SINGING-VOICE SEPARATION FROM MONAURAL RECORDINGS USING ROBUST PRINCIPAL COMPONENT ANALYSIS

Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign
Department of Electrical and Computer Engineering
405 North Mathews Avenue, Urbana, IL 61801 USA
{huang146, chen124, paris, jhasegaw}@illinois.edu

ABSTRACT

Separating singing voices from music accompaniment is an important task in many applications, such as music information retrieval, lyric recognition and alignment. Music accompaniment can be assumed to be in a low-rank subspace, because of its repetition structure; on the other hand, singing voices can be regarded as relatively sparse within songs. In this paper, based on this assumption, we propose using robust principal component analysis for singing-voice separation from music accompaniment. Moreover, we examine the separation result by using a binary time-frequency masking method. Evaluations on the MIR-1K dataset show that this method can achieve around 1∼1.4 dB higher GNSDR compared with two state-of-the-art approaches without using prior training or requiring particular features.

Index Terms— Robust Principal Component Analysis, Music/Voice Separation, Time-Frequency Masking

1. INTRODUCTION

A singing voice provides useful information for a song, as it embeds the singer, the lyrics, and the emotion of the song. There are many applications using this information, for example, lyric recognition [1] and alignment [2], singer identification [3], and music information retrieval [4]. However, these applications encounter problems when music accompaniment exists, since music accompaniment is as noise or interference to singing voices. An automatic singing-voice separation system is used for attenuating or removing the music accompaniment.

Human auditory system has extraordinary capability in separating singing voices from background music accompaniment. Although this task is effortless for humans, it is difficult for machines. In particular, when spatial cues acquired from two or more microphones are not available, monaural singing-voice separation becomes very challenging.

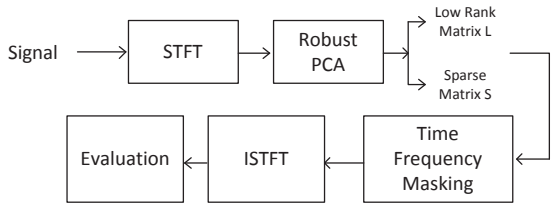


Fig. 1. Proposed framework

Previous work on singing-voice separation systems can be classified into two categories: (1) Supervised systems, which usually first map signals onto a feature space, then detect singing voice segments, and finally apply source separation techniques such as non-negative matrix factorization [5], adaptive Bayesian modeling [6], and pitch-based interference [7, 8]. (2) Unsupervised systems, which requires no prior training or particular features, such as the source/filter model [9] and the autocorrelation-based method [10].

In this paper, we propose to model accompaniment based on the idea that repetition is a core principle in music [11]; therefore we can assume the music accompaniments lie in a low-rank subspace. On the other hand, the singing voice has more variation and is relatively sparse within a song. Based on these assumptions, we propose to use Robust Principal Component Analysis (RPCA) [12], which is a matrix factorization algorithm for solving underlying low-rank and sparse matrices.

The organization of this paper is as follows: Section 2 introduces RPCA. Section 3 discusses binary time frequency masks for source separation. Section 4 presents the experimental results using the MIR-1K dataset. We conclude the paper in Section 5.

This research was supported in part by U.S. ARL and ARO under grant number W911NF-09-1-0383. The authors thank Chao-Ling Hsu [8] and Zafar Rafii [10] for detailed discussions on their experiments, and Dr. Yi Ma [12, 13] and Arvind Ganesh for discussions on robust principal component analysis.

2. ROBUST PRINCIPAL COMPONENT ANALYSIS

Candès et al. [12] proposed RPCA, which is a convex program, for recovering low-rank matrices when a fraction of their entries have been corrupted by errors, i.e., when the matrix is sufficiently sparse. The approach, Principal Component Pursuit, suggests solving the following convex optimization problem:

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

where $M \in \mathbb{R}^{n_1 \times n_2}$, $L \in \mathbb{R}^{n_1 \times n_2}$, $S \in \mathbb{R}^{n_1 \times n_2}$. $||\cdot||_*$ and $||\cdot||_1$ denote the nuclear norm (sum of singular values) and the L1-norm (sum of absolute values of matrix entries), respectively. $\lambda > 0$ is a trade-off parameter between the rank of L and the sparsity of S. As suggested in [12], using a value of $\lambda = 1/\sqrt{\max(n_1, n_2)}$ is a good rule of thumb, which can then be adjusted slightly to obtain the best possible result. We explore $\lambda_k = k/\sqrt{\max(n_1, n_2)}$ with different values of k, thus testing different tradeoffs between the rank of L and the sparsity of S.

