sensors

MDPI

Article

Unsupervised Single-Channel Singing Voice Separation with Weighted Robust Principal Component Analysis Based on Gammatone Auditory Filterbank and Vocal Activity Detection

Feng Li [1,2], Yujun Hu [1] and Lingling Wang [1,*]

[1] Department of Computer Science and Technology, Anhui University of Finance and Economics, Bengbu 233030, China
[2] School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China
* Correspondence: wll@aufe.edu.cn

Abstract: Singing-voice separation is a separation task that involves a singing voice and musical accompaniment. In this paper, we propose a novel, unsupervised methodology for extracting a singing voice from the background in a musical mixture. This method is a modification of robust principal component analysis (RPCA) that separates a singing voice by using weighting based on gammatone filterbank and vocal activity detection. Although RPCA is a helpful method for separating voices from the music mixture, it fails when one single value, such as drums, is much larger than others (e.g., the accompanying instruments). As a result, the proposed approach takes advantage of varying values between low-rank (background) and sparse matrices (singing voice). Additionally, we propose an expanded RPCA on the cochleagram by utilizing coalescent masking on the gammatone. Finally, we utilize vocal activity detection to enhance the separation outcomes by eliminating the lingering music signal. Evaluation results reveal that the proposed approach provides superior separation outcomes than RPCA on ccMixter and DSD100 datasets.

Keywords: single channel; singing voice; source separation; robust principal component analysis; gammatone filterbank; vocal activity detection

Citation: Li, F.; Hu, Y.; Wang, L. Unsupervised Single-Channel Singing Voice Separation with Weighted Robust Principal Component Analysis Based on Gammatone Auditory Filterbank and Vocal Activity Detection. *Sensors* **2023**, *23*, 3015. https://doi.org/10.3390/s23063015

Academic Editor: Pablo Angueira

Received: 4 January 2023
Revised: 27 February 2023
Accepted: 9 March 2023
Published: 10 March 2023

1. Introduction

Singing voice separation (SVS) has drawn a lot of interest and consideration in many downstream applications [1–4]. It deals with the technique of separating a singing voice or background from a mix of music, which is a crucial strategy for singer identification [5,6], music information retrieval [7,8], lyric recognition and alignment [9–12], song language identification [13,14], and chord recognition [15–17]. The recent separation techniques, however, fall well short of the capabilities of human hearing. It is challenging to resolve the existing SVS because of the instruments utilized and the spectral overlap between the speech and background music [11,18–21]. In daily life, human listeners generally have the remarkable ability to distinguish sound streams from a mixture of sounds, but this continues to be a difficult task for machines, particularly in the monaural case because it lacks the spatial cues that can be learned when two or more microphones are used. Additionally, the singing separation feeling could not directly translate from spoken separation. Speaking and singing voices have many similarities with one another but also differ in important ways. Because singing and speaking have distinct histories, there are significant challenges involved in separating them. The nature of the other accompanying sounds is the key distinction between singing and speech in terms of their independence from a background. The background noise that mingles with speech may be harmonic or nonharmonic, narrowband or broadband, and often unrelated to the speech. However, the musical accompaniment to a song is typically harmonic and wideband, associated

with the singing, and does not adhere to the common misconceptions about noise, such as its whiteness or stationarity. Consequently, conventional noise-suppression techniques are inappropriate.

Furthermore, singing voices typically have distinct and powerful harmonic structures as well as harmonics that change quickly, such as vibratos or slides, and musical accompaniment may be thought of as the combination of percussive and harmonic elements. Because the extraction findings are erroneous because harmonic instruments (other than singing) also include harmonics, simple harmonic extraction techniques are still not suitable for polyphonic mixes and rapidly changing harmonics. Because singing voices and musical accompaniments typically start and stop at the same time, onset and offset cues—which are ordinarily helpful in auditory scene classification because distinct sounds typically start and end at various times—are also ineffective. Moreover, lyrics are conveyed through singing by altering notes in accordance with the melody, making singing an interpretation of a predetermined musical score. As a result, singing has a piecewise constant pitch with rapid pitch shifts and other sorts of variations. Until recently, various research strategies and algorithms have been introduced to improve the separation results in SVS tasks [22,23]. The deep learning techniques [24–27] are perhaps the most widely used for SVS. Yu et al. [24] proposed a new feature-extraction module based on UNet++ for SVS. An enhanced encoder–decoder is first created to initially extract multiscale information from a magnitude spectrogram of the mixed music. As a last step, soft masks are constructed for the separation of each source after more fine characteristics are obtained by using the feature-extraction module at various sizes. By utilizing the parameters it has trained, the suggested network can capture the key characteristics of the multiscale spectrogram. Hu et al. [25] proposed a hierarchic temporal convolutional network with cross-domain encoder for SVS. The model uses the hierarchic temporal convolutional network for the separation of different musical sources and integrates the complexed spectrogram domain feature and time-domain feature via a cross-domain encoder. With the help of the cross-domain encoder, the network will be able to encode interactive information for time-domain and complexed spectrogram domain characteristics. Guizzo et al. [26] proposed an antitransfer learning with convolutional neural network (CNN) for speech processing. Ni et al. [27] proposed a novel deep neural network based on UNet for SVS. The time-invariant, completely linked layers are built along the frequency axis of the network, which is a two-level nested U-structure. Due to its form, it is possible to record long-distance speech signal associations along the frequency alignment in addition to local and global context information. Also presented is a unique loss function that combines binary and ratio masks. The estimated voices channel is cleaner and carries fewer accompanying signals thanks to this technique.

Although they have been successful for SVS, these models present challenges when dealing with minimal audio data because a significant amount of training data is required beforehand [28–32]. Learning entails a lot of observation of the world's objects and judgments made about their classifications, along with sporadic encounters with guided learning. In other words, knowledge gained from data may expand that learned from labeled data and may lead to the development of certain underlying assumptions or rules. Additionally, the limitations of previous neural networks become evident between the mismatch of training and testing samples [33]. The separation results lead to a decrease due to overfitting. In light of this, the unsupervised method is nevertheless appealing for mono source separation, especially when there is a lack of data or no other information. One of them is to assume and use the fundamental characteristics of singing voice and musical accompaniment. The RPCA strategy for SVS [34], which separates a mixture music into a singing voice (sparse) and background music (low rank), respectively. RPCA is a method that does not need any training or labeled data, and hence, it is convenient to use. It is the basis for or an inspiration for a number of unsupervised algorithms. Yang [35] developed two novel improvements of the matrix decomposition of the magnitude spectrogram by fusing the harmonicity priors information. Later, by breaking down a mixted spectrogram into a multiple low-rank representation (MLRR) will be introduced [36]. Despite being

