LOW-RANK REPRESENTATION OF BOTH SINGING VOICE AND MUSIC ACCOMPANIMENT VIA LEARNED DICTIONARIES

Yi-Hsuan Yang
Research Center for IT Innovation, Academia Sinica, Taiwan
yang@citi.sinica.edu.tw

ABSTRACT

Recent research work has shown that the magnitude spectrogram of a song can be considered as a superposition of a low-rank component and a sparse component, which appear to correspond to the instrumental part and the vocal part of the song, respectively. Based on this observation, one can separate singing voice from the background music. However, the quality of such separation might be limited, because the vocal part of a song can sometimes be low-rank as well. Therefore, we propose to learn the subspace structures of vocal and instrumental sounds from a collection of clean signals first, and then compute the low-rank representations of *both* the vocal and instrumental parts of a song based on the learned subspaces. Specifically, we use online dictionary learning to learn the subspaces, and propose a new algorithm called multiple low-rank representation (MLRR) to decompose a magnitude spectrogram into two low-rank matrices. Our approach is flexible in that the subspaces of singing voice and music accompaniment are both learned from data. Evaluation on the MIR-1K dataset shows that the approach improves the source-to-distortion ratio (SDR) and the source-to-interference ratio (SIR), but not the source-to-artifact ratio (SAR).

1. INTRODUCTION

A musical piece is usually composed of multiple layers of voices sounded simultaneously, such as human vocal, melody line, bass line and percussion. These components are mixed in most songs sold in the market. For many music information retrieval (MIR) problems, such as predominant instrument recognition, artist identification and lyrics alignment, separating one source from the others is usually an important pre-processing step [6, 9, 13].

Many algorithms have been proposed for blind source separation in monaural music signals [21, 22]. For the particular case of separating singing voice from music accompaniment, it has been found that characterizing the music accompaniment as a repeating structure on which varying vocals are superimposed leads to good separation quality [8, 16, 17, 23]. For example, Huang *et al.* [8] found that, by decomposing the magnitude spectrogram of a song into a low-rank matrix and a sparse matrix, the sparse component appears to correspond to the singing voice. Evaluation on the MIR-1K data set [7] shows that such a low-rank decomposition (LRD) method outperforms sophisticated, pitch-based inference methods [7, 22].

However, the low-rank and sparsity assumptions about the music accompaniment and singing voice have not been carefully studied so far. From mathematical point of view, the low-rank component corresponds to a succinct representation of the observed data in a lower dimensional subspace, whereas the sparse component corresponds to the (small) fraction of the data samples that are far away from the subspace [2, 11]. Without any prior knowledge of the data, it is not easy to distinguish between data samples originated from the subspace of music accompaniment and those from the subspace of singing voice. Therefore, the low-rank matrix resulting from the aforementioned decomposition might be actually a mixture of the subspaces of vocal and instrumental sounds, and the sparse matrix might contain a portion of the instrumental sounds such as the main melody or the percussion sounds [23].

Because MIR-1K comes with "clean" vocal and instrumental sources recorded separately at the left and right channels, in our pilot study we tried LRD using principal component analysis (PCA) [2] for the two clean sources, respectively. Result shows that, contrary to the sparsity assumption, the vocal channel can also be well approximated by a low-rank matrix. As Figure 1 exemplifies, we are able to reduce the rank of the singing voice and the music accompaniment matrices (by PCA) from 513 to 50 and 10, respectively, with less than 40% loss in the source-to-distortion ratio (SDR) [20].

Motivated by the above observation, in this paper we investigate the quality of separation as a result of decomposing the magnitude spectrogram of a song into "two" low-rank matrices plus one sparse matrix. The first two matrices represent the singing voice and music accompaniment in the subspaces of vocal and instrumental sounds, respectively, whereas the last matrix contains data samples deviated from the subspaces. Therefore, unlike existing methods, the vocal part of a song is also modeled as a low-rank signal. Moreover, different subspaces are explicitly used for vocal and instrumental sounds.

