Separation of Singing Voice from Music Background

Harshada Burute
Research Student, Electronics Department
AISSMS's, IOIT,
Pune, India

P.B. Mane, PhD
Principal
AISSMS's, IOIT,
Pune, India

ABSTRACT
Songs are representation of audio signal and musical instruments. An audio signal separation system should be able to identify different audio signals such as speech, background noise and music. In a song the singing voice provides useful information regarding pitch range, music content, music tempo and rhythm. An automatic singing voice separation system is used for attenuating or removing the music accompaniment. The paper presents survey of the various algorithm and method for separating singing voice from musical background. From the survey it is observed that most of researchers used Robust Principal Component Analysis method for separation of singing voice from music background, by taking into account the rank of music accompaniment and the sparsity of singing voices.

Keywords
Music Accompaniment, pitch, music tempo, rhythm.

1. INTRODUCTION
Singing is used to produces musically relevant sounds by the human voice, and it is employed in most cultures for entertainment or self-expression. The singing voice becomes immediately the main focus of attention when we listen to musical pieces with a voice part. Now a days, in multimedia technology various audio editor software's are available. Mixture of singing voice and music accompaniment known as a song. Music recording are either monaural (single channel) or stereo (two channel) basis. Speech is an acoustic signal produced from a speech production system. Sound is a representation of an audio signal. 20 Hz to 20 kHz are the audio frequency range. The human auditory system has a better capability in separating sounds from different sources [1].

Speech separation is a very challenging task in signal processing. An Audio signal classification system detecting the audio type of a signal (speech, background noise and musical genres). A singing voice separation system has its applications in areas such as automatic lyrics recognition and alignment, singer identification, musical information retrieval, karaoke, musical genre classification, melody extraction, audio signal classification[1], [6], [7], [14], [18] etc. An audio signal separation system should be able to identify different audio signals such as speech, background noise. Audio signal classification system analyzes the input audio signal and describes the signal at the output. Typically in case of songs, these are used to characterize both music and singing voice signals. Aim of this research work is to separate out the singing voice from music background. System consists of the singing voice detection stage and separation stage to separate out the singing voice.

An automatic singing voice separation system is used for removing or attenuating the music accompaniment. Robust Principal Component Analysis (RPCA), which is a matrix factorization algorithm for solving low-rank and sparse matrices. Music accompaniment in a low-rank subspace. Repetition of music is a main parameter in a song [16], [19]. Singing voice is relatively sparse due to its variations or different pitch ranges within the songs. In system use Binary frequency mask for quality of separation results. Inverse Short Time Fourier Transform (ISTFT) is applied, in order to obtain the waveform of the estimated results and recover the original signal.

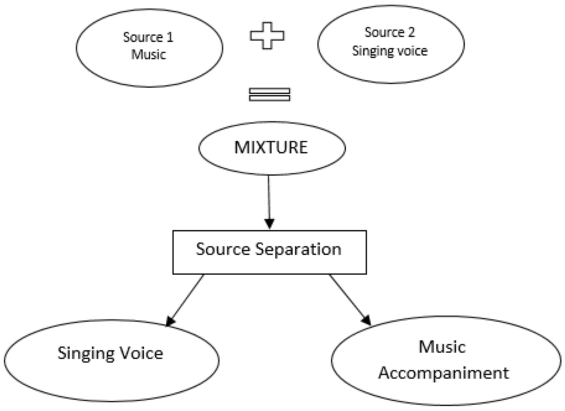

Fig.1. Overview Diagram of Separation system

Single channel source separation is the challenging task. Separation of vocals from polyphonic audio recordings [20] also available with some limitations. Motivation for the research work analysis can be taken from the human auditory system, which have a powerful ability to segregate and separate incoming sounds. Human beings can identify more than one sound source at a time like sounds of nearby passing of a car, distant voices and nearby music played in the background. Thus significant number of complex tasks is performed by human auditory signals, given only two input streams i.e. from left and right ears. Even though this task is natural and simple for human but it is difficult for machine.

