A COMPRESSED SENSING APPROACH FOR UNDERDETERMINED BLIND AUDIO SOURCE SEPARATION WITH SPARSE REPRESENTATION

Tao Xu and Wenwu Wang

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, United Kingdom
Emails:[t.xu; w.wang]@surrey.ac.uk

ABSTRACT

The problem of underdetermined blind audio source separation is usually addressed under the framework of sparse signal representation. In this paper, we develop a novel algorithm for this problem based on compressed sensing which is an emerging technique for efficient data reconstruction. The proposed algorithm consists of two stages. The unknown mixing matrix is firstly estimated from the audio mixtures in the transform domain, as in many existing methods, by a K-means clustering algorithm. Different from conventional approaches, in the second stage, the sources are recovered by using a compressed sensing approach. This is motivated by the similarity between the mathematical models adopted in compressed sensing and source separation. Numerical experiments including the comparison with a recent sparse representation approach are provided to show the good performance of the proposed method.

***Index Terms*—** Underdetermined Blind Source Separation, Sparse Representation, Compressed Sensing

1. INTRODUCTION

Human perception of acoustic mixtures, classically referred to as the cocktail party problem, results from the vibration of the ear drum by superposition of different air-pressure signals that are emitted from diverse audio sources at the same time. Human auditory systems can distinguish the sources from the mixtures quite efficiently. It is, however, a difficult task for machines to separate them robustly with no or very limited prior information about the sources and the mixing environment. Several techniques including blind source separation (BSS) have been employed to address this problem. A widely used technique for BSS is independent component analysis (ICA) which has good performence for an over- or exactly-determined case. In practice, however, an underdetermined problem is usually encountered, where the number of the sources is greater than that of the mixtures. An effective method for this problem is to use the so-called sparse signal representation [1], assuming that the sources are sparse or can be decomposed into combinations of sparse components.

Based on such a representation, a two-stage approach is typically employed to recover the source signals. First, the unknown mixing matrix is estimated from the audio mixture data by using a clustering algorithm, operating in the time domain or the transform domain. Second, the source signals are recovered by using a nonlinear optimization algorithm. In this paper, we adopt a similar strategy for the underdetermined problem. Different from existing approaches, however, we propose a new method for the source recovery in the second stage based on the emerging technique of compressed sensing (CS). The CS, which has attracted growing interests in signal processing, is an efficient technique for data acquisition and reconstruction [2]. It can randomly sample signals under Nyquist rate and then reconstruct the signal with a high probability. It provides potentially a powerful framework for computing a sparse representation of signals. In this work, we analyse the similarity between the fundamental models for CS and BSS, and then develop an algorithm for source recovery based on their relations. The next section describes the proposed method in detail. Numerical results are given in Section 3, followed by conclusions in Section 4.

2. THE PROPOSED METHOD

We employ a two-stage sparse representation approach, with the first stage devoted for the estimation of the mixing matrix and the second stage for the source recovery using the estimated mixing matrix. Our main contributions focus on the second stage.

2.1. Estimating the mixing matrix by clustering

In this stage, we estimate the mixing matrix using the K-means clustering algorithm based on the short-time Fourier transform (STFT) coefficients. Considering a noise-free model, the underdetermined BSS problem can be described as

$$\mathbf{X} = \mathbf{A}\mathbf{S} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{M\times N}$ is the unknown mixing matrix assumed to be of full rank with $M < N$, $\mathbf{X} \in \mathbb{R}^{M\times T}$ is the observed

This work was supported by the CSC of China and the EPSRC of the UK.