To achieve the above decomposition, we propose a new algorithm called multiple low-rank representation (MLRR),

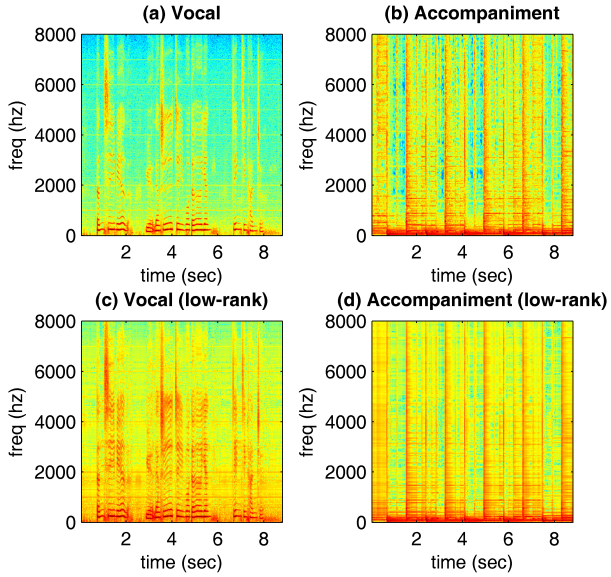

Figure 1. (a) (b) The original, full-rank magnitude spectrograms (in log scale) of the vocal and instrumental parts of the clip 'Ani_1_01' in MIR-1K [7]. (c) (d) The low-rank matrices of the vocal part (rank=50) and the instrumental part (rank=10) obtained by PCA. Such low-rank approximation only incurs 40% loss in signal-to-distortion ratio.

which involves an iterative optimization process that seeks the lowest rank representation [2, 10, 11]. Moreover, instead of decomposing a signal from scratch, we employ an online dictionary learning algorithm [12] to learn the subspace structures of the vocal and instrumental sounds in advance from an external collection of clean vocal and instrumental signals. In this way, we are able to incorporate prior knowledge about the nature of vocal and instrumental sounds to the decomposition process.

The paper is organized follows. Section 2 reviews related work on LRD its application to singing voice separation. Section 3 describes the proposed algorithms. Section 4 presents the evaluation and Section 5 concludes.

2. REVIEW ON LOW-RANK DECOMPOSITION

It has been shown that many real-world data can be well characterized by low-dimensional subspaces [11]. That is, if we put n m-dimensional data vectors in the form of a matrix $X \in \Re^{m \times n}$, X should have rank $r \ll \min(m, n)$, meaning few linearly independent columns [2]. The goal of LRD is to obtain a low-rank approximation of X in the presence of outliers, noises, or missing values [11].

The classical principal component analysis (PCA) [2] seeks a rank-r estimate A of the matrix X by solving

$$\begin{aligned} \min_{A} \quad & \|X - A\| \\ \text{subject to} \quad & \text{rank}(A) \leq r \,, \end{aligned} \tag{1}$$

where $\|X\|$ denotes the spectral norm, or the largest singular value of X. This problem can be efficiently solved via singular value decomposition (SVD) by using the r largest singular values [2].

It is well-known that PCA is sensitive to outliers. To remedy this issue, robust PCA (RPCA) [2] uses the l_1 norm to characterize sparse corruptions and solves

$$\min_{A} \quad \|A\|_* + \lambda \|X - A\|_1 \,, \tag{2}$$

where $\|\cdot\|_*$ denotes the nuclear norm (the sum of its singular values), $\|\cdot\|_1$ is the l_1 norm that sums the absolute values of matrix entries, and λ is a positive weighting parameter. The use of nuclear norm as a surrogate of the rank function makes it possible to solve (2) by convex optimization algorithms such as accelerated proximal gradient (APG) or augmented Lagrange multipliers (ALM) [10].

