



Singing Voice separation from Polyphonic Music Accompaniment using Compositional Model

Priyanka Umap¹, Kirti Chaudhari²

PG Student [Microwave], Dept. of Electronics, AISSMS Engineering College, Pune, Maharashtra, India¹

Assistant Professor, Dept. of Electronics, AISSMS Engineering College, Pune, Maharashtra, India²

ABSTRACT: There are abundant real time applications for singing voice separation from mixed audio. By means of Robust Principal Component Analysis (RPCA) which is a compositional model for segregation, which decomposes the mixed source audio signal into low rank and sparse components, where it is presumed that musical accompaniment as low rank subspace since musical signal model is repetitive in character while singing voices can be treated as moderately sparse in nature within the song. We propose an efficient optimization algorithm called as Augmented Lagrange Multiplier designed to solve robust low dimensional projections. Performance evaluation of the system is verified with the help of performance measurement parameter such as source to distortion ratio(SDR),source to artifact ratio(SAR), source to interference ratio(SIR) and Global Normalized source to Distortion Ratio (GNSDR).

KEYWORDS: Robust Principle Component Analysis (RPCA),Singing Voice Separation, Augmented Lagrange Multiplier (ALM),low rank matrix,sparse matrix.

I.INTRODUCTION

Numerous classes of information are composed as constructive mixtures of portions. Constructive combination, is additive combination that do not result in deduction or diminishment of any of the portion of information, this is referred to as “compositional” data. To characterize such information, various mathematical models are developed. Such models have provided new standards to solve audio processing problems, such as blind and supervised source separation and robust recognition. Compositional models are used in audio processing systems to advance the state of the art on many difficulties that deal with audio data involving of multiple sources, for example on the analysis of polyphonic music and recognition of noisy speech. Thus we use robust principal component analysis (RPCA) as a compositional model in this paper.

Robust Principal Component Analysis (RPCA) method is extensively used in the field of image processing for image segmentation, surveillance video processing, batch image alignment etc. This procedure has received recent prominence in the field of audio separation for the application of singer identification, musical information retrieval, lyric recognition and alignment.

A song usually comprises of mixture of human vocal and musical instrumental audio pieces from string and percussion instruments etc. Our area of interest is segregating vocal line from music which is complex and vital musical signal element from song, thus we can treat musical as intrusion or noise with respect to singing voice. Human auditory system has incredible potential in splitting singing voices from background music accompaniment. This task is natural and effortless for humans, but it turns out to be difficult for machines [15].

Compositional model based RPCA has emerged as a potential method for singing voice separation based on the notion that low rank subspace can be assumed to comprise of repetitive musical accompaniment, whereas the singing voice is relatively sparse in time frequency domain. Basic audio voice separation systems can be divided into two categories that are supervised system and unsupervised system. In supervised systems, training data is required to train the system. On the contrary, in unsupervised systems, no prior training or particular feature extraction is required.

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

The challenges for singing voice separation from background music accompaniment are as follows. In general, the auditory scene created by a musical composition can be viewed as a multi-source background, where varied audio sources from several classes of instruments are momentarily active, some of them only sparsely. The music sources could be of different instrumental type (so exhibiting altered timbral perceptions), which is played at various pitches and loudness, and even the spatial location of a given sound source may differ with respect to time. Regularly individual sources repeat during a musical piece, one or the other way in a different musical environment. Therefore, the section can be considered as a time-varying schedule of source activity encompassing both novel and recurring patterns, representing changes in the spectral, temporal, and spatial complexity of the mixture. Moreover the singing voice has fluctuating pitch frequency for male and female singer which may at some instant overlap with background frequency arrangement of musical instruments. To solve these tasks a compositional model designed using a novel technique Robust Principal Component Analysis (RPCA) is proposed using Augmented Lagrange Multiplier (ALM) as a optimization algorithm for better convergence. Robust Principal Component Analysis[2], which is a matrix factorization algorithm for solving low rank matrix and sparse matrix. Here in our proposed system we assume that the music accompaniment lie in low rank subspace while the singing voice is relatively sparse due to its more variability within the song.