effectively applied to SVS, RPCA fails when one singular value, such as drums, is significantly greater than others, which lowers the separation results, particularly for drums included in the combined music signal. Although all approaches can produce effective separation results, they all ignore the characteristics of auditory system, which is crucial for enhancing the quality of separation outcomes. To overcome this problem, in previous studies, we proposed a novel unsupervised approached that extends RPCA exploiting rank-1 constraint for SVS tasking [37]. A recent study found that the cochleagram, a different time frequency (T-F) masking technique with gammatone, is more effective for audio signal separation than the spectrogram [38–40]. In the cochleagram representation, a gammatone filter is used to simulate the cochleagram representation's frequency components, which are based on the human cochlea's frequency selectivity ability. Yuan et al. [38] proposed a data augmentation method by using chromagram-based and pitch-aware methods for SVS. A popular and effective method for synchronizing and aligning music is the use of chromagrams or chroma-based characteristics. The twelve distinct pitch classes and chromagram are closely connected. In order to create a 1-D vector that represents how the harmonic content of the representation inside the timeframe is distributed throughout the 12 chroma bands, the fundamental concept is to aggregate each pitch class across octaves for a specific local time window. A 2-D time chroma representation is produced as the time frame is moved across the song. As a result of its great resilience to timbral fluctuations and tight relationship to the musical harmony, we employ a chromagram correlation across song sections as a metric by which to evaluate song commonalities. Gao et al. [39] proposed a novel machine learning method, the optimized nonnegative matrix factorization (NMF) for SVS. The suggested cost function was created specifically for factorization of nonstationary signals with temporally dependent frequency patterns. Moreover, He et al. [40] suggested a method that will be able to get around all of the sparse nonnegative matrix factorization (SNMF) 2-D's previously mentioned drawbacks. The suggested model allows for many spectral and temporal changes, which are not inherent in the NMF and SNMF models. This allows for overcomplete representation. In order to provide distinctive and accurate representations of the nonstationary audio signals, sparsity must be imposed.

Additonally, the singing voice performances element on the cochleagram is rather distinct from the background music. For a singing voice, the spectral energy concentrates in a small number of time frequency units, so we may thus presume that it is sparse [41]. Additionally, the cochleagram's accompaniment to music exhibits comparable patterns and structures that can be represented in the basis spectral vectors. As a result, an example of blind monaural SVS system is described in Figure 1. The underlying low-rank and sparsity hypotheses, however, could not always hold true. Both the decomposed low-rank matrix and the decomposed sparse matrix may include vocal sounds in addition to instrumental sounds (such as percussion). There is still some background music audible while listening to the separated singing voice. Similar to this, a portion of the singing voice is mistakenly categorized as background music. In order to improve the separation accuracy, additional approaches or techniques must be used to categorize the RPCA output.

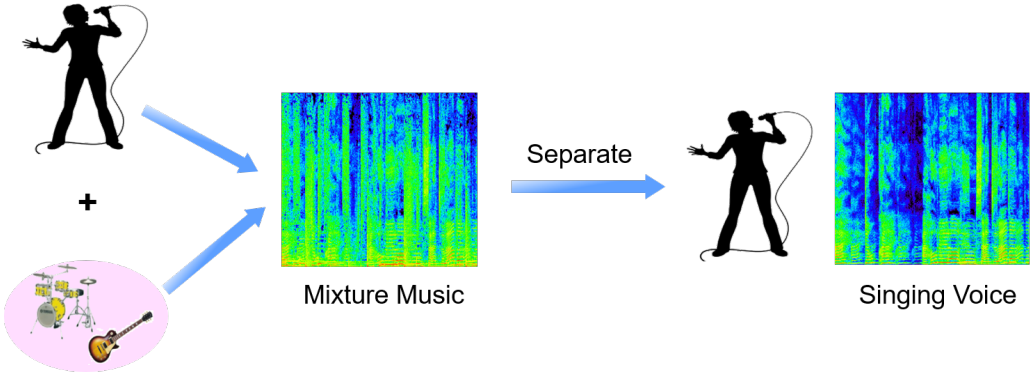


Figure 1. Blind monaural SVS system.

Therefore, in this paper, to address the existing problems in RPCA for SVS, in our work, we provide the varying values approach to characterize low-rank and sparse matrices. This method is referred as weighted RPCA (WRPCA) [41], and it selects various weighted values from the separated by low-rank and sparse matrices. Meanwhile, as the first step of WRPCA, we simulate the human auditory system by using the gammatone filterbank. To further remove the nonseparated background music, we combined the harmonic masking and T-F masking [42–44]. The mixed signal's time-frequency (T-F, or spectrogram) representation has been employed in the majority of prior speech-separation techniques. This signal is approximated from the waveform by using the short-time Fourier transform (STFT). The goal of speech separation technologies in the T-F domain is to approach the mixed spectrogram's clean spectrogram of the separate sources. Nonlinear regression techniques may be used to directly approximate each source's spectrogram description from the combination in this procedure. Finally, we utilize the vocal activity detection (VAD) [45–47] to get rid of any remaining background music. In a word, the key contributions of the work are outlined below as a summary.

- We offer the WRPCA addition to RPCA, which uses various weighted RPCA to achieve the improved separation performance.
- We combine gammatone auditory filterbank with vocal activity detection for SVS. Gammatone filterbanks are designed to imitate the human auditory system.
- We build the coalescent masking by fusing the harmonic masking and T-F masking, which can remove nonseparated background music. Additionally, we restrict the temporal segments that can include the singing voice part by using VAD.
- The extensive monaural SVS experiments reveal that the proposed approach can achieve greater separation performance than the RPCA method.

The remainder of this paper is arranged as follows. Section 2 provides RPCA and RPCA for SVS tasks. The proposed WRPCA on the cochleagram with VAD is illustrated in Section 3. The proposed approach is assessed with two different datasets in Section 4. Finally, we draw some conclusions from the paper and ideas for the further study in Section 5.

2. Related Work

This section discusses the RPCA and its application in SVS.

2.1. Overview of RPCA

RPCA was first introduced by Candés et al. [48] to divide the $M \in \mathbb{R}_{m\times n}$ into $L \in \mathbb{R}_{m\times n}$ plus $S \in \mathbb{R}_{m\times n}$. Thus, the optimization model is defined as

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

where $|L|_*$ stands for the sum of singular values and $|S|_1$ is the sum of absolute values of the matrix. According to the previous study, we set $\lambda = 1/\sqrt{max(m,n)}$. Meanwhile, we solve the convex program by accelerated proximal gradient (APG) or augmented Lagrange multiplier (ALM) [49] algorithms. In our work, a baseline experiment was conducted by using an inexact version of ALM.

2.2. RPCA for SVS

Music is typically made up of a variety of blended sounds, including both human singing voice and background music. The conventional RPCA approach can solve the task of SVS [34]. The magnitude spectrogram of a song may be broken down by using RPCA and can be thought of as the superposition of a sparse matrix and a low-rank matrix. The low-rank decomposition matrix and sparse matrix seem to match to the singing voice and background music. In light of these assumptions, RPCA may be used to

solve the singing/accompaniment separation problem. The assumptions are that singing corresponds to sparse matrices and low rank to accompaniment.

Due to the fact that musical instruments may replicate the same sounds again in music, the low-rank structure is used to conceptualize the spectrogram. In summary, the harmonic structure element of the singing voice part causes it to vary widely and to have a sparse distribution, producing in a spectrogram with the sparse matrix structure. Figure 2 shows the separation process of SVS with the low-rank and sparse model. Music is a low-rank signal because musical instruments can recreate the same sounds each time a piece is performed and music generally has an underlying recurring melodic pattern. Contrarily, voices are relatively scarce in the temporal and frequency domains but have greater diversity (higher rank). The singing voices can thus be seen as elements of the sparse matrix. By RPCA, we anticipate that the sparse matrix S will contain voice signals and the low-rank matrix L will include backing music.