RPCA has been successfully applied to singing voice separation [8]. Researchers found that the resulting sparse component (i.e., $X - A$) appears to correspond to the vocal part and the low-rank one (i.e., A) corresponds to the music accompaniment. More recently, Yang [23] found that the sparse component often contains percussion sounds and proposed a back-end drum removal procedure to enhance the quality of the separated singing voice. Sprechmann *et al.* [17] considered both A and $X - A$ to be non-negative and employed multiplicative algorithms to solve the resulting robust non-negative matrix factorization (RNMF) problem. Efficient, supervised or semi-supervised variants have also been proposed [17]. Although promising result is obtained, none of the reviewed methods justified the assumption of considering singing voice as sparse.

Durrieu *et al.* [3] proposed a non-negative matrix factorization (NMF)-based method for singing voice separation that regards the vocal spectrogram as an element-wise multiplication of an excitation spectrogram and a filter spectrogram. Many other NMF-based methods that do not rely on the sparse assumption have also been proposed [14]. However, we tend to focus on LRD-based methods that have similar form as RPCA in this work. The comparison with NMF-based methods is left as a future work.

Finally, low-rank representation (LRR) [11] seeks the lowest rank estimate of data X with respect to $D \in \Re^{m \times k}$, a "dictionary" that is assumed to linearly span the space of the data being analyzed. Specifically, it solves

$$\min_{Z} \quad \|Z\|_* + \lambda \|X - DZ\|_1 \,, \tag{3}$$

where $Z \in \Re^{k \times n}$ and k denotes the dictionary size. Since $\text{rank}(DZ) \leq \text{rank}(Z)$, DZ is also a low-rank recovery to X. As discussed in [11], by properly choosing D, LRR can recover data drawn from a mixture of several low-rank subspaces. By setting $D = I_m$, the $m \times m$ identify matrix, the formulation (3) reduces to (2). Although it is possible to use dictionary learning algorithms such as K-SVD [1] to learn a dictionary from data, Liu *et al.* [11] simply set $D = X$, using the data matrix itself as the dictionary. In contrast, we extend LRR to the case of multiple dictionaries and employ online dictionary learning (ODL) [12] to learn the dictionaries, as described below.

3. PROPOSED ALGORITHMS

By extending formulation (3), we are able to obtain the low-rank representations of X with respect to multiple dic-

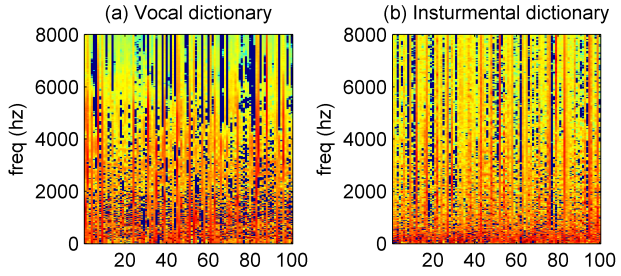

Figure 2. The spectra (in log scale) of the learned dictionaries (with 100 codewords) for (a) vocal and (b) instrumental spectra, using online dictionary learning.

tionaries $D_1, D_2, \ldots, D_\kappa$, where κ denotes the number of dictionaries. Although it is possible to use a dictionary for each musical component (e.g., human vocal, melody line, bass line and percussion), we consider the case $\kappa = 2$ and use one dictionary for human vocal and the other for the music accompaniment.

3.1 Multiple Low-Rank Representation (MLRR)

Given an input data X and two pre-defined (or pre-learned) dictionaries $D_1 \in \Re^{m\times k_1}$ and $D_2 \in \Re^{m\times k_2}$ (k_1 and k_2 can take different values), MLRR seeks the lowest rank matrices Z_1 and Z_2 by solving

$$\min_{Z_1,Z_2} \|Z_1\|_* + \beta \|Z_2\|_* + \lambda \|X - D_1Z_1 - D_2Z_2\|_1 , \quad (4)$$

where β is a positive parameter. This optimization problem can be solved by the method of ALM [10], by first reformulating (4) as

$$\begin{aligned} \min_{Z_1,Z_2,J_1,J_2,E} \quad & \|J_1\|_* + \beta \|J_2\|_* + \lambda \|E\|_1 \\ \text{subject to} \quad & X = D_1Z_1 + D_2Z_2 + E , \\ & Z_1 = J_1 , \; Z_2 = J_2 , \end{aligned} \quad (5)$$

