Lai and Wang *EURASIP Journal on Audio, Speech, and Music Processing* (2022) 2022:4
https://doi.org/10.1186/s13636-022-00236-9

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

RPCA-DRNN technique for monaural singing voice separation

Wen-Hsing Lai[1*] and Siou-Lin Wang[2]

Abstract

In this study, we propose a methodology for separating a singing voice from musical accompaniment in a monaural musical mixture. The proposed method uses robust principal component analysis (RPCA), followed by postprocessing, including median filter, morphology, and high-pass filter, to decompose the mixture. Subsequently, a deep recurrent neural network comprising two jointly optimized parallel-stacked recurrent neural networks (sRNNs) with mask layers and trained on limited data and computation is applied to the decomposed components to optimize the final estimated separated singing voice and background music to further correct misclassified or residual singing and background music in the initial separation. The experimental results of MIR-1K, ccMixter, and MUSDB18 datasets and the comparison with ten existing techniques indicate that the proposed method achieves competitive performance in monaural singing voice separation. On MUSDB18, the proposed method reaches the comparable separation quality in less training data and lower computational cost compared to the other state-of-the-art technique.

Keywords: Singing separation, Robust principal component analysis, Deep recurrent neural network, Stacked recurrent neural network

1 Introduction

In a natural environment rich in sound emanating from multiple sources, a target sound reaching our ears is usually mixed with other acoustic interference. The sources of background acoustic interference, including car noise, street noise, music, other people's voices [1], and even reverberations [2], corrupt the target sound, complicate signal processing, pose severe challenges for the hearing impaired, and degrade the performance of automatic sound recognition systems. In musical pieces, instead of background noise, singing voices are often mixed with musical accompaniments. Generally, a song is a combination of human vocal singing and music played using string and percussion instruments. Vocal melody has a unique pitch contour, whereas background music is a repetitive rhythm created using a variety of instruments. With respect to a singing voice, which is generally the focus, musical accompaniment can be considered interference or noise because in most cases, the singing voice in a song is the most impressive part to listeners and it conveys abundant important information useful in a wide variety of research; for instance, determining the lyrics [3], language [4], singer [5, 6], and emotion [7] conveyed by a song. Therefore, techniques for separating singing voices from accompaniments are important for various music information retrieval (MIR) applications, such as automatic lyric recognition [3] and alignment [8], melody extraction [9], song language identification [4], singer identification [5, 6], content-based music retrieval, and annotation [10]. Such applications are indispensable in systems such as karaoke gaming, query-by-humming, active music listening [11], and audio remixing.

However, the separation of singing voice from musical accompaniment is genuinely challenging. Human listeners generally have the remarkable ability to segregate sound streams from a mixture of sounds in day-to-day life, but this remains a highly demanding job for

* Correspondence: lwh@nkust.edu.tw
[1]Department of Computer and Communication Engineering, National Kaohsiung University of Science and Technology, Kaohsiung 824005, Taiwan
Full list of author information is available at the end of the article

machines, especially in the monaural case because it lacks the spatial cues that can be acquired when two or more microphones are used. Furthermore, the experience of speech separation may not straightforwardly apply to singing separation. Singing voice and speech, both human sounds, have many similarities, but they are also dissimilar. Therefore, the difficulties encountered in the separation of singing and speech from their respective backgrounds are different. The most important difference between singing and speech in terms of their separation from a background is the nature of the other coexisting sounds. The background interference mixed with speech may be harmonic or nonharmonic, narrowband, or broadband and generally uncorrelated to the speech. However, the musical accompaniment in a song is usually harmonic and broadband, correlated to the singing, and does not fit the general assumptions of noise, such as whiteness or stationarity. Hence, traditional noise-suppression methods are unsuitable.

Additionally, singing voices usually contain clear and strong harmonic structures and rapidly changing harmonics, such as vibratos or slides, and musical accompaniment can be considered the sum of percussive sounds and harmonics. Simple harmonic extraction techniques are not useful for polyphonic mixtures and rapidly changing harmonics because the extraction results are inaccurate and harmonic instruments (not only singing) also contain harmonics. Moreover, onset and offset cues, which are generally useful in auditory scene analysis because different sounds normally start and end at different times, are not useful either because the starting and ending times of singing voices and musical accompaniments usually coincide. In addition, in singing, lyrics are expressed by changing notes according to the melody, which makes singing an interpretation of a predefined musical score. Therefore, pitch in singing tends to be piecewise constant, with abrupt pitch changes and different types of fluctuations. The pitch range of singing could be as high as 1000 or 1400 Hz for soprano singers [12], compared with the normal range of 80 to 400 Hz for speech. Hence, pitch-extraction techniques are commonly inaccurate, and in songs, distinguishing between voiced and unvoiced is problematic.

The separation of singing can be classified into several categories on the basis of the underlying methodologies, namely probabilistic [13, 14], spectrogram factorization [15–24], pitch-based [25–27], repetition-based [28, 29], low-rank and sparse decomposition [30, 31], harmonic/percussive sound separation (HPSS) [32–34], deep neural network (DNN)-based [35–48], and hybrid or fusion approaches.

Among them, by assuming and utilizing the underlying properties of singing and musical accompaniment, Huang et al. [30] performed a robust principal component analysis (RPCA) to decompose the magnitude spectrogram of a song into low-rank and sparse matrices, which correspond to the music accompaniment and singing voice, respectively. Studies have demonstrated that such decomposition methods outperform sophisticated pitch-based methods [31]. However, the assumptions of low-rank and sparsity may not be true in all cases. For example, the sound of drums is sparse but not low-rank, and the vocal part of a song can sometimes be low-rank [31].

The state-of-the-art method is to use a DNN, which learns a model from a large amount of data and has been demonstrated to be particularly successful in the separation of singing voice [35–48]. The research on DNNs includes joint optimization of deep recurrent neural networks (DRNNs) with a masking layer [35–37], the combination of DNNs with spatial covariance matrices [38], deep regression neural networks [39], proximal DRNN (P-DRNN) [40], bi-directional deep LSTM [41], enhanced feature network (EFN) [42], convolutional recurrent neural network (CRNN) [43], and variants of the convolutional neural network (CNN), such as improved MMDenseNet [44], deep U-Net [45], Wave-U-Net [46], evolving multi-resolution pooling CNN (E-MRP-CNN) [47], and extended multiscale DenseNet [48].

In addition to the aforementioned categories of methods, hybrids or fusions of existing building blocks have emerged. Among them, some integrate pitch or F0 information to improve separation. For example, Virtanen et al. [49] proposed a hybrid method that combines pitch-based inference and nonnegative spectrogram factorization. Rafii et al. [50] combined the repetition-based method — repeating pattern extraction technique with a similarity matrix (REPET-SIM), which is a generalization of REPET and uses a similarity matrix to identify the repeating elements of the background music, with a pitch-based method. Ikemiya et al. [51, 52] utilized the mutual dependency of F0 estimation and source separation to improve singing voice separation by combining a time-frequency mask based on RPCA with a mask based on harmonic structures. Other cascading [53] and fusion methods [54] have been proposed as well.

In the real world, learning involves observing large numbers of objects in the world and drawing inferences about their categories, and this is coupled with occasional experiences of supervised learning [55]. In other words, information gleaned from data may create some underlying assumptions or rules and extends the knowledge obtained from labeled data, which helps to improve category learning [56]. Although the manner in which humans combine different ways of learning and jointly exploit different data remains unclear [55], we may assume that humans use underlying knowledge

derived from observation and inferences plus supervised learning for pattern recognition. In our daily experiences, we glean information from a large amount of data to arrive at a reasonable central tendency and draw boundaries between different categories. The hidden structure discovered by the process can be leveraged to obtain a deep insight into the informational content, and the insight can lead to assumptions of the underlying properties, and as a result, no prior training is required. The RPCA approach [30] is a famous example. We, therefore, chose RPCA as a pre-processing method. Then, the following supervised learning can further adjust the results of RPCA to increase their accuracy.

Therefore, our intention is to effectively combine assumptions of the underlying properties with supervised learning to improve the separation of singing voice and background music given by a monaural mixture. Because labeled data are always difficult to obtain and are usually insufficient, employing approaches without prior training for initial separation and then employing supervised learning with limited data based on the initial separation rather than on the original input can help improve the separation quality. Benefit from the initial separation without prior training, our method may achieve good results without data augmentation if the amount of data is not too low and therefore can greatly reduce the computational load. Hence, we propose using RPCA based on the underlying low-rank and sparse properties of accompaniments and vocals, respectively, to achieve the initial separation and apply supervised DRNN to limited data to further separate the results of RPCA in order to further correct misclassified or residual singing and background music from the initial separation.

The remainder of this paper is organized as follows. Section 2 introduces the proposed RPCA-DRNN model, including RPCA, postprocessing (median filter, morphology, and high-pass filter), and the architecture of the DRNN. Section 3 describes the datasets, objective and subjective measures, and experiment results. A comparison of the proposed method with the reference methods is given as well. Finally, conclusions are provided in the final section.

2 Proposed RPCA-DRNN method

Music is usually composed of multiple mixed sounds, such as human vocals and various instrumental sounds. Huang et al. [30] reported that the magnitude spectrogram of a song can be regarded as the superposition of a sparse matrix and a low-rank matrix and can be decomposed by RPCA. The sparse matrix and low-rank matrix decomposed appear to be corresponding to the singing voice and accompaniment. Hence, based on the assumptions of the correspondence of singing with sparse matrix, and low-rank with accompaniment, RPCA can be applied to the singing/accompaniment separation problem. Without any pretraining, its results are superior to those of sophisticated pitch-based methods [31].

However, the underlying low-rank and sparsity assumptions may not be true in all cases. The decomposed sparse matrix may contain instrumental sounds (e.g., percussion) besides singing voice, and the decomposed low-rank matrix may contain vocal besides instrumental sounds. Upon listening to the separated singing voice, it is apparent that there is some residual background music. Likewise, some part of the singing voice is misclassified as background music. Therefore, additional methods or techniques are needed to reclassify the RPCA output to increase the separation accuracy.

We propose an RPCA-DRNN method that employs an RPCA with postprocessing to perform the initial separation and a supervised DRNN to perform the subsequent separation. The mixed signal is input into the RPCA and separated into the sparse and low-rank matrices. Then, postprocessing, including median filter, morphology, and high-pass filter, is applied. The DRNN that follows comprises two jointly optimized parallel-stacked recurrent neural networks (sRNNs) with mask layers. The resulting sparse and low-rank matrices obtained after RPCA and postprocessing are sent to their corresponding sRNNs. One sRNN further separates the sparse matrix into the estimated singing and musical accompaniment parts because there is a residual background music component in the initial separated sparse matrix. Similarly, the other sRNN further separates the low-rank matrix into the estimated singing and musical accompaniment parts because there is a residual singing vocal component in the low-rank matrix. The final estimated singing is the sum of the singing part estimated from the sparse matrix and the residual singing part estimated from the low-rank matrix, and it is compared with the original clean singing voice. Correspondingly, the final estimated musical accompaniment is the sum of the residual musical accompaniment part estimated from the sparse matrix and the musical accompaniment part estimated from the low-rank matrix, and it is compared with the original clean musical accompaniment. By reducing the error between the estimated singing and clean singing parts and that between the estimated musical accompaniment and clean musical accompaniment parts, we can jointly optimize the DRNN and obtain the final model. The time-domain waveform of singing/music is reconstructed by applying inverse short-time Fourier transform (ISTFT) to the estimated magnitude spectrum of singing/music along with the phase spectrum of the sparse/low-rank matrix.

In the following subsections, details of the techniques associated with each part of the proposed method are discussed.

2.1 RPCA

The convex program RPCA was proposed by Candès et al. [57] to recover a low-rank matrix L from highly corrupted measurements $C = L + S$, where S is a sparse matrix and has arbitrary magnitude. The convex optimization problem is defined as

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

where $\|\cdot\|_*$ and $\|\cdot\|_1$ denote the nuclear norm and $l1$-norm (i.e., the sum of the singular values and the sum of absolute values of the matrix entries, respectively). The dimensions of L, S, and C are $m \times n$, and λ is a positive tradeoff parameter that can be selected based on prior knowledge about solutions to practical problems [57]. Generally, $\lambda = 1/\sqrt{\max(m,n)}$, which works well for incoherent matrices [54].

Musical accompaniment generally has an underlying repeating structure and can be considered a low-rank signal L. By contrast, singing voices with more variation have a higher rank and are comparatively sparse in the time and frequency domains; they can be considered sparse signals S [30]. Then, C, L, S, m, and n can be considered as the spectrum of mixture, accompaniment, singing, and the number of frequency bins and frames. Therefore, RPCA can be used to separate singing vocals from a mixture without training.

The separation is performed as follows. First, short-time Fourier transform (STFT) is used to obtain the spectrogram of the mixture C. Then, the inexact augmented Lagrange multiplier (ALM) method [58] is used to obtain L and S.