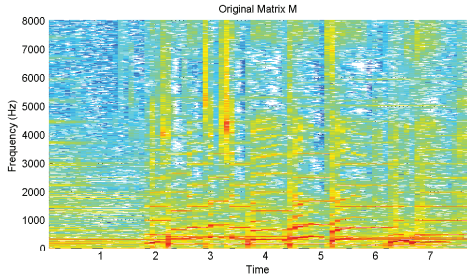
Since music instruments can reproduce the same sounds each time they are played and music has, in general, an underlying repeating musical structure, we can think of music as a low-rank signal. Singing voices, on the contrary, have more variation (higher rank) but are relatively sparse in the time and frequency domains. We can then think of singing voices as components making up the sparse matrix. By RPCA, we expect the low-rank matrix L to contain music accompaniment and the sparse matrix S to contain vocal signals.

We perform the separation as follows: First, we compute the spectrogram of music signals as matrix M, calculated from the Short-Time-Fourier Transform (STFT). Second, we use the inexact Augmented Lagrange Multiplier (ALM) method [13], which is an efficient algorithm for solving RPCA problem, to solve $L + S = |M|$, given the input magnitude of M. Then by RPCA, we can obtain two output matrices L and S. From the example spectrogram, Figure 2, we can observe that there are formant structures in the sparse matrix S, which indicates vocal activity, and musical notes in the low-rank matrix L.

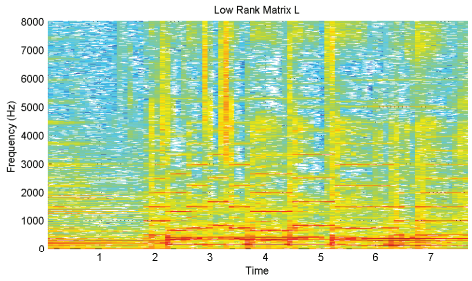
Note that in order to obtain waveforms of the estimated components, we record the phase of original signals $P = \text{phase}(M)$, append the phase to matrix L and S by $L(m,n) = Le^{jP(m,n)}$, $S(m,n) = Se^{jP(m,n)}$, for $m = 1 \ldots n_1$ and $n = 1 \ldots n_2$, and calculate the inverse STFT (ISTFT). The source codes and sound examples are available at http://mickey.ifp.uiuc.edu/wiki/Software.

3. TIME-FREQUENCY MASKING

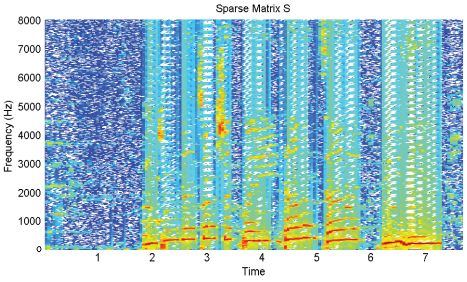
Given the separation results of the low-rank L and sparse S matrices, we can further apply binary time-frequency masking methods for better separation results. We define *binary time frequency masking* M_b as follows:



(a) Original Matrix M



(b) Low-Rank Matrix L



(c) Sparse Matrix M

Fig. 2. Example RPCA results for yifen_2_01 at SNR=5 for (a) the original matrix, (b) the low-rank matrix, and (c) the sparse matrix.

$$M_b(m,n) = \begin{cases} 1 & |S(m,n)| > \text{gain} * |L(m,n)| \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

for all $m = 1 \ldots n_1$ and $n = 1 \ldots n_2$.

Once the time-frequency mask M_b is computed, it is applied to the original STFT matrix M to obtain the separation matrix X_{singing} and X_{music}, as shown in Equation (2).

$$\begin{cases} X_{\text{singing}}(m,n) &= M_b(m,n)M(m,n) \\ X_{\text{music}}(m,n) &= (1 - M_b(m,n))M(m,n) \end{cases} \quad (2)$$

for all $m = 1 \ldots n_1$ and $n = 1 \ldots n_2$.

To examine the effectiveness of the binary mask, we assign $X_{\text{singing}} = S$ and $X_{\text{music}} = L$ directly as the case with *no mask*.

