

An Improved Technique for Blind Audio Source Separation

Namgook Cho, Yu Shiu and C.-C. Jay Kuo

Department of Electrical Engineering and Integrated Media Systems Center
University of Southern California, Los Angeles, CA 90089-2564
namgookc@usc.edu, yshiu@usc.edu, cckuo@sipi.usc.edu

Abstract

A blind audio source separation technique with an ill-posed mixing matrix and additive noise is proposed in this work. With this technique, we divide the solution into two steps. The first step is to estimate the ill-posed mixing matrix and the second step is to separate original sources. To estimate the ill-posed mixing matrix, an enhanced soft-assignment method is used in the first step. Then, the generalized p -norm optimization method is adopted in the second step, which can yield a solution sparser than the l_1 -norm minimization technique. Experimental results on synthetic mixtures and real-world mixtures are used to demonstrate the efficiency of the proposed technique in the presence of an ill-posed mixing matrix and additive noise.

1. Introduction

The blind source separation problem has received a lot of attention in recent years due to its wide applications in various signal processing fields such as the enhancement of acoustic, audio, medical and wireless communication signals. Here, we are concerned with the separation of speech and musical sound sources. The objective is to separate the speech signal from the musical background from synthetic mixtures of vocal signals and musical sounds and real-world mixtures from commercial music CD excerpts.

The blind source separation problem has been examined by researchers in recent years, *e.g.* [1, 5]. Most of previous work focused on a well-posed mixing system. The independent component analysis (ICA) can be applied under the assumption that sources are statistically independent [2]. While ICA-like algorithms are fast and reliable, they demand that the number of sensors be no less than the number of sources. To relax this constraint, an optimization approach was adopted in [1, 5] to maximize the MAP (maximum *a posteriori* probability) func-

tion for an under-determined system, and it yields good results. However, the normal system for the problem addressed in [1, 5] is well-conditioned. For real-world mixed speech and music data, the mixing system is actually ill-posed. Thus, the approach used in [1, 5] does not work properly for the problem of our interest.

For an ill-posed mixing system, it is difficult to estimate the mixing matrix and the signal sources simultaneously. Thus, we propose a technique that divides the solution into two steps. The first step is to estimate the ill-posed mixing matrix in a noisy environment while the second step is to separate original sources. To estimate the ill-posed mixing matrix, an enhanced soft-assignment method is used in the first step. Then, the generalized p -norm optimization method is adopted in the second step, which can yield a solution sparser than the l_1 -norm minimization technique. Experimental results using synthetic mixtures and real-world mixtures are used to validate the proposed technique in the presence of an ill-posed mixing matrix and additive noise.

2. Ill-posed Mixing System

A practical audio source mixing model can be expressed as

$$\underline{x} = A \cdot \underline{s} + \underline{n}, \quad (1)$$

where \underline{x} is the output vector consisting N mixed signals, \underline{s} is the input vector consisting of M sources, A is a matrix of size $N \times M$ and \underline{n} is an N -dimensional noise vector. The system in (1) is under-determined if $N < M$. Furthermore, it is ill-posed if A has a singular value close to zero and the ratio between its largest and smallest nonzero singular values is large. The condition number of matrix A is defined as [4].

$$\text{cond}(A) \equiv \|A\| \cdot \|A^+\| \quad (2)$$

where $\|A^+\|$ is the Moore-Penrose pseudo-inverse. If A is ill-posed, its condition number is large. The large condition number implies that column vectors of A are

close to linearly dependent. In a blind source separation problem, we attempt to estimate mixing matrix A and source signal \underline{s} based on observed mixture vector \underline{x} . The noise term \underline{n} is often small so that it can be neglected as an approximation.

When the input sources are sparse in time or frequency, the scatter plot of coefficients of observations would constitute a mixture of lines [1]. Each oriented line is related to a single source and, thus, it corresponds to a column of the mixing matrix. When A is well-conditioned, two oriented lines intersects with each other at a large angle so that they can be easily identified.