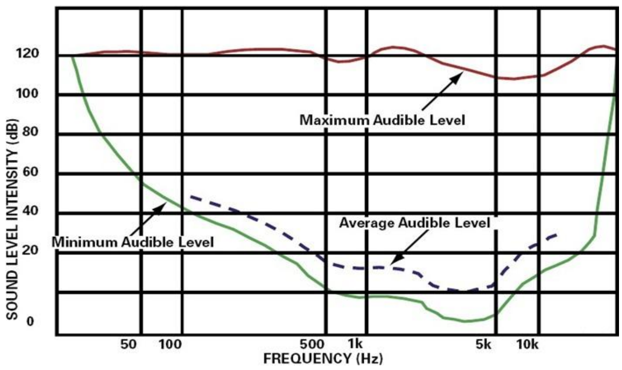

Fig.2 The range of audibility of the human ear

Human beings are basically able to understanding of the audio signals perceived which contain multiple frequency and time overlapping sources. When the interfering energy is close to or exceeds the energy of the target source humans are identified, this phenomenon of the human capability to focus on a specific source from within the mixture is known as Auditory Scene Analysis (ASA) [2].

The challenges for singing voice separation from background music accompaniment as follows-

- In general, the auditory scene produced by a musical composition can be regarded as a multisource environment, where different sound sources from various class of instruments are temporarily active, some of them only sparsely.
- The music sources may be of different instrumental type, may be played at various pitches and loudness, and even the spatial position of a given sound source may vary with respect to time.
- Often individual sources recur during a musical piece, either in a different musical context or by revisiting already established phrases. Thus, the scene can be regarded as a time-varying schedule of source activity containing both novel and repeated patterns, indicating changes in the spectral, temporal, and spatial complexity of the mixture.
- Also the singing voice has varying pitch frequency for male and female singer which may at some instant overlap with background frequency pattern of musical instruments.

Audio mixtures are categories as music or speech and live recording or synthetic mixture. Goal of source separation is to make clear modeling assumption and performed issues of each approaches. Number of techniques are used for audio signal separation with some advantages and disadvantages. Audio source separation has a large number of potential applications. Depending on the application, one could be interested in each individual extracted source or may be just in extracting one source from the mixture. Previous research study has been carried out for speech separation, but only a small number of researchers are devoted to separating singing voice from music accompaniment. Adaptive Bayesian model [3], Tandem Algorithm [4], Harmonic/Percussive Sound Separation (HPSS) [5], combined model for singing voice and music [8], Adaptive filter [9], and Non-negative Matrix Factorization (NMF) [21] are the various singing voice separation techniques.

Singing voice bears many similarities to speech. They both consist of voiced and unvoiced sounds. While the dissimilarity between them is also major. A well-known difference is the presence of an additional formant, called the singing formant, in the frequency range of 2000Hz-3000Hz in operatic singing; this singing formant makes the voice of a singer to stand out from the accompaniment. Another difference also lies to the method the singing voice and the speech is uttered that is a singer is intentionally stretches the voiced sound and shrinks the unvoiced sound to match other musical instruments which has direct two consequences that it changes that it changes the number of voiced and unvoiced sound in singing.

Singing voice mainly consists of voiced parts while that of speech primarily consists of unvoiced parts. Analyze the pitch content of singing voice, observed that there are abrupt changes in the pitch in between while in contrast the pitch of natural speech smoothly changes in an utterance. Likewise singing voice has much wider frequency ranges while the pitch range of normal speech is between 80-400Hz.

The organization of this paper is as follows: Section II describes literature survey of sound separation techniques. Section III concludes the paper.

2. LITERATURE SURVEY

Li and Wang [1] "Separation of Singing Voice from music Accompaniment for Monaural Recordings" proposed a computational auditory scene analysis (**CASA**) system to separate singing voice from music accompaniment for monaural (single channel) recordings. System consist of singing voice detection stage, pitch detection stage used Hidden Markov Model (**HMM**) and separation stage. Singing voice separation from monaural recordings where only one channel is available i.e. Mono signal. Large majority of sounds generated during singing is voiced (about 90%), while speech has a larger amount of unvoiced sounds Wang described.

Bregman [2] proposed that the auditory system employs a process called auditory scene analysis (ASA) for different sound sources.

Ozerov et al. [3] "Adaptation of Bayesian models for single channel source separation and its application to voice/music separation in popular songs" introduce a general formalism for source model adaptation which is expressed in the framework of Bayesian models. Particular cases of the proposed approach are then investigated experimentally on the problem of separating voice from music in popular songs. The obtained results show that an adaptation scheme can improve consistently and significantly the separation performance in comparison with nonadapted models. The difference of spectral distribution (timbre) of singing voice and instruments, and modelled them by Gaussian Mixture Model (**GMM**). In their method, the GMM was trained in advance in a supervised way, and tuned adaptively for each input.

