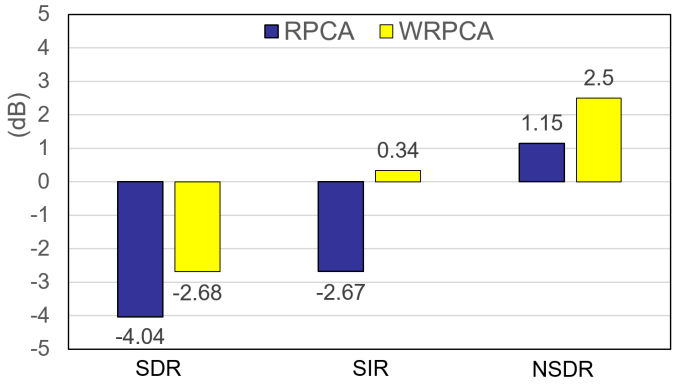
Fig. 3: Comparison of the separation results by using conventional RPCA and our proposed WRPCA on ccMixter dataset. Note that SDR for the original dataset, ccMixter, is -5.19 dB.

SDR represents the quality of the separated target sound signals, and SIR represents the degree of separation between the target and other sound signals. All the metrics are expressed in dB.

C. Experiment Results

We evaluate WRPCA method on ccMixter dataset. Figure 3 describe the experiment results of SDR, SIR and NSDR between WRPCA and RPCA on ccMixter dataset. From the experiment results, we can see clearly that our proposed method gets better results on this dataset.

Furthermore, we compare with the conventional RPCA on DSD100 dataset. Figure 4: (a) is the separation results of SDR, SIR and NSDR on ***dev*** data (left); (b) is the separation results of SDR, SIR and NSDR on ***test*** data (right). From the above two figures, we can see clearly the proposed WRPCA method also yields promising experimental results than the conventional RPCA method on DSD100 dataset.



(b)

Fig. 4: Comparison of the separation results by using conventional RPCA and our proposed WRPCA on DSD100 dataset. (a) is the set of DSD100/***dev*** data; (b) is the set of DSD100/***test*** data. Note that SDRs for the original datasets, DSD100/***dev*** and DSD100/***test***, are -5.98dB and -5.18dB, respectively.

V. Conclusions and Future Works

In this paper, an extension of RPCA with weighted values for singing voice separation was proposed. From the experimental results on the datasets (ccMixter and DSD100), we can see clearly that the proposed method outperforms the conventional RPCA on singing voice separation task. In future work, since prior information and spatial information are very significant for separate music signal, prior information and spatial information are fused and expected to improve the separation performance.

Acknowledgment

This work was supported in part by the Natural Science Foundation of the Higher Education Institutions of Anhui Province under grant No. KJ2020A0011, the Science Research Project of Anhui University of Finance and Economics under grant No. ACKYB20012, the Natural Science Foundation of China under grants No. 61704001, Anhui Provincial Natural Science Foundation under grant No.1808085QF196.

References

[1] Kawai D, Yamamoto K, Nakagawa S, "Lyric recognition in monophonic singing using pitch-dependent DNN", 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 326-330.

[2] Stoller, Daniel, Simon Durand, and Sebastian Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model."ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[3] Murthy, YV Srinivasa, and Shashidhar G. Koolagudi, "Content-based music information retrieval (cb-mir) and its applications toward the music industry: A review." ACM Computing Surveys (CSUR) 51.3 (2018): 1-46.

[4] Sharma B, Das R K, Li H. "On the Importance of Audio-Source Separation for Singer Identification in Polyphonic Music", INTERSPEECH. 2019: 2020-2024.
[5] McFee B, Bello J P.,"Structured Training for Large-Vocabulary Chord Recognition,"ISMIR. 2017: 188-194.
[6] Rafii, Zafar, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, Derry FitzGerald, and Bryan Pardo,"An overview of lead and accompaniment separation in music." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26, no. 8 (2018): 1307-1335.
[7] Luo Y, Mesgarani N.,"Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation", IEEE/ACM transactions on audio, speech, and language processing, 2019, 27(8): 1256-1266.
[8] Stöter, Fabian-Robert, Antoine Liutkus, and Nobutaka Ito,"The 2018 signal separation evaluation campaign." In International Conference on Latent Variable Analysis and Signal Separation, pp. 293-305. Springer, Cham, 2018.
[9] Cano, Estefania, Derry FitzGerald, Antoine Liutkus, Mark D. Plumbley, and Fabian-Robert Stöter,"Musical source separation: An introduction." IEEE Signal Processing Magazine 36, no. 1 (2018): 31-40.
[10] Z. Rafii, B. Pardo,"Repeating pattern extraction technique (REPET): A simple method for music/voice separation", IEEE transactions on audio, speech, and language processing 21.1 (2013): 73-84.
[11] P. S. Huang, S. D. Chen, P. Smaragdis, M. H. Johnson,"Singing-voice separation from monaural recordings using robust principal component analysis", Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012.
[12] Y. H. Yang,"On sparse and low-rank matrix decomposition for singing voice separation", in proceedings of the 20th ACM international conference on Multimedia. ACM, 2012.
[13] Y. H Yang,"Low-rank representation of both singing voice and music accompaniment via learned dictionaries", ISMIR. 2013.
[14] Fotios K. Pantazoglou, Georgios P. Kladis, Nikolaos K. Papadakis,"A Greek Voice Recognition Interface for ROV Applications, Using Machine Learning Technologies and the CMU Sphinx Platform", WSEAS Transactions on Systems and Control, pp. 550-560, Volume 13, 2018.
[15] Anton V. Kvasnov, Vyacheslav P. Shkodyrev, Dmitry G. Arsenyev,"Method of Recognition the Radar Emitting Sources based on the Naive Bayesian Classifier", WSEAS Transactions on Systems and Control, pp. 112-120, Volume 14, 2019.
[16] E. J. Candés, X. Li, Y. Ma, J. Wright,"Robust principal component analysis?", Journal of the ACM (JACM) 58.3 (2011): 11.
[17] Z. Lin, M. Chen, Y. Ma,"The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices", arXiv preprint arXiv:1009.5055 (2010).
[18] E. J. Candés, M. B. Wakin, S. Boyd,"Enhancing sparsity by reweighted l_1 minimization", Journal of Fourier analysis and applications 14.5 (2008): 877-905.
[19] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, L. Zhang,"Weighted nuclear norm minimization and its applications to low level vision", International journal of computer vision 121.2 (2017): 183-208.
[20] E. Vincent, R. Gribonval, C. Févotte,"Performance measurement in blind audio source separation", IEEE transactions on audio, speech, and language processing 14.4 (2006): 1462-1469.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Feng Li implemented the separation models, performed the experiments, analyzed the experiment data and wrote the paper.

Hao Chang conceived and fine-tuned the paper, and gave some precious advances.