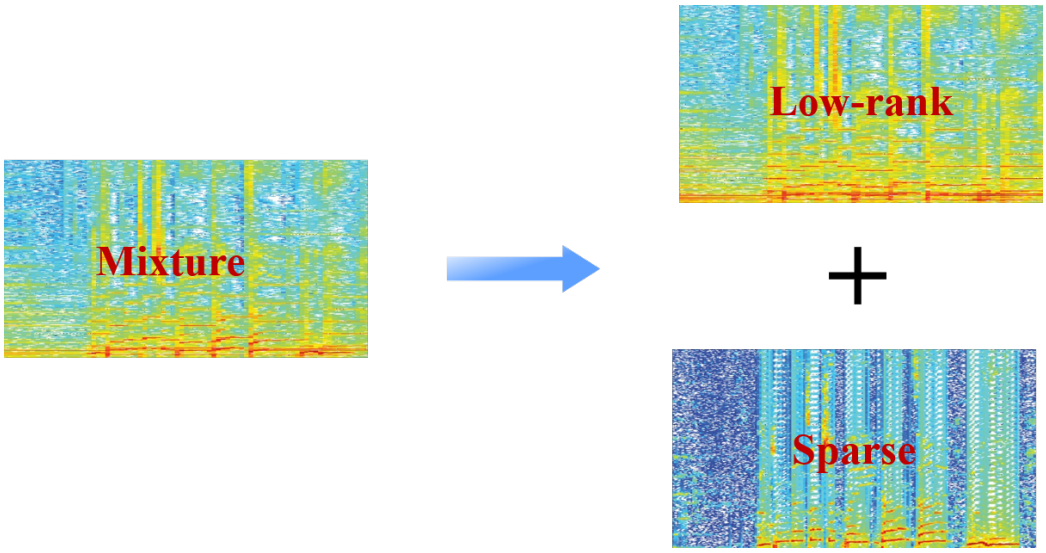


Figure 2. The separation process of SVS with low-rank and sparse model.

Consequently, in this study, we may divide an input matrix by using the RPCA approach into a low-rank and sparse matrices. Nevertheless, it does make significant assumptions. Drums, for instance, could not be low rank but rather lie in the sparse subspace, which lowers the results, especially for drums included in mixed music.

3. Proposed Method

This section firstly presents the WRPCA approach. Then, the gammatone filterbank and vocal acivity detection are utilized as the postprocessing for SVS. Finally, we provide the architecture of the proposed SVS approach.

3.1. Overview of WRPCA

WRPCA is an extension of RPCA, which has different scale values between sparse and low-rank matrices. The corresponding model can be defined as follows,

$$\begin{aligned} &minimize \ \ |L|_{w,*} + \lambda |S|_1, \\ &subject\ to\ M = L + S. \end{aligned} \tag{2}$$

where $|L|_{w,*}$ is the different weighted values in the matrix of low rank, and the S is sparse. $M \in \mathbb{R}_{m\times n}$ is made up $L \in \mathbb{R}_{m\times n}$ and $S \in \mathbb{R}_{m\times n}$, and the parameter $\lambda = 1/\sqrt{max(m,n)}$ is indicated [48]. Thus, we define the function of $|L|_{w,*}$ as follows,

$$|L|_{w,*} = |w_i \sigma_i(M)|, \tag{3}$$

where w_i denotes the weight assigned to singular value $\sigma_i(M)$.

In this paper, we also adopted an efficient, inexact version of the augmented Lagrange multiplier (ALM) [49] to solve this convex model. The corresponding augmented Lagrange function is defined as follows:

$$\begin{aligned} J(M, L, S, \mu) = \ & |L|_{w,*} + \lambda \, |S|_1 + < J, M - L - S > \\ & + \frac{\mu}{2} |M - L - S|_F^2. \end{aligned} \tag{4}$$

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

Algorithm 1 WRPCA for SVS

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

3.2. Weighted Values

The standard nuclear norm minimization regularizes each singular value equally to pursue the convexity of the objective function. However, the RPCA method simply ignores the differences between the scales of the sparse and low-rank matrices. In order to solve this problem, and inspired by the success of weighted nuclear norm minimization [50], we adopted different weighted value strategies to trim the low-rank matrix during the SVS processing. This enables the features of the separated matrices to be better represented.

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

$$\sum = \begin{pmatrix} diag(\delta_1(M), \delta_2(M), ..., \delta_n(M)) \\ 0 \end{pmatrix}, \tag{5}$$

and $\delta_i(M)$ represents the i-th singular value of M.

Thus, we define the weight function as follows,

$$W_i^{l+1} = \frac{C}{\delta_i(L_l) + \varepsilon}, \tag{6}$$

where ε is a small positive number to avoid dividing by zero and C is a compromising constant. $C > 0$ and $0 < \varepsilon < min(\sqrt{C}, \frac{C}{\delta_1(M)})$. According to the enhancing sparsity by reweighed l_1 minimization [51], the reweighted model is described as follows,

$$L^* = U \sum{}' V^T, \tag{7}$$

where

$$\sum' = \begin{pmatrix} diag(\delta_1(L^*), \delta_2(L^*), ..., \delta_n(L^*)) \\ 0 \end{pmatrix}, \tag{8}$$

and

$$\delta_i(L^*) = \begin{cases} 0 \\ \dfrac{c_1 + \sqrt{c_2}}{2}, \end{cases} \tag{9}$$

where $c_1 = \delta_i(M) - \varepsilon$ and $c_2 = (\delta_i(M) + \varepsilon)^2 - 4C$. In this paper, the maximum matrix size was determined empirically by the regularization parameter $C = max(m, n)$ [50].

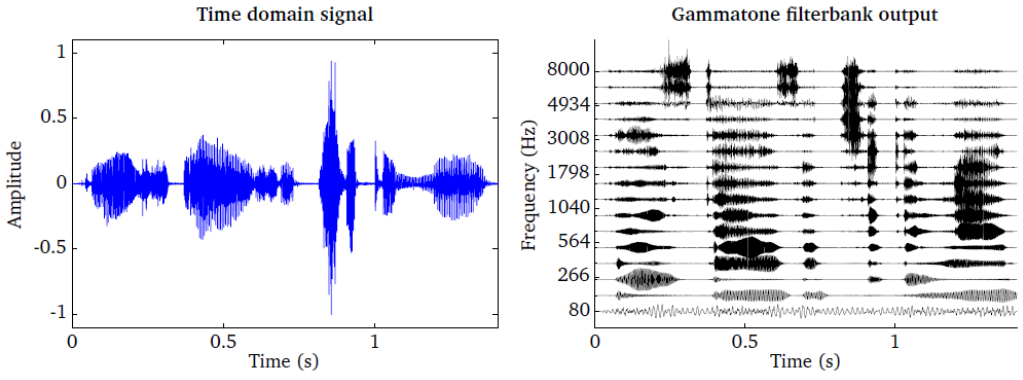
3.3. Gammatone Filterbank

The information obtained from primary auditory fibers is characterized by the gammatone function [52]. It characterizes physiological impulse-response data gathered from primary auditory fibers in the cat. Gammatone filter banks were designed to model the human auditory system. The modeling process mimics the organization of the peripheral sound processing step cites using a physiologically strategy [53]. In our work, we first pass a mixture music signal into the gammatone filterbank. Thus, the impulse response function is defined as follows,

$$h(t) = At^{N-1} \exp(-2\pi bt) \cos(2\pi f_c t + \varphi) \quad (t \geq 0, N \geq 1), \tag{10}$$

where A is an arbitrary factor, N is the filter order, b is the between the impulse functions'length and the filter's bandwidth, f_c is the center frequency, and φ is the tone phase.