Hsu et al. [4] "A Tandem algorithm for singing pitch extraction and voice separation from music accompaniment" proposed algorithm that estimates the singing pitch and separates the singing voice jointly and iteratively. Tandem Algorithm detects multiple pitch contours and separates the singer by estimating the Ideal Binary Mask (IBM). System having trend estimation algorithm first estimates the pitch ranges of the singing voice.

Tachibana et al. [5] "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms" focused on the fluctuation of a singing voice and considered to detect it by exploiting two differently resolved spectrograms. This system is based on pitch estimation parameter. Proposed two stage harmonic/percussive sound separation system on multiple resolution spectrograms.

Zhu et al. [6] "Multi-stage non negative matrix factorization for monaural singing voice separation" developed a new algorithm for monaural singing voice separation. The algorithm applies Non negative Matrix Factorization (**NMF**) to decompose long window and short window mixture spectrograms and employs a spectral discontinuity and a temporal discontinuity thresholding method to select components for the two NMFs respectively. Comparative evaluation parameter performance with different songs are performed.

Umesh and Sinha [7] “A study of filter bank smoothing in MFCC features for recognition of children’s speech” addressed during vocal tract length normalization, the Bandwidth (BW) of the MFCC (Mel Frequency Cepstral Coefficient) filters should not be scaled, only the center frequencies should be scaled to get improved performance.

In real-world sound sources are usually mixed with different audio signals. The process in which individual sources are estimated from the mixture signal is called **Sound Source Separation.** Musical accompaniment is an interference in singing due to its harmonics changes and repetition in a nature. Background noise is an interference in speech signal.

Rafii et al. [8] “Combining modelling of singing voice and background music for automatic separation of musical mixtures” proposed the combining models for singing voice and music for automatic separation of musical mixtures. Singer –independent model based on NMF (Non-negative Matrix Factorization) techniques and music model using similarity matrix and median filter.

Adaptive filter is important in the signal processing field. Adaptive filter is used to remove unwanted signal and take original signal. An adaptive filter has an adaptation algorithm that is meant to monitor the environment situation and according to that vary the filter transfer function. Based in actual signal received attempts to find optimum filter design [9].

The rejection of unwanted signals through the use of optimization (minimization) theory is becoming popular in the area of adaptive filtering. Adaptive filters have a self-adjusting ability [9]. It can eliminate unwanted signals from useful information. These filters minimize the mean square of the error signal.

Audio source separation received first importance in mid-20th century when the need came for the analysis of audio signal from black box of aircrafts, where large amount of noise is accompanied with audio signal from pilot. The main traditional approaches to the source separation problem have always been **Beamforming** and Independent Component Analysis (**ICA**).

Huang et al. [10] “Singing voice separation from monaural recordings using robust principal component analysis” proposed Robust Principal Component Analysis algorithm for singing voice separation from monaural recordings. This method used decomposed of low rank matrix and sparse matrix of input data matrix. Singing voice separation systems can be divided into two main parts. First Supervised system in which training data required. Second Unsupervised system in which not required training data, all observations are assumed to do continuous improvement. Proposed method should be unsupervised approach.

Lin et al. [11] “The augmented Langrange multiplier method exact recovery of corrupted low-rank matrices” addressed algorithm the Augmented Lagrange Multiplier for exact recovery of corrupted low-rank matrices. Method include optimization techniques and fast convergence rate. This algorithm also known as matrix recovery method. In exact ALM is better than exact ALM techniques.

Candes et al. [12] “Robust principal component analysis” addressed about Robust Principal Component Analysis method with formulation and detailed derivation with explanation. Recover the principal component of data matrix result are achieve using this method. This paper discussed an algorithm for solving optimization problem.

Wright et al. [13] “Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices by Convex Optimization” addressed about Robust Principal Component Analysis method with formulation. This paper proves that using this technique result matrices can be efficiently and exactly recovered. This method is used to solve a simple convex program.

Salamon et al. [14] “Melody Extraction from Polyphonic Music Signals” thesis addressed about melody extraction application from polyphonic music signals. Discussed about general information of music signal and its properties and also explained the important definitions.

