HARMONIC SINGLE CHANNEL SOURCE SEPARATION

Ishaan Rao
University of Rochester
ishaan.rao@rochester.edu

ABSTRACT

Sound source separation has become a popular research topic in computer audition due to its wide range of applications in the analysis and manipulation of audio data. Applications of audio source separation include music information retrieval, automatic transcription of music, and sampling of musical sounds for electronic music composition amongst many others. Through this paper, different algorithms proposed by researchers were analyzed with the aim of comparing different methods based on evidence provided in the papers. The literature review gives a better understanding of the various methods used in differentiating different instruments and helps in picking the most prudent method or in optimizing a current algorithm for classifying different instruments.

I. INTRODUCTION

In real world audio signals, several sources are usually mixed together and separating out and identifying the different sources is called sound source separation. Separation can be *Blind* when no prior information of the sound sources is given. Blind source separation of audio sources is based on the assumption that the different sources are independent. In contrast *Non-blind or supervised* separation methods are provided with prior information of the sources usually in the form of solo excerpts in order to train the separation model.

According to [1], source separation methods can be classified as *over-determined* or *under-determined* according to the number of sensors and sources. In over-determined cases, the number of sensors outnumbers the number of sources and vice versa for the case of under-determined methods. Single channel source separation can be considered to be a drastic form of the under-determined case.

The preponderance of recent algorithms for sound source separation put forth by researchers can be broadly classified into three different categories – Spectral Decomposition based Methods, CASA based methods and Model Based Methods [1].

1.1 Spectral Decomposition based Methods

In spectral decomposition methods the spectral representation of a mixed audio signal, in the form of a spectrogram, are modeled as a combination of a set of spectral components. In recent years, Independent Component Analysis (ICA) and its extension Independent Subspace Analysis (ISA) as well as Nonnegative Matrix Factorization (NMF) methods have received a lot of attention of the purpose of source separation. However their iterative nature results in a high computational cost.

1.2 CASA based Methods

CASA based methods model the human auditory system and perform source separation by grouping “Time-Frequency” signal components with similar source attributes into auditory streams [2]. The psychoacoustical cues for grouping the signal components, to give the separated sounds, are usually harmonicity, onset/offset times, timbre etc. However these methods are inefficient as sources with the same pitch or common harmonic partials tend to remain undetected.

1.3 Model based Methods

According to [3], model based approaches consist of developing a model that describes a particular source separation problem. The parameters may be as simple as a mixing matrix and set of source signals or may be much more complicated including positions, orientations and interactions with sources. Then either Hidden Markov Models (HMM) [4-5] or Bayesian Methodology [3] is used to train solo excerpts and obtain a solution to the source separation problem. These methods work well only on specific separation problems and require training of a large number of parameters, i.e., they cannot be used in an unsupervised fashion.

The paper deals with only harmonic single channel source separation and the outline is as follows. Section II deals with the Spectral Decomposition based methods for source separation. Section III gives an overview of CASA based methods and Section IV deals with Model based methods. A discussion on the merits/demerits of each method is provided in Section V.

II. SPECTRAL BASED DECOMPOSITION METHODS

1.1 Independent Component Analysis (ICA)

Independent Component Analysis (ICA), discussed in [7-9], is one of the most widely used techniques for solving Blind Source Separation (BSS) problems. ICA is used in source separation under the assumption that the individual source signals are mutually independently distributed. The second fundamental assumption is that the individual sources must have a non Gaussian distribution.

ICA assumes a statistical model where the observation signal is observed as a product of the mixing matrix and a vector of statistically independent signals (sources). From the discussions in [8],

$$x = As \tag{1}$$

where $A = [a_1,\dots,a_p]$ is a nxp invertible mixing matrix, $s = [s_1 \dots s_p]^T$ is the vector of p statistically independent sources and $x = [x_1 \dots x_n]^T$ is the n-dimensional observation vector with n >= p

The objective is thus to estimate the original source signal in the vector *s* from the observation vector *x.* This can be accomplished by finding an unmixing matrix $W \approx A^{-1}$ so that the estimated source signals *u* are as independent as possible and using a multiplicative update rule for minimizing the error between *s* and *u*.

$$u = Wx = WAs \tag{2}$$

However, ICA is limited in its use to only over-determined cases where the number of sources has to be less than or equal to the number of input variables (length of the observation vector *x*). Therefore for single channel source separation, an extension of the ICA method, called Independent Subspace Analysis (ISA), is used to remove the limitation. The ISA problem is solved by a simple ICA followed by a grouping of the ICA components.

