A BLOCK-BASED COMPRESSED SENSING METHOD FOR UNDERDETERMINED BLIND SPEECH SEPARATION INCORPORATING BINARY MASK

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

A block-based compressed sensing approach coupled with binary time-frequency masking is presented for the underdetermined speech separation problem. The proposed algorithm consists of multiple steps. First, the mixed signals are segmented to a number of blocks. For each block, the unknown mixing matrix is estimated in the transform domain by a clustering algorithm. Using the estimated mixing matrix, the sources are recovered by a compressed sensing approach. The coarsely separated sources are then used to estimate the time-frequency binary masks which are further applied to enhance the separation performance. The separated source components from all the blocks are concatenated to reconstruct the whole signal. Numerical experiments are provided to show the improved separation performance of the proposed algorithm, as compared with two recent approaches. The block-based operation has the advantage in improving considerably the computational efficiency of the compressed sensing algorithm without degrading its separation performance.

Index Terms— Underdetermined blind source separation (BSS), sparse representation, compressed sensing (CS), block-based processing, binary time-frequency mask

1. INTRODUCTION

The problem of underdetermined blind source separation (BSS) has been studied extensively in recent years. The objective of underdetermined speech separation is to estimate the unknown speech sources from the mixtures without (or with limited) prior knowledge about the mixing channels, where the number of the sources is greater than that of the mixtures. Considering a noise-free instantaneous model, the problem can be described as:

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

where $\mathbf{A} \in R^{M \times N}$ is the unknown mixing matrix assumed to be of full rank with $M < N$, $\mathbf{X} \in R^{M \times T}$ is the observed data matrix whose row vector $\mathbf{x}_i$ is the ith microphone signal with each having T samples at discrete time instants $t = 1, ..., T$, and $\mathbf{S} \in R^{N \times T}$ is the unknown source matrix containing N source vectors $\mathbf{s}_j$, $j = 1, ..., N$.

An effective method for this problem is to use the so-called sparse signal representation, assuming that the sources are sparse or can be decomposed into the combination of sparse components [1] [2]. Using such a representation, we have recently proposed a novel algorithm based on compressed sensing (CS) [3] in which the BSS model is reformulated to a signal recovery model. However, the optimization process for source estimation is computationally demanding as the microphone signals of full length are stacked into a single vector, resulting in a large dimension of the measurement matrix, as well as the signal dictionary. In this paper, we propose to improve its computational efficiency using a block-based method. Another contribution of this work is to use the binary masking technique to enhance the separation performance of the compressed sensing algorithm. As a result, the proposed approach is a multi-stage system. We will evaluate its systematic performance using the metrics and datasets in [4], as compared with two recent methods. The next section describes the proposed method in detail. Numerical results are given in Section 3, followed by conclusions in Section 4.

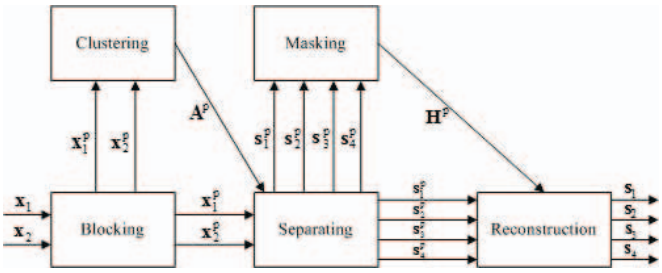


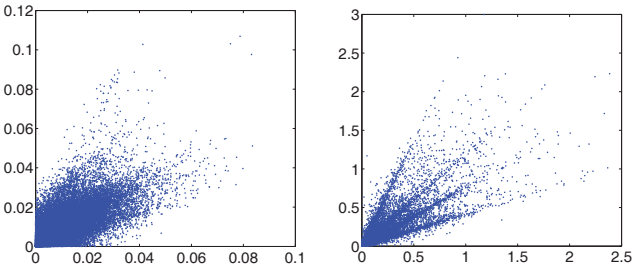
Fig. 1. Flow chart of the proposed system for separating four speech sources from two mixtures.