978-1-4244-2710-9/09/$25.00 © 2009 IEEE

data matrix whose row vector $\mathbf{x}_i$ is the ith sensor signal with each having T samples at discrete time instants $t = 1, ..., T$, and $\mathbf{S} \in \mathbb{R}^{N \times T}$ is the unknown source matrix containing N source vectors. For convenience, we consider the case of $M = 2$ and $N = 3$ in this paper. Assuming the sources are sparse, i.e. ideally only one source has nonzero value at each time instant, the scatter plot of the mixtures will show clear directions correspond to the columns of $\mathbf{A}$. For example, when $M = 2$, at any time instant, the point on the scatter plot of $\mathbf{x_1}$ versus $\mathbf{x_2}$ should lie on the line that can be represented by one of the column vectors in $\mathbf{A}$, since there is only one source data in this time instant. The vector of the plotted points is a product of a scalar and one of column vectors in $\mathbf{A}$. When all the data points are plotted, some lines in the coordinate plane can be clearly identified, and the number of lines equals to that of the columns of $\mathbf{A}$. In practice, however, the sparseness assumption is seldom satisfied nicely, due to the observation noises in real applications. The lines are usually broadened especially in the time domain, as shown in Figure 1(a). It has been observed that the audio mixtures become sparser if they are transformed into the frequency domain. As a result, it becomes easier to observe the distributions of the data points in the scatter plot, as shown in Figure 1(b). Therefore, to obtain more accurate estimate of the mixing matrix, we apply the K-means algorithm to the audio data in the transform domain obtained by the STFT. In summary, the algorithm for estimating $\mathbf{A}$ (in the case M=2, N=3) is described as follows:

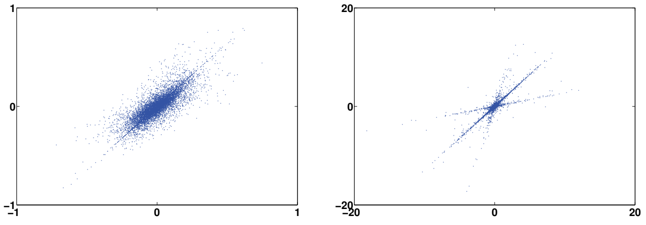
Algorithm 1:

- Step 1. Compute the sparse coefficients of the two mixture vectors by the STFT, and obtain $\tilde{\mathbf{X}}$, i.e. the time-frequency representation of $\mathbf{X}$.
- Step 2. Normalize the vectors in $\tilde{\mathbf{X}}$ to move all points to a unit semi-circle for the application of the K-means algorithm.
- Step 3. Choose the starting points for the K-means algorithm, and divide $\tilde{\mathbf{X}}$ to three parts (equals to the number of sources) and compute the mean values of each part as the initial centres.
- Step 4. Run a K-means clustering algorithm to update iteratively the three centres until convergence, and compute the column vectors of the estimated mixing matrix $\hat{\mathbf{A}}$ as the final cluster centres.

2.2. Separating sources by compressed sensing

In this stage, we formulate the signal recovery problem as a compressed sensing model with the mixing matrix $\hat{\mathbf{A}}$ being obtained in the previous section. For $M = 2$ and $N = 3$, Equation (1) can be expanded as:

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \mathbf{s}_3 \end{pmatrix} \qquad (2)$$



(a) Time domain (b) Transform domain

Fig. 1. An example of scatter plots of two linear audio mixtures in the time (a) and frequency (b) domain.

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the mixtures, $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ are the sources, and a_{ij} is the ij-th element of the mixing matrix $\mathbf{A}$. We can formulate the above equation as follows

$$\underbrace{\begin{pmatrix} x_1(1) \\ \vdots \\ x_1(T) \\ x_2(1) \\ \vdots \\ x_2(T) \end{pmatrix}}_{\mathbf{b}} = \underbrace{\begin{pmatrix} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} \end{pmatrix}}_{\mathbf{M}} \underbrace{\begin{pmatrix} s_1(1) \\ \vdots \\ s_1(T) \\ s_2(1) \\ \vdots \\ s_2(T) \\ s_3(1) \\ \vdots \\ s_3(T) \end{pmatrix}}_{\mathbf{f}} \qquad (3)$$

where $\Lambda_{ij} \in \mathbb{R}^{T \times T}$ is a diagonal matrix whose diagonal elements are all a_{ij}. Let $\mathbf{b} = vec(\mathbf{X})$, $\mathbf{f} = vec(\mathbf{S})$, where vec is an operator stacking the column vectors of a matrix into a single vector. Equation (3) can be written in a compact form as:

$$\mathbf{b} = \mathbf{Mf} \qquad (4)$$

We can clearly observe that this equation is similar to the model of CS in which $\mathbf{b}$ is the compressed vector of signals in $\mathbf{f}$ and $\mathbf{M}$ is the measurement matrix. Therefore, a sparse representation in the transform domain can be employed for $\mathbf{f}$:

$$\mathbf{f} = \mathbf{\Phi y} \qquad (5)$$

where $\mathbf{\Phi}$ is a transform dictionary and $\mathbf{y}$ contains the weighting coefficients in the $\mathbf{\Phi}$ domain. Combining (4) and (5), we have

$$\mathbf{b} = \mathbf{M\Phi y} \qquad (6)$$

According to compressed sensing theories, if both $\mathbf{M}$ and $\mathbf{\Phi}$ satisfy certain conditions and also $\mathbf{y}$ is sparse, the signal $\mathbf{f}$ can be recovered by measurements $\mathbf{b}$ using an optimization process. This indicates that separation of the sources in our underdetermined problem can be achieved by computing $\mathbf{y}$ in Equation (6) using any signal recovery algorithm in CS. The vector $\mathbf{f}$ which consists of the three separated sources can be obtained by simply multiply the dictionary $\mathbf{\Phi}$ with $\mathbf{y}$ using Equation (5). Consequently, the undetermined BSS is transformed to the problem of signal recovery in CS, and many CS algorithms can therefore be used for recovering the sources in the underdetermined system. A straightforward approach for estimating $\mathbf{f}$ from $\mathbf{b} = \mathbf{Mf} = \mathbf{M\Phi y}$ is to solve the following l_0 minimization problem

$$min \parallel \mathbf{y} \parallel_0 \qquad s.t. \qquad \mathbf{b} = \mathbf{M\Phi y} \tag{7}$$

where $\parallel \mathbf{y} \parallel_0$ is the l_0 norm which measures the sparseness of $\mathbf{y}$. The solution to the optimisation of the above cost function is an NP-hard problem, which is not a good choice in practice. However, it has been shown in [3] that the solution to the l_0 minimisation is essentially equivalent to solving the following l_1 minimisation problem

$$min \parallel \mathbf{y} \parallel_1 \qquad s.t. \qquad \mathbf{b} = \mathbf{M\Phi y}. \tag{8}$$

Therefore, we can use the Basis Pursuit (BP) method [3] to solve the problem and to find the sparsest representation of $\mathbf{b}$. This can be achieved based on the following iterative process. First, we choose an initial basis matrix $\mathbf{B}$ wihch is a squared matrix having the same rank as $\mathbf{M\Phi}$ and consists of the selected columns of $\mathbf{M\Phi}$, i.e. the smallest possible complete dictionary. Then, we improve the basis by swapping a column of $\mathbf{B}$ with an unselected column in $\mathbf{M\Phi}$. When the basis cannot be further improved, the optimal solution is reached. Finally, $\mathbf{y}$ can be readily computed by $\mathbf{B}^{-1}\mathbf{b}$.

Apart from the BP method used above, there are alternative methods for recovering the source signals, such as the FOCUSS algorithm, the EM algorithm, and the Matching Pursuit algorithm [4]. All these approaches require certain conditions on the dictionary $\mathbf{M\Phi}$ be satisfied in order to recover the source signal successfully. Successful signal recovery can be achieved when $\mathbf{M\Phi}$ obeys a uniform uncertainty principle [5]. It means that every submatrix of $\mathbf{M\Phi}$ has to be well designed to obey a Restricted Isometry Property (RIP) [5]. In other words, if $(\mathbf{M\Phi})_\Omega$ represents the submatrix obtained by extracting the columns of $\mathbf{M\Phi}$ and $\Omega \subset \{1, \cdots, k\}$ is a set of indices of the columns, there is a restricted isometry constant $\delta_p = \delta_p(\mathbf{M\Phi})$, which is the smallest number such that the following inequality holds for all subsets Ω with $k \leq p$.