Min et al. [15] “Decomposing background topics from keywords by principal component pursuit” explained about Principal Component Pursuit that can effectively decomposes the low rank and sparse matrices of low dimensional data. In this paper, explained this method in image processing data analysis application.

Rafii and Pardo [16] “A simple music/voice separation method based on the extraction of the repeating musical structure” proposed a method for separation of singing voice and music. Method based on the extracting of the repeating musical structures from the song. Repetition is a core principle of music parameter.

A desirable dataset for singing voice separation should meet the following criteria [17] “On the improvement of singing voice separation for monaural recordings using the MIR-1K datasets”,

- The singing voice and the music accompaniment should be recorded separately, so that the performance of the separation result can be evaluated by comparing it with the premixed singing voice.
- The manual explanation (such as lyrics, pitch range, unvoiced types, variations and repetition of music note) for each clip should be as sufficient as possible for all kinds of possible evaluations for singing voice separation.
- The dataset should be publicly available without copyright issues.

MIR-1K dataset meets all these criteria. MIR-1K Dataset [17] (Multimedia Information Retrieval lab, 1000 song clips), which is extracted from 110 recordings of Chinese pop songs recorded at 16 - kHz sampling rate with 16 - bit resolution. The duration of each clip ranges from 4 to 13sec.

Salamon et al. [18] “Melody extraction from polyphonic music signals: Approaches, applications and challenges” addressed approaches, applications, challenges and case studies for melody extraction from Polyphonic Music signals.

In a musical context different types of mixtures of recording or live concerts are available. In that context some portion is either the singing voice or the normal speech for entertainment purpose. Main parameter is the pitch ranges.

Rafii and Pardo [19] “REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation” addressed new improvement in older method with large no. new datasets. They present new technique for separation of singing voice from music accompaniment i.e. REpeating Pattern Extraction Technique (REPET).

Vembu and Baumann [20] "Separation of vocals from polyphonic audio recordings" proposed a method to the extraction of vocal tracks from polyphonic audio recordings. This paper presented additional method for identify vocal portions in a song and design a classifier for vocal-nonvocal parameters. This method is mostly proposed for Non-stationary signals like vocal part. Limitation of this method is poor quality of source separation.

Hu and Liu [21] "Separation of Singing Voice Using Nonnegative Matrix Partial Co-Factorization for Singer Identification" proposed a method for separation of singing voice from music. Method consist of two steps. First Non-negative Matrix Partial Co-Factorization and second is singer identification.

Table 1. Summary of singing voice separation techniques

Reference No.	Singing Voice Separation Techniques	Performance Comparision
[1]	GMM Based Source Separation	Three steps – • GMM sources modling • Separation by adaptive Wiener filtering • Inverse Fourier Transform Drawbacks - • This requires large set of Gussian functions with GMMs. • Performance poor.
[3], [8], [21]	Source Adapted Models	Steps – • Vocal and non vocal frames • Adaptive wiener filtering • Inverse Fourier Transform Drawbacks – • Quality of sound is poor • Manually song separated into vocal and non vocal portions
[3]	Baysian Models	Steps – • Vocal and nonvocal parts. • Adapted model Drawback – • Poor quality voice separation.
[1], [4]	Pitch Based Inference	Steps – • Mixture of song • Identify voiced and unvoiced dominant T-F units • Resynthesized Drawback – • This method has their pros and cons.
[10], [11], [12], [13], [15], [17]	Robust Principal Component Analysis (RPCA)	Steps – • STFT • RPCA • ISTFT Advantages – • Training data not required. • Easy to understand. • Many researchers used for separation. • This method can achieve highrer GNSDR.
[8], [16], [19]	Based on Repetition	• Repetition is a basic requirement of music. • Main parameter is to findout the repeating nature of a music in a song. Disadvantage – • Required to find Repeating structure of music.
[5]	Two-Stage Harmonic percussive Sound Separation	• Two different resolved spectrograms. Drawback – • Inefficient for singing voice and music variations.
[6], [20], [21]	Non-negative Matrix Factorization (NMF)	Disadvantage – • Imposes only the non-negativity constraint.

3. CONCLUSION & FUTURE SCOPE

The paper presents survey of the various algorithm and method for separating singing voice from musical background. From the survey it is observed that Robust Principal Component Analysis method is better for separation of singing voice from music background, by taking into account the rank of music accompaniment and the sparsity of singing voices. The separation technique uses repetition of musical structure property. The developed algorithm and methods can be applied for intelligent music separation application platform, singer identification, musical information retrieval, melody extraction and music instrument separation.