and then minimizing the augmented Lagrangian function

$$\begin{aligned} \mathcal{L} = \; & \|J_1\|_* + \text{tr}(Y_1^T(Z_1 - J_1)) + \frac{\mu}{2}\|Z_1 - J_1\|_F^2 \\ + \; & \beta \|J_2\|_* + \text{tr}(Y_2^T(Z_2 - J_2)) + \frac{\mu}{2}\|Z_2 - J_2\|_F^2 \\ + \; & \lambda \|E\|_1 + \text{tr}(Y_3^T(X - D_1Z_1 - D_2Z_2 - E)) \\ + \; & \frac{\mu}{2}\|X - D_1Z_1 - D_2Z_2 - E\|_F^2 , \end{aligned} \quad (6)$$

where $\|\cdot\|_F$ denotes the Frobenius norm (square root of the sum of the squares of its elements) and μ is a positive penalty parameter. We can minimize (6) with respect to Z_1, Z_2, J_1, J_2, E, respectively, by fixing the other variables and then updating the Lagrangian multipliers Y_1, Y_2 and Y_3. For example, J_2 can be updated by

$$J_2^* = \text{argmin}\; \beta \|J_2\|_* + \frac{\mu}{2}\|J_2 - (Z_2 + \mu^{-1}Y_2)\|_F^2 , \quad (7)$$

which can be solved via the singular value thresholding (SVT) operator [2], whereas Z_1 can be updated by

$$Z_1^* = \Sigma_1 \left(D_1^T(X - D_2Z_2 - E) + J_1 + \mu^{-1}(D_1^TY_3 - Y_1)\right) , \quad (8)$$

where $\Sigma_1 = (I + D_1^TD_1)^{-1}$. The update rule for the other variables can be obtained in a similar way as described in [10, 11], mainly by taking the first-order derivative of the augmented Lagrangian function $\mathcal{L}$ with respect to the variable. By using a non-decreasing sequence of $\{\mu_t\}$ as suggested in [10] (i.e., using μ_t in the t-th iteration), empirically we observe that the optimization usually converges in 100 iterations. After the decomposition, we consider D_1Z_1 and D_2Z_2 as the vocal and instrumental parts of the song and discard the intermediate matrices E, J_1 and J_2.

3.2 Learning the Subspace Structures of Singing and Instrumental Sounds

The goal of dictionary learning is to find a proper representation of data by means of reduced dimensionality subspaces, which are adaptive to both the characteristics of the observed signals and the processing task at hand [19]. Many dictionary learning algorithms have been proposed, such as $k$means and K-SVD [1, 19]. In this work, we adopt the online dictionary learning (ODL) [12], a first-order stochastic gradient descent algorithm, for its low memory consumption and computational cost. ODL has been used in many MIR tasks such as genre classification [24].

Given N signals $p_i \in \Re^m$, ODL learns a dictionary D by solving the following joint optimization problem,

$$\begin{aligned} \min_{D,Q} \quad & \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\|p_i - Dq_i\|_2^2 + \eta\|q_i\|_1\right) , \\ \text{subject to} \quad & d_j^Td_j \le 1 , q_i \ge 0 , \end{aligned} \quad (9)$$

where $\|\cdot\|_2$ denotes the Euclidean norm for vectors, Q denotes the collection of the (unknown) nonnegative encoding coefficients $q_i \in \Re^k$, and η is a regularization parameter. The dictionary D is composed of k codewords $d_j \in \Re^m$, whose energy is limited to be less than one. Formulation (9) can be solved by updating D and Q in an alternating fashion. The optimization of q_i involves a typical sparse coding problem that can be solved by the LARS-lasso algorithm [4]. Our implementation of ODL is based on the SPAMS toolbox [12]. [1]

Figure 2 shows the dictionaries for vocal and instrumental spectra we learned from a subset of MIR-1K, using $k_1 = k_2 = 100$. It can be found that the vocal dictionary contains voices of higher fundamental frequency. In addition, we see more energy in the so-called "singer's formant" (around 3 khz) from the vocal dictionary [18], showing that the two dictionaries capture distinct characteristics of the signals. Finally, we also observe some atoms that span almost the whole spectra in both dictionaries (e.g., the 12th codeword in the instrumental dictionary), possibly because of the need to reconstruct a signal by a sparse subset of the dictionary atoms, by virtue of the l_1-based sparsity constraint in formulation (9).