By examining real-world mixture data carefully, we find that the oriented lines intersect with each other at a small angle which implies that mixing matrix A is ill-posed. An example is shown in Fig. 1, where we show the scatter plot from a commercial music CD excerpt, "Let It Be," by Beatles. The x- and y-axes are real parts of Fourier coefficients of audio signals from two stereo channels. The small intersection angle prevents the line orientation algorithm in [5] from estimating the mixing matrix correctly. Thus, its source separation performance degrades. Our objective is to estimate ill-posed mixing matrix A and source vector \underline{s} .

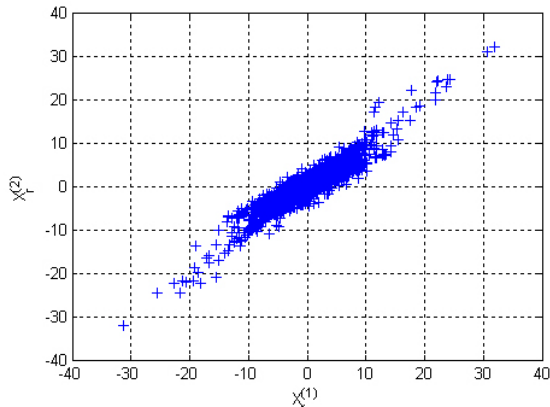


Figure 1. The scatter plot of a commercial music CD excerpt.

3. Proposed Algorithm

3.1. Mixing Matrix Estimation

The basic idea is to estimate oriented lines from the scatter plot, which form columns of the mixing matrix. A method to determine oriented lines using soft data assignment, which was proposed in [5], is stated below.

Let k and N be the numbers of oriented lines and data points, respectively. To estimate oriented lines, each observed data point \underline{x}_j is assigned to line \underline{v}_i according to a weight

$$\hat{z}_{ij} = \frac{\exp(-\beta \cdot z_{ij})}{\sum_r \exp(-\beta \cdot z_{rj})}, \quad (3)$$

where $1 \leq i \leq k$ and $1 \leq j \leq N$, β controls the softness, and $z_{ij} = \|\underline{x}_j - \langle \underline{v}_i, \underline{x}_j \rangle \cdot \underline{v}_i\|^2$ measures the distance between a data point and a line.

The above soft assignment method works well for a well-posed mixing system and clean observations. However, for an ill-posed mixing system with noisy observations, it fails to identify oriented lines correctly. The main reason is that the difference between weights \hat{z}_{i1} and $\hat{z}_{(i+1),1}$ of data \underline{x}_1 is too small. We consider a modification to (4), *i.e.*

$$\hat{z}_{ij} = \frac{z_{ij}^{-m}}{\sum_r z_{rj}^{-m}}, \quad (4)$$

where m is a control parameter. Please note that, as compared with (3), the difference between a data point close to and far away from a line is enlarged by the modified weight in (4). As a result, it works better for an ill-posed mixing matrix.

3.2. Source Signal Estimation

After estimating the mixing matrix, sparse sources can be separated using the generalized p -norm optimization method. That is, the sparse source estimation is obtained by

$$\min \lambda \cdot \|\underline{s}\|_p^p \quad \text{subject to } \underline{x} = \hat{A} \cdot \underline{s} \quad (0 < p < 1) \quad (5)$$

where λ is a regularization parameter and \hat{A} is the estimated mixing matrix. The above optimization problem can be interpreted as a MAP estimation where the prior probability of a source is in form of $p(s) \propto \exp(-|s|^p)$. The regularization parameter λ can make the solution more robust with respect to additive noise.

When compared to the l_1 -norm minimization technique or the linear programming approach given in [1, 5], the generalized p -norm optimization method can produce a sparser solution. That is, there are more vanishing components (or zeros) in the estimated source vector \underline{s} for $\underline{x} = \hat{A} \cdot \underline{s}$. Thus, for an ill-posed mixing system with noisy data, the above generalized p -norm optimization method can produce a robust and sparser solution.

3.3. Application to Audio Source Separation

When applied to stereo audio channels, the proposed blind source separation algorithm can be described below.