II.PROPOSED SYSTEM ALGORITHM

A song clip is superimposition of singing voice and background musical instruments which can be considered in terms data matrix(audio signal) which is combination of low rank component(musical accompaniment) and sparse components(singing voice).We assume that such data have low intrinsic dimensionality as they lie on several low dimensional subspace, are sparse also in few basis[8].We perform separation of singing voice as follows seen in figure 1.The steps are as follows:

1) We compute Short-Time Fourier Transform (STFT) of the targeted audio signal where signal is represented in time frequency domain. In the separation method, STFT of the input audio signal is computed using overlapping hamming window with N=1024 samples at sampling rate of 16Khz.

2) After calculation of STFT, RPCA is applied by means of Augmented Langrange Multiplier (ALM) as optimization technique which deciphers the computational problem of RPCA [2]. After applying RPCA we get two output matrices ‘L’ low rank matrix and ‘S’ sparse matrix. Binary frequency mask is later applied for quality of separation end result.

3)Inverse Short Time fourier transform (ISTFT) is latter applied, in order to obtain the waveform of the estimated results followed by evaluation of the results.

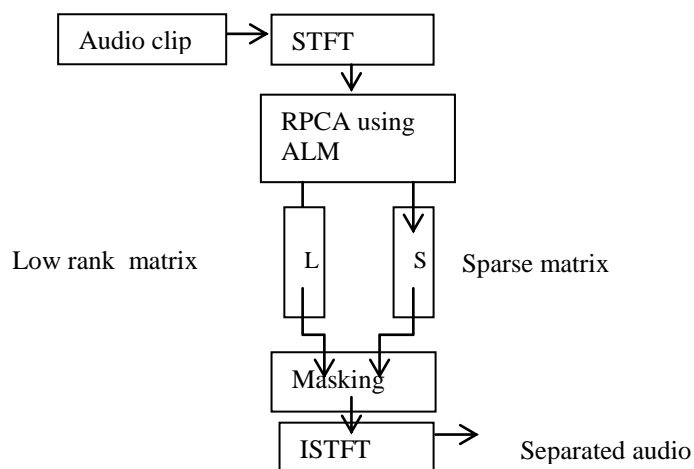


Fig.1. Proposed System

In real world data if n m-dimensional data vectors is put in the form of a matrix $A \in R^{m \times n}$, where A should have a rank $\ll \min(m,n)$, which means that there are some linearly independent columns[5].The objective is to obtain low rank



International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

approximation of A in the existence of noises and outliers. The classical principal component analysis approach which assume the given high dimensional data lie near a much lower dimensional subspace [11]. The method seeks a rank r estimate M of the matrix A by solving,

$$\min_X \|A - M\|$$

Subject to $\text{rank}(X) \leq r$, (1)

where $\|A\|$ denotes the largest singular value decomposition value of A or the spectral norm. The above problem can be solved via singular value decomposition (SVD), by using r largest singular values. But PCA is subtle to outliers and performance declines under bulky corruption. To solve this issue, robust PCA (RPCA) [2,6] is used to render PCA robust to outliers and gross corruption.

A data matrix $M \in R^{m \times n}$ can be uniquely and exactly be decomposed into a low rank component A and a sparse component E, also retrieval of low rank matrix by convex programming. The convex optimization problem can be put forth as follows in terms of objective function and a constraint function [8].

$$\text{minimize } \|A\|_* + \alpha \|E\|_1$$

subject to $A + E = M$ (2)

where $\|\cdot\|_*$ denote the nuclear norm i.e the sum of singular values and $\|\cdot\|_1$ denote the L1-norm that is the sum of the absolute values of matrix entries is a valuable surrogate for L0 pseudo norm, the number of non-zero entries in the matrix. α is the trade off parameter between the rank of A and sparsity of E [6].

$$\alpha_k = k / \sqrt{\max(m, n)} \quad (3)$$

where for $\alpha > 0$ is a regularization parameter and for $k=1$ we get best quality separation result and the results are tested for different values of k. Proficient optimization scheme the Augmented langrange multiplier method is used for solving the above RPCA problem which has higher convergence property. ALM algorithm is iterative converging scheme which works by repeatedly minimizing the rank of A and E matrices simultaneously [4]. ALM is optimization technique for noise reduction.