In the human auditory system, there are around 3000 inner hair cells along the 35-mm spiral path cochlea. Each hair cell could resonate to a certain frequency within a suitable critical bandwidth. This means that there are approximately 3000 bandpass filters in the human auditory system. This high resolution of filters can be approximated by specifying certain overlapping between the contiguous filters. The impulse response of each filter follows the gammatone function shape. The bandwidth of each filter is determined according to the auditory critical band, which is the bandwidth of the human auditory filter at different characteristic frequencies along the cochlea path [54]. The speech signal shown in the left panel is passed through a bank of 16 gammatone filters spaced between 80 Hz and 8000 Hz. The output of each individual filter is shown in the right panel. As a result, Figure 3 depicts the gammatone filterbank.



(**a**) (**b**)

Figure 3. (**a**) Time domain signal. (**b**) The corresponding output of gammatone.

3.4. T-F Masking

After obtaining the separation results of sparse S and low-rank matrices L by using WRPCA, we applied T-F masking to further improve the separation performance. Thus, we define the ideal binary mask (IBM) and ideal ratio mask (IRM) as follows,

$$IBM = \begin{cases} 1 & S_{i,j} \geq L_{i,j} \\ 0 & S_{i,j} < L_{i,j}, \end{cases} \tag{11}$$

and

$$IRM = \frac{S_{i,j}}{S_{i,j} + L_{i,j}} \tag{12}$$

where $S_{i,j}$ and $L_{i,j}$ denote the complex spectral values of singing voice and accompaniment, respectively.

3.5. F0 Estimation

In this work, we use F0 to enhance the effectiveness of separation results. Due to the fact that F0 varies over time and is a property of the parts played by various singing voice and background accompaniment, it can greatly improve separation quality by removing the spectral components of nonrepeating instruments (e.g., bass and guitar). The salience function is defined as

$$H(t,s) = \sum_{n=1}^{N} h_n P(t, s + 1200 \log_2(n)), \tag{13}$$

where t is the sequence index and s is the logarithmic frequency. The number of harmonic components is N and the decaying fact is h_n.

The function of C can be calculated as

$$C = argmax \sum_{t=1}^{T-1} (\log a_t H(t, s_t) + \log T(s_t, s_{t+1})), \tag{14}$$

where a_t is the factor in normalization that brings the salience values to a sum of 1, and $T(s_t, s_{t+1})$ is a transition probability that denotes the likelihood of current F0 moving to the next F0 in the following sequence. Additionally, by utilizing the Viterbi search approach, the melody contour C value is optimized.

3.6. Harmonic Masking

As a result of our prior study [55], the harmonic masking is defined as

$$M_h(t,f) = \begin{cases} 1 & nF_t - \frac{w}{2} < f < nF_t + \frac{w}{2} \\ 0 & others, \end{cases} \tag{15}$$

where w is the frequency width used to extract the energy surrounding each harmonic, n is the harmonic's index, and the vocal F0 is represented by F_t at sequence t.

3.7. Coalescent Masking

We are interested in constructing coalescent masking by using harmonic masking M_h and IBM. It is possible to define the corresponding formulation M_c as follows,

$$M_c = IBM \otimes M_h, \tag{16}$$

where the elementwise multiplication operator is indicated by $\otimes$, and the time frequency masking and harmonic masking are denoted by IBM and M_h, respectively.

3.8. Vocal Activity Detection

To remove the residual music signal and restrict the values of voice and accompaniment, we are using a VAD approach. The output results s_o can be described as follows,

$$s_o = \begin{cases} s_v & \Omega > k \\ s_a & others, \end{cases} \tag{17}$$

where s_v is the state of the singing voice, s_a is the state of the background music, and k is the threshold. According to the vocal F0 estimation methods [56], the definition of the function Ω is as follows,

$$\Omega = \sum_f \{ \frac{1}{H_f} \sum_{n=1}^{H_f} P(t, s + 1200 \log_2^n) \}^{1.8}, \tag{18}$$

where $P(t, s)$ denotes the the value of power and H_f is the sum of harmonics for each frequency.

The architecture of our proposed method for the blind monaural SVS system is illustrated in Figure 4. We first apply a gammatone filterbank to the test dataset's mixture music signal to obtain the cochleagram, and then utilize proposed WRPCA approach to separate it into the L and S. By combining T-F masking and harmonic masking, we also create coalescent masking to eliminate nonseparated music. To enhance separation performance, VAD is being used. Finally, we synthesize the voice and background music. According to Wang et al. [57], the separated signal can be synthesized.

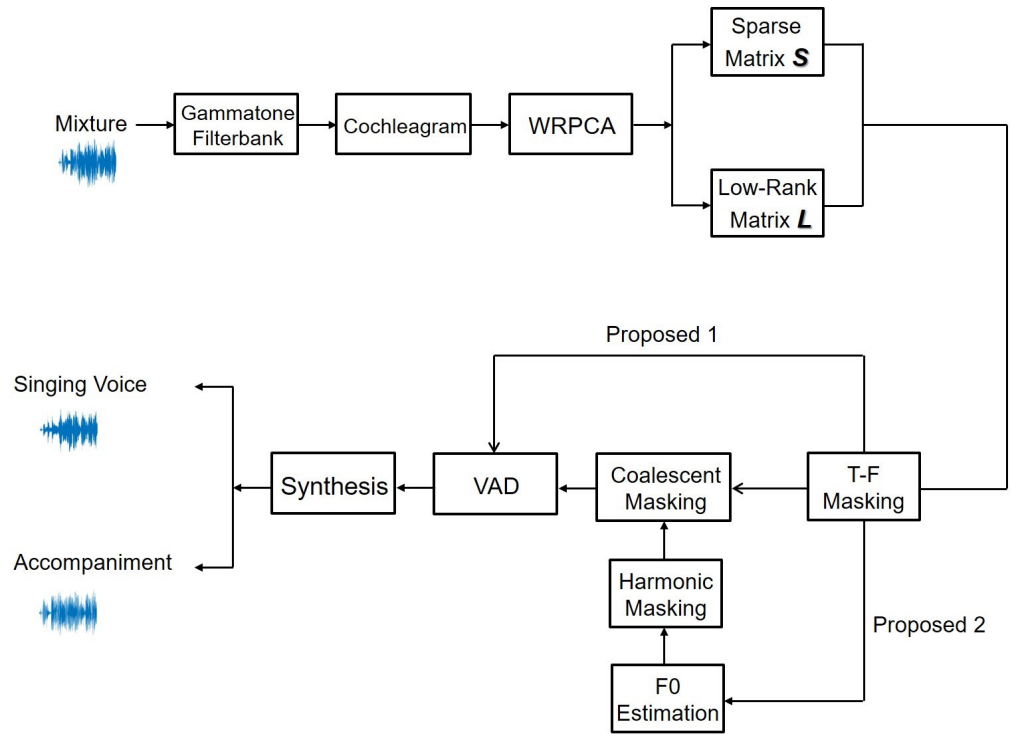


Figure 4. The architecture of the proposed SVS approach.