4. EXPERIMENTAL RESULTS

4.1. Dataset

We evaluate our system using the MIR-1K dataset[1]. There are 1000 song clips encoded with a sample rate of 16 kHz, with a duration from 4 to 13 sec. The clips were extracted from 110 Chinese karaoke pop songs performed by both male and female amateurs. The dataset includes manual annotations of the pitch contours, lyrics, indices and types for unvoiced frames, and indices of the vocal and non-vocal frames.

4.2. Evaluation

Following the evaluation framework in [8, 10], we create three sets of mixtures using the 1000 clips of the MIR-1K dataset. For each clip, the singing voice and the music accompaniment were mixed at -5, 0, and 5 dB SNRs, respectively. Zero indicates that the singing voice and the music are at the same energy levels, negative values indicate the energy of the music accompaniment is larger than the singing voice, and so on.

For source separation evaluation, in addition to evaluating the Global Normalized Source to Distortion Ratio (GNSDR) as [8, 10], we also evaluate our performance in terms of Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR), and Source to Distortion Ratio (SDR) by BSS-EVAL metrics [14]. The Normalized SDR (NSDR) is defined as

$$\text{NSDR}(\hat{v}, v, x) = \text{SDR}(\hat{v}, v) - \text{SDR}(x, v) \qquad (3)$$

where $\hat{v}$ is the resynthesized singing voice, v is the original clean singing voice, and x is the mixture. NSDR is for estimating the improvement of the SDR between the preprocessed mixture x and the separated singing voice $\hat{v}$. The GNSDR is calculated by taking the mean of the NSDRs over all mixtures of each set, weighted by their length.

$$\text{GNSDR}(\hat{v}, v, x) = \frac{\sum_{n=1}^{N} w_n \text{NSDR}(\hat{v}_n, v_n, x_n)}{\sum_{n=1}^{N} w_n} \qquad (4)$$

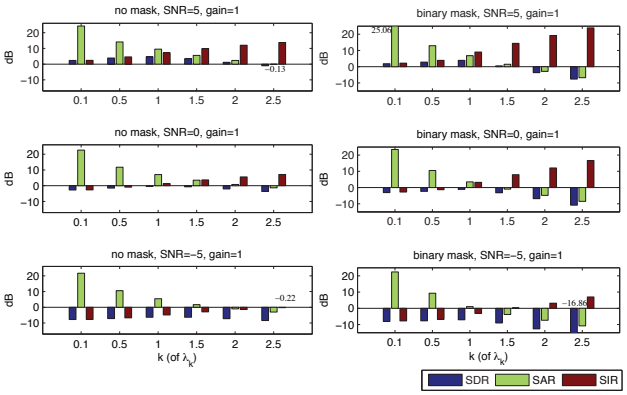
where n is the index of a song and N is the total number of the songs, and w_n is the length of the nth song. Higher values of SDR, SAR, SIR, and GNSDR represent better separation quality[2].

4.3. Experiments

In the separation process, the spectrogram of each mixture is computed using a window size of 1024 and a hop size of 256 (at Fs=16,000). Three experiments were run: (1) an evaluation of the effect of λ_k for controlling low rankness and sparsity, (2) an evaluation of the effect of the gain factor with a binary mask, and (3) a comparison of our results with the previous literature [8, 10] in terms of GNSDR.

(1) The effect of λ_k

The value $\lambda_k = k/\sqrt{(\max(n_1, n_2)}$ can be used for trading off the rank of L with the sparsity of S. The matrix S is sparser with higher λ_k, and vice versa. Intuitively, for the source separation problem, if the matrix S is sparser, there is less interference in the matrix S; however, deletions of original signal components might also result in artifacts. On the other hand, if S matrix is less sparse, the signal contains less artifacts, but there is more interference from the other sources that exist in matrix S. Experimental results (Figure 3) show these trends for both the case of *no mask* and the case of *binary mask* (gain=1) at different SNR values.



(a) no mask (b) binary mask

Fig. 3. Comparison between the case using (a) no mask and (b) a binary mask at SNR={-5,0,5}, k (of λ_k)={0.1, 0.5, 1, 1.5, 2, 2.5}, and gain=1.

(2) The effect of the gain factor with a binary mask

The gain factor adjusts the energy between the sparse matrix and the low-rank matrix. As shown in Figure 4, we examine different gain factors {0.1, 0.5, 1, 1.5, 2} at λ_1 where SNR={-5, 0, 5}. Similar to the effect of λ_k, a higher gain factor results in a lower power sparse matrix S. Hence, there is larger interference and fewer artifacts at high gain, and vice versa.