For better separation outcomes masking can be applied to the separation results of ALM that are low rank A and sparse E matrices by using binary time frequency masking [6]. We need to accurately segregate the components as singing voice mostly lines the music accompaniment during beat instances in order to match with rhythmic structure of the song and hence we apply masking for enhanced separation outcomes.

Binary time frequency masking J_m as follows:

$$J_m(m, n) = \begin{cases} 1, & |E(m, n)| > \text{gain} * |A(m, n)| \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

After application of time frequency masking it is applied to the original audio signal M in order to obtain the separation matrix as singing voice and music respectively.

III. RESULTS AND DISCUSSION

We have worked on this algorithm using MIR-1K database, comprising of male singer and female singer with a sample rate of 16Khz and the duration of the audio clip is 10-14seconds. We create three clips, first consisting of mixed song, second consisting of singing voice and third consisting of musical accompaniment from the stereo database by converting it to mono channel using Audacity software, for the evaluation of the results. The separated audio files are compared with these files.

For the separation and evaluation purpose, spectrograms of each mixture is computed for input audio signal and separated audio signals i.e. the singing voice and music accompaniment. We have taken audio clips consisting of two or more musical instruments in the background and studied its impact on separation. Figures 2 and 3 show the spectrograms for respective audio signal separately for different values of k (of α_k) and on merging the spectrogram of singing voice and music accompaniment we get spectrogram of mixed song. For construction of spectrogram results the low rank and sparse matrix and multiplied by the initial phase of the audio signal. We can examine the varying pitch pattern of separated vocal from song in the spectrograms obtained. In figure 1 spectrogram consists of larger

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

voiced part than that of figure 2 spectrogram for singing voice. From spectrogram results of separated synthesized song musical accompaniments harmonic structure of instruments can be verified.

