International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 03| June-2015 www.irjet.net p-ISSN: 2395-0072

# Perceptually Motivated Robust Principal Component Analysis Based Separation of Singing Voice from Music

**Madhuri A. Patil[1], S. P. Bhosale[2]**

[1] *Post graduate Student, Electronics Engineering, AISSMS COE Pune, Maharashtra, India*
[2] *Professor, Electronics Engineering, AISSMS COE Pune, Maharashtra, India*

---

**Abstract -** ***Audio signal is an acoustic in the signal processing. Audio signals result from the mixing of several sound sources. During singing, singer stretched the voice sound & shrinking unvoiced sound. Components of singing voice are not as smooth as those of harmonic instruments. Audio signal classification system analyzes the input audio signal and creates a label that describes the signal at the output. These are used to characterize both music and speech signals. The categorization can be done on the basis of pitch, music content, music tempo and rhythm. The signal classifier analyzes the content of the audio format thereby extracting information about the content from the audio data. This is also called audio content analysis, which extends to retrieval of content information from signals. Basically principal component analysis technique is used for the unsupervised singing voice separation from music. The separated singing voice and estimated pitches are used to improve each other iteratively. During singing, singer stretched the voice sound & shrinking unvoiced sound. Components of singing voice are not as smooth as those of harmonic instruments. Singing pitch estimation and singing voice separation are challenging due to the presence of music accompaniments that are often non-stationary and harmonic. A perceptually motivated robust principal component analysis (PRPCA) method is represented to accomplish the challenging singing voice separation. Cochleagram is used as input to the PRPCA method. Music accompaniment can be assumed to be in low-rank subspace, because of its repetition structure and singing voice can be recorded as relatively space within songs. Hence, separate the singing voice from music and audio signal such as speech, background noise and musical instrument.***

***Key Words:*** *Singing voice separation, Cochleagram, Robust principal component analysis (RPCA), and PRPCA*

## 1. INTRODUCTION

Audio signals have frequencies in the audio frequency range of roughly 20 to 20,000 Hz. It is well known that the human auditory system has a remarkable capability in separating sounds from different sources [7]. Singing pitch estimation and singing voice separation are challenging due to the presence of music accompaniments that are often non-stationary and harmonic[8]. 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 and alignment, singer identification, and music information retrieval [2]. An automatic singing-voice separation system is used for attenuating or removing the music accompaniment, since music accompaniment is consider as noise or interference to singing separate the singer's voice from pop music recordings [1].

Although songs today are often recorded in stereo, hence focus on singing voice separation for the monaural recording where only one channel is available. Before applying any techniques, it is instructive to compare singing voice and speech. Singing voice bears many similarities to speech [7]. In recent years, some new methods emerged and have shown great potential for the supervised or unsupervised separation, such as nonnegative matrix factorization (NMF), support vector machine (SVM), robust principal component analysis (RPCA) [1] etc. In which, RPCA is quite attractive method. The spectrum of singing voice is sparse and the spectrum of accompaniment music is low-rank, they are separable in the time-frequency (T-F) domain derived from short-time Fourier transform (STFT) [1]. In this paper, use Perceptually Motivated Robust Principal Component Analysis (PRPCA), which is a matrix factorization algorithm for solving underlying low-rank and sparse matrices. Actually, the Itakura-Saito (IS) measure shows more consistency to the human hearing properties than the Frobenius norm. This is because it is motivated by the auditory masking phenomenon that the human's ear has limited ability to detect noises in frequency bands where the voice signal has high energy.

## 2. METHODOLOGY

The nonlinearity for the frequency perception of the basilar membrane is a remarkable characteristic in the human's auditory system, which is usually modeled with a bank of gammatone filters. Cochleagram is derived from the non-uniform T-F transform, and the T-F units in the

sensitive low frequency regions have higher resolution than that in the high-frequency regions. Hence, the monaural mixed audio signal shows more separable on cochleagram than spectrogram [1].