To improve the quality of the separation results, a binary time-frequency masking is further applied in [30]. The binary time-frequency mask M_b is defined as

$$M_b(i,j) = \begin{cases} 1, & |S(i,j)| > \kappa * |L(i,j)| \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$i = 1 \ldots m$ and $j = 1 \ldots n$. κ is the threshold of the magnitude ratio of sparse to low-rank. When the ratio is greater than the threshold, the binary mask is classified as 1.

However, the use of a soft mask in REPET can marginally improve the quality of the overall results (only statistically significant for the source-to-artifact ratio (SAR) of a singing voice), except for the source-to-interference ratio (SIR) of the singing voice [59]. Moreover, the experiments of [29] revealed that the use of a soft mask is perceptually better than the use of a binary mask. Therefore, the following soft mask M_s is adopted in the proposed method:

$$M_s(i,j) = \frac{|S(i,j)|}{|S(i,j)| + gain * |L(i,j)|} \tag{3}$$

RPCA is a method that does not need any training or labeled data, and hence, it is convenient to use. Nevertheless, the sparse and low-rank assumptions are rather strong assumptions, and they may not be suitable for every situation. For example, the sound of drums is sparse and can be classified as a singing voice, and the vocal part can sometimes be classified as low-rank. The decomposed low-rank matrix might be a mixture of singing and instrumental sounds, and the decomposed sparse matrix might be a mixture of vocal and percussion sounds [31]. It is even more bothering for separation when the low-rank matrix contains a non-vocal, harmonic instrument (such as electric guitar and string instrument). From the study of Yang [60], the sparse signals generated by RPCA also often contain percussive components. It is because percussive sound can be considered as a periodic stream that is sparse in time-domain [61]. Moreover, RPCA does not consider other information, such as pitch or structure information. The output of RPCA—the separated singing—still contains some background music, and the separated music still contains singing. Thus, the quality of such RPCA separation is limited, and other methods must be employed to improve the results.

2.2 Postprocessing

The soft mask generated from RPCA is postprocessed to improve the separation performance. The postprocessing is applied to the soft mask instead of L and S, so the sum of the obtained low-rank and sparse matrices are still equal to the mixture. The postprocessing includes median filtering, morphology, and high-pass filtering, as depicted in Fig. 1. Between the spectrum of clean singing and the resultant singing spectrum obtained using RPCA, the former is more spotless and has a clearer structure. By contrast, the spectrum of estimated singing contains noise, has a broken structure, and has very-low-frequency parts that seldom appear in vocals. Hence, postprocessing is needed to further improve the mask.

2.2.1 Median filter

Median filter, a widely used nonlinear digital filtering technique in image processing and sound separation,

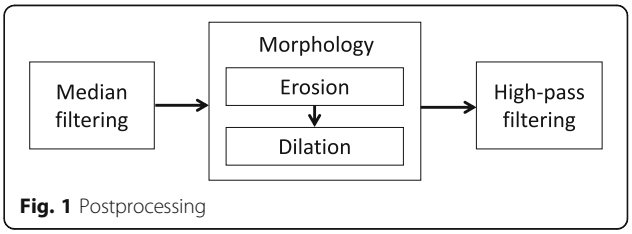

Fig. 1 Postprocessing

especially for separating harmonic from percussive sounds [62], is applied to remove noise from the soft mask M_s. As the low-rank hypothesis barely holds for drum sounds, a median filter is applied to enhance the separation. The procedure of the two-dimensional median filter is to run through the time-frequency unit of the mask unit-by-unit and replace each unit with the median value of the neighboring d_m by d_n units belonging to a window sliding over the mask unit-by-unit. The soft mask after the processing of the median filter is M_{sm}.

2.2.2 Morphology

Morphology [63] is a set of popular operations in image processing that is employed to process images based on predefined kernels or structuring elements. Two very common morphology operators, erosion and dilation [63], with predefined structuring elements, are applied to the soft mask M_{sm} to enhance the possible singing spectrum pattern. By creating a structuring element of a certain size and shape, operations sensitive to specific shapes can be constructed. A structuring element defines common shapes, such as lines or circles, and is represented by a matrix of 0 s and 1 s, where 1 symbolizes the neighborhood. The center of the structuring element slides through the pixel being processed. First, normalization is performed to transform the soft mask into a grayscale image. Then, grayscale erosion is performed, followed by grayscale dilation. The erosion operation outputs the minimum neighborhood value of the pixels that fall within the predefined structuring elements. The dilation operation, by contrast, outputs the maximum neighborhood value. Considering the original singing spectrogram contains horizontal line-like structures as in Fig. 2a, and the horizontal line structures in the singing spectrogram after RPCA with soft mask and median filter as in Fig. 2b are broken, a line-structuring element of length *len* and degree θ, as shown in Fig. 3a, is applied to both the erosion and dilation operations. Figure 3b is a schematic diagram of erosion and dilation by using a line-structuring element of length 10 and degree 5 on a binary example. From Fig. 3b, observing the places circled, after the erosion and dilation, the small gap on the horizontal line is patched up. With D_K representing the domain of the kernel (the structuring element) K, grayscale erosion is defined and performed as follows:

$$(M_{sm} \ominus K)(i,j) = \min\{M_{sm}(i + i', j + j') | (i', j') \in \mathrm{D}_K\}, \quad (4)$$

which is equivalent to a local-minimum operator. $\ominus$ is an erosion operator.

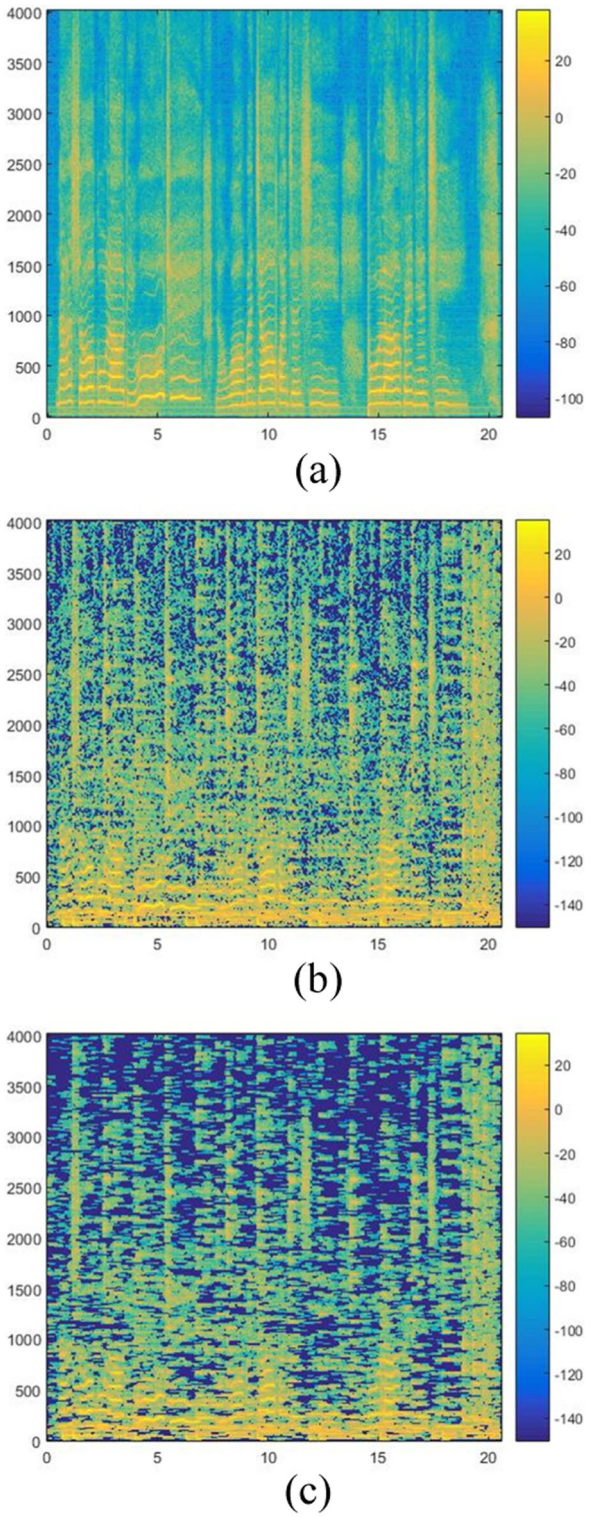

Fig. 2 **A** An example of original singing spectrogram. **B** Singing spectrogram after RPCA with soft mask and median filter. **c** Singing spectrogram after applying morphology to the soft mask M_{sm}

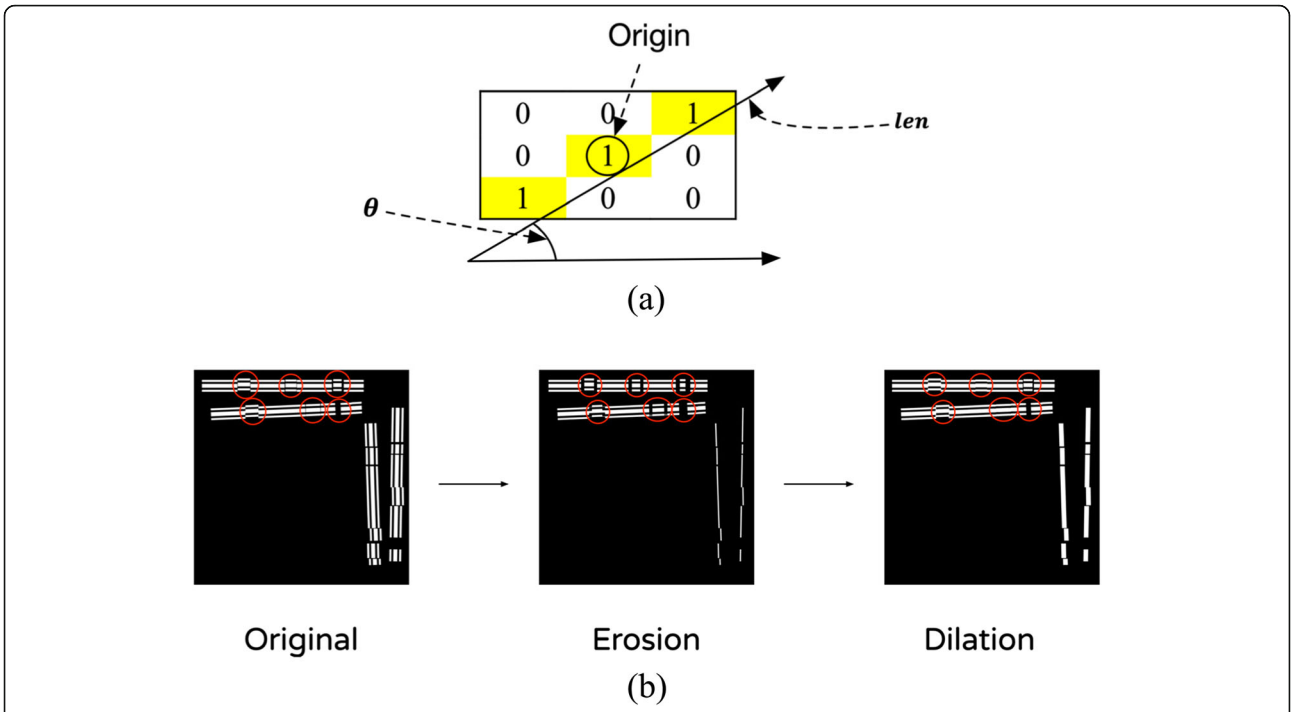

Fig. 3 Application of **a** a line-structuring element to **b** the erosion and dilation operations

By contrast, grayscale dilation is equivalent to a local-maximum operator and is defined as follows:

$$(M_{sm} \oplus K) \times (i,j) = \max\left\{ M_{sm}(i-i', j-j') | (i', j') \in D_K \right\}. \quad (5)$$

$\oplus$ is a dilation operator.

By using the erosion and dilation operations, the skeleton of the horizontal line structure of the singing spectrum can be reconstructed. Observing the singing spectrogram after applying morphology as in Fig. 2c, the line structures are rebuilt.

2.2.3 High-pass filter

Liutkus et al. [59] demonstrated that the application of a high-pass filter at the cutoff frequency of 100 Hz to the estimated singing voice yields overall statistically superior results, except for SAR. Therefore, we adopt the same filtering scheme in the postprocessing of the vocal estimate because the frequency of a singing voice is rarely lower than 100 Hz.

2.3 DRNN

Although the RPCA with postprocessing can separate a mixture into singing voice and background music through the processed mask, the estimated singing voice is doped with background music. Similarly, the estimated background music is doped with vocal melody. Therefore, it is necessary to use other techniques to generate a model for suitably reclassifying the doped part as either singing voice or background music.