In this work, we randomly selected 30-s audio sample data at random in the ccMixter. The spectrograms of the isolated singing voice and background portions from mixed musical signals are illustrated in Figure 5. The original spectrograms are contrasted employing different separation approaches. As seen in the figures, the original clean singing voice and music spectrograms are shown in (a) and (b), whereas (c) and (d) exhibit the separated signal divided by RPCA. The WRPCA has split the signals (e) and (f) in the **Proposed 1**. Similarly, (g) and (h) show the separation results by the **Proposed 2**.

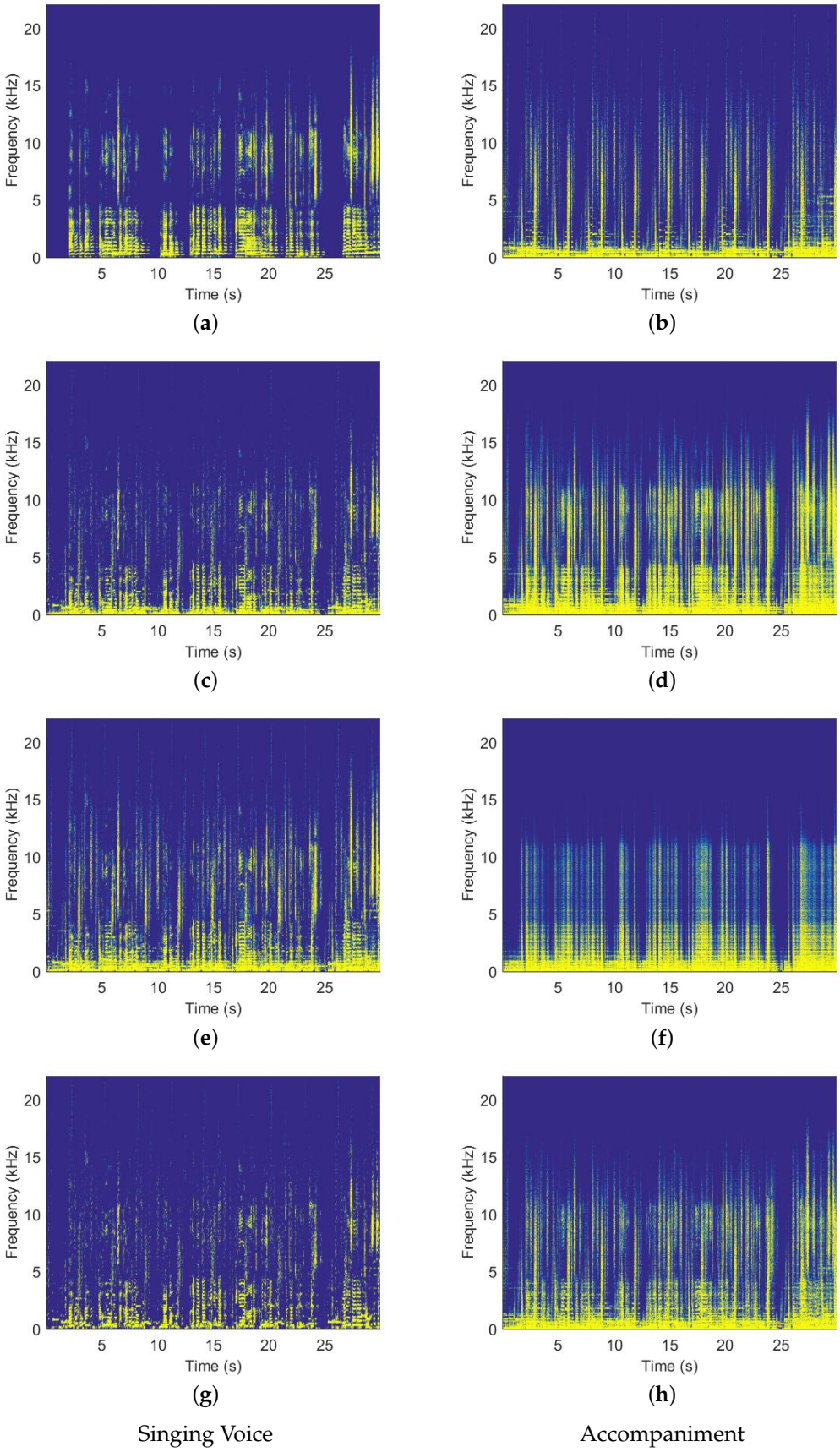


Figure 5. Examples are taken from the ccMixter dataset's spectrograms. The singing voice is represented by the left four spectrograms, whereas the equivalent musical accompaniment is represented by the right four.

From the abovementioned spectrograms, we can see that Figure 5c has the strongest background music signal (accompaniment), whereas Figure 5g has the lowest constraint. In other words, the latter is therefore preferable than the former in processing of SVS.

4. Experimental Evaluation

This section discusses the two experiments for SVS, including datasets we used on ccMixter [58] and DSD100 [59], respectively. We also present a comparison and analysis of the experiment results.

4.1. Datasets

One was the ccMixter, for which we selected 43 full stereo tracks with only 30 s (from 30 s to 1 min) at the same time of each track, and every piece of music can only contain voice for so long. Three components make up each mixture song: voice, background and the combination.

Another was the DSD100 dataset, which consisted of 36 development data and 46 test data. To reduce dimensionality and speed up speech processing, we also utilized 30-s fragments (from 1 min 45 s to 2 min 15 s) simultaneously for all data.

4.2. Settings

In our work, we focused on single-channel source separation, which is more difficult and complex than multichannel source separation because only less useful information is available. The two-channel stereo mixture datasets we used were downmixed to be mono by averaging two channels.

To evaluate our proposed approach, the spectrogram was computed by STFT by using 1024 points, and the size of hope is 256 samples. The experimental data was converted to mono after being sampled at 44.1 kHz. We established 128 channels, the frequency length ranged from 40 to 11,025 Hz, and a 256 frequency length for cochleagram analysis.

To confirm the effectiveness of our proposed algorithm, we assessed its quality of separation in terms of the source-to-distortion ratio (SDR) and the source-to-artifact ratio (SAR) by using the BSS-EVAL 3.0 metrics [60,61] and the normalized SDR (NSDR). Therefore, we define the estimated value $\hat{S}$(t) as follows,

$$\hat{S}(t) = S_{target}(t) + S_{interf}(t) + S_{artif}(t), \tag{19}$$

where $S_{target}(t)$ is the target audio's permissible distortion, $S_{interf}(t)$ denotes the allowable length change of source information to account for disturbances from unwanted sources, and $S_{artif}(t)$ denotes a potential artifact connected to the artifact of the separation technique. We therefore categorize them as follows,

$$SDR = 10log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t \{S_{interf}(t) + S_{artif}(t)\}^2}, \tag{20}$$

$$SAR = 10log_{10}\frac{\sum_t \{S_{target}(t) + S_{interf}(t)\}^2}{\sum_t S_{artif}(t)^2}, \tag{21}$$

$$NSDR(\hat{v}, v, x) = SDR(\hat{v}, v) - SDR(x, v), \tag{22}$$

where $\hat{v}$ is the estimated signal, v stands for the reference isolated, and x for the mixed music. The NSDR takes into account the SDR's overall increase from x to $\hat{v}$. The measurement units used are all dB.