### 2.1. Perceptually Motivated Robust Principal Component Analysis (PRPCA)

The T-F representation of the audio signal on either spectrogram or on cochleagram $\boldsymbol{Y} \in \mathbf{R}^{m \times n}$ could be decomposed to two parts: sparse voice term $\boldsymbol{S}$, low-rank accompaniment term $\boldsymbol{L}$,

$$\mathbf{Y} = \mathbf{S} + \mathrm{L} \qquad (1)$$

Then, the distance between $\boldsymbol{Y}$ and $\boldsymbol{L}+\boldsymbol{S}$, i.e., $\mathfrak{D}(\boldsymbol{Y}, \boldsymbol{L}+\boldsymbol{S})$ is minimized to derive $\boldsymbol{S}$ and $\boldsymbol{L}$ we are interested in. Conventionally, the Frobenius norm is chosen as the measure,

$$\arg\min_{L,S} \quad \|Y - S - L\|_F^2 \qquad (2)$$
$$\text{Subject to} \quad rank(\mathbf{L}) \le \gamma_L, card(\mathbf{S}) \le c_S$$

where $rL$, $cs$ denotes the rank of $\boldsymbol{L}$ and the cardinality of $\boldsymbol{S}$, respectively. However, the IS measure is more significant for auditory processing, which is defined as,

$$\mathrm{IS}(y, x) = \frac{y}{x} - \log\frac{y}{x} - 1 \qquad (3)$$

In its matrix form, the IS measure is defined as,

$$\mathrm{IS}(\boldsymbol{Y}, \boldsymbol{X}) = \sum_{j=1}^{n}\sum_{i=1}^{m} \mathrm{IS}\,(y_{ij}, x_{ij}) \qquad (4)$$

The objective function for PRPCA is to minimize the IS measure between $\boldsymbol{Y}$ and $\boldsymbol{L}+\boldsymbol{S}$, i.e., $\mathrm{IS}(\boldsymbol{Y}, \boldsymbol{L}+\boldsymbol{S})$. Different from the classical RPCA, $\boldsymbol{Y}$ , $\boldsymbol{S}$ and $\boldsymbol{L}$ here are all constrained to be nonnegative to respect the non-negativity
of the elements in spectrogram or cochleagram. Therefore, the PRPCA problem can be formulated as follows,

$$\arg\min_{L,S} \mathrm{IS}\,(\boldsymbol{Y}, \boldsymbol{L}+\boldsymbol{S}) + \lambda \|\boldsymbol{S}\|_1 + \beta\|\boldsymbol{L}\|_*$$
$$\text{Subject to} \quad \boldsymbol{L} \ge \mathbf{0}, \boldsymbol{S} \ge \mathbf{0}. \qquad (5)$$

### 2.2. Separation via PRPCA

The overall framework of the proposed PRPCA method for singing voice separation is illustrated in below figure 1. Audio signal of the music clip which is dataset for the separation framework.

The separating procedure consists of two stages: PRPCA on cochleagram and singing voice separation. Audio signal is passing through the gammatone filter which filters the unwanted low or high frequencies. At first performed cochlear analysis with gammatone filter and calculate cochleagram of the mixed audio signal. Follow the alternative direction method of multipliers (ADMM) for solving the optimization problem of the PRPCA.

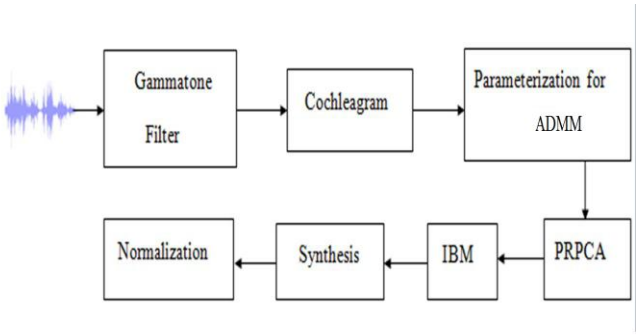



**Fig -1**: Separation of singing voice from music via PRPCA

The cochleagram is then decomposed to sparse singing voice term S, low rank accompaniment music term L. Hence, the ideal binary mask (IBM) $M_{ij}$ could be estimated as follows,

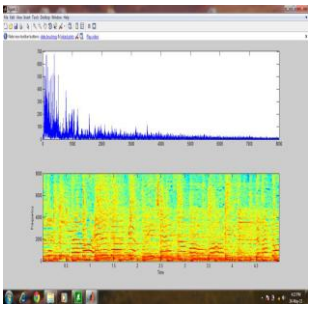
$$M_{ij} = \begin{cases} 1 & Sij \ge \hat{L}ij \\ 0 & Sij < \hat{L}ij \end{cases} \qquad (6)$$

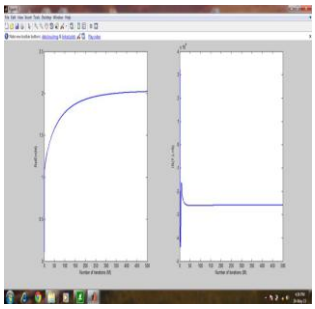
Finally, the singing voice and accompaniment music could be separated and synthesized by weighting the mixed cochleagram with IBM.