In principle, we can improve the reconstruction accuracy (i.e., smaller $\|p_i - Dq_i\|_2$ in (9)) by using larger k [12], at the expense of increasing the computational cost in solving both (9) and (5). However, as Section 4.1 shows, larger

[1] http://spams-devel.gforge.inria.fr/

k does not necessarily lead to better separation quality, possibly because of the mismatch between the goals of reconstruction and of separation.

The source codes, sound examples, and more details of this work are available online. [2]

4. EVALUATION

Our evaluation is based on the MIR-1K dataset collected by Hsu & Jang [7]. [3] It contains 1,000 song clips extracted from 110 Chinese pop songs released in karaoke format, which consists of a clean music accompaniment track and a mixture track. A total number of eight female and 11 male amateur singers were invited to sing the songs, thereby creating the clean singing voice track for each clip. Each clip is 4 to 13 seconds in length and sampled at 16 khz. Although MIR-1K also comes with human-labeled pitch values, unvoiced sounds and vocal/nonvocal segments, lyrics, and the speech recordings of the lyrics for each clip [7], these information are not exploited in this work.

Following [17], we reserved 175 clips sang by one male and one female singers ('abjones' and 'amy') for training (i.e., learning the dictionaries D_1 and D_2), and used the remaining 825 clips of 17 singers for testing the performance of separation. For the test clips, we mixed the two sources v and a linearly with equal energy (i.e., 0 db signal-to-noise ratio) to generate x, the mixture of sounds similar to the one available from commercial CDs. The goal is to recover v and a from x for each test clip separately.

Given a music clip, we first computed its short-time Fourier transform (STFT) by sliding a Hamming window of 1024 samples and 1/4 overlapping (as in [8]) to obtain the spectrogram, which consists of the magnitude part X and the phase part P. We applied matrix decomposition using X to get the separated sources. To synthesize the time-domain waveforms $\hat{v}$ and $\hat{a}$, we performed inverse STFT using the magnitude spectrogram of the separated source and the phase P of the original signal [5]. Because the separated spectrogram may contain negative values, we converted negative values to zero before inverse STFT.

The quality of separation is assessed in terms of the following measures [20], which are computed for the vocal part v and the instrumental part a, respectively,

- Source-to-distortion ratio (SDR), which measures the energy ratio between the source and the distortion (e.g., v to $v - \hat{v}$).
- Source-to-artifact ratio (SAR), which measures the amount of artifacts of the source separation algorithm such as musical noise.
- Source-to-interference ratio (SIR), which measures the interference from other sources.

Higher values of these ratios indicate better separation quality. We computed these ratios by using the BSS Eval toolbox v3.0, [4] assuming that the admissible distortion is a time-invariant filter [20]. As in [7], we compute the *normalized* SDR (NSDR) by $\text{SDR}(\hat{v}, v) - \text{SDR}(x, v)$. Moreover, we aggregate the performance over all the test clips by taking the weighted average, with weight proportional to the length of each clip [7]. The resulting measures are denoted as GNSDR, GSAR, and GSIR, respectively (the later two are not normalized). [5]