In [8], a spectrogram based subspace separation is used where a spectrogram is decomposed into independent subspaces and then inverted to give the separated source signals. According to [6], "the factorization of the spectrogram can be seen as a separation of phase independent features into invariant feature subspaces" and the separated source signals are obtained by inverting the transformation.

1.2 Nonnegative Matrix Factorization (NMF)

In NMF based methods, first proposed by Lee and Seung [11], the audio spectrogram *V* can be approximated as a product of two non-negative matrices *W* and *H* and the decomposition is achieved by minimizing the error between V and WH

$$V \approx WH \tag{3}$$

W can be described a matrix of basis vectors containing dictionary components. In other words, *W* contains the spectral bases for the different pitch components. The matrix *H* captures the gain of the basis vectors, i.e. H specifies a matrix of pitch content vs. time.

Two measures were used for minimizing the reconstruction error: the square of the Euclidean distance and the K-L Divergence.

The square of the Euclidean distance is given by

$$\textstyle\sum_{k,t} ([V]_{k,t} - [WH]_{k,t})^2 \tag{4}$$

and the K-L Divergence *D* is defined as

$$D(V||WH) = \textstyle\sum_{k,t} [V]_{k,t} \log([V]_{k,t}/[WH]_{k,t}) - [V]_{k,t} + [WH]_{k,t} \tag{5}$$

where

$$V \in R_{\geq}^{0,\ mxn} \tag{6}$$

$$W \in R_{\geq}^{0,\ mxr} \tag{7}$$

$$H \in R_{\geq}^{0,\ rxn} \tag{8}$$

and $r \leq \min\{m,n\}$ is the rank of matrix V.

The algorithm for implementing the NMF decomposition is as follows:

Using D(V||WH), considering K-L Divergence, perform iterations to

1. Update W using multiplicative update

$$W \leftarrow W.*(((V/WH)*H^T)/\ W^T1 \tag{9}$$

2. Update H using multiplicative update

$$H \leftarrow H.*((W^T*(V/WH))/\ 1H^T \tag{10}$$

3. Check D(V||WH) for convergence

By assuming the audio spectrogram *V* to be equal to the sum of the individual source spectrograms $[V_1 \dots V_n]$, the source dictionaries $[W_1 \dots W_n]$ can be used to separate sound sources in the mixture signal via NMF decomposition.

$$V \approx V_1 + V_2 + \dots V_n \tag{11}$$

$$V \approx W_1 H_1 + W_2 H_2 + \dots W_n H_n \tag{12}$$

$$V = [W_1, W_2,..W_n]\ \begin{matrix}[H_1]\\ [H_2]\\ [H_n]\end{matrix} \tag{13}$$

The individual source signals could then be reconstructed as W_iH_i where i = 1…n.

Separation via NMF decomposition could be both supervised and un-supervised.

Even though the papers do not discuss the performance of the NMF and ICA algorithms in the presence of noise, it is generally believed that the performance of the NMF is better than that of ICA when noise (especially Gaussian noise) is present. NMF may not perfectly suppress the noise, yet it can still separate the sources.

1.3 Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria

In [6], the author presents an algorithm for monaural sound source separation that combines NMF with temporal continuity and sparseness objectives. Experimental evaluations showed that the proposed algorithm had better accuracy in source separation compared to the ISA and basic NMF methods.

Estimation of the *W* and *H* matrices were realized by minimizing a cost function consisting of a weighted sum of the reconstruction error, temporal continuity and sparseness term.

$$c(W,H) = c_r(W,H) + \alpha c_t(H) + \beta c_s(H) \qquad (14)$$

Where *c(W,H)* is the cost function, $c_r(H)$ the reconstruction error term, $\alpha c_t(H)$ the temporal continuity term and $\beta c_s(W,H)$ the sparseness term. α and β are the weights attached to the respective terms.

The reconstruction error term was minimized using Divergence discussed earlier. Temporal continuity was addressed by assigning a cost to large changes in gain in between successive frames.

$$c_t(H) = \sum_{j=1}^{J} 1/\sigma^2_j \sum_{t=2}^{T} (h_{t,j} - h_{t-1,j})^2 \qquad (15)$$

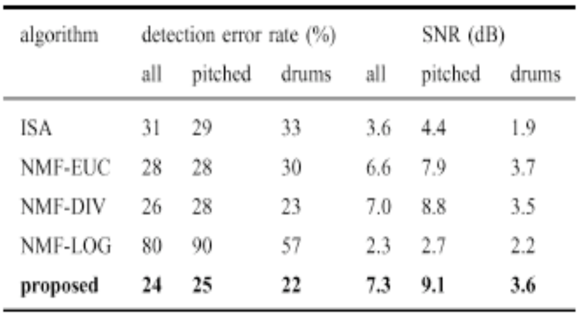
$h_{t,j}$ and $h_{t-1,j}$ are the gains in the adjacent frames and σ_j is the standard deviation used to normalize the gains.