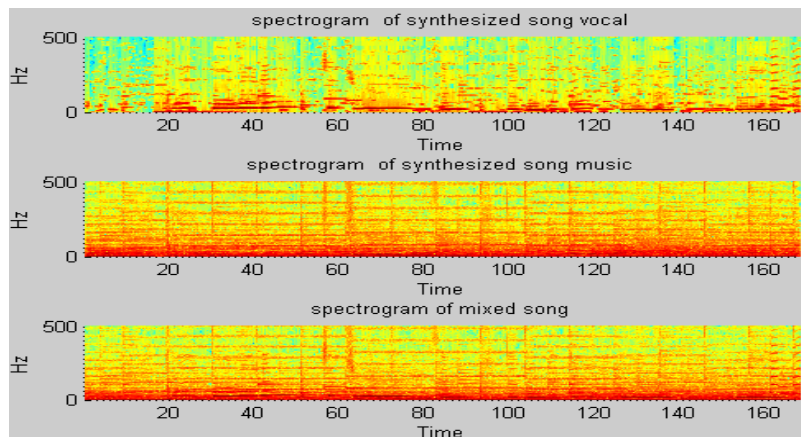


Fig.2. RPCA results of spectrogram for song 1

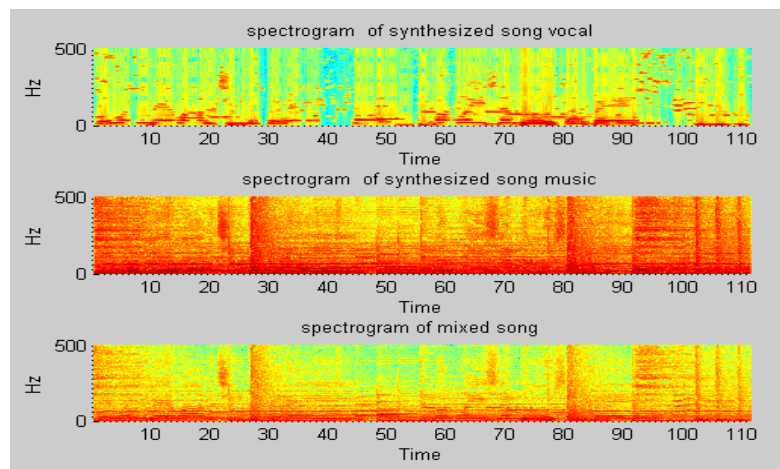


Fig.3. RPCA results of spectrogram for song 2

The value of $\alpha_k = k / \sqrt{\max(m, n)}$ which is an adjustment parameter with respect to rank of A (low rank component) and with the scarcity of E (sparse matrix). From investigational outcomes it has been observed that

- If the matrix E is sparser which means that there is less interference in the matrix E (sparse matrix) but due to this deletion of original components may result in artifacts which is unwanted for the proposed system.
- If E matrix is less sparse, then audio signal will contain, then the signal contains less artifacts which implies that there is more intrusion from the sources which exist in matrix E.

Thus from this we can say that matrix E (sparse matrix) is sparser with higher α_k value and vice versa. We can spot this difference for value of k (of) = {0.1, 0.25, 0.50, 0.75, 1, 2, 3, 4} from the above array we can notice that for values above 1 in the array separation does not take place.

For assessment of the performance of separation results in terms of Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR) and Source to Distortion Ratio (SDR) with assistance of BSS-EVAL metrics. We also evaluate the performance in terms of Global Normalized Source to Distortion Ratio which takes into account the resynthesized singing voice (\bar{v}), original clean voice (v) and the mixture (x).

The Normalized SDR (NSDR) is defined as

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

$$\text{NSDR}(\bar{v}, v, x) = \text{SDR}(\bar{v}, v) - \text{SDR}(x, v) \quad (6)$$

Global Normalized Source to Distortion Ratio:

$$\text{GNSDR}(\bar{v}, v, x) = \frac{\sum_{n=1}^N w_n \text{NSDR}(\bar{v}_n, v_n, x_n)}{\sum_{n=1}^N w_n} \quad (7)$$

k(of λ_k)	SDR	SIR	SAR	GNSDR
0.10	0.0927	0.1031	29.2480	0.0177
0.25	0.1926	0.2662	20.8194	0.1176
0.50	0.6007	0.8967	14.9970	0.5257
0.75	1.3295	2.0241	11.7475	1.2545
1.00	2.7360	5.0676	7.7281	2.6610
1.25	3.5370	12.6067	4.3434	3.4620
1.50	1.3043	19.9380	1.4080	1.2293
1.75	-1.5218	34.6055	-1.5193	-1.5968
2	-3.7622	45.0948	-3.7620	-3.8372
3	-11.4609	56.5936	-11.4608	-11.5358
4	-17.2891	45.0206	-17.2890	-17.3641

Table 1: Results for different values of k(of α_k) for song1

k(of λ_k)	SDR	SIR	SAR	GNSDR
0.10	0.0115	0.0387	25.0416	0.0397
0.50	1.3941	2.2545	10.8753	1.4223
0.75	2.4416	5.2213	6.8367	2.4699
1.00	2.5535	8.8622	4.2416	2.5818
1.25	1.3038	12.3903	1.8993	1.3320
1.50	-0.4133	14.7784	-0.1377	-0.3851
1.75	-2.1408	17.4368	-2.0150	-2.1126
2	-3.6884	20.0092	-3.6267	-3.6602
3	-9.1585	32.6209	-9.1558	-9.1302
4	-21.318	20.6010	-21.280	-21.29