(3) Comparison with previous systems

From previous observations, moderate values for λ_k and the gain factor balance the separation results in terms of SDR, SAR, and SIR. We empirically choose λ_1 (also suggested in [12]) and gain $= 1$ to compare with previous literature on singing-voice separation in terms of GNSDR using the MIR-1K dataset [8, 10].

Hsu and Jang [8] performed singing-voice separation using a pitch-based inference separation method on the MIR-1K dataset. Their method combines the singing-voice separation method [7], the separation of the unvoiced singing-voice frames, and a spectral subtraction method. Rafii and Pardo proposed a singing-voice separation method by extracting the

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

[2] The suppression of noise is reflected in SIR. The artifacts introduced by the denoising process are reflected in SAR. The overall performance is reflected in SDR.

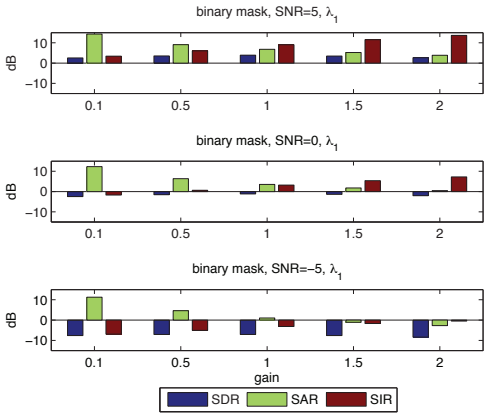


Fig. 4. Comparison between various gain factors using a binary mask with λ_1

repeating musical structure estimated by the autocorrelation function [10] .

As shown in Figure 5, both our approach using *binary mask* and our approach using *no mask* achieve better GNSDR compared with previous approaches [8, 10], with the *no mask* approach providing the best overall results. These methods are compared to an ideal situation where an ideal binary mask is used as the upper-bound performance of the singing-voice separation task with algorithms based on binary masking technique.

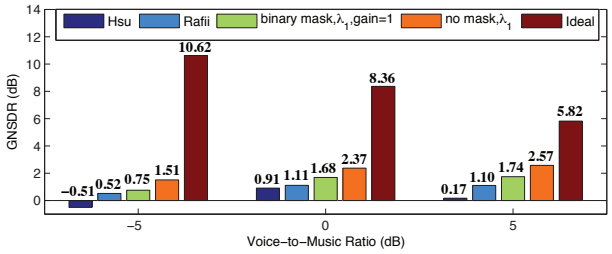


Fig. 5. Comparison between Hsu [8], Rafii [10], the *binary mask*, the *no mask*, and the ideal binary mask cases in terms of GNSDR.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed an unsupervised approach which applies robust principal component analysis on singing-voice separation from music accompaniment for monaural recordings. We also examined the parameter λ_k in RPCA and the gain factor with a binary mask in detail. Without using prior training or requiring particular features, we are able to achieve around 1∼1.4 dB higher GNSDR compared with two state-of-the-art approaches, by taking into account the rank of music accompaniment and the sparsity of singing voices.

There are several ways in which this work could be extended, for example, (1) to investigate dynamic parameter selection methods according to different contexts, or (2) to expand current work to speech noise reduction, since in many situations, the noise spectrogram is relatively low-rank, while the speech spectrogram is relatively sparse.

6. REFERENCES

[1] C.-K. Wang, R.-Y. Lyu, and Y.-C. Chiang, "An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker," in *Proc. of Interspeech*, 2003, pp. 1197–1200.

[2] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals," in *Proc. of ISM*, 12 2006, pp. 257–264.

[3] A. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," in *AES 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, 2002.

[4] H. Fujihara and M. Goto, "A music information retrieval system based on singing voice timbre," in *ISMIR*, 2007, pp. 467–470.

[5] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *ISMIR*, 2005, pp. 337–344.

[6] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1564 –1578, July 2007.

[7] Yipeng Li and DeLiang Wang, "Separation of singing voice from music accompaniment for monaural recordings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1475 –1487, May 2007.

[8] C.-L. Hsu and J.-S.R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 2, pp. 310 –319, Feb. 2010.

[9] J.-L. Durrieu, G. Richard, B. David, and C. Fevotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 564–575, March 2010.

[10] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *ICASSP*, May 2011, pp. 221– 224.

[11] H. Schenker, *Harmony*, University of Chicago Press, 1954.

[12] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, pp. 11:1–11:37, Jun. 2011.

[13] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," Tech. Rep. UILU-ENG-09-2215, UIUC, Nov. 2009.

[14] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462 –1469, July 2006.