Neural networks can effectively perform this separation. Among neural networks, recurrent neural networks (RNNs), which introduce the memory from previous time steps, are widely used to model the temporal information in time-series signals, such as audio or speech [64]. However, in the current time step, there is only one layer between the input information and the output. If hierarchical processing or multiple time scales are needed for processing the time series, RNNs do not support such operations. To solve the problem, a DRNN is proposed for performing hierarchical processing and capturing the structure of the time series [65].

The architecture of the DRNN [37], which is conceptually a combination of DNN and RNN, is a multilayer perceptron in which each layer is equivalent to an RNN—each layer has temporal feedback loops. Hermans and Schrauwen [65] demonstrated that a DRNN generates diverse time scales at different levels. Therefore, it can capture a time series more inherently. The architecture of the DRNN can be represented with hidden layers. Temporally recurrent connections can happen in all layers, as in the sRNN [66], or in a single layer. We used an sRNN in the experiments conducted herein.

However, using recurrent connections at a single layer is a possible choice as well.

2.3.1 sRNN

sRNNs, which have multiple levels of transition functions, can be presented as follows:

$$h_t^l = f_h\left(h_t^{l-1}, h_{t-1}^l\right) = \varnothing_l\left(U^l h_{t-1}^l + W^l h_t^{l-1}\right), \tag{6}$$

where h_t^l is the hidden activation of the lth layer at time t, and h_t^0 is equal to the input; f_h is a state transition function; $\varnothing_l(\cdot)$ is an element-wise nonlinear function in the lth layer; W^l is the weight matrix of the lth layer, which is multiplied with the activation of the l-1 layer h_t^{l-1}; and U^l is the weight matrix of the recurrent connection in the lth layer, which is multiplied with the activation of layer l at time t-1 h_{t-1}^l.

2.3.2 Gated recurrent unit

Instead of using a traditional nonlinear function unit, such as sigmoid, tanh, or rectified linear unit (ReLU) in the sRNN, we use the gated recurrent unit (GRU) as the hidden unit. The GRU [67] is a variant of the long short-term memory (LSTM) unit, and it combines the forget and input gates into a single update gate and is simpler to compute and implement. Chung et al. [68] reported that in polyphonic music modeling and speech signal modeling, the performance of the GRU is comparable to that of an LSTM unit and superior to that of the traditional unit tanh.

2.3.3 Proposed model architecture

The architecture of the proposed method is depicted in Fig. 4. A DRNN comprising two jointly optimized parallel sRNNs with mask layers, which are not trainable and just arithmetic operations on the network outputs, is used to further improve the results of RPCA with postprocessing. The inputs to the DRNN are S_t and L_t, which are the sparse and low-rank magnitude spectra obtained from the RPCA output after postprocessing at time t.

The output predictions $\hat{V}_{St}\epsilon\mathbb{R}^F$ and $\hat{B}_{St}\epsilon\mathbb{R}^F$ represent the predicted magnitude spectra of the singing voice and residual background music separated from S, respectively, and $\hat{B}_{Lt}\epsilon\mathbb{R}^F$ and $\hat{V}_{L,t}\epsilon\mathbb{R}^F$ represent the predicted magnitude spectra of the background music and residual singing voice separated from L, respectively, where F is the dimension of the magnitude spectra. The ReLU is used as the activation function in the output layer owing to its advantages of efficient computation, scale invariance, superior gradient propagation, biological plausibility, and sparse activation.

Huang et al. [35] used a time-frequency mask to further smooth the separation outcomes and enforce the constraint that the sum of the separated components is equal to the original unseparated signal. They incorporated the mask as a layer in the neural network to make sure that the DRNN is optimized based on the masked

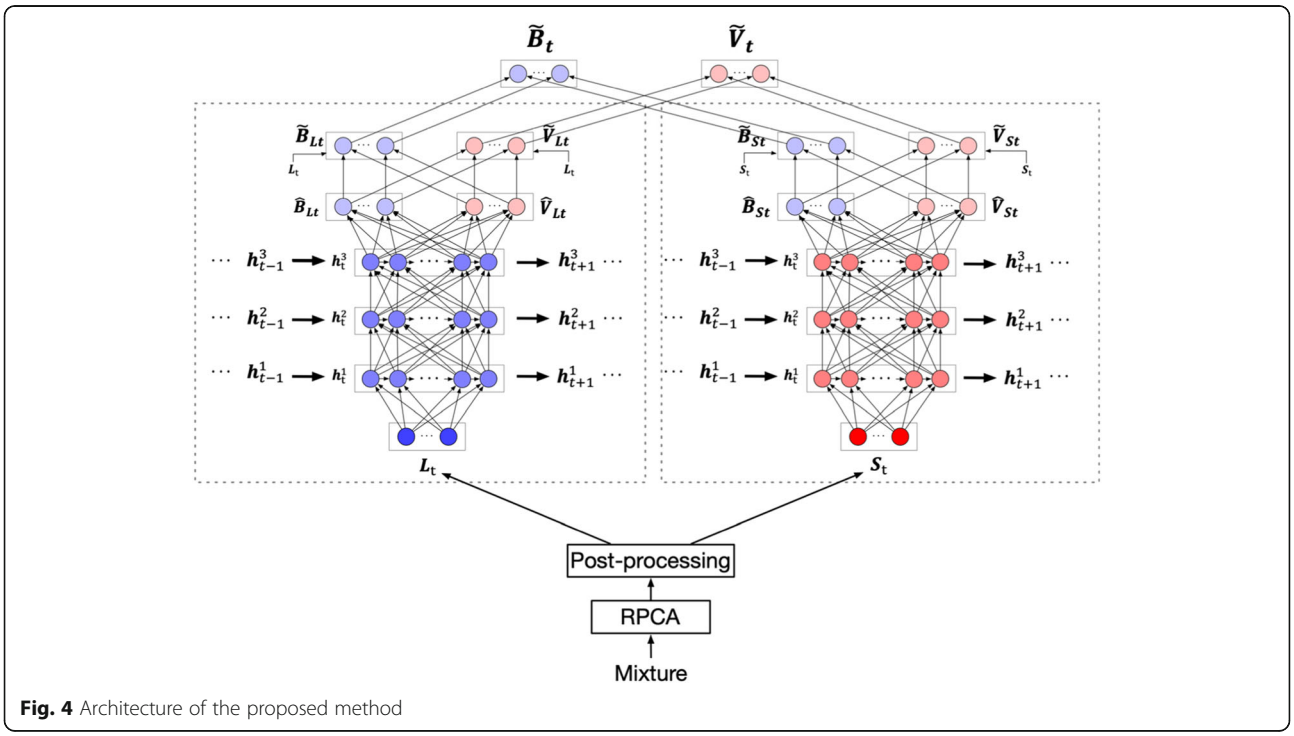

Fig. 4 Architecture of the proposed method

output. Hence, a mask layer is added to each sRNN, as in the architecture depicted in Fig. 4. The predicted magnitude spectra of the singing voice $\tilde{V}_{St}$ and the residual background music $\tilde{B}_{St}$ separated from S_t by incorporating the mask concept can be respectively expressed as follows:

$$\tilde{V}_{St} = \frac{|\hat{V}_{St}|}{|\hat{B}_{St}| + |\hat{V}_{St}|} \odot S_t, \tag{7}$$

$$\tilde{B}_{st} = \frac{|\hat{B}_{st}|}{|\hat{B}_{st}| + |\hat{V}_{st}|} \odot S_t. \tag{8}$$

Likewise, the predicted magnitude spectra of the background music $\tilde{B}_{Lt}$ and the residual singing voice $\tilde{V}_{Lt}$ separated from L_t by incorporating the mask concept can, respectively, be expressed as follows:

$$\tilde{V}_{Lt} = \frac{|\hat{V}_{Lt}|}{|\hat{B}_{Lt}| + |\hat{V}_{Lt}|} \odot L_t, \tag{9}$$

$$\tilde{B}_{Lt} = \frac{|\hat{B}_{Lt}|}{|\hat{B}_{Lt}| + |\hat{V}_{Lt}|} \odot L_t. \tag{10}$$

The final estimated singing $\tilde{V}_t$ is the sum of $\tilde{V}_{St}$ and $\tilde{V}_{Lt}$, as expressed in (11), and it is compared with the original clean singing voice. The final estimated musical accompaniment is the sum of $\tilde{B}_{St}$ and $\tilde{B}_{Lt}$, as expressed in (12), and it is compared with the original clean musical accompaniment. Therefore, an extra layer to perform the summation is added, as depicted in Fig. 4.

$$\tilde{V}_t = \tilde{V}_{St} + \tilde{V}_{Lt} \tag{11}$$

$$\tilde{B}_t = \tilde{B}_{St} + \tilde{B}_{Lt} \tag{12}$$

2.3.4 Discriminative training

The neural network is optimized by minimizing the sum of the squared errors between the estimated and clean singing voices and those between the estimated and clean background musical accompaniments. Moreover, two discriminative terms [35] are added to further penalize the interferences from other sources.

The loss function is defined as follows:

$$J_{DISCRIM} = \frac{1}{2}\sum_{t=1}^{T} \begin{pmatrix} \left\|\tilde{V}_t - V_t\right\|^2 - \omega_{dis}\left\|V_t - \tilde{B}_t\right\|^2 \\ + \left\|\tilde{B}_t - B_t\right\|^2 - \omega_{dis}\left\|B_t - \tilde{V}_t\right\|^2 \end{pmatrix} \tag{13}$$

where T represents the length of an input sequence, and $0 \le \omega_{dis} \le 1$. The output targets $V_t \in \mathbb{R}^F$ and $B_t \in \mathbb{R}^F$ represent the clean magnitude spectra of the singing voice and background music at time t, respectively. In (13), $\left\|\tilde{V}_t - V_t\right\|^2$ and $\left\|\tilde{B}_t - B_t\right\|^2$ are subloss terms to penalize the deviation between the final estimated and clean singing voices and that between the final estimated and clean background musical accompaniments; moreover, $-\omega_{dis}\left\|V_t - \tilde{B}_t\right\|^2$ and $-\omega_{dis}\left\|B_t - \tilde{V}_t\right\|^2$ are discriminative terms to further penalize the interference from other sources. The term ω_{dis} is a weight to control the prominence of the discriminative terms.

3 Experiment results and evaluations

Three datasets, including MIR-1K [15], an amateur Chinese karaoke set, ccMixter [69], gathered from ccmixter.org, and MUSDB18 [70], a professionally produced set, were used, and ten existing source separation techniques were evaluated and compared in our experiments.

3.1 Dataset

MIR-1K, ccMixter, and MUSDB18 are used in our experiments. The MIR-1K dataset was developed by Jyh-shing Roger Jang [15]. This dataset consists of 110 Chinese karaoke songs performed by 11 male and 8 female amateurs. These songs are split into 1000 song clips with durations ranging from 4 to 13 s. The sampling rate is 16 kHz, and each sample occupies 16 bits. Each clip is composed of singing voices and background music in different channels. The mixture is generated by mixing the singing voice and background music with the same energy—with a signal-to-noise ratio equaling 0 dB. One hundred and seventy-five clips with a total length of 23 min 36 s sung by one male and one female singer were used as the training set. The remaining 825 clips with a total length of 1 h 49 min 49 s and sung by ten male and seven female singers were used as the test set. Because the training dataset was not large and may have lacked variety, we repeatedly shifted the background music by 10,000 samples in each instance and mixed the shifted background music with the singing voice to create a more diverse mixture. After the above-described circular shift, the total length of the clips in the training dataset was 5 h 8 min 7 s.

The ccMixter dataset [69] contains 50 full-length tracks with many different musical genres. Each track ranges from 1 m 17 s to 7 m 36 s. The total length is 3 h 12 m 48 s. In training and testing, 40 and 10 tracks are used, and the lengths are 2 h 33 m 34 s and 39 m 13 s, respectively. The sampling rate 16 kHz, down-sampled from 44.1 kHz, is used.

The MUSDB18 dataset was developed by Rafii et al. [70] and released by the 2018 community-based signal separation evaluation campaign (SiSEC 2018), which aims to compare the performance of source separation systems on the same data and metrics. The songs in MUSDB18 are stereophonic mixture. Each song contains four instrumental categories, namely vocals, bass, drums, and others. The MUSDB18 dataset contains a total of 150 songs of different genres, with 100 of them used for

training and 50 for testing. The total length of MUSDB18 is 11 h 14 min 15 s, of which the total length of training is 7 h 46 min 24 s and the total length of testing is 3 h 27 min 50 s. The sampling rate is 44.1 Hz. In monaural singing voice separation, sources other than vocal will be treated as accompaniment. We estimate vocal and accompaniment from left and right channels of the mixtures of MUSDB18, respectively.