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

4.3. Experiment Results

The following two approaches are presented based on the proposed WRPCA; we take them as Proposed 1 and Proposed 2, respectively. More specifically, the Proposed 1 is utilize WRPCA and T-F masking, whereas Proposed 2 adopts WRPCA and coalescent masking. Both algorithms are use VAD technology:

- **Proposed 1**: WRPCA with T-F masking
- **Proposed 2**: WRPCA with coalescent masking.

We evaluated them by using the ccMixter. The comparative outcomes for RPCA, MLRR, WRPCA, CRPCA, RPCA with IRM, WRPCA using IRM, CRPCA using IRM, RPCA using IBM, WRPCA using IBM, CRPCA using IBM, Proposed 1, and Proposed 2 are shown in Figure 6. To further completely confirm the efficacy of our proposed methodology, we designed multiple comparative experiments. The RPCA, MLRR, WRPCA, CRPCA, RPCA using IRM, RPCA using IBM, CRPCA using IRM, and CRPCA using IBM are evaluated on the spectrogram, whereas the WRPCA using IRM, WRPCA using IBM, Proposed 1, and Proposed 2 are evaluated on the cochleagram. We can find from the SDR and SAR experiment results that WRPCA performs better, especially for VAD on the cochleagram. The standard RPCA, in contrast, performed less well than the others.

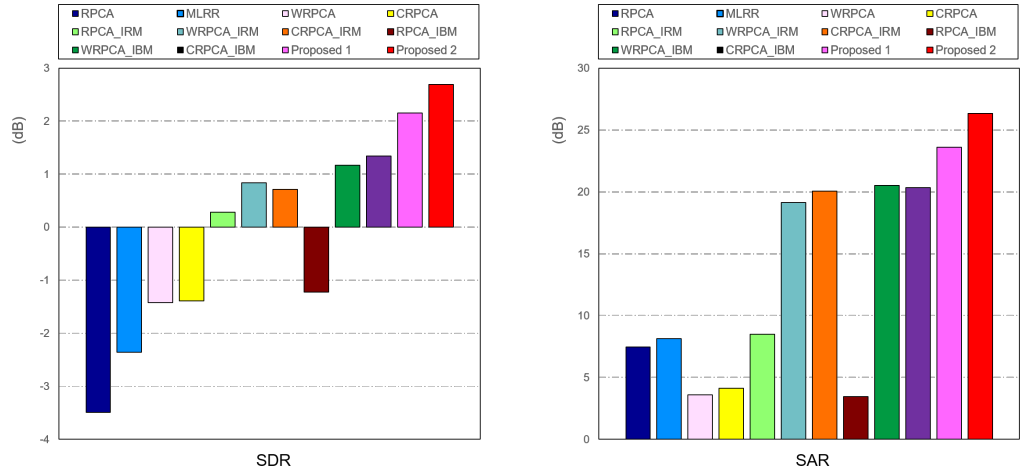


Figure 6. Comparison of SVS on **ccMixter** dataset for each of the following: RPCA, MLRR, WRPCA, CRPCA, RPCA using IRM, WRPCA using IRM, WRPCA using IRM, RPCA using IBM, WRPCA using IBM, CRPCA using IBM, Proposed 1, and Proposed 2, respectively.

Additionally, we assessed WRPCA by using the DSD100 dataset. The comparative outcomes for RPCA, MLRR, WRPCA, CRPCA, RPCA using IRM, WRPCA using IRM, CRPCA using IRM, RPCA using IBM, WRPCA using IBM, CRPCA using IRM, Proposed 1, and Proposed 2 are each shown in Figure 7. Similarly, RPCA, WRPCA, CRPCA, RPCA using IRM, RPCA using IBM, CRPCA using IRM, and CRPCA using IBM are evaluated on the spectrogram, whereas the WRPCA using IRM, WRPCA using IBM, Proposed 1, and Proposed 2 are evaluated with the cochleagram. We can find from the SDR and SAR experiment results that WRPCA performs better, especially for the VAD on the cochleagram. The standard RPCA, in contrast, performed less well than the others in Figure 6 and Figure 7, respectively.

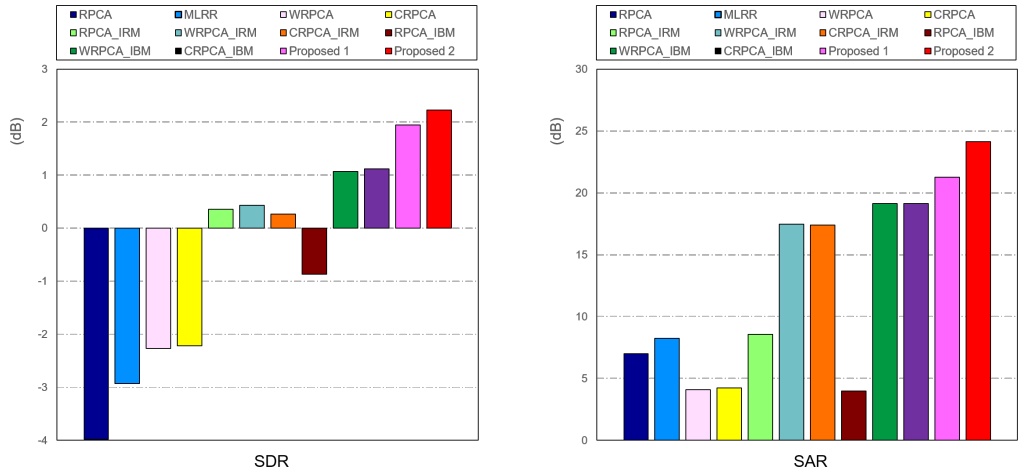


Figure 7. Comparison of SVS on **DSD100** dataset for each of the following: RPCA, MLRR, WRPCA, CRPCA, RPCA using IRM, WRPCA using IRM, WRPCA using IRM, RPCA using IBM, WRPCA using IBM, CRPCA using IBM, Proposed 1, and Proposed 2, respectively.

Figure 8 exhibits the NSDR results from the ccMixter and DSD100 datasets that we obtained by using WRPCA. In other words, the NSDR gives improved removal efficiency in SVS and overall optimizes the SDR. The results demonstrated that our Proposed 2, which was used, produced the best results.

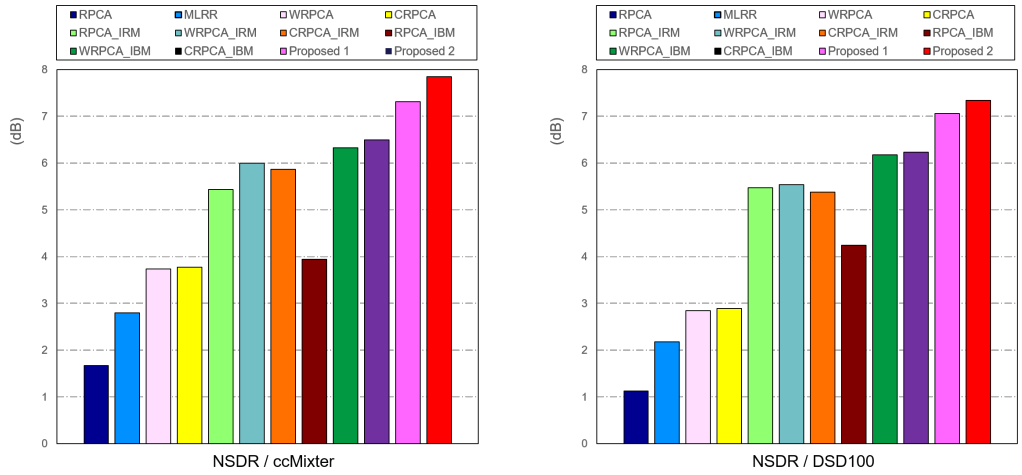


Figure 8. Comparison of SVS between **ccMixter** and **DSD100** datasets for each of the following: RPCA, MLRR, WRPCA, CRPCA, RPCA using IRM, WRPCA using IRM, WRPCA using IRM, RPCA using IBM, WRPCA using IBM, CRPCA using IBM, Proposed 1, and Proposed 2, respectively.