The sparseness term, derived from MAP estimation of the sources was defined as:

$$c_t(H) = \sum_{j=1}^{J} \sum_{t=2}^{T} (h_{t,j}/\sigma_j) \qquad (16)$$

f(-) was defined as a function that penalizes non-zero gains and suggested functions that could be incorporated are $f(x) = log(x^2 + 1)$, $f(x) = x$ and $f(x) = -exp(-x^2)$. However, the authors preferred to use $f(x) = |x|$ as it was found to be less sensitive to the weight β.

After exhaustive experimentation, the authors found the proposed model using temporal continuity and sparseness model has higher accuracy and a much better error detection rate compared to the ISA and basic NMF algorithms. Source separation using basic NMF algorithms was also found to be far superior than ISA based methods.



algorithm	detection error rate (%)			SNR (dB)		
	all	pitched	drums	all	pitched	drums
ISA	31	29	33	3.6	4.4	1.9
NMF-EUC	28	28	30	6.6	7.9	3.7
NMF-DIV	26	28	23	7.0	8.8	3.5
NMF-LOG	80	90	57	2.3	2.7	2.2
proposed	**24**	**25**	**22**	**7.3**	**9.1**	**3.6**

Table 1. Simulation Results obtained from [6]

The ISA implementation was performed using the algorithm proposed in [8]. NMF was tested based on algorithms proposed in [11]. NMF-EUC denotes the NMF algorithm by minimizing the reconstruction error based on the square of the Euclidean Distance. NMF-DIV minimizes the Divergence. NMF-LOG is based on nonnegative sparse coding. As it has the lowest SNR and highest detection error, nonnegative sparse coding is not discussed in the paper

A problem associated with most source separation methods is its inefficiency in handling overlapping harmonics. This is particularly common in Western music that favors the twelve-tone equal temperament scale. As a result, common musical intervals have pitch relationships that are very close to integer ratios – 3/2, 4/3, 5/3, etc. Therefore a large number of sources have harmonics that are overlapped with the harmonics of another source. Source separation methods based on spectral decomposition are able to handle overlapping harmonics to a great deal as they operate in the magnitude domain and rely on the observed magnitudes in overlapped T-F regions to recover individual harmonics. However they ignore relative phases of the overlapping harmonics, which play a critical role in the harmonic spectrum [12]. Hence they performance of spectral based decomposition methods are not considered to be optimal, even though they account for harmonic overlapping.

III. CASA BASED METHODS

CASA tries to explain the astonishing abilities of the human auditory system in selective attention where perceived auditory events are grouped intro auditory streams according to common psycho-acoustical cues.

It is generally agreed that CASA algorithms are divided into four steps [16]: Transforming the mixture into a front-end representation such as a correlogram or STFT magnitude for simplicity; Extracting a collection of sinusoidal partials according to according to the period or principal component magnitude information; Grouping the extracted partials iteratively according to some pre defined grouping rules and lastly extracting the sources by binary masking.

Figure 1 shows the schematic diagram of the CASA Model proposed by Wang and Brown. The input first passes through a model of the auditory periphery (cochlear filtering and hair cells) that simulates auditory nerve activity. Midlevel auditory representations are then formed (correlogram and cross-channel correlation map). Next, a two-layer neural oscillator network performs grouping of acoustic components. A final resynthesis path facilitates computation of signal-to-noise ratio. Although the model was proposed for speech separation, it can easily be extended for harmonic source separation.

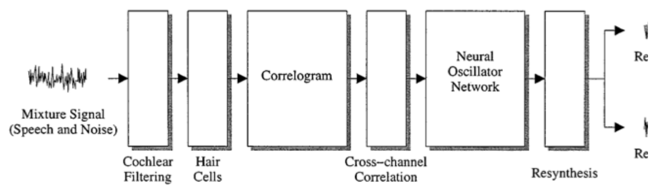


Figure 1. Schematic diagram of the Wang and Brown CASA model

However, CASA based methods are particularly incompetent in separating instruments playing in the same pitch-range into different streams. Proximity of spectral centroids, matching of timbre features learnt on solo excerpts (spectral envelope, onset duration, vibrato amplitude) and location similarities were proposed as supplementary cues to improve source separation [21].