3.2 Evaluated techniques

In the experiment of MIR-1K, ten source separation techniques were evaluated, including the proposed method, RPCA with the proposed postprocessing, and eight popular reference methods, namely RPCA [30], sRNN, multiple low-rank representation (MLRR) [31], robust low-rank nonnegative matrix factorization (RNMF) [24], modified group delay (MOD-GD) [37], U-Net [45], EFN [42], and CRNN with an attention (CRNN-A) [43]. In the experiment of MUSDB18, three source separation techniques were evaluated, including the proposed methods RPCA-DRNN and its light and feather versions, Open-Unmix [41], and E-MRP-CNN [47]. The thirteen evaluated techniques and versions are defined as follows.

RPCA_b: implemented reference method; the RPCA technique with a binary time-frequency mask (with κ = 1). The window size and hop size in STFT are 1024 and 256. $\lambda = 1/\sqrt{\max(m,n)}$.

sRNN: implemented reference method; the sRNN technique with model architecture, as shown in Fig. 5. In the last two layers of sRNN, two branches are used, so we can optimize the output of the network to be as close as possible to clean vocals and clean accompaniment at the same time. The GRU is used as the hidden unit. Joint mask optimization, as presented in (7–10), and discriminative training ($\omega_{dis} = 0.5$), as described in Section 2, are applied as well. The sRNN architecture contains three hidden layers, with 1000 neurons per layer. The input spectrum is calculated by carrying out a 1024-point STFT with a hop size of 256. The Adam optimization [68] is used. The batch size is 64. The learning rate is 0.0001. The global step 100,000 is used as a stop criterium.

MLRR: reference method proposed in [31]. MLRR considers both the vocal and instrumental spectrograms as low-rank matrices and uses the learned dictionaries for decomposition. The results were directly reported from the literature.

RNMF: reference method proposed in [24]. RNMF is a nonnegative variant of RPCA. The results were directly reported from [35].

MOD-GD: reference method proposed in [37]. MOD-GD function for learning the time-frequency masks of the sources is used. The results were directly reported from the literature of the 2-DRNN architecture.

U-Net: reference method proposed in [45]. U-Net is a convolutional network initially developed for biomedical image. The results were directly reported from [43].

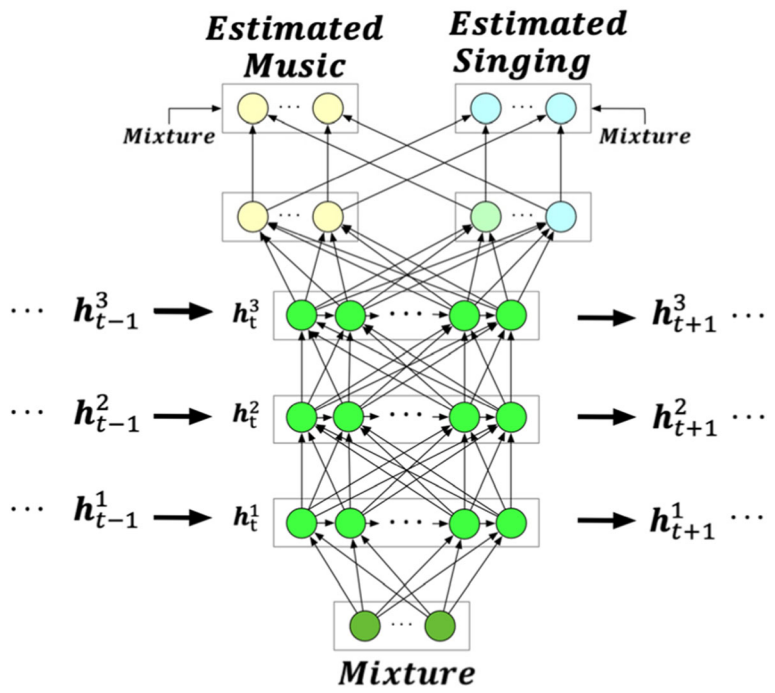

Fig. 5 The sRNN model

EFN: reference method proposed in [42]. EFN is an end-to-end framework to extract effective representations of the magnitude spectra. The results were directly reported from the literature of the model with GRU.

CRNN-A: reimplemented reference method proposed in [43]. CRNN-A uses a CNN as the front-end of an RNN. The objective result was directly reported from the literature and the subjective result was obtained from our implementation.

Open-Unmix: reference method proposed in [41]. Open-Unmix is based on the bi-directional LSTM model. It used stronger data augmentation methods on MUSDB18 than E-MRP-CNN [47]. Normalization and input/output scalar is also used. The results were directly reported from the literature.

E-MRP-CNN: reference method proposed in [47]. E-MRP-CNN automatically searches for effective MRP-CNN structures using genetic algorithms. Gain and sliding data augmentation is used on MUSDB18. The ratio of the augmentation is four times of the original data. The results were directly reported from the literature of model S-17-1-MUS.

RPCA-DRNN: the proposed method that uses RPCA with soft mask, medium filter, morphology, and high-pass filter followed by a DRNN that contains two parallel sRNNs to further correct the residual singing voice and music output. The window size and hop size in STFT are 1024 and 256, respectively. $\lambda = 1/\sqrt{\max(m,n)}$, and κ = 1 in RPCA. The expression d_m = d_n = 3 is set in the median filter. A line-structuring element of length 10 and degree 5 is applied to both the erosion and dilation operations to recover the horizontal line structures. The choice is made from experiments. The cutoff frequency of the high-pass filter is set to 100 Hz. The weight ω_{dis} in discriminative training equals 0.5. To enable a fair comparison with the sRNN reference method, there are three hidden layers with 500 neurons per layer for each sRNN. The Adam optimization [71] is used. The batch size is 64. The learning rate is 0.0001. The global step 100,000 is used as a stop criterium. In the experiment of MIR-1K, 1024-point STFT with a hop size of 256 and circular shifting data augmentation are applied. In the experiment of MUSDB18, to be compared with Open-Unmix and E-MRP-CNN, the window size and hop size of STFT are 5644 and 1411. No data augmentation is applied.

RPCA-DRNN$_l$: a light version of RPCA-DRNN, which uses only 20 songs of MUSDB18 without data augmentation for training. The total length of the training data was 53 min 45 s, about 0.115 times of the length of the original training dataset of MUSDB18.

RPCA-DRNN$_f$: a feather version of RPCA-DRNN, which uses only 5 songs of MUSDB18 with data augmentation for training. After the circular shift augmentation, the total length was 54 min 28 s.

3.3 Objective measures

There are four fundamental metrics, namely source-to-distortion ratio (SDR), image-to-spatial distortion ratio (ISR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR) [72–74], and a derived metric, namely normalized SDR (NSDR) [13]. To compare the performance with other existing source separation systems on the same data and metrics, we used the overall performance metrics, namely global NSDR (GNSDR), global SIR (GSIR), and global SAR (GSAR) [35, 47], to objectively measure the performance of the evaluation methods considered in the experiment of MIR-1K and ccMixter. Based on the same reason, BSS Eval version 4 [74] of SDR, ISR, SIR, SAR, the evaluation metrics released by SiSEC, is used in the experiment of MUSDB18, which is a dataset also released by SiSEC.

Assume an estimated source signal in separation is $\hat{s}(t)$. This signal is the same as the clean source signal $s(t)$ plus the spatial distortion $e_{spat}(t)$ [75], the interference error $e_{interf}(t)$, and the artifact error $e_{artif}(t)$, as presented in (14).

$$\hat{s}(t) = s(t) + e_{spat}(t) + e_{interf}(t) + e_{artif}(t). \tag{14}$$

The metrics SDR, SIR, and SAR are defined as follows:

$$SDR(\hat{s}, s) = 10\log_{10}\frac{\|s\|^2}{\left\|e_{spat} + e_{interf} + e_{artif}\right\|^2} \tag{15}$$

$$ISR(\hat{s}, s) = 10\log_{10}\frac{\|s\|^2}{\left\|e_{spat}\right\|^2}, \tag{16}$$

$$SIR(\hat{s}, s) = 10\log_{10}\frac{\left\|s + e_{spat}\right\|^2}{\left\|e_{interf}\right\|^2}, \tag{17}$$

$$SAR(\hat{s}, s) = 10\log_{10}\frac{\left\|s + e_{spat} + e_{interf}\right\|^2}{\left\|e_{artif}\right\|^2}, \tag{18}$$

where $\|\cdot\|$ represents the Euclidean norm.

The other metrics derived from SDR, SAR, and SIR are NSDR, GNSDR, GSIR, and GSAR. NSDR [13] calculates the difference in SDR between the mixture and the separated singing voice as in (19). It can be considered as the improvement of SDR owing to the adoption of the separation technique. GNSDR, GSIR, and GSAR [35] are the length-weighted means of NSDR, SIR, and SAR, respectively, as expressed in (19–22).

$$NSDR(\hat{v}_k, v_k, c_k) = SDR(\hat{v}_k, v_k) - SDR(c_k, v_k), \tag{19}$$

$$GNSDR = \frac{\sum_k l_k NSDR(\hat{v}_k, v_k, c_k)}{\sum_k l_k}, \tag{20}$$

$$GSAR = \frac{\sum_k l_k SAR(\hat{v}_k, v_k)}{\sum_k l_k}, \tag{21}$$

$$GSIR = \frac{\sum_k l_k SIR(\hat{v}_k, v_k)}{\sum_k l_k}, \tag{22}$$

where k is the song index and $\hat{v}_k$, v_k, c_k, and l_k are the estimated singing voice, clean singing voice, mixture, and song length of the kth song, respectively. GNSDR, GSIR, and GSAR are adopted as objective measures in the experiment. From (15–22), the higher the values of SDR, ISR, SIR, SAR, NSDR, GNSDR, GSIR, and GSAR are, the better is the separation performance.

3.4 Subjective measures

The objective measures GNSDR, GSIR, and GSAR help to objectively compare the separation quality of our proposed methods and the reference methods. Higher values of GNSDR, GSIR, and GSAR for the estimated separated singing voice represent closeness to the original clean singing voice. However, the estimated separated singing voice with the highest objective measure scores is not necessarily perceived as the cleanest separation. Therefore, subjective measures, by asking listeners to consider interference (the residue of background music) and artifacts in the separated singing voice, are applied. Two subjective measures, namely mean opinion score (MOS) [76] and comparison mean opinion score (CMOS) [76], are adopted to this end.

The MOS is commonly used in audio and video analysis. The absolute category rating scale is often used, typically in the range of 1–5, which represents the ratings of bad, poor, fair, good, and excellent. Given the difficulty of absolute grading of subjective perceptions of the MOS for separated singing, the CMOS measure prescribed in Annex E of the ITU-T Recommendation P.800 [76] is additionally used as another subjective measure to evaluate the separation quality. In CMOS, listeners listen to and compare the target separated singing voice with the reference voice. Scores ranging from − 3 to + 3, totaling seven levels of assessment, can be assigned. To reduce the difficulty associated with distinguishing subtle differences and grading for evaluators, we reduced the number of assessment levels to 5, where the scores − 2, − 1, 0, 1, and 2 for the singing voice represent the ratings of worse, slightly worse, equal, slightly better, and better, respectively, than the reference sound.

3.5 Experiment results of ablation study

To compare the performances of each part of RPCA-DRNN, an ablation study that removes some part of the system was built. Three datasets including MIR-1K, ccMixter, and MUSDB18 are used. Table 1 lists all the combinations of different parts of RPCA-DRNN. These combinations will be included in our ablation experiment and the experiment results are shown in Table 2.

Observing the experiment results of Table 2 (parts a and b), adding any postprocessing helps reduce the total error (the sum of the interference and artifact error). Among the three steps of postprocessing (median filter, morphology, and high-pass filter), high-pass filter reduces the total error most. It can be observed that RPCA_s_h performed better in GNSDR than RPCA_s_m and RPCA_s_M, and RPCA_s_h-DRNN performed better in GNSDR than RPCA_s_m-DRNN and RPCA_s_M-DRNN. In addition, the combinations without high-pass filter performed worst in reducing the total error. It can be observed that RPCA_s_m_M was worse in GNSDR than RPCA_s_m_h and RPCA_s_M_h, and RPCA_s_m_M-DRNN was worse in GNSDR than RPCA_s_m_h-DRNN and RPCA_s_M_h-DRNN. sRNN outperformed all the combinations without DRNN in GNSDR. At last, the proposed RPCA-DRNN beats all the combinations in GNSDR and GSAR, and beats RPCA_b-DRNN and RPCA_s-DRNN in all the objective measures. Therefore, RPCA-DRNN performs better than conventional RPCA and sRNN, and taking the postprocessing on the soft mask does improve the separated quality.

Besides, observing the experiment results of Table 2 (part c) on MUSDB18, the combinations without high-pass filter performed worst in SDR, ISR, SIR, and SAR. It can be observed that RPCA_s_m_M-DRNN was worse in all of the four measures than RPCA_s_m_h-DRNN and RPCA_s_M_h-DRNN. At last, the proposed RPCA-DRNN beats all in all of the four measures. Therefore, on MUSDB18, it is confirmed again that RPCA-DRNN performs better than sRNN and high-pass filter is the most influential part in postprocessing.