As a consequence, we confirm that WRPCA on a cochleagram offers higher sensitivity and selectivity than RPCA under similar circumstances with or without T-F masking based on the results of Figure 6, Figure 7, and Figure 8, respectively. Additionally, WRPCA delivered superior outcomes to RPCA by using the gammatone and T-F masking. We show that, across all evaluation modalities, our suggested strategies offer improved separation outcomes.

5. Conclusions

In this work, we proposed an extension of RPCA by using weighting on the cochleagram. The mixing signal's cochleagram was segmented into low-rank and sparse matrices by WRPCA, and the coalescent masking was constructed by integrating the harmonic and T-F masking. Finally, we constrained the temporal segments that could include the

singing voice part utilizing VAD. Evaluations on ccMixter and DSD100 datasets reveal that WRPCA performs better than RPCA for SVS, especially for WRPCA on cochleagram using gammatone and VAD approach.

In future work, to further expand the functionality of our system, we will research the vocal augmentation option. Additionally, unsupervised training that depends on the complementary nature of these two tasks will be tried because of the modest size of the public datasets that comprise both pure vocal samples and their related F0 annotations.

Author Contributions: F.L., conceptualization, methodology, formal analysis, project administration, experiment, writing. Y.H., investigation, data curation, validation, and visualization. L.W., experimental data processing and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant No. 62202001 and Innovation Support Program for Returned Overseas Students in Anhui Province under Grant No. 2021LCX032.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during and analysed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this paper:

SVS	singing voice separation
RPCA	robust principal component analysis
MLRR	multiple low-rank representation
WRPCA	weighted robust principal component analysis
VAD	voice activity detection
IBM	ideal binary mask
IRM	ideal ratio mask
ALM	augmented lagrange multiplier
APG	accelerated proximal gradient
SVD	singular value decomposition
T-F	time-frequency
CNN	convolutional neural network
NMF	nonnegative matrix factorization
SNMF	sparse nonnegative matrix factorization
SDR	source-to-distortion ratio
SAR	source-to-artifact ratio
NSDR	normalized SDR

References

1. Schulze-Forster, K.; Doire, C.S.; Richard, G.; Badeau, R. Phoneme level lyrics alignment and text-informed singing voice separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2382–2395. [CrossRef]
2. Gupta, C.; Li, H.; Goto, M. Deep Learning Approaches in Topics of Singing Information Processing. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2422–2451. [CrossRef]
3. Yu, S.; Li, C.; Deng, F.; Wang, X. Rethinking Singing Voice Separation With Spectral-Temporal Transformer. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 884–889.
4. Basak, S.; Agarwal, S.; Ganapathy, S.; Takahashi, N. End-to-end Lyrics Recognition with Voice to Singing Style Transfer. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–12 June 2021; pp. 266–270.

5. Zhang, X.; Qian, J.; Yu, Y.; Sun, Y.; Li, W. Singer identification using deep timbre feature learning with knn-net. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–12 June 2021; pp. 3380–3384.
6. Hu, S.; Liang, B.; Chen, Z.; Lu, X.; Zhao, E.; Lui, S. Large-scale singer recognition using deep metric learning: An experimental study. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–6.
7. da Silva, A.C.M.; Silva, D.F.; Marcacini, R.M. Multimodal representation learning over heterogeneous networks for tag-based music retrieval. *Expert Syst. Appl.* **2022**, *207*, 117969. [CrossRef]
8. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf. Fusion* **2022**, *83–84*, 19–52. [CrossRef]
9. Stoller, D.; Dur, S.; Ewert, S. End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 181–185.
10. Gupta, C.; Yılmaz, E.; Li, H. Automatic lyrics alignment and transcription in polyphonic music: Does background music help? In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 496–500.
11. Huang, J.; Benetos, E.; Ewert, S. Improving Lyrics Alignment Through Joint Pitch Detection. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 451–455.
12. Gupta, C.; Sharma, B.; Li, H.; Wang, Y. Lyrics-to-audio alignment using singing-adapted acoustic models and non-vocal suppression. *Music Inf. Retr. Eval. Exch. Audio-Lyrics Alignment Chall.* **2022**. Available online: https://www.music-ir.org/mirex/abstracts/2018/GSLW3.pdf (accessed on 2 January 2023).
13. Srinivasa Murthy, Y.V.; Koolagudi, S.G.; Jeshventh Raja, T.K. Singer identification for Indian singers using convolutional neural networks. *Int. J. Speech Technol.* **2021**, *24*, 781–796. [CrossRef]
14. Tuncer, T.; Dogan, S.; Akbal, E.; Cicekli, A.; Rajendra Acharya, U. Development of accurate automated language identification model using polymer pattern and tent maximum absolute pooling techniques. *Neural Comput. Appl.* **2022**, *34*, 4875–4888. [CrossRef]
15. Chen, T.P.; Su, L. Attend to chords: Improving harmonic analysis of symbolic music using transformer-based models. *Trans. Int. Soc. Music. Inf. Retr.* **2021**, *4*, 1–13. [CrossRef]
16. Chen, T.P.; Su, L. Harmony Transformer: Incorporating chord segmentation into harmony recognition. *Neural Netw.* **2019**, *12*, 15.
17. Byambatsogt, G.; Choimaa, L.; Koutaki, G. Data generation from robotic performer for chord recognition. *IEEE Trans. Electron. Inf. Syst.* **2021**, *141*, 205–213. [CrossRef]
18. Mirbeygi, M.; Mahabadi, A.; Ranjbar, A. Speech and music separation approaches—A survey. *Multimed. Tools Appl.* **2022**, *81*, 21155–21197. [CrossRef]
19. Ju, Y.; Rao, W.; Yan, X.; Fu, Y.; Lv, S.; Cheng, L.; Wang, Y.; Xie, L.; Shang, S. TEA-PSE: Tencent-ethereal-audio-lab personalized speech enhancement system for ICASSP 2022 DNS CHALLENGE. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 9291–9295.
20. Mitsufuji, Y.; Fabbro, G.; Uhlich, S.; Stöter, F.R.; Défossez, A.; Kim, M.; Choi, W.; Yu, C.Y.; Cheuk, K.W. Music demixing challenge 2021. *Front. Signal Process.* **2022**, *1*, 18. [CrossRef]
21. Ji, X.; Han, J.; Jiang, X.; Hu, X.; Guo, L.; Han, J.; Shao, L.; Liu, T. Analysis of music/speech via integration of audio content and functional brain response. *Inf. Sci.* **2015**, *297*, 271–282. [CrossRef]
22. Chen, K.; Yu, S.; Wang, C.I.; Li, W.; Berg-Kirkpatrick, T.; Dubnov, S. Tonet: Tone-octave network for singing melody extraction from polyphonic music. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 621–625.
23. Petermann, D.; Wichern, G.; Wang, Z.Q.; Le Roux, J. The cocktail fork problem: Three-stem audio separation for real-world soundtracks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 526–530.
24. Yu, Y.; Peng, C.; Tang, Q.; Wang, X. Monaural Music Source Separation Using Deep Convolutional Neural Network Embedded with Feature Extraction Module. In Proceedings of the 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), Shanghai, China, 17–19 March 2022; pp. 546–551.
25. Hu, Y.; Chen, Y.; Yang, W.; He, L.; Huang, H. Hierarchic Temporal Convolutional Network With Cross-Domain Encoder for Music Source Separation. *IEEE Signal Process. Lett.* **2022**, *29*, 1517–1521. [CrossRef]
26. Guizzo, E.; Weyde, T.; Tarroni, G. Anti-transfer learning for task invariance in convolutional neural networks for speech processing. *Neural Netw.* **2021**, *142*, 238–251. [CrossRef]
27. Ni, X.; Ren, J. FC-U 2-Net: A Novel Deep Neural Network for Singing Voice Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 489–494. [CrossRef]
28. Xu, Y.; Wang, W.; Cui, H.; Xu, M.; Li, M. Paralinguistic singing attribute recognition using supervised machine learning for describing the classical tenor solo singing voice in vocal pedagogy. *EURASIP J. Audio Speech Music Process.* **2022**, *2022*, 1–16. [CrossRef]