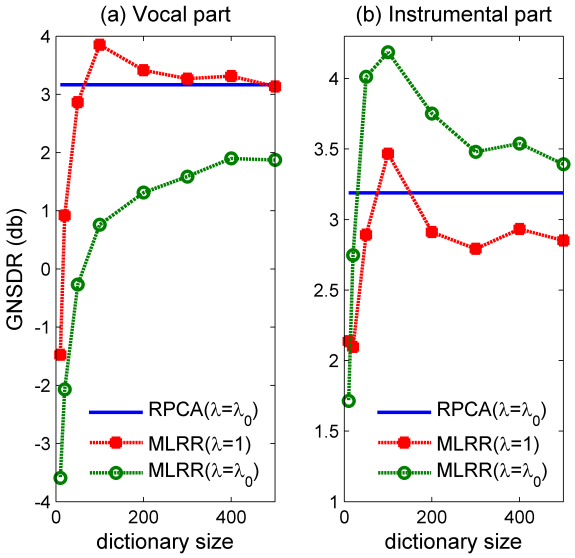

Figure 3. The quality of the separated (a) vocal and (b) instrumental parts of the 825 clips in MIR-1K in terms of global normalized source-to-distortion ratio (GNSDR).

4.1 Result

We first compared the performance of MLRR with RPCA, one of the state-of-the-art algorithms for singing voice separation [8]. We used ALM-based algorithm for both MLRR and RPCA [10]. For MLRR, we learned dictionaries from the training set and evaluate separation on the test set of MIR-1K. Although it is interested to use different dictionary sizes for the vocal and instrumental dictionaries, we set $k_1 = k_2 = k$ in this study. For RPCA, we simply evaluated it on the test set, without using the training set. The value of λ was set to either $\lambda_0 = 1/\sqrt{\max(m, n)}$, according to [2] (recall that (m, n) is the size of the input matrix X), or 1, as suggested in [11]. We only use λ_0 for RPCA because using 1 did not work. Moreover, we simply set β to 1 for MLRR. For future work it would be interesting to use different β to investigate whether we want to penalize the rank of one particular source more. [6]

Figure 3 shows the quality (in terms of GNSDR) of the separated vocal and instrumental parts using different algorithms, different values of the parameter λ and different values of the dictionary size k. We found that MLRR attains the best result when $k = 100$ for both parts (3.85 db and 4.19 db). The performance difference in GNSDR be-

[2] http://mac.citi.sinica.edu.tw/mlrr

[3] https://sites.google.com/site/unvoicedsoundseparation/

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

[5] Please note that in some previous work the older version BSS Eval toolbox v2.1 was used [7, 8, 23], assuming that the admissible distortion is purely a time-invariant gain.

[6] In fact, when $\beta = 1$ one can combine Z_1 and Z_2, reducing (4) to (3), and used an LRR-based algorithm to solve the problem as well.

Table 1. Separation quality (in db) for the singing voice

Method	GNSDR	GSIR	GSAR
RPCA (λ=λ_0) [8]	3.17	4.43	11.1
RPCAh (λ=λ_0) [23]	3.25	4.52	11.1
RPCAh+FASST [23]	3.84	6.22	9.19
MLRR (k=100, λ=1)	3.85	5.63	10.7

Table 2. Separation quality for the music accompaniment

Method	GNSDR	GSIR	GSAR
RPCA (λ=λ_0) [8]	3.19	5.24	9.23
RPCAh (λ=λ_0) [23]	3.27	5.31	9.30
RPCAh+FASST [23]	3.21	5.24	9.30
MLRR (k=100, λ=λ_0)	4.19	7.80	8.22

tween MLRR (when $k = 100$) and RPCA is significant, either for the vocal or instrumental part, under one-tailed t-test (p-value$<$0.001; d.f.=1648). [7]

From Figure 3, several observations can be made. First, it can be found that using larger k does not always lead to better performance, as discussed in Section 3.2. Second, for the instrumental part, using $k = 20$ ($\lambda = \lambda_0$) already yields high GNSDR (2.74 db), whereas for the vocal part we need to use at least $k = 50$ ($\lambda = 1$). This result shows that we need more dictionary atoms to represent the space of the singing voice, possibly because the subspace of singing voice is of higher rank (cf. Figure 1). The separation quality of the singing voice is worse (i.e., lower than zero) when k is too small. Third, we saw that the vocal and instrumental parts favor different values of λ for MLRR, which deserves future study. [8]