## 3. RESULTS & DISCUSSION

At first stage, randomly select 50 song clips from the MIR-1K dataset. These audio signals are sampled at 16 kHz and clip them with durations from 4 to 5 seconds. Without loss of generality, for each audio clip, singing voice and accompaniment music are mixed with the Signal-to-Noise Ratio (SNR) at -5 dB, 0 dB, and 5 dB.

(a) Input Signal (b) Error and Costing





(c) Output by using PRPCA (d) Output by using RPCA

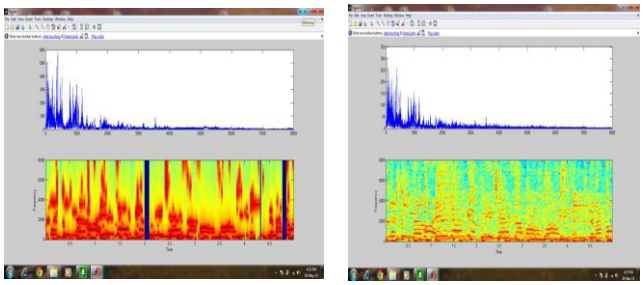


**Fig-.2**: Separation of singing voice from music (a) input audio music signal (b) Error and costing graph of music signal (c) separated singing voice by using PRPCA (d) separated singing voice by using RPCA.

In the figure 2 separation of singing voice from music is shows the cochleagram or spectrogram by using the PRPCA and RPCA method. In above figure part (a) shows the spectrogram of the music audio signal clip which has sampling frequency 16 KHz and clip duration up to 5 to 6 sec., part (b) shows iteration error and costing of the audio signal clip, part (c) gives output of the input audio signal by using the PRPCA method in the form of cochleagram, and part (d) gives output by using the PRPCA method in the form of spectrogram. Hence, if compared the output of both methods, so it clearly shows that PRPCA method gives more efficient and clear singing voice separation.

## 4. CONCLUSIONS

In this paper, an unsupervised method is approach which applies perceptually motivated robust principal component analysis (PRPCA) on separation of singing voice from music. This method decomposes cochleagram of the mixed audio signal into sparse and low-rank components, which corresponds to singing voice and accompaniment music. From result examined the two graphs such as: first explain initial error in sample iteration, and second graph shows vertical line path is of normalization and smooth horizontal line path is created only after output. Also shows the separation of singing voice result by using the RPCA and PRPCA methods. If compared output of both methods that is PRPCA & RPCA, so it clearly shows that PRPCA method gives more efficient and clear singing voice separation.

## REFERENCES

[1] Gang Min, Jibin Yang, Meng Sun, Li Li, Xia Zou, Xiongwei Zhang, "Unsupervised Singing Voice Separation Using Perceptually Motivated Robust Principal Component Analysis".

[2] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, Mark Hasegawa-Johnson, "Singing Voice Separation From Monaural Recording Using Robust Principal Component Analysis", 978-1-4673-0046-9/12©2012 IEEE (ICASSP).

[3] C. L. Hsu, J. S. R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recording Using the MIR-1K Dataset", IEEE Transaction on Audio, Speech, Language Process, 18(2):310-319, February 2010.

[4] Li and Wang, "Separation of Singing Voice from Music Accompaniment for Monaural Recordings", Journal, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.4, May 2007, pp.1475-1487

[5] Ozerov et al., "Adaptation of Bayesian models for single channel source separation and its application to voice/music separation in popular songs", IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no.5, July 2007, pp. 1564-1578.

[6] B. Gao, W. L. Woo, and S. S. Dlay. Unsupervised single-channel separation of nonstationary signals using Gammatone filter-bank and Itakura-Saito nonnegative matrix two-dimensional factorizations. IEEE Trans. on Circuits and Syst.I, 60(3):662–675, March 2013.

[7] Molla et al., "Separation of mixed audio signals by source localization and binary masking with Hilbert spectrum", Springer, ICA 2006, pp. 641-648.

[8] Han and Wang, "Towards generalizing classification Based Speech Separation", IEEE Transactions on audio, speech, and language processing, vol.21, no.1, January 2013, pp.166-175.

## BIOGRAPHIES

Miss Madhuri A. Patil, born on 21 August 1990, is a PG student at AISSMS, COE, Pune, Savitribai Phule Pune University. Her area of interest is Signal Processing.

Mr. S.P. Bhosale is working as assistant professor in Electronics Engg. Department of AISSMSCOE, Pune. His teaching experience is 18 years and area of interest digital image processing.