In [22], the authors proposed an algorithm for source separation by grouping the tracks of the same instrument based on common onset of partials and pre trained timbre models describing the evolution of the spectral envelope.

A timbre model can be described as a time-frequency (T-F) template, showing the evolution of the spectral shape with time. Methods based on sinusoidal model extract sinusoidal tracks from some T-F representations of the signal, and then apply grouping rules to assign these tracks to different sources. They are typically used for monaural signals, and may adopt some psychoacoustic cues like loudness [16]. Sinusoidal modeling techniques use sinusoids with time varying frequencies and amplitudes to represent harmonic signals. In [23], the authors proposed a system capable of separating harmonic sounds, using synchronicity and harmonic relations of sinusoidal spectral components. The system is able to yield respectable results, however the amplitude estimations for overlapping partials are substandard. The system is also unable to account for sounds having same onset times.

IV. MODEL BASED METHODS

The Spectral Decomposition methods discussed in section II (ICA and NMF based source separation) were claimed to be inept in separating low intensity notes, according to [21] and produced spurious notes with short durations. Model based methods cans solve these issues by learning accurate priors of the log-spectra of the sources on solo data and by setting priors on event durations. Therefore separation using model-based methods can be performed only using the priori knowledge and is futile for requirements of Blind Source Separation. Model based separations are usually addressed in a Bayesian framework or by using Hidden Markov Models.

In [17], the authors use a probabilistic model of the mixture combining generic priors for harmonicity, spectral envelope, note duration and continuity. However the importance is only given to decomposing the audio signal into harmonic components and not in grouping disparate source streams.

In [20], a three-layer probabilistic generative model, combining ISA, localization models and segmentation models is employed for source estimation in a Bayesian framework.

The main advantage of model based methods lie in the generality of the Bayesian network formalism. The proposed models in [17] and [20] may be improved by modifying only some parts of the layer models and the estimation algorithms depending on the kind of mixture and on the wanted tradeoff between performance and computational ease.

V. DISCUSSION AND CONCLUSION

Despite a wide variety of methods and techniques, music source separation is still largely an unsolved problem. There are some clear shortcomings in existing algorithms that were briefly discussed in previous sections.

CASA based methods are incapable of separating mixed audio signals that have sources within the same pitch range or have a large number of overlapping partials Spectral based decomposition methods are able to overcome this shortcoming to a certain extent and even though not considered to be optimal, they are able to account for harmonic overlapping. Also they were claimed in be inefficient in separating low intensity notes. However being an iterative process, the computational time required for such methods are fairly large.

Model based methods work well only on specific separation problems and requires training of a large number of parameters. Therefore they cannot be sued for separating sources in a BSS fashion.

A problem encountered in most source separation algorithm, particularly blind source separation scenarios, is that the mixed audio signal is usually separated into more signals than active sources [13]. Therefore clustering is required to address this shortcoming of blind source separation.

Two blind clustering algorithms are proposed in [13] that are based on source-filter modeling on an NMF separation method proposed in [6] and are shown is Figures 2 and 3.

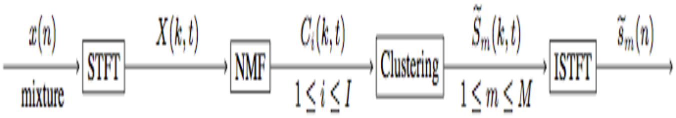


Figure 2. Signal flow of the separation algorithms proposed in [13]

The clustering algorithms (MFCC based and NMF based) are relatively undemanding and are depicted by a signal flow diagram in Figure 3.

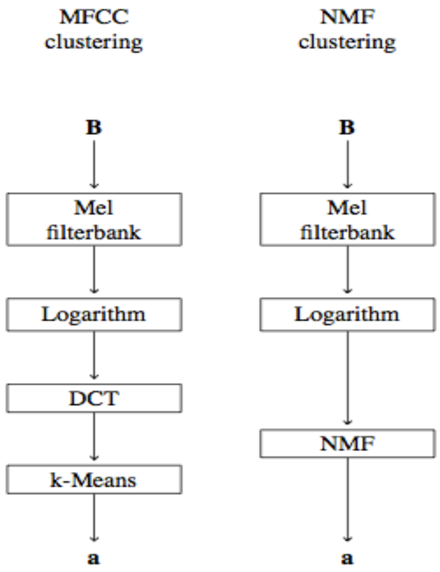


Figure 3. Signal flow of the blind clustering algorithms proposed in [13]

Most of the papers reviewed, make no attempt to discuss noise suppression or discuss the merits of their algorithms in the presence of noise. A further understanding and insight into the merits or demerits of each method would be gained if all the algorithms discussed thus far were evaluated against a common performance index. In [19], the authors propose a method to evaluate the performance of Blind Audio Source Separation (BASS) by taking into account different distortions between estimated sources and the required true source. The amount of interferences, sensor noise and artifacts were also evaluated under a MATLAB toolbox and is available for distribution online.