1. The observed stereo audio data are transformed from the time domain to the Fourier domain to reveal the sparse property better.
2. We assign weights to each data point using (4) and define

$$\Sigma_i = \frac{\sum_j \hat{z}_{ij} \cdot \underline{x}_j \cdot \underline{x}_j^T}{\sum_j \hat{z}_{ij}} \quad (6)$$

Then, we perform the eigen value decomposition of Σ_i ,

$$\Sigma_i = U_i \cdot \Lambda_i \cdot U_i^{-1},$$

which provides the estimated column vector $\underline{v}_i^{new} = \underline{u}_{max}$ of the mixing matrix.

3. The sparse input sources are estimated by (5).
4. The source estimates are transformed from the frequency domain back to the time domain.

4. Experimental Results

Experiments were conducted to evaluate the performance of the proposed blind source separation algorithm with synthetic mixtures of vocal and musical sounds and real-world mixtures from commercial music CD excerpts. For synthetic mixtures, two cases were examined and compared. They are a clean mixture and a mixture with additive white noise. Furthermore, to emulate the ill-posed mixing system, synthetic mixing matrices with a high condition number were used. The speech data were randomly chosen from the TIMIT speech database while musical signals were musical instrument sounds from the Acoustical Society of America (<http://asa.aip.org/sound.html>).

Time-domain signals were transformed to the frequency domain using a 512-point windowed FFT with 256-point overlap between windows and real coefficients were used for the scatter plot as in Fig. 1. Similar results are obtained when imaginary coefficients are used. In our experiments, parameter m for the improved weight in (4) was set to 2. The source-to-distortion ratio (SDR) [6] in the unit of dB is used to measure the performance of estimated sources.

The regularization parameter λ in (5) can be set independently for each vector as in [3], *i.e.*

$$\lambda_j = \frac{\|\underline{x}_j\| - \|\underline{e}_j\|}{\|\underline{x}_j\|} \cdot \lambda_{max} \quad (7)$$

where $\lambda_{max} = 2 \times 10^{-3}$ and $\underline{e}_j = \underline{x}_j - \hat{A} \cdot \underline{s}_j$.

The performance of three source separation algorithms was compared in Tables 1-3 using synthetic mixtures of vocal and musical signals. They are the LOST algorithm [5], the proposed method with $p = 0.5$ and fICA [2]. The test conditions were:

- Case 1: two sources and two observed mixtures without additive noise;
- Case 2: two sources and two observed mixtures with additive white noise (SNR = 10 dB);
- Case 3: three sources and two observed mixtures with additive white noise (SNR = 15 dB).

Two mixing matrices A_1 and A_2 with a different condition number were tested in each case. Please note that fICA can be applied to Cases 1 and 2, but not Case 3 (which is an under-determined case).

We observe a similar performance for all algorithms for well-posed matrix A_1 . In Table 1, the LOST algorithm with two β values were considered to illustrate the softness effect. We see that the performance of the LOST algorithm with $\beta = 2$ degrades significantly for ill-posed mixing matrix A_2 . For ill-posed matrix A_2 with additive noise, the generalized p -norm optimization method proposed in this work has the best performance. This can be explained as follows. As the noise level and condition number become higher, the soft-assignment in (3) fails to identify the oriented lines correctly. Furthermore, when the condition number increases, the noise is amplified as well.

Table 1. Test Case 1: two sources and two observed mixtures without additive noise.

	A_1 (1.4561)		A_2 (11.0)	
	\hat{s}_1	\hat{s}_2	\hat{s}_1	\hat{s}_2
LOST ($\beta = 2$)	43.01	43.37	-7.29	7.54
LOST ($\beta = 21$)	39.66	45.65	30.87	34.11
$l_{0.5}$ -norm	42.77	45.65	32.22	33.69
fICA	46.96	42.06	37.91	48.39

For experiments with the real-world mixtures, we considered stereo channels from the commercial music CD excerpt, which was downsampled to 11,025 kHz

Table 2. Test Case 2: two sources and two observed mixtures with additive white noise (SNR = 10 dB).