3.6 Experiment results of MIR-1K

Songs from the test set were used in the objective and subjective tests. Both the separated singing voice and accompaniment are evaluated. Ten techniques, namely RPCA_b, RPCA_p, sRNN, MLRR, RNMF, MOD-GD, U-Net, EFN, CRNN-A, and RPCA-DRNN, were compared in our objective experiment. Accordingly, ten varieties of separated singing voices were evaluated in the objective voice quality assessment. The comparison of the

Table 1 Combinations of different parts of RPCA-DRNN

Method	Definition
RPCA_s	RPCA with soft mask
RPCA_s_m	RPCA with soft mask and medium filter
RPCA_s_M	RPCA with soft mask and morphology
RPCA_s_h	RPCA with soft mask and high-pass filter
RPCA_s_m_M	RPCA with soft mask, medium filter, and morphology
RPCA_s_m_h	RPCA with soft mask, medium filter, and high-pass filter
RPCA_s_M_h	RPCA with soft mask, morphology, and high-pass filter
RPCA_p	RPCA with soft mask, medium filter, morphology, and high-pass filter
RPCA_b-DRNN	RPCA with binary mask and DRNN
RPCA_s-DRNN	RPCA with soft mask and DRNN
RPCA_s_m-DRNN	RPCA with soft mask, medium filter, and DRNN
RPCA_s_M-DRNN	RPCA with soft mask, morphology, and DRNN
RPCA_s_h-DRNN	RPCA with soft mask, high-pass filter, and DRNN
RPCA_s_m_M-DRNN	RPCA with soft mask, medium filter, morphology, and DRNN
RPCA_s_m_h-DRNN	RPCA with soft mask, medium filter, high-pass filter, and DRNN
RPCA_s_M_h-DRNN	RPCA with soft mask, morphology, high-pass filter, and DRNN

proposed method RPCA-DRNN with RPCA_b, RPCA_p, sRNN, MLRR, RNMF, Mod-GD, U-Net, EFN, and CRNN-A in terms of the objective measures GNSDR, GSIR, and GSAR is summarized in Table 3. The results indicate that the proposed method RPCA-DRNN is superior to all of the reference methods in GNSDR and GSAR. Therefore, RPCA-DRNN can reduce the total error most, but respectively speaking, it is more successful in reducing artifact than interference error. The box plots of the comparison of the proposed method RPCA-DRNN with RPCA_b, RPCA_p, sRNN, and CRNN-A are presented in Fig. 6, which shows a clearer statistical insight.

RPCA-DRNN, CRNN-A, and sRNN were further compared in a subjective assessment. In the subjective test, there were ten listeners. All of them are music enthusiasts but are not familiar with source separation or audio engineering. Each listener was allotted ten sets of RPCA-DRNN, CRNN-A, and sRNN separated singing voice clips from different songs in the test set. In total, there were 100 testing sets of singing voice clips. The ordering of the target and reference voices was changed randomly and was not revealed to the evaluators. The listeners were asked to evaluate the separation performance and provide MOS scores. Table 4 shows the percentage distribution of the MOS scores assigned to the singing voices separated using sRNN, CRNN-A, and the proposed RPCA-DRNN. Sixty-seven percent, 51%, and 43% of the singing voice clips separated using RPCA-DRNN, CRNN-A, and sRNN were rated good or excellent, respectively. Thus, it is clear that the percentages of good and excellent scores of RPCA-DRNN were higher than those of CRNN-A and sRNN (67% vs. 51% and 43%). Furthermore, the average MOS of RPCA-DRNN was 3.79, whereas those of CRNN-A and sRNN were 3.54 and 3.46. Therefore, the subjective performance of the proposed RPCA-DRNN method in terms of the MOS scores was superior to that of the CRNN-A and sRNN.

A further analysis of the separated singing voice clips was then conducted by performing a CMOS test to measure the subjective quality of separation. Two target–reference pairs, RPCA-DRNN vs. sRNN and RPCA-DRNN vs. CRNN-A, were used. The same ten listeners and 100 testing sets as those in the MOS test were used, but different testing sets were allocated to each listener. The percentages of RPCA-DRNN-separated clips assigned "worse," "slightly worse," "equal," "slightly better," and "better" CMOS scores compared with the percentages of each class of score for sRNN and CRNN-A separated clips assigned are listed in Table 5. The results in Table 5 indicate that RPCA-DRNN was preferred compared to sRNN and CRNN-A. For sRNN, based on the vote percentages, RPCA-DRNN was preferred (slightly better and better) for 70% of testing pairs, but 18% of the listeners perceived its output to be indistinguishable from that of sRNN. By contrast, sRNN was preferred by only 12% of the listeners. For CRNN-A, RPCA-DRNN was voted 75% equal, slightly better or better.

To ensure that the results of our subjective auditory tests in Table 5 are statistically significant and in support of our argument, we examined the p-values of the binomial statistic in addition to the preferred percentage. Given that five options were available to the listeners

Table 2 Experiment results of the ablation study using (a) MIR-1K, (b) ccMixter, and (c) MUSDB18

	Vocal			Accompaniment		
(a)						
Method	**GNSDR**	**GSIR**	**GSAR**	**GNSDR**	**GSIR**	**GSAR**
RPCA_b	2.72	3.10	5.90	2.87	10.02	4.88
RPCA_s	2.94	1.55	9.97	2.80	7.80	11.21
RPCA_s_m	3.12	1.85	9.63	4.01	5.25	10.71
RPCA_s_M	3.18	6.82	3.24	4.41	10.21	5.25
RPCA_s_h	4.16	3.53	9.36	5.24	6.67	9.67
RPCA_s_m_M	3.41	2.56	8.54	5.02	9.40	5.70
RPCA_s_m_h	4.36	3.83	9.26	5.87	7.47	10.66
RPCA_s_M_h	4.33	5.52	6.71	5.21	9.71	8.02
RPCA_p	4.76	4.44	8.19	4.91	**14.03**	5.80
RPCA_b-DRNN	5.17	5.88	6.10	5.54	10.98	5.02
RPCA_s-DRNN	5.43	7.38	7.01	5.57	9.23	9.54
RPCA_s_m-DRNN	6.44	6.81	7.56	6.37	7.40	9.88
RPCA_s_M-DRNN	6.01	**8.16**	6.96	6.04	9.70	7.61
RPCA_s_h-DRNN	7.14	7.46	9.41	6.85	8.13	11.83
RPCA_s_m_M-DRNN	7.07	6.29	9.37	6.26	6.88	10.44
RPCA_s_m_h-DRNN	7.26	6.45	8.37	6.94	8.30	9.25
RPCA_s_M_h-DRNN	7.40	6.88	9.91	7.28	11.48	9.88
sRNN	6.43	7.69	6.82	6.20	8.76	6.46
RPCA-DRNN	**8.46**	7.72	**10.83**	**8.02**	12.32	**11.99**
(b)						
RPCA_b	1.89	1.02	6.52	3.48	9.12	4.62
RPCA_s	2.04	1.10	9.22	4.28	9.47	8.31
RPCA_s_m	2.99	1.57	8.91	5.13	9.12	9.47
RPCA_s_M	2.98	5.77	3.66	4.65	11.71	7.97
RPCA_s_h	4.35	2.01	6.31	5.81	8.81	11.44
RPCA_s_m_M	4.06	3.30	8.32	5.02	6.69	12.10
RPCA_s_m_h	4.54	2.19	7.90	5.96	8.92	12.72
RPCA_s_M_h	4.29	7.17	3.84	6.19	12.49	8.42
RPCA_p	5.53	4.98	6.23	7.15	8.59	11.48
RPCA_b-DRNN	2.75	3.68	7.81	4.03	10.68	5.01
RPCA_s-DRNN	2.94	3.91	8.51	4.19	12.29	4.85
RPCA_s_m-DRNN	4.87	5.66	9.20	6.09	7.22	11.35
RPCA_s_M-DRNN	5.01	2.55	5.92	6.28	11.06	9.31
RPCA_s_h-DRNN	6.53	3.76	7.27	6.82	9.28	10.95
RPCA_s_m_M-DRNN	5.15	4.57	6.45	7.31	12.03	9.34
RPCA_s_m_h-DRNN	6.76	4.10	7.24	7.70	10.73	11.04
RPCA_s_M_h-DRNN	6.72	5.51	4.26	7.72	11.02	10.82
sRNN	6.86	8.11	6.56	7.23	10.15	10.74
RPCA-DRNN	**7.57**	**8.69**	**9.77**	**8.55**	**12.77**	**12.92**
(c)						
Method	**SDR**	**ISR**	**SIR**	**SAR**	**SDR**	**ISR**	**SIR**	**SAR**
RPCA_s_m_M-DRNN	5.80	10.57	17.13	5.36	6.98	13.44	19.16	11.72

Table 2 Experiment results of the ablation study using (a) MIR-1K, (b) ccMixter, and (c) MUSDB18 *(Continued)*

	Vocal			Accompaniment			
RPCA_s_m_h-DRNN	6.07	11.36	17.93	5.38	7.79	14.51	19.85	14.93
RPCA_s_M_h-DRNN	5.95	11.12	18.14	5.91	7.15	14.19	20.21	14.71
sRNN	5.58	10.48	15.58	5.42	6.69	11.04	16.74	11.66
RPCA-DRNN	**6.41**	**12.32**	**19.53**	**6.87**	**8.70**	**18.37**	**24.77**	**15.78**

(worse, slightly worse, equal, slightly better, better), we assumed the probability of choosing any answer as 1/5 to calculate the p-values. For example, in 48 of 100 trials, RPCA-DRNN was voted as better than sRNN. Consequently, the p-value of the binomial statistic was less than 5.3e−10. The p-value of the preferred rates of RPCA-DRNN is considerably lower than 0.05, which is, by convention [77], considered statistically significant with a confidence level of 95%. This means that there is an extremely low chance that the observed differences among the listeners' choices were due to chance, and the listeners did have preferences. For comparing with CRNN-A, RPCA-DRNN was voted as 31% better than CRNN-A. Consequently, the p-value of the binomial statistic was less than 0.0084, and the p-values of the preferred rates of RPCA-DRNN for individual listeners are also considerably lower than 0.05.

Table 3 Comparison of the GNSDR, GSIR, and GSAR scores of the proposed RPCA-DRNN method with RPCA_b, RPCA_p, sRNN, MLRR, RNMF, MOD-GD, U-Net, EFN, and CRNN-A on MIR-1K

Vocal			
Model method	**GNSDR**	**GSIR**	**GSAR**
RPCA_b	2.72	3.10	5.90
RPCA_p	4.76	4.44	8.19
sRNN	6.43	7.69	6.82
MLRR	3.85	5.63	10.70
RNMF	4.97	7.66	10.03
MOD-GD	7.50	13.73	9.45
U-Net	7.43	11.79	10.42
EFN	7.76	12.97	10.16
CRNN-A	7.89	**13.75**	10.17
RPCA-DRNN	**8.46**	7.72	**10.83**
Accompaniment			
Model method	**GNSDR**	**GSIR**	**GSAR**
RPCA_b	2.87	10.02	4.88
RPCA_p	4.91	**14.03**	5.80
sRNN	6.20	8.76	6.46
MLRR	4.19	7.80	8.22
RNMF	-	-	-
MOD-GD	-	-	-
U-Net	-	-	-
EFN	7.86	13.54	9.87
CRNN-A	7.12	9.62	11.97
RPCA-DRNN	**8.02**	12.32	**11.99**

The estimated training duration for the training set, inference duration for the testing set, and carbon footprint of training of the proposed and reference methods are shown in Table 6. The power of the graphics processing unit (GPU) (one 1080Ti) used is about 250 W. The carbon footprint is obtained under the assumption that 1 kWh of electricity discharges 0.62 kg of carbon dioxide and only the GPU power consumption is considered.

3.7 Experiment results of MUSDB18

In the experiment of MUSDB18, sounds other than vocal, such as bass, drums, and others, are considered as accompaniment. Vocal and accompaniment are estimated respectively from the left-channel and right-channel. Both the separated singing voice and accompaniment are evaluated.