2. THE PROPOSED METHOD

We consider the case of $M = 2$ and $N = 4$ in this paper. The proposed separation system is depicted in Figure 1. First, the

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

speech mixtures $\mathbf{x}_i$, $i = 1, 2$, are segmented into P blocks $\mathbf{x}_i^p$, $p = 1, ..., P$, with each block having L samples. Then, the mixture signals at each block $\mathbf{x}_i^p$, $i = 1, 2$ are processed in the subsequent clustering, separating, and masking steps. Finally, the separated sources are reconstructed from the source components within each block. For simplifying notations, we omit p in $\mathbf{x}^p$, $\mathbf{s}^p$, $\mathbf{A}^p$, and $\mathbf{H}^p$ in the following descriptions. As shown in Section 3, compared with processing the whole signal, the block-based processing considerably improves the computational efficiency of the algorithm without degrading its separation performance.



(a) Time domain (b) Frequency domain

Fig. 2. An example of the scatter plots of two mixtures of four speech sources in the time (a) and the frequency (b) domain.

In the clustering step, we estimate the mixing matrix using the K-means clustering algorithm based on the short-time fourier transform (STFT) coefficients. Assuming the sources are sparse, i.e. ideally only one source has a nonzero value at each time instant, some lines in the scatter plot of the mixtures can be clearly identified, and the number of lines is equal to that of 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 $\mathbf{A}$. The vector of the plotted points is a product of a scalar and one of column vectors in $\mathbf{A}$. In practice, however, the sparseness assumption is seldom satisfied nicely, due to the observation noises in real data. The lines are usually broadened especially in the time domain, as shown in Figure 2(a). It has been observed that the audio mixtures become sparser if they are transformed into the frequency domain, as shown in Figure 2(b). As a result, it becomes easier to observe the distributions of the data points in the scatter plot. Therefore, to estimate the mixing matrix, we apply the K-means algorithm to the audio data in the transform domain obtained by the STFT, and details can be found in [3].

In the separating step, with the estimated mixing matrix $\hat{\mathbf{A}}$, we formulate the signal recovery problem as a compressed sensing [5] model. For $M = 2$ and $N = 4$, (1) can be expanded as:

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{pmatrix} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \mathbf{s}_3 \\ \mathbf{s}_4 \end{pmatrix} \quad (2)$$

where a_{ij} is the ij-th element of the mixing matrix $\mathbf{A}$. We can formulate the above equation as follows,

$$\underbrace{\begin{pmatrix} x_1(1) \\ \vdots \\ x_1(L) \\ x_2(1) \\ \vdots \\ x_2(L) \end{pmatrix}}_{\mathbf{b}} = \underbrace{\begin{pmatrix} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} & \Lambda_{14} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} & \Lambda_{24} \end{pmatrix}}_{\mathbf{M}} \underbrace{\begin{pmatrix} s_1(1) \\ \vdots \\ s_1(L) \\ \vdots \\ s_4(1) \\ \vdots \\ s_4(L) \end{pmatrix}}_{\mathbf{f}} \quad (3)$$

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

$$\mathbf{b} = \mathbf{M}\mathbf{f} \quad (4)$$

The above equation can be interpreted as a CS model in which $\mathbf{M}$ is the measurement matrix and $\mathbf{b}$ is the compressed vector of samples in $\mathbf{f}$. Therefore, a sparse representation in the transform domain can be employed for $\mathbf{f}$:

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

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

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

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 the measurement $\mathbf{b}$ using an optimization process. This indicates that source estimation in our underdetermined problem can be achieved by computing $\mathbf{y}$ in (6) using the CS based signal recovery algorithms. An approach for estimating $\mathbf{f}$ from $\mathbf{b} = \mathbf{M}\mathbf{f} = \mathbf{M}\mathbf{\Phi}\mathbf{y}$ is to solve the following l_0 minimization problem

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

where $\parallel \mathbf{y} \parallel_0$ is the l_0 norm measuring 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 [6] that the solution to the l_0 minimisation problem is essentially equivalent to the solution of the following l_1 minimisation problem

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

The basis pursuit (BP) algorithm [6] can be used to solve the problem, according to the following iterative process. First, we choose an initial basis matrix $\mathbf{B}$ which is a squared matrix having the same rank as $\mathbf{M}\mathbf{\Phi}$ and consists of the selected columns of $\mathbf{M}\mathbf{\Phi}$, 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, we reach the optimal solution. Finally, $\mathbf{y}$ can be readily computed by $\mathbf{B}^{-1}\mathbf{b}$. The vector $\mathbf{f}$ consisting of the separated sources $\mathbf{s}_j$ can be obtained by simply multiplying the dictionary $\mathbf{\Phi}$ with $\mathbf{y}$ using (5).