4. REFERENCES

[1] Yipeng Li and DeLiang, “Separation of Singing Voice From Music Accompaniment for Monaural Recordings”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1475 – 1487, May 2007.

[2] A.S.Bregman, *Auditory scene analysis*, Cambridge, MA: MIT press, 1990, pp.1-45,455-528.

[3] Alexey Ozerov, Pierrick Philippe, Frederic Bimbot and RemiGribonval, “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, pp. 1564-1578, July 2007.

[4] Chao-Ling Hsu, DeLiang Wang, Jyh-Shing Roger Jang and Ke Hu, “A Tandem Algorithm for Singing Pitch Extraction and voice Separation from Music accompaniment”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol.20, no.5, pp.1482-1491, July 2012.

[5] Hideyuki Tachibana, Nobutaka Ono and Shigeki Sagayama, “Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms” , *IEEE Transactions on Audio, Speech, and Language Processing*, vol.22, no.1, pp.228-237, January 2014.

[6] Bilei Zhu, Wei Li, Ruijiang Li and XiangyangXue, “Multi-stage non negative matrix factorization for monaural singing voice separation”, *IEEE Transactions on Audio, Speech, and Language Processing,* vol.21, no.10, pp.2096-2107, October 2013.

[7] S. Umesh and Rohit Sinha, “A study of filter bank smoothing in MFCC features for recognition of children’s speech”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol.15, no.8, pp.2418-2430, November 2007.

[8] Zafar Rafii, Francois G. Germain, Dennis L. Sun and Gautham J. Mysore, “Combining modelling of singing voice and background music for automatic separation of musical mixtures”, *ISMIR*, 2013.

[9] Ifeachor and Jervis, “*Digital signal processing: a practical approach*”, second edition, Pearson educations, pp.645-680.

[10] Po-.Sen. Huang, Scott Deeann Chen, Paris Smaragdis and Mark Hasegawa-Johnson, “Singing voice separation from monaural recordings using robust principal component analysis”, *ICASSP*, 2012.

[11] Zhouchen Lin, Minming Chen, Leqin Wu and Yi Ma, “The augmented Langrange multiplier method exact recovery of corrupted low-rank matrices”, *Tech. Rep.UILU-ENG-09- 2215*, UIUC, Nov.2009.

[12] Emmanuel J. Candes, Xiaodong Li, Yi Ma and John Wright, “Robust principal component analysis ?”, *Journal of the ACM*, vol.58, no. 3, article 11, pp.11:1-11:37, May 2011.

[13] John Wright, Yigang Peng, Yi Ma, Arvind Ganesh and Shankar Rao, “Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices by Convex Optimization”, pp. 1-9.

[14] Justin J. Salamon, “*Melody Extraction from Polyphonic Music Signals”* , Ph.D. thesis, Department of Information and Communication Technologies University Pompeu Fabra, Barcelona, Spain, 2013.

[15] Kerui Min, Zhengdong Zhang, John Wright and Yi Ma, “Decomposing background topics from keywords by principal component pursuit”, *CIKM*, 2010.

[16] Zafar Rafii and Bryan Pardo, “A simple music/voice separation method based on the extraction of the repeating musical structure”, *ICASSP*, May 2011, pp.221-224.

[17] Chao-Ling Hsu and Jyh-Shing Roger Jang, “On the improvement of singing voice separation for monaural recordings using the MIR-1K datasets”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, Issue 2, pp. 310-319, February 2010.

[18] Justin Salamon, Emilia Gómez, Daniel P. W. Ellis and Gael Richard, “Melody extraction from polyphonic music signals: Approaches, applications and challenges,” *IEEE Signal Processing Magazine*, pp. 118-134, March 2014.

[19] Zafar Rafii and Bryan Pardo, “REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 71 – 82, January 2013.

[20] Shankar Vembu and Stephan Baumann, “Separation of Vocal from Polyphonic Audio Recordings”, *Proceedings of 6th international conference of Music Information Retrieval*, 2005.

[21] Ying Hu and Guizhong Liu, “Separation of Singing Voice Using Nonnegative Matrix Partial Co-Factorization for Singer Identification”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 643 – 653, April 2015.