Table 7 compares the results of RPCA-DRNN, RPCA-DRNN$_l$, RPCA-DRNN$_f$, sRNN, Open-Unmix, and E-MRP-CNN. For readers interested in the separation performance of other techniques using MUSDB18 up to 2018, they can reference the results of the 2018 signal separation evaluation campaign [74]. RPCA-DRNN is superior to sRNN both in vocal and accompaniment separation. Besides, in vocal separation, RPCA-DRNN is superior to Open-Unmix and E-MRP-CNN in SIR (19.53 to 12.19 and 13.40) and slightly better in SAR (6.87 to 5.98 and 6.32) and SDR (6.41 to 5.57 and 6.36), but slightly worse in ISR (12.32 to 14.07 and 13.61). In accompaniment separation, RPCA-DRNN is superior to Open-Unmix and E-MRP-CNN in SIR (24.77 to 19.62 and 16.18) and slightly better in SAR (15.78 to 12.54 and 14.41), but worse in SDR (8.70 to 11.06 and 12.99) and ISR (18.37 to 19.06 and 23.00). Since our proposed method is for monaural separation, the spatial distortion is not under consideration. The worst performance on reducing spatial distortion makes low ISR. For separated vocals, from the definition of SDR, when the spatial distortion of RPCA-DRNN is bigger, it can be observed that the reduced sum of interference and artifact error by RPCA-DRNN is more than the reduced sum by

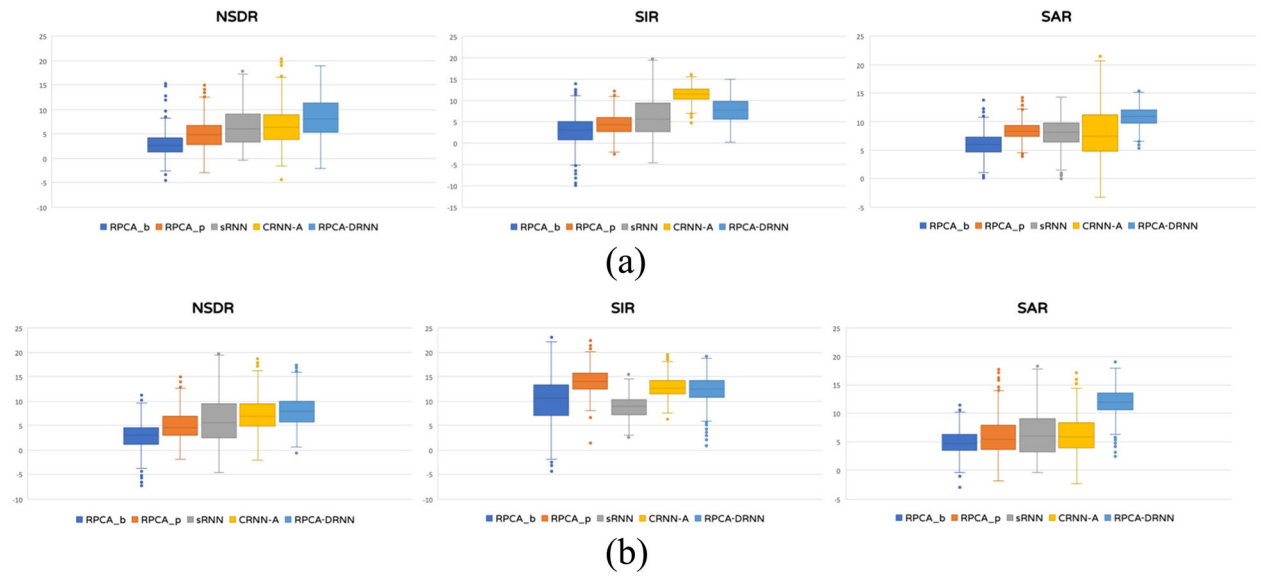

Fig. 6 Box plots of the NSDR, SIR, and SAR of the separated **a** vocal and **b** accompaniment of RPCA-DRNN and RPCA_b, RPCA_p, sRNN, and CRNN-A on MIR-1K

Open-Unmix and E-MRP-CNN. For separated accompaniment, since the spatial distortion of RPCA-DRNN is more than the other two, the performance of SDR of RPCA-DRNN is also low. Note that E-MRP-CNN used gain and sliding augmentation method, and the ratio of the augmented data of E-MRP-CNN to the original data is 1:4. Open-Unmix used even stronger data augmentation methods. It also used normalization and input/output scalar, while the proposed RPCA-DRNN did not use any data augmentation in the experiment of MUSDB18. Besides, the performance of the light and feather versions of RPCA-DRNN is still very competitive even in much less, and very limited training data. Including the augmented data, RPCA-DRNN only uses 0.25 times, and RPCA-DRNN$_l$ and RPCA-DRNN$_f$ only use about 0.03 times of the data amount E-MRP-CNN used. If we exclude the augmented data, RPCA-DRNN$_f$ only uses 0.05 times the original training data. The box plots of the comparison of the proposed method RPCA-DRNN with E-MRP-CNN is presented in Fig. 7, which presents more statistical details. The data of E-MRP-CNN is from [47]. From Fig. 7, compared to E-MRP-CNN, RPCA-DRNN is better in SIR and slightly better in SDR, ISR, and SAR in vocal separation, and better in SIR, slightly better in SAR, worse in SDR, and slightly worse in ISR in accompaniment separation.

RPCA-DRNN was further compared with Open-Unmix and E-MRP-CNN in a subjective assessment. Twenty listeners were attended. All of them are music enthusiasts but are not familiar with source separation or audio engineering. Four of them were former band members. Each listener was randomly assigned 5 songs from 6 songs. For each song, listeners were allocated the three separated singing from Open-Unmix, E-MRP-CNN, and RPCA-DRNN in random order. Table 8 shows the percentage distribution of the MOS scores. All the three methods are evaluated as good and

Table 4 Percentages of bad, poor, fair, good, and excellent MOS scores assigned to singing voice clips separated using sRNN, CRNN-A, and RPCA-DRNN on MIR-1K

sRNN					
MOS	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**
Percentage	0	13%	**44%**	27%	16%
CRNN-A					
MOS	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**
Percentage	0	11%	**38%**	37%	14%
RPCA-DRNN					
MOS	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**
Percentage	0	11%	22%	**44%**	23%

Table 5 Percentages of RPCA-DRNN-separated clips assigned "worse," "slightly worse," "equal," "slightly better," and "better" CMOS scores compared with percentages of such scores for sRNN and CRNN-A on MIR-1K

vs. sRNN					
	Worse	**Slightly worse**	**Equal**	**Slightly better**	**Better**
Percentage	0	12%	18%	22%	**48%**
vs. CRNN-A					
	Worse	**Slightly worse**	**Equal**	**Slightly better**	**Better**
Percentage	5	20%	28%	16%	**31%**

Table 6 The training duration, inference duration, and carbon footprint of training on MIR-1K of RPCA-DRNN and reference methods

	Training duration (h)	Inference duration (h)	Carbon footprint of training (kg of carbon dioxide)
sRNN	112	0.5	17
CRNN-A	144	0.5	22
RPCA-DRNN	224	0.5	35

excellent. The percentages of excellent scores of RPCA-DRNN were higher than those of Open-Unmix and E-MRP-CNN (38% vs. 32% and 36%). The average MOS of RPCA-DRNN, Open-Unmix, and E-MRP-CNN were 4.38, 4.32, and 4.36, respectively. RPCA-DRNN obtained the highest average MOS.

A further analysis of the separated singing voice was then conducted by performing a CMOS test. The RPCA-DRNN vs. Open-Unmix and RPCA-DRNN vs. E-MRP-CNN were evaluated. The same 20 listeners and 6 songs as those in the MOS test were used, and 5 songs (two pairs in each song) were allocated to each listener randomly. Table 9 lists the CMOS results. RPCA-DRNN was voted as 88% and 91% equal or slightly better than Open-Unmix and E-MRP-CNN. For both the pairs (RPCA-DRNN vs. Open-Unmix and RPCA-DRNN vs. E-MRP-CNN), the p-value of the binomial statistic was less than 2.2e−16. It is considerably lower than 0.05, which is considered statistically significant with a confidence level of 95%. Compared to Open-Unmix, the percentage of slightly better of RPCA-DRNN is higher than slightly worse, and compared to E-MRP-CNN, the percentage of slightly better and slightly worse of RPCA-DRNN are the same. Therefore, RPCA-DRNN achieves better performance compared to Open-Unmix and competitive performance with E-MRP-CNN in CMOS in monaural singing voice separation.

Table 7 Comparison of the SDR, ISR, SIR, and SAR scores of the proposed RPCA-DRNN methods with sRNN, Open-Unmix, and E-MRP-CNN on MUSDB18

Vocal				
Method	**SDR**	**ISR**	**SIR**	**SAR**
sRNN	5.58	10.48	15.58	5.42
Open-Unmix	5.57	**14.07**	12.19	5.98
E-MRP-CNN	6.36	13.61	13.40	6.32
RPCA-DRNN$_f$	5.96	13.56	15.84	5.65
RPCA-DRNN$_l$	6.26	11.47	18.19	6.24
RPCA-DRNN	**6.41**	12.32	**19.53**	**6.87**
Accompaniment				
Method	**SDR**	**ISR**	**SIR**	**SAR**
sRNN	6.69	11.04	16.74	11.66
Open-Unmix	11.06	19.06	19.62	12.54
E-MRP-CNN	**12.99**	**23.00**	16.18	14.41
RPCA-DRNN$_f$	7.70	16.27	20.60	13.33
RPCA-DRNN$_l$	8.53	17.26	22.62	14.77
RPCA-DRNN	8.70	18.37	**24.77**	**15.78**

The training duration for the training set, inference duration for the testing set, the GPU (power) used, and carbon footprint of training on MUSDB18 of RPCA-DRNN, RPCA-DRNN$_l$, RPCA-DRNN$_f$, sRNN, and E-MRP-CNN are shown in Table 10. The power of one 1080Ti and one 3090 is about 250 W and 350 W. The carbon footprint of training is under the assumption that 1 kWh of electricity discharges 0.62 kg of carbon dioxide and only the GPU power consumption is considered. The estimation of E-MRP-CNN counts only the most time-consuming evolution process and is based on the total power consumption 1560 W of 6 GPUs they used, including two 1080Ti, one 2080Ti, one Titan RTX, one Titan V, and one Titan XP, under the condition of 100 generations with 2 h running for each evolution, while RPCA-DRNN counts the total computation. In such case, the carbon footprint of E-MRP-CNN is about 2.5 times of RPCA-DRNN on one 1080Ti, 5.7 times of RPCA-DRNN on one 3090, and 32.2 times of RPCA-DRNN$_l$ and RPCA-DRNN$_f$ on one 3090. Therefore, the proposed RPCA-DRNN provides competitive performance at a lower training cost. Besides, the light and feather versions of RPCA-DRNN, which achieve better separation quality than sRNN, have only half the carbon footprint than sRNN.

4 Conclusions

We proposed a method based on our daily learning experiences that first uses the underlying knowledge and characteristics gleaned or inferenced and adopts method without prior training to separate sources on the basis of reasonable tendencies and assumptions and then uses supervised learning to jointly exploit labeled data to further improve the separation results. A method combining RPCA and supervised DRNN was employed in an experiment to improve the separation of singing voice from musical accompaniment in monophonic mixtures. First, RPCA was used to roughly separate the mixture into sparse voice and low-rank music. Second, postprocessing, including median filtering, morphology, and high-pass filtering, was performed to smooth and