Next, we compared MLRR with the two algorithms presented in [23], in terms of more performance measures. RPCAh is an APG-based algorithm that uses harmonicity priors to take into account the similarity between sinusoidal elements [23]; RPCAh+FASST employs Flexible Audio Source Separation Toolbox for removing the drum sounds in the vocal part [15]. Because FASST involves a heavy computational process, we set the maximal number of iterations to 100 in this evaluation. [9]

Result shown in Tables 1 and 2 indicates that, except for the GSIR for singing voice, MLRR outperforms all the evaluated RPCA-based methods [8,23] in terms of GNSDR and GSIR, especially for the music accompaniment. However, we also found that MLRR introduces some artifacts and leads to slightly lower GSAR. This is possibly because the separated sounds are linear combination of the dictionary atoms, which may not be comprehensive enough to capture every nuance of music signals.

Finally, to provide a visual comparison, Figure 4 shows the separation result for RCA (λ=λ_0), RCAh+FASST, and MLRR (k=100, λ=1) for the clip 'Ani_1_01,' focusing on low frequency parts 0–4 khz. We saw that the recovered vocal signal well captures the main vocal melody, and that components with strong harmonic structure are present in the recovered instrumental part. We also observed undesirable artifacts in the higher frequency components of MLRR, which should be the subject of future research.

[7] We have tried imposing a nonnegative constraint on the dictionary D (c.f. Eq. 9) but this did not further improve the result.

[8] It is fair to use different λ for the two sources; for example, if the application is about analyzing singing voice, one can use λ=1.

[9] We did not compare our result with another two state-of-the-art methods [17] and [16], because somehow we cannot reproduce the result for the former and because the latter did not evaluate on MIR-1K. Moreover, please note that the evaluation here is performed on 825 clips (excluding those used for dictionary learning) instead of the whole MIR-1K.

5. CONCLUSION AND DISCUSSION

In this paper, we have presented a time-frequency based source separation algorithm for music signals that considers both the vocal and instrumental spectrograms as low-rank matrices. The technical contributions we have brought to the field include the use of dictionary learning algorithms to estimate the subspace structures of music sources and the development of a novel algorithm MLRR that uses the learned dictionaries for decomposition. The proposed method is advantageous in that potentially more training data can be harvested to improve the result of separation. Although it might not be fair to directly compare the performance of MLRR and RPCA (because the former uses an external dictionary), our result shows that we can still get similar separation quality without the sparse assumption on the singing voice. However, because the separated sounds are linear combination of the atoms in the pre-learned dictionaries, there are some unwanted artifacts that are audible, which should be the subject of future work.

6. ACKNOWLEDGMENTS

This work was supported by the National Science Council of Taiwan under Grants NSC 101-2221-E-001-017, NSC 102-2221-E-001-004-MY3 and the Academia Sinica Career Development Award.

7. REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, 54(11):4311–4322, 2006.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

[3] J.-L. Durrieu, G. Richard, and B. David. An iterative approach to monaural musical mixture de-soloing. In *Proc. ICASSP*, pages 105–108, 2009.

[4] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[5] D. Ellis. A phase vocoder in Matlab, 2002. [Online] http://www.ee.columbia.edu/ dpwe/resources/matlab/pvoc/.

[6] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno. Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics. *J. Sel. Topics Signal Processing*, 5(6):1252–1261, 2011.