Table 2: Results for different values of k(of α_k) for song2

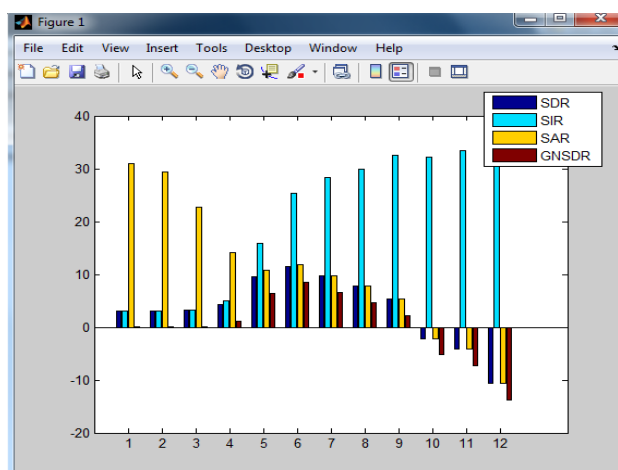


Fig 4: Bar graph for SDR, SIR, SAR, GNSDR for k(of α_k)



International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2015

Better separation results are obtained for value of k (of α_k) less than 1.5 as seen in table 1 and 2 numerical values as well as the bar graph seen in figure 4 respectively. It has been observed that greater the value of SDR, SAR, SIR and GNSDR improved separation results are achieved, the above values can be compared for various values of k (of α_k) in the above table and bar graph respectively.

IV. CONCLUSION

Robust Principal Component Analysis (RPCA) is used as a audio separation technique in this paper. We have enhanced singing voice separation using Augmented Lagrange Multiplier (ALM) for numerous values of α_k . Separation results all depend on trade of parameter α_k . Quality separation outcomes are proven for value of α_k less than 2 and the outcomes are justified through spectrogram outcomes also can be verified perceptually through separated audio files acquired for singing voice and music accompaniment.

REFERENCES

1. Yipeng Li and Deliang Wang, "Separation of singing voice from music accompaniment for monaural recordings," *Audio, Speech Language Processing, IEEE Transaction on*, vol. 15, no. 4, pp. 1475-1487, May 2007.
2. Emmanuel J. Candes, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, pp. 11:1-11:37, Jun 2011.
3. 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 Language Processing, IEEE Transaction on*, vol. 15, no. 5, pp. 1564-1578, July 2007.
4. Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method exact recovery of corrupted low-rank matrices," Tech. Rep. UILU-ENG-09-2215, UIUC, Nov. 2009.
5. Y. H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *ISMIR*, 2013.
6. P. S. Huang, S. D. Chen, P. Smaragdīs, and M. Hasegawa Johnson, "Singing voice separation from monaural recordings using robust principal component analysis," in *ICASSP*, 2012.
7. B. Zhu, W. Li, R. Li, and X. Xue, "Multi-satge non-negative matrix factorization for monaural singing voice separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2096-2107, 2013.
8. E. J. Candès, and B. Recht, "Exact matrix completion via convex optimization," *Found Comput. Math.*, vol. 9, no. 6, pp. 717-772, 2009.
9. J. Salamon, Melody Extraction from Polyphonic Music Signals, Ph.D. thesis, Department of Information and Communication Technologies Universitat Pompeu Fabra, Barcelona, Spain, 2013.
10. K. Min, Z. Zhang, J. Wright, and Y. Ma, "Decomposing background topics from keywords by principal component pursuit," in *CIKM*, 2010.
11. Y. Peng, A. Ganesh, J. Wright, and Y. Xu, W. and Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233-2246, 2012.
12. J. Salamon, E. Gómez, D. P. W. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications and challenges," *IEEE Signal Process. Mag.*, 2013.
13. E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4 pp. 1462-1469, July 2006.
14. Bregman, A. S. (1990), *Auditory Scene Analysis: The Perceptual Organization of Sound*.
15. [15] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K datasets," *Audio, Speech and Language Processing, IEEE Transactions on*, vol. 18, no. 2, pp. 310-319, Feb 2010.
16. [16] 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.