29. Zhou, Y.; Lu, X. HiFi-SVC: Fast High Fidelity Cross-Domain Singing Voice Conversion. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6667–6671.
30. Kum, S.; Lee, J.; Kim, K.L.; Kim, T.; Nam, J. Pseudo-Label Transfer from Frame-Level to Note-Level in a Teacher-Student Framework for Singing Transcription from Polyphonic Music. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 796–800.
31. Wang, Y.; Stoller, D.; Bittner, R.M.; Bello, J.P. Few-Shot Musical Source Separation. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 121–125.
32. Zhang, X.; Wang, J.; Cheng, N.; Xiao, J. Mdcnn-sid: Multi-scale dilated convolution network for singer identification. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padova, Italy, 18–23 July 2022; pp. 1–7.
33. Wang, D.; Chen, J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1702–1726. [CrossRef]
34. Huang, P.S.; Chen, S.D.; Smaragdis, P.; Hasegawa-Johnson, M. Singing-voice separation from monaural recordings using robust principal component analysis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 57–60.
35. Yang, Y.-H. On sparse and low-rank matrix decomposition for singing voice separation. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012; pp. 757–760.
36. Yang, Y.-H. Low-Rank Representation of Both Singing Voice and Music Accompaniment Via Learned Dictionaries. In Proceedings of the ISMIR, Curitiba, Brazil, 4–8 November 2013; pp. 427–432.
37. Li, F.; Akagi, M. Unsupervised singing voice separation based on robust principal component analysis exploiting rank-1 constraint. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 1920–1924.
38. Yuan, S.; Wang, Z.; Isik, U.; Giri, R.; Valin, J.M.; Goodwin, M.M.; Krishnaswamy, A. Improved singing voice separation with chromagram-based pitch-aware remixing. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 111–115.
39. Gao, B.; Woo, W.L.; Ling, B.W.K. Machine learning source separation using maximum a posteriori nonnegative matrix factorization. *IEEE Trans. Cybern.* **2013**, *44*, 1169–1179.
40. Gao, B.; Woo, W.L.; Tian, G.Y.; Zhang, H. Unsupervised diagnostic and monitoring of defects using waveguide imaging with adaptive sparse representation. *IEEE Trans. Ind. Inform.* **2015**, *12*, 405–416. [CrossRef]
41. Li, F.; Akagi, M. Weighted robust principal component analysis with gammatone auditory filterbank for singing voice separation. In Proceedings of the Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, 14–18 November 2017; Springer: Cham, Switzerland, 2017; pp. 849–858.
42. Li, Y.P.; Wang, D.L. On the optimality of ideal binary time-frequency masks. *Speech Commun.* **2009**, *51*, 230–239. [CrossRef]
43. Healy, E.W.; Vasko, J.L.; Wang, D. The optimal threshold for removing noise from speech is similar across normal and impaired hearing—A time-frequency masking study. *J. Acoust. Soc. Am.* **2019**, *145*, EL581–EL586. [CrossRef]
44. Luo, Y.; Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266. [CrossRef]
45. Fujihara, H.; Goto, M.; Ogata, J.; Okuno, H.G. Lyric Synchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 1252–1261. [CrossRef]
46. Lehner, B.; Widmer, G.; Bock, S. A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks. In Proceedings of the 2015 23rd European signal processing conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 21–25.
47. Ramona, M.; Richard, G.; David, B. Vocal detection in music with support vector machines. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 1885–1888.
48. Candès, E.J.; Li, X.; Ma, Y.; Wright, J. Robust principal component analysis? *J. ACM (JACM)* **2011**, *58*, 1–37. [CrossRef]
49. Lin, Z.; Chen, M.; Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv* **2010**, arXiv:1009.5055.
50. Gu, S.H.; Xie, Q.; Meng, D.Y.; Zuo, W.M.; Feng, X.C.; Zhang, L. Weighted nuclear norm minimization and its applications to low level vision. *Int. J. Comput. Vis.* **2017**, *121*, 183–208. [CrossRef]
51. Candes, E.J.; Wakin, M.B.; Boyd, S.P. Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Appl.* **2008**, *14*, 877–905. [CrossRef]
52. Johannesma, P.L.M. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In *Symposium on Hearing Theory*; IPO: Bristol, UK, 1972.
53. Abdulla, W.H. Auditory based feature vectors for speech recognition systems. *Adv. Commun. Softw. Technol.* **2002**, 231–236.
54. Zhang, Y.; Abdulla, W.H. Gammatone auditory filterbank and independent component analysis for speaker identification. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006; pp. 2098–2101.
55. Li, F.; Akagi, M. Blind monaural singing voice separation using rank-1 constraint robust principal component analysis and vocal activity detection. *Neurocomputing* **2019**, *350*, 44–52. [CrossRef]

56. Salamon, J.; Gomez, E. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 1759–1770. [CrossRef]
57. Wang, D.L.; Brown, G.J. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*; Wiley-IEEE Press: Hoboken, NJ, USA, 2006.
58. Liutkus, A.; Fitzgerald, D.; Rafii, Z. Scalable audio separation with light kernel additive modelling. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 76–80.
59. Liutkus, A.; Stoter, F.R.; Rafii, Z.; Kitamura, D.; Rivet, B.; Ito, N.; Ono, N.; Fontecave, J. The 2016 signal separation evaluation campaign. In Proceedings of the Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, 21–23 February 2017; pp. 323–332.
60. Stöter, F.R.; Liutkus, A.; Ito, N. The 2018 signal separation evaluation campaign. In Proceedings of the Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Guildford, UK, 2–5 July 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 293–305.
61. Vincent, E.; Gribonval, R.; Févotte, C. Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1462–1469. [CrossRef]