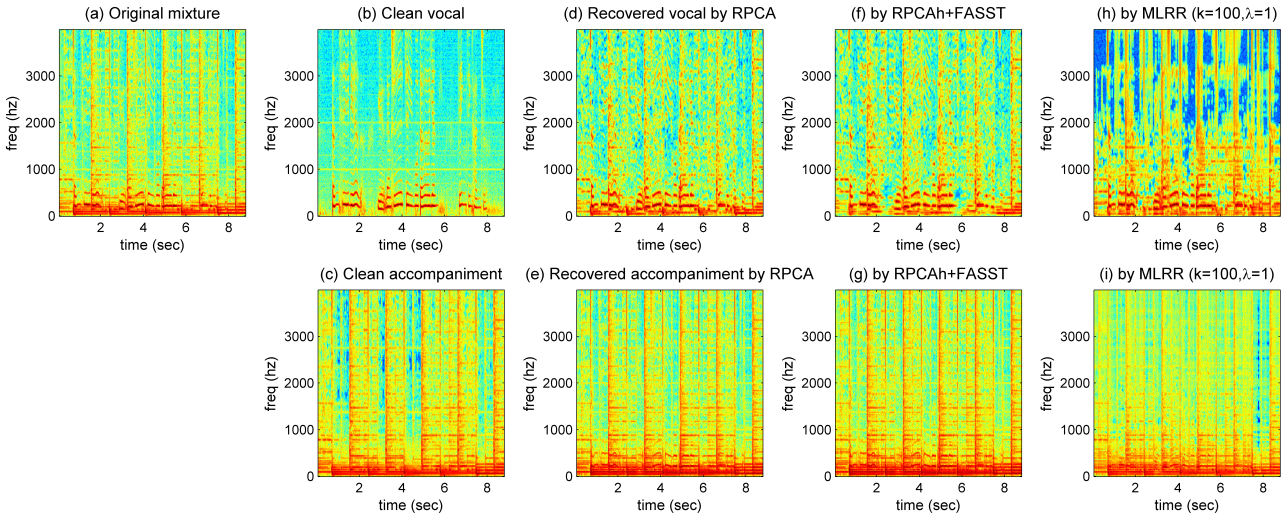

Figure 4. (a) The magnitude spectrogram (in log scale) of the mixture of singing and music accompaniment for the clip 'Ani_1_01' in MIR-1K [7]; (b) (c) The groundtruth spectrograms for the two sources; the separation result for (d) (e) RPCA [8], (f) (g) RPCAh+FASST [23], and (h) (i) the proposed method MLRR (k=100, λ=1) for the two sources, respectively.

[7] C.-L. Hsu and J.-S. R. Jang. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Trans. Audio, Speech & Language Processing*, 18(2):310–319, 2010.

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

[9] M. Lagrange, A. Ozerov, and E. Vincent. Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning. In *Proc. ISMIR*, pages 595–560, 2012.

[10] Z. Lin, M. Chen, L. Wu, and Yi Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. Technical Report UILU-ENG-09-2215, 2009.

[11] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. & Machine Intel.*, 35(1):171–184, 2013.

[12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proc. Int. Conf. Machine Learning*, pages 689–696, 2009.

[13] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *J. Sel. Topics Signal Processing*, 5(6):1088–1110, 2011.

[14] G. Mysore, P. Smaragdis, and B. Raj. Non-negative hidden Markov modeling of audio with application to source separation. In *Int. Conf. Latent Variable Analysis and Signal Separation*, pages 829–832, 2010.

[15] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Trans. Audio, Speech & Language Processing*, 20(4):1118–1133, 2012.

[16] Z. Rafii and B. Pardo. REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation. *IEEE Trans. Audio, Speech & Language Processing*, 21(2):73–84, 2013.

[17] P. Sprechmann, A. Bronstein, and G. Sapiro. Real-time online singing voice separation from monaural recordings using robust low-rank modeling. In *Proc. ISMIR*, pages 67–72, 2012.

[18] J. Sundberg. *The science of the singing voice*. Northern Illinois University Press, 1987.

[19] I. Tošić and P. Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2):27–38, 2011.

[20] E. Vincent, R. Gribonval, and C. Fèvotte. Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech & Language Processing*, 16(4):766–778, 2008.

[21] T. Virtanen. Unsupervised learning methods for source separation in monaural music signals. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music Transcription*, pages 267–296. Springer, 2006.

[22] D. Wang and G. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[23] Y.-H. Yang. On sparse and low-rank matrix decomposition for singing voice separation. In *Proc. ACM Multimedia*, pages 757–760, 2012.

[24] C.-C. M. Yeh and Y.-H. Yang. Supervised dictionary learning for music genre classification. In *Proc. ACM Int. Conf. Multimedia Retrieval*, pages 55:1–55:8, 2012.