	A_1 (1.4561)		A_2 (11.0)	
	\hat{s}_1	\hat{s}_2	\hat{s}_1	\hat{s}_2
LOST ($\beta = 21$)	9.56	10.28	-2.71	-5.81
$l_{0.5}$ -norm	9.73	10.25	7.86	5.21
fICA	9.56	10.27	2.56	-4.38

Table 3. Test Case 3: three sources and two observed mixtures with additive white noise (SNR = 15 dB).

	A_1 (3.88)			A_2 (7.92)		
	\hat{s}_1	\hat{s}_2	\hat{s}_3	\hat{s}_1	\hat{s}_2	\hat{s}_3
LOST	4.45	13.25	8.39	-0.01	4.83	3.04
$l_{0.5}$ -norm	6.57	13.06	10.4	5.76	11.7	5.64

with 16 bits resolution and of 7 second long. The stereo channels contained a vocal sound with the piano background. The estimated sound y_1 is the vocal sound and y_2 the background piano sound.

Fig. 2 compares the spectrograms of estimated sources using the l_1 -norm minimization and the generalized p -norm optimization methods. We see that the generalized p -norm optimization method ($p = 0.5$) has better performance. With the l_1 -norm minimization technique, the estimated sound y_1 has some component of the residual background piano sound (the horizontal lines) while the estimate musical sound y_2 still contains some residual vocal sound. In contrast, with the generalized $l_{0.5}$ -norm optimization method, the vocal component in y_1 is much more apparent and the residual musical component almost disappears completely. Similarly, y_2 has only the background music sound. The above claim is clearly confirmed by the hearing test.

5. Conclusion and Future Work

The blind audio source separation problem with ill-posed mixing system and noisy observation data was studied. We proposed an improved algorithm that used an enhanced soft-assignment method to estimate the mixing matrix and the generalized p -norm optimization technique to separate the sparse sources. It

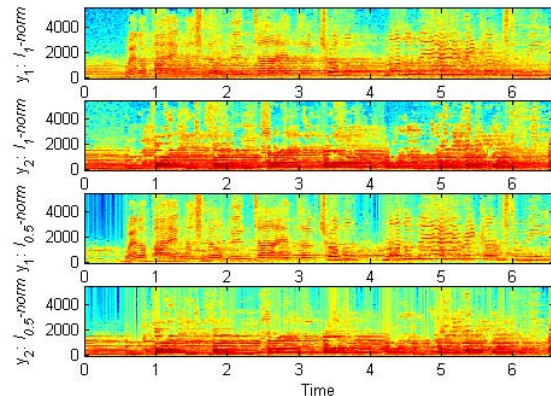


Figure 2. Spectrograms of estimated sources from an excerpt of “Let It Be” by Beatles (from top to bottom): y_1 and y_2 with the l_1 -norm minimization technique and y_1 and y_2 with the generalized p -norm optimization method proposed in this work.

was shown by experimental results that the proposed method is more robust against noisy observations and ill-posedness of the mixing matrix. Even though our current study is insightful, it is basically an algorithmic approach. Some in-depth theoretic analysis is still lacking. In the near future, we will conduct a more thorough analysis so as to understand the performance of various blind source separation algorithms more rigorously.

References

- [1] P. Bofill and M. Zibulevsky. Blind separation of more sources than mixtures using sparsity of their short-time fourier transform. In *Int. Conf. Independent Component Anal.*, pages 87–92, Helsinki, Finland, June 2000.
- [2] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10:626–634, 1999.
- [3] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15:349–396, February 2003.
- [4] T. K. Moon and W. C. Stirling. *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall, Upper Saddle River, NJ, 1999.
- [5] P. D. O’Grady and B. A. Pearlmutter. Soft-lost: Em on a mixture of oriented lines. In *Int. Conf. Independent Component Anal.*, pages 428–435, Granada, Spain, Sept. 2004.
- [6] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Speech and Audio Processing*, 14:1462–1469, July 2006.