$$(1 - \delta_p) \parallel \mathbf{y} \parallel_2^2 \leq \parallel (\mathbf{M\Phi})_\Omega \mathbf{y} \parallel_2^2 \leq (1 + \delta_p) \parallel \mathbf{y} \parallel_2^2 \tag{9}$$

According to Theorem (1.4) in [5], if $\mathbf{M\Phi}$ satisfies

$$\delta_\Omega(\mathbf{M\Phi}) + \delta_{2\Omega}(\mathbf{M\Phi}) + \delta_{3\Omega}(\mathbf{M\Phi}) < 1 \tag{10}$$

$\mathbf{f}$ can be exactly recovered by the l_1 norm reconstruction. Some certain type of random matrices (e.g. Gaussian, Bernoulli) strongly satisfies the uniform uncertainty principle. The mixing matrix estimated by the clustering algorithm also approximately satisfies such conditions, based on a recent study in [6], where the relation between $\mathbf{M}$, $\mathbf{\Phi}$, and $\delta_\Omega(\mathbf{M\Phi})$, together with a bound of errors in the estimation of $\mathbf{M}$ has been derived.

We have focused on finding the best representation of a signal using a fixed overcomplete dictionary. However, we can also use the overcomplete dictionary for decomposing the signals since overcomplete representations have greater flexibility in matching the structure of signals and can form more compact representations. Our model provides the feasibility when the product of the mixing matrix and the overcomplete dictionary satisfies the uniform uncertainty principle. The overcomplete dictionary is typically composed of a set of complete bases (e.g. Fourier, Wavelet and Gabor basis), or the atoms added to a complete basis (e.g. adding harmonic frequencies to the Fourier basis). In the overcomplete dictionary, signals are described by linear combinations of these atoms which can be very sparse. This is an advantage of using the compressed sensing model. Based on the above discussions, the algorithm for recovering the sources (in the case M=2, N=3) is summarised below.

Algorithm 2:

- Step 1. Transform the observed data matrix to a column vector as the measurement vector in CS.
- Step 2. Obtain the product of the transformation of the estimated mixing matrix $\mathbf{M}$ and the dictionary $\mathbf{\Phi}$ by the library of the linear operators in Sparco framework [7].
- Step 3. Use the BP algorithm to find the sparsest coefficients in the dictionary computed in Step 2.
- Step 4. Compute an inverse transform of these coefficients to the original (time) domain.
- Step 5. Split the resultant single vector into multiple source vectors.

3. EXPERIMENTAL RESULTS

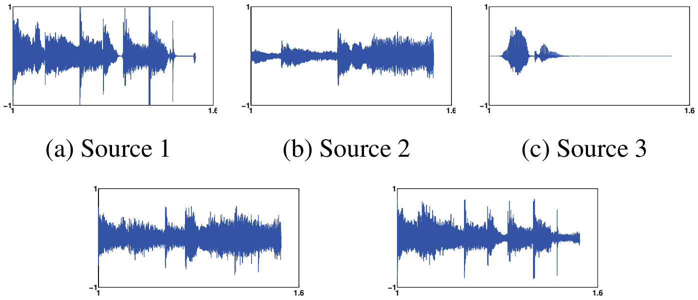
In our simulations, we generate two mixture signals by mixing together three audio sources using the following mixing matrix. The sources include a guitar signal, a piano signal and an English utterance, shown in Figure 2 (a), (b) and (c) respectively. The mixtures are plotted in Figure 2 (d) and (e).

$$\mathbf{A} = \begin{pmatrix} 0.6118 & 0.9648 & 0.2360 \\ 0.7910 & 0.2629 & 0.9718 \end{pmatrix} \tag{11}$$

Firstly, we perform Algorithm 1 for estimating the mixing matrix $\hat{\mathbf{A}}$, shown in (12). The Hamming window and a 50% overlap are used for computing the STFT.

$$\hat{\mathbf{A}} = \begin{pmatrix} 0.6165 & 0.3100 & 0.9247 \\ 0.7841 & 0.9447 & 0.3547 \end{pmatrix} \tag{12}$$