Future research in the field of audio source separation may include estimating the number of components, automatic clustering and a better estimation of overlapping partials. Since model based techniques provide more information of the sources, mixed approaches using model based methods and unsupervised learning are currently being used and are shown to be more robust and precise.

VI. REFERENCES

[1] Zhiyao Duan, Yungang Zhang, Changshui Zhang and Zhenwei Shi, “Unsupervised Single-Channel Music Source Separation by Average Harmonic Structure Modeling”, IEEE Transactions on Audio, Speech and Language Processing – May 2008

[2] Laura Drake, Aggelos K. Katsaggelos, Janet Rutledge and Jun Zhang “Sound Source Separation via Computational Auditory Scene Analysis-Enhanced Beamforming”, IEEE 2002

[3] Kevin H. Knuth “A Bayesian Approach to Source Separation”, Albert Einstein School of Medicine, Department of Neuroscience

[4] S.T. Roweis, “One Microphone Source Separation”, in Proc. NIPS, 2001

[5] J. Hersheya and M. Casey, “Audio-Visual Source Separation via Hidden Markov Models”, in Proc. NIPS, 2002

[6] Tuomas Virtanen, “Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria”, IEEE Transactions on Audio, Speech and Language Processing – March 2007

[7] Nikolaos Mitianoudis, “Audio Source Separation using Independent Component Analysis”, PhD Thesis, University of London

[8] M.A. Casey and A. Westner, “Separation of mixed audio sources by independent subspace analysis”, int Comp. Music Conference, Berlin, Germany, 2000

[9] Ganesh R. Naik and Dinesh K. Kumar, “An Overview of Independent Component Analysis and its Applications”

[10] Paris Smaragdis, Judith C. Brown: “Non-Negative Matrix Factorization for Polyphonic Music Transcription”, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003.

[11] D.D. Lee and H.S. Seung, “Algorithms for nonnegative matrix factorization”, in Neural Inf. Process Syst., Denver, CO, 2001

[12] Yipeng Li, John Woodruff and DeLiang Wang, “Monaural Musical Sound Source Separation based on Pitch and Common Amplitude Modulation”, IEEE Transactions on Audio, Speech and Language Processing – September 2009

[13] Martin Spiertz, Volker Gnann, “Source Based Clustering For Monaural Blind Source Separation”, Proc. Of the 12th Int. Conference on Digital Audio Effects (DAFx-09) – September 2009

[14] Andre J. W. van der Kouwe, DeLiang Wang, Guy J. Brown, ”A Comparison of Auditory and Blind Separation Techniques for Speech Segregation”, IEEE Transactions on Speech and Audio Processing – March 2001

[15] Peter Jancovic and Munevver Kokuer, “Separation of Harmonic and Speech Signals Using Sinusoidal Modeling”, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics – October 2007

[16] Ruolun Liu, Suping Li, “A Review on Music Source Separation”, Shandong University, Weihai, Shandong, China

[17] Emmanuel Vincent, Mark D. Plumbey “Single-channel Mixture Decomposition using Bayesian Harmonic Models”, Proc. ICA06

[18] Sam T. Roweis, “One Microphone Source Separation”, Proc. NIPS’00, Denver, USA

[19] Emmanuel Vincent, Rémi Gribonval, Cédric Févotte, “Performance Measurement in Blind Audio Source Separation”, IEEE Transactions on Audio, Speech and Language Processing – 2006

[20] Emmanuel Vincent, “Musical Source Separation Using Time-Frequency Source Priors”, IEEE Transactions on Audio, Speech and Language Processing – April, 2005

[21] Burred and T. Sikora, "Monaural Source Separation from Musical Mixtures Based on Time-Frequency Timbre Models." Proc. ISMIR2007, Vienna, Austria, Sept. 2007.

[22] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," Proc. IEEE ICASSP, Istanbul, Turkey, June 2000.