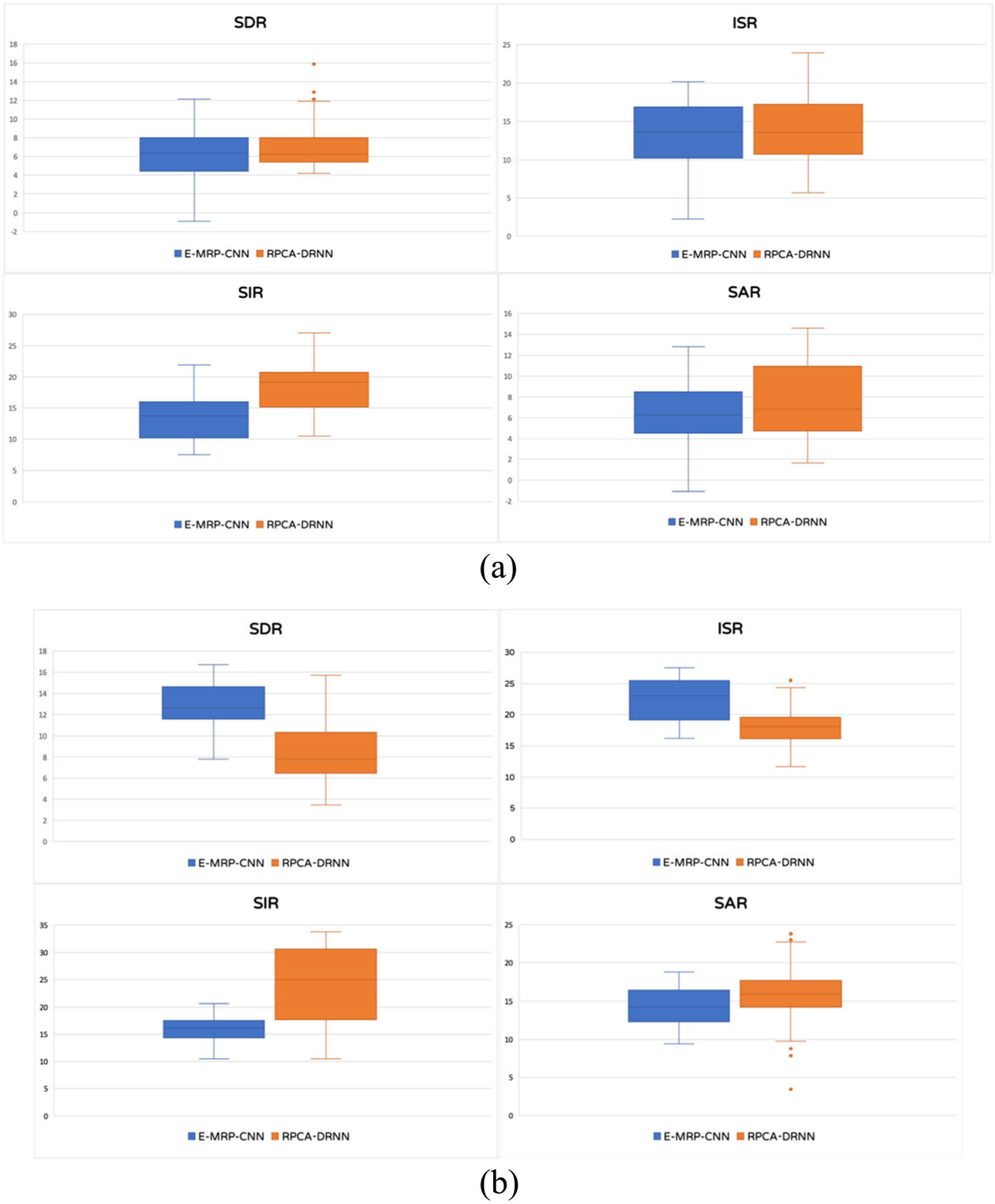

Fig. 7 Box plots of SDR, ISR, SIR, and SAR of the separated **a** vocal and **b** accompaniment of RPCA-DRNN and E-MRP-CNN on MUSDB18

enhance the spectral structure of estimated singing and filter out unnecessary parts. Then, supervised DRNN was utilized to achieve further separation. The misclassified or residual singing and background music from the initial separation was further corrected to improve the results.

Table 8 Percentages of bad, poor, fair, good, and excellent MOS scores assigned to singing voice clips separated using Open-Unmix, E-MRP-CNN, and RPCA-DRNN on MUSDB18

Open-Unmix					
MOS	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**
Percentage	0	0	0	68%	32%
E-MRP-CNN					
MOS	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**
Percentage	0	0	0	64%	36%
RPCA-DRNN					
MOS	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**
Percentage	0	0	0	62%	**38%**

Based on the objective scores on MIR-1K, the proposed method was found to be superior to RPCA, sRNN, MLRR, RNMF, MOD-GD, U-Net, EFN, and CRNN-A in terms of GNSDR and GSAR. Moreover, when the total numbers of neurons are the same, RPCA-DRNN with two smaller nets outperformed sRNN with one larger net in the subjective tests in terms of MOS and CMOS scores, because RPCA-DRNN is a solution combining underlying knowledge and supervised learning, and the DRNN of RPCA-DRNN is only for further correcting the residual singing voice and music from the output of RPCA with soft mask, medium filter, morphology, and high-pass filter. The variation of the inputs of the DRNNs of RPCA-DRNN is relatively small compared to the input of sRNN, which is the original sound mixture. RPCA-DRNN was also voted 75% equal, slightly better or better than CRNN-A. In addition, based on the objective scores on MUSDB18, RPCA-DRNN is superior to Open-Unmix and E-MRP-CNN in SDR, SIR, and SAR in vocal separation, and superior to Open-Unmix and E-MRP-CNN in SIR and SAR in accompaniment separation. This result is obtained under the condition of no data augmentation applied in the proposed RPCA-DRNN, while both Open-Unmix and E-MRP-CNN use data augmentation. The subjective test also confirms the preference of RPCA-DRNN. Besides, the performance of the light and feather versions of RPCA-DRNN is still very competitive even in very few and limited training data.

Table 9 Percentages of RPCA-DRNN-separated clips assigned "worse," "slightly worse," "equal," "slightly better," and "better" CMOS scores compared with percentages of such scores for Open-Unmix and E-MRP-CNN on MUSDB18

vs. Open-Unmix					
	Worse	**Slightly worse**	**Equal**	**Slightly better**	**Better**
Percentage	0	12%	72%	16%	0%
vs. E-MRP-CNN					
	Worse	**Slightly worse**	**Equal**	**Slightly better**	**Better**
Percentage	0	9%	82%	9%	0%

Therefore, the combination of the underlying properties inferenced and supervised learning, which is characteristic of humans' daily learning experiences, improved the separation of a singing voice from background music in the case of a monaural mixture. Benefitting from the initial RPCA separation without prior training, the proposed method achieves competitive results even with limited data or without data augmentation and hence can greatly reduce the computational load.

The main limitation of RPCA-DRNN is that at least one GPU card with 1080Ti or higher is needed for training. The database with less training data (e.g., MIR-1K with 23 min 36 s training data) is recommended to use data augmentation, and the database with enough training data (e.g., MUSDB18 with 7 h 46 min 24 s training data or 53 min 45 s in RPCA-DRNN$_l$) can still get a good result without data augmentation.

In the future, we will try with other neural network architectures, data augmentation methods that generate realistic mixtures, and use the proposed method in applications such as singing voice analysis and resynthesis systems. The proposed system can also be revised for more source separation by adding additional DRNNs. Moreover, adapting the method to stereo source separation by handling the spatial relation of the sound in different channels is also an interesting future work.

Abbreviations

ALM: Augmented Lagrange multiplier; CNN: Convolutional neural network; CRNN: Convolutional recurrent neural network; CRNN-A: CRNN with an attention; DNN: Deep neural network; DRNNs: Deep recurrent neural networks; EFN: Enhanced feature network; E-MRP-CNN: Evolving multi-resolution pooling CNN; GRU: Gated recurrent unit; GNSDR: Global normalized source-to-distortion ratio; GSAR: Global source-to-artifact ratio; GSIR: Global source-to-interference ratio; GPU: Graphics processing unit; HPSS: Harmonic/percussive sound separation; ISR: Image-to-spatial distortion ratio; ISTFT: Inverse short-time Fourier transform; LSTM: Long short-term memory; MOD-GD: MODified Group Delay; MLRR: Multiple low-rank representation; MIR: Music information retrieval; NSDR: Normalized source-to-distortion ratio; ReLU: Rectified linear unit; REPET-SIM: REpeating Pattern Extraction Technique with a SIMilarity matrix; RNMF: Robust low-rank non-negative matrix factorization; RPCA: Robust principal component analysis; STFT: Short-time Fourier transform; SiSEC: Signal separation evaluation campaign; SAR: Source-to-artifact ratio; SDR: Source-to-distortion ratio; SIR: Source-to-interference ratio; sRNNs: Stacked recurrent neural networks

5 Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13636-022-00236-9.

Additional file 1. Separated accompaniment.
Additional file 2. Separated vocal.

Acknowledgements

Not applicable.

Table 10 The training duration, inference duration, GPU (power), and carbon footprint of training on MUSDB18 of RPCA-DRNN methods, sRNN and E-MRP-CNN

	Training duration (h)	Inference duration (h)	GPU (power)	Carbon footprint of training (kg of carbon dioxide)
E-MRP-CNN	200	-	Two 1080Ti + one 2080Ti + one Titan RTX + one Titan V + one Titan XP (1560 W)	193
RPCA-DRNN	496	1	One 1080Ti (250 W)	77
RPCA-DRNN	156	0.3	One 3090 (350 W)	34
RPCA-DRNN$_l$	28	0.3	One 3090 (350 W)	6
RPCA-DRNN$_f$	28	0.3	One 3090 (350 W)	6
sRNN	54	0.3	One 3090 (350 W)	12

Authors' contributions

WH led and supervised the research, proposed the methodology, analyzed and interpreted the result, and was the contributor in writing the manuscript. SL performed the experiments. All authors read and approved the final manuscript.

Funding

This work was supported in part by the Ministry of Science and Technology of Taiwan under Contract MOST 105-2221-E-327-040.

Availability of data and materials

The datasets used during the current study are available in the MIR-1K, ccMixter, and MUSDB18 repositories, https://sites.google.com/site/unvoicedsoundseparation/mir-1k, https://members.loria.fr/ALiutkus/kam/, and https://sigsep.github.io/datasets/musdb.html. The separated singing and accompaniment supporting the conclusions of this article are included within the additional file.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computer and Communication Engineering, National Kaohsiung University of Science and Technology, Kaohsiung 824005, Taiwan. [2]Ph.D. Program in Engineering Science and Technology, College of Engineering, National Kaohsiung University of Science and Technology, Kaohsiung 824005, Taiwan.

Received: 8 August 2021 Accepted: 17 January 2022
Published online: 05 February 2022

References

1. K. Hu, D. Wang, An unsupervised approach to cochannel speech separation. IEEE Trans. Audio. Speech. Lang. Process. **21**(1), 122–131 (2013). https://doi.org/10.1109/TASL.2012.2215591
2. Z. Jin, D. Wang, Reverberant speech segregation based on multipitch tracking and classification. IEEE Trans. Audio. Speech. Lang. Process. **19**(8), 2328–2337 (2011). https://doi.org/10.1109/TASL.2011.2134086
3. D. Kawai, K. Yamamoto, S. Nakagawa, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speech analysis of sung-speech and lyric recognition in monophonic singing (IEEE, Shanghai, 2016), pp. 271–275
4. M. Mehrabani, J.H.L. Hansen, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Language identification for singing (IEEE, Prague, 2011), pp. 4408–4411
5. Y. Hu, G. Liu, in *2013 IEEE International Conference on Multimedia and Expo (ICME)*. Automatic singer identification using missing feature methods (IEEE, San Jose, 2013), pp. 1–6. https://doi.org/10.1109/ICME.2013.6607641
6. Y. Hu, G. Liu, Separation of singing voice using nonnegative matrix partial co-factorization for singer identification. IEEE/ACM Trans Audio Speech Lang Process **23**(4), 643–653 (2015). https://doi.org/10.1109/TASLP.2015.2396681
7. Y. Shi, X. Zhou, in *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*. Emotion recognition in singing using convolutional neural networks (IEEE, Xi'an, 2021), pp. 576–579
8. B. Sharma, C. Gupta, H. Li, Y. Wang, in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models (Brighton, 2019), pp. 396–400
9. H. Chou, M. Chen, T. Chi, in *in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A hybrid neural network based on the duplex model of pitch perception for singing melody extraction (IEEE, Calgary, 2018), pp. 381–385
10. P. Jao, Y. Yang, Music annotation and retrieval using unlabeled exemplars: correlation and sparse codes. IEEE Signal Process. Lett. **22**(10), 1771–1775 (2015). https://doi.org/10.1109/LSP.2015.2433061
11. M. Goto, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Active music listening interfaces based on signal processing (Honolulu, 2007), pp. IV-1441–IV–1444. https://doi.org/10.1109/ICASSP.2007.367351
12. J. Sundberg, Perception of singing. Dept. for Speech, Music and Hearing of KTH computer science and communication. STL-QPSR **20**(1), 001–048 (1979)
13. A. Ozerov, P. Philippe, F. Bimbot, R. Gribonval, Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. IEEE Trans Audio Speech Lang Proces **15**(5), 1564–1578 (2007). https://doi.org/10.1109/TASL.2007.899291
14. L. Benaroya, F. Bimbot, in *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*. Wiener based source separation with HMM/GMM using a single sensor (Granada, 2003), pp. 957–961
15. C.-L. Hsu, J.-S.R. Jang, On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. IEEE Trans. Audio. Speech. Lang. Process. **18**(2), 310–319 (2010). https://doi.org/10.1109/TASL.2009.2026503
16. B. Zhu, W. Li, R. Li, X. Xue, Multi-stage non-negative matrix factorization for monaural singing voice separation. IEEE Trans. Audio. Speech. Lang. Process. **21**(10), 2096–2107 (2013). https://doi.org/10.1109/TASL.2013.2266773
17. A. Chanrungutai, C. A. Ratanamahatana, in *2008 International Conference on Advanced Technologies for Communications*. Singing voice separation for mono-channel music using Non-negative Matrix Factorization (Hanoi, 2008), 243–2246. https://doi.org/10.1109/ATC.2008.4760565.
18. A. Chanrungutai, C.A. Ratanamahatana, *2008 International Symposium on Communications and Information Technologies. Singing voice separation in mono-channel music* (Vientiane, 2008), pp. 256–261. https://doi.org/10.1109/ISCIT.2008.4700194
19. T. Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio. Speech. Lang. Process. **15**(3), 1066–1074 (2007). https://doi.org/10.1109/TASL.2006.885253