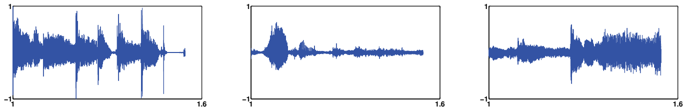
We see that $\hat{\mathbf{A}}$ is reasonably close to $\mathbf{A}$ except the permutation ambiguity within the second and third column. Secondly, we recover the three sources by applying Algorithm 2. The result is shown in Figure 3. The separation quality is further



(a) Source 1 (b) Source 2 (c) Source 3

(d) Mixture 1 (e) Mixture 2

Fig. 2. The three audio sources (a), (b), (c) and the two mixtures (d), (e) used in the experiment. The horizontal and vertical axises are the time instants (in seconds) and amplitude respectively, same for Figure 3.



(a) Estimation 1 (b) Estimation 2 (c) Estimation 3

Fig. 3. The three estimated audio source signals.

measured using the Signal-to-Noise Ratio (SNR), defined as follows:

$$SNR = 10log_{10}\left(\frac{\sum_{t=1}^{T}\mathbf{s}(t)^2}{\sum_{t=1}^{T}(\mathbf{s}(t)-\hat{\mathbf{s}}(t))^2}\right) \tag{13}$$

where $\mathbf{s}(t)$ is the original source and $\hat{\mathbf{s}}(t)$ is the estimated source. Table 1 shows the comparison between our proposed method which is called Clustering-CS (CCS) and a recent sparse representation approach in [8], denoted by NUIRLS, based on the single experiment for the above mixture signals in terms of SNR measurements for the three sources.

	source1	source2	source3
CCS	13.26	16.77	6.17
NUIRLS	3.89	5.82	1.08

Table 1. The SNR (in dB) of the three sources.

The SNR in Table 1 is measured for the sources recovered by Algorithm 2 based on the estimated mixing matrix $\hat{\mathbf{A}}$. If the original mixing matrix $\mathbf{A}$ is used instead, the SNR of the three sources are measured as 14.56dB, 17.65dB, and 7.66dB respectively. It means that our method is reasonably robust to the error in estimating the mixing matrix. Using the same sources, we have also created 20 pairs of mixtures by using 20 randomly generated mixing matrices. We have applied both CCS and NUIRLS algorithms to these mixtures and measured the averaged separation performance, in terms of the mean and the standard deviation of SNRs of these experiments. The result is shown in Table 2. From this table, we can observe that the proposed method has considerably better performance than the NUIRLS algorithm.

	source1	source2	source3
CCS	11.60 ± 4.47	16.48 ± 4.09	5.14 ± 3.38
NUIRLS	3.10 ± 2.86	5.69 ± 3.73	2.11 ± 1.97

Table 2. The mean and standard deviation of the SNR (in dB) in 20 experiments.

4. CONCLUSIONS AND FUTURE WORK

We have presented a new two-stage method for the signal recovery in underdetermined blind audio source separation using compressed sensing. Numerical experiments have shown that the proposed method outperforms considerably a recent sparse representation based approach. An advantage of the proposed method resides in its potential to accommodate more efficient algorithms in compressed sensing or to design more accurate basis dictionaries, so as to enhance further the performance of the underdetermined separation systems. This, together with extensions of the proposed approach for the separation of reverberant mixtures, will be further investigated in our future work.

5. REFERENCES

[1] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, Nov. 2001.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[4] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[5] E. Candes and T. Tao, "The dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, pp. 2313–2351, 2007.

[6] T. Blumensath and M. E. Davies, "Compressed sensing and source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation*, 2007, pp. 341–348.

[7] E. van den Berg, M. P. Friedlander, G. Hennenfent, F. Herrmann, R. Saab, and Ö. Yılmaz, "Sparco: A testing framework for sparse reconstruction," Tech. Rep. TR-2007-20, Dept. Computer Science, University of British Columbia, Vancouver, Oct. 2007.

[8] P. D. O'Grady and S. T. Rickard, "Compressive sampling of non-negative signals," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing*, Cancun, Mexico, Oct. 2008, pp. 133–138.