20. S. Koundinya, A. Karmakar, Homotopy optimisation based NMF for audio source separation. IET Signal Process **12**(9), 1099–1106 (2018). https://doi.org/10.1049/iet-spr.2018.5093
21. J.-T. Chien, P.-K. Yang, Bayesian factorization and learning for monaural source separation. IEEE/ACM Trans Audio Speech Lang Process **24**(1), 185–195 (2016). https://doi.org/10.1109/TASLP.2015.2502141
22. S. Vembu, S. Baumann, in *6th International Conference on Music Information Retrieval (ISMIR 2005)*. Separation of vocals from polyphonic audio recordings (London, 2005), pp. 337–344
23. J.-T. Chien, H.-L. Hsieh, Bayesian group sparse learning for music source separation. EURASIP J Audio Speech Music Process **2013**(1), 18 (2013). https://doi.org/10.1186/1687-4722-2013-18
24. P. Sprechmann, A. Bronstein, G. Sapiro, in *13th International Society for Music Information Retrieval Conference (ISMIR 2012)*. Real-time online singing voice separation from monaural recordings using robust low-rank modeling (Portugal, 2012), p. 6
25. X. Zhang, W. Li, B. Zhu, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Latent time-frequency component analysis: a novel pitch-based approach for singing voice separation (South Brisbane, 2015), pp. 131–135. https://doi.org/10.1109/ICASSP.2015.7177946
26. Y. Li, D. Wang, Separation of singing voice from music accompaniment for monaural recordings. IEEE Trans. Audio. Speech. Lang. Process. **15**(4), 1475–1487 (2007). https://doi.org/10.1109/TASL.2006.889789
27. C.L. Hsu, D. Wang, J.S.R. Jang, K. Hu, A tandem algorithm for singing pitch extraction and voice separation from music accompaniment. IEEE Trans. Audio. Speech. Lang. Process. **20**(5), 1482–1491 (2012). https://doi.org/10.1109/TASL.2011.2182510
28. Z. Rafii, B. Pardo, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A simple music/voice separation method based on the extraction of the repeating musical structure (Prague, 2011), pp. 221–224. https://doi.org/10.1109/ICASSP.2011.5946380
29. Z. Rafii, B. Pardo, REpeating Pattern Extraction Technique (REPET): a simple method for music/voice separation. IEEE Trans. Audio. Speech. Lang. Process. **21**(1), 73–84 (2013). https://doi.org/10.1109/TASL.2012.2213249
30. P.-S. Huang, S.D. Chen, P. Smaragdis, M. Hasegawa-Johnson, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Singing-voice separation from monaural recordings using robust principal component analysis (Kyoto, 2012), pp. 57–60. https://doi.org/10.1109/ICASSP.2012.6287816
31. Y.-H. Yang, in *International Society for Music Information Retrieval Conference (ISMIR 2013)*. Low-rank representation of both singing voice and music accompaniment via learned dictionaries (Brazil, 2013), pp. 427–432
32. H. Tachibana, N. Ono, S. Sagayama, Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms. IEEE/ACM Trans. Audio. Speech. Lang. Process. **22**(1), 228–237 (2014). https://doi.org/10.1109/TASLP.2013.2287052
33. I.-Y. Jeong, K. Lee, Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints. IEEE Signal Process. Lett. **21**(10), 1197–1200 (2014). https://doi.org/10.1109/LSP.2014.2329946
34. H. Tachibana, T. Ono, N. Ono, S. Sagayama, in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source (Dallas, 2010), pp. 425–428. https://doi.org/10.1109/ICASSP.2010.5495764
35. P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans. Audio. Speech. Lang. Process. **23**(12), 2136–2147 (2015). https://doi.org/10.1109/TASLP.2015.2468583
36. P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, in *15th International Society for Music Information Retrieval Conference (ISMIR 2013)*. Singing-voice separation from monaural recordings using deep recurrent neural networks (Taipei, 2013), p. 6
37. J. Sebastian, H.A. Murthy, in *2016 International Conference on Signal Processing and Communications (SPCOM)*. Group delay based music source separation using deep recurrent neural networks (Bangalore, 2016), pp. 1–5. https://doi.org/10.1109/SPCOM.2016.7746672
38. A.A. Nugraha, A. Liutkus, E. Vincent, in *2016 24th European Signal Processing Conference (EUSIPCO)*. Multichannel music separation with deep neural networks (Budapest, 2016), pp. 1748–1752. https://doi.org/10.1109/EUSIPCO.2016.7760548
39. S. Yang, W.-Q. Zhang, in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. Singing voice separation based on deep regression neural network (Ajman, 2019), pp. 1–5. https://doi.org/10.1109/ISSPIT47144.2019.9001837
40. W. Yuan, S. Wang, X. Li, M. Unoki, W. Wang, in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Proximal deep recurrent neural network for monaural singing voice separation (Brighton, 2019), pp. 286–290. https://doi.org/10.1109/ICASSP.2019.8682879
41. F.-R. Stöter, S. Uhlich, A. Liutkus, Y. Mitsufuji, Open-Unmix - a reference implementation for music source separation. J. Open Source Softw. **4**(41), 1667 (2019). https://doi.org/10.21105/joss.01667
42. W. Yuan, B. He, S. Wang, J. Wang, M. Unoki, Enhanced feature network for monaural singing voice separation. Speech Commun. **106**, 1–6 (2019). https://doi.org/10.1016/j.specom.2018.11.004
43. C. Sun, M. Zhang, R. Wu, J. Lu, G. Xian, Q. Yu, X. Gong, R. Luo, A convolutional recurrent neural network with attention framework for speech separation in monaural recordings. Sci. Rep. **11**(1), 1434 (2021). https://doi.org/10.1038/s41598-020-80713-3
44. N. Takahashi, N. Goswami, Y. Mitsufuji, MMDenseLSTM: an efficient combination of convolutional and recurrent neural networks for audio source separation. arXiv:1805.02410 (2018) [Online]. Available: http://arxiv.org/abs/1805.02410. Accessed: 30 June 2021
45. A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, T. Weyde, in *18th International Society for Music Information Retrieval Conference (ISMIR 2017)*. Singing voice separation with Deep U-NET convolutional networks (Suzhou, 2017), p. 7
46. A. Cohen-Hadria, A. Roebel, G. Peeters, Improving singing voice separation using deep U-Net and Wave-U-Net with data augmentation. arXiv **1903.01415** (2019) [Online]. Available: http://arxiv.org/abs/1903.01415. Accessed: 30 June 2021
47. W. Yuan, B. Dong, S. Wang, M. Unoki, W. Wang, Evolving multi-resolution pooling CNN for monaural singing voice separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. **29**, 807–822 (2021). https://doi.org/10.1109/TASLP.2021.3051331
48. M. Huber, G. Schindler, C. Schörkhuber, W. Roth, F. Pernkopf, H. Fröning, in *2020 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Towards real-time single-channel singing-voice separation with pruned multi-scaled densenets (2020), pp. 806–810. https://doi.org/10.1109/ICASSP40776.2020.9053542
49. T. Virtanen, A. Mesaros, M. Ryynänen, in *Statistical and Perceptual Audition (SAPA)*. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music (Brisbane, 2008), pp. 17–22
50. Z. Rafii, Z. Duan, B. Pardo, Combining rhythm-based and pitch-based methods for background and melody separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**(12), 1884–1893 (2014). https://doi.org/10.1109/TASLP.2014.2354242
51. Y. Ikemiya, K. Yoshii, K. Itoyama, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*. Singing voice analysis and editing based on mutually dependent F0 estimation and source separation (Brighton, 2015), pp. 574–578. https://doi.org/10.1109/ICASSP.2015.7178034
52. Y. Ikemiya, K. Itoyama, K. Yoshii, Singing voice separation and vocal F0 estimation based on mutual combination of robust principal component analysis and subharmonic summation. IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**(11), 2084–2095 (2016). https://doi.org/10.1109/TASLP.2016.2577879
53. J. Driedger, M. Müller, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Extracting singing voice from music recordings by cascading audio decomposition techniques (South Brisbane, 2015), pp. 126–130. https://doi.org/10.1109/ICASSP.2015.7177945
54. X. Jaureguiberry, E. Vincent, G. Richard, Fusion methods for speech enhancement and audio source separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**(7), 1266–1279 (2016). https://doi.org/10.1109/TASLP.2016.2553441
55. B.R. Gibson, T.T. Rogers, X. Zhu, Human semi-supervised learning. Top. Cogn. Sci **5**(1), 132–172 (2013). https://doi.org/10.1111/tops.12010
56. O. Chapelle, B. Schölkopf, A. Zien (eds.), *Semi-supervised learning* (MIT Press, Cambridge, 2006)

57. E.J. Candes, X. Li, Y. Ma, J. Wright, Robust principal component analysis. arXiv **0912.3599** (2009) [Online]. Available: http://arxiv.org/abs/0912.3599. Accessed: 30 June 2021
58. Z. Lin, M. Chen, Y. Ma, The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. J. Struct. Biol. **181**(2), 116–127 (2013). https://doi.org/10.1016/j.jsb.2012.10.010
59. A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, G. Richard, *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Adaptive filtering for music/voice separation exploiting the repeating musical structure* (Kyoto, 2012), pp. 53–56. https://doi.org/10.1109/ICASSP.2012.6287815
60. Y. Yang, in *20th ACM international conference on Multimedia*. On sparse and low-rank matrix decomposition for singing voice separation (New York, 2012), pp. 757–760
61. S. Scholler, H. Purwins, Sparse approximations for drum sound classification. IEEE J. Sel. Top. Sig. Process. **5**(5), 933–940 (2011)
62. D. FitzGerald, in *13th International Conference on Digital Audio Effects (DAFx-10)*. Harmonic/percussive separation using median filtering (Graz, 2010), pp. 1–4
63. R.M. Haralick, S.R. Sternberg, X. Zhuang, Image analysis using mathematical morphology. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-9**(4), 532–550 (1987). https://doi.org/10.1109/TPAMI.1987.4767941
64. B. Lehner, G. Widmer, S. Bock, in *2015 23rd European Signal Processing Conference (EUSIPCO)*. A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks (Nice, 2015), pp. 21–25
65. M. Hermans, B. Schrauwen, in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. Training and analyzing deep recurrent neural networks (Red Hook, 2013), pp. 190–198
66. R. Pascanu, C. Gulcehre, K. Cho, Y. Bengio, How to construct deep recurrent neural networks. arXiv **1312.6026** (2014) [Online]. Available: http://arxiv.org/abs/1312.6026. Accessed: 30 June 2020
67. K. Cho, K. Cho, B. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv **1406.1078** (2014) [Online]. Available: http://arxiv.org/abs/1406.1078. Accessed: 30 June 2021
68. J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv **1412.3555** (2014) [Online]. Available: http://arxiv.org/abs/1412.3555. Accessed: 30 June 2021
69. A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, L. Daudet, Kernel additive models for source separation. IEEE Trans. Sig. Process. **62**(16), 4298–4310 (2014)
70. R. Zafar, L. Antoine, F.-R. Stöter, M.S. Ioannis, B. Rachel, MUSDB18 - a corpus for music separation. Zenodo (2017). https://doi.org/10.5281/ZENODO.1117372
71. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv **1412.6980** (2017) [Online]. Available: http://arxiv.org/abs/1412.6980. Accessed: 03 Aug 2020
72. C. Févotte, R. Gribonval, E. Vincent, BSS_EVAL Toolbox user guide – revision 2.0 (2005). https://hal.inria.fr/inria-00564760/document. Accessed 6 June 2018.
73. E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation. IEEE Trans. Audio, Speech Lang. Process. **14**(4), 1462–1469 (2006). https://doi.org/10.1109/TSA.2005.858005
74. F.-R. Stöter, A. Liutkus, N. Ito, in *Latent Variable Analysis and Signal Separation*. The 2018 signal separation evaluation campaign (Cham, 2018), pp. 293–305. https://doi.org/10.1007/978-3-319-93764-9_28
75. E. Vincent, H. Sawada, P. Bofill, S. Makino, J.P. Rosca, in *Independent Component Analysis and Signal Separation*. First stereo audio source separation evaluation campaign: data, algorithms and results (Berlin, 2007), pp. 552–559. https://doi.org/10.1007/978-3-540-74494-8_69
76. P.800: Methods for subjective determination of transmission quality. https://www.itu.int/rec/T-REC-P.800-199608-I. Accessed 09 Jan 2021.
77. R.E. Livezey, W.Y. Chen, Statistical field significance and its determination by Monte Carlo techniques. Monthly Weather Rev. **111**(1), 46–59 (1983)

Publisher's Note