According to our listening tests, however, the separated sources $\mathbf{s}_j$ obtained from the above algorithm may still contain a certain amount of interference from other sources $\mathbf{s}_u$, where $u \neq j$. Recent studies in [7] show that binary time-frequency (T-F) masking can be used to further enhance the performance of BSS algorithms, due to its superior performance in interference rejection. Therefore, in this step, we further employ binary masks to improve the quality of the separated sources. First, the separated sources $\mathbf{s}_j$ are transformed to the T-F representations such as spectrograms. The binary masks are estimated by comparing the energy of the "target" source with that of the summation of other sources at each T-F unit. The element of the mask matrix $\mathbf{H}_j$ is assigned 1 if the energy of the target is stronger than that of the interference from all the other sources at that T-F unit, and 0 otherwise. Specifically, $H_j(m,k)$ is computed as follows:

$$H_j(m,k) = \begin{cases} 1 & \mid S_j(m,k) \mid \; > \; \frac{1}{3} \sum\limits_{u \neq j} \mid S_u(m,k) \mid, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

where $S_j(m,k)$ is the STFT of the jth source, m and k are the time frame index and frequency bin index respectively. Then we use the mask matrix $\mathbf{H}_j$ and the STFT of the mixtures to obtain the original sources $\mathbf{s}_j$ as,

$$\mathbf{s}_j = ISTFT\left(\mathbf{X}_i \odot \mathbf{H}_j\right) \tag{10}$$

where $\mathbf{X}_i$ is the the T-F representation of the ith mixture, $\odot$ denotes element-wise multiplication, and $ISTFT$ is the inverse STFT. For the two-mixture case, both $\mathbf{X}_1$ and $\mathbf{X}_2$ can be used in the above equation. In practice, we choose $\mathbf{X}_i$ based on a_{ij} which is already obtained from the clustering step. If $a_{1j} > a_{2j}$, we choose $\mathbf{X}_1$ for the above equation, and vice versa. The above process (including clustering, separating and masking steps) is repeated for each block of the mixture signals. The whole sources $\mathbf{s}_j$, $j = 1, ..., 4$ are reconstructed by concatenating together all the blocks of the estimated source components $\mathbf{s}_j^p$, $p = 1, ..., P$.

3. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed algorithm using simulations. The speech signals used in our experiments were taken from the database of the source separation evaluation campaign [4]. We used two groups of speech sources in evaluations, with one group containing four male speech signals and the other group four female speech signals. Each signal has a duration of 10s and is obtained in a meeting room using omnidirectional microphones with spacing 5cm [4], sampled at 16kHz. We evaluate the results using three objective criteria, i.e. the source image to spatial distortion ratio (ISR), the source to interference ratio (SIR), the source to artifacts ratio (SAR), which measure the relative amounts of the spatial distortion, the interference and the artifacts, defined in [4]. The total error was also measured by the signal to distortion ratio (SDR), defined as [4],

$$SDR_j = 10log_{10}\left(\frac{\sum\limits_{i=1}^{I}\sum\limits_{t=1}^{T}\mathbf{s}_{ij}^{img}(t)^2}{\sum\limits_{i=1}^{I}\sum\limits_{t=1}^{T}(\mathbf{s}_{ij}^{spat}(t)+\mathbf{s}_{ij}^{interf}(t)+\mathbf{s}_{ij}^{artif}(t))^2}\right) \tag{11}$$

where $s_{ij}^{img}(t)$ is the true source image of source j and $s_{ij}^{spat}(t)$, $s_{ij}^{interf}(t)$, and $s_{ij}^{artif}(t)$ are the distinct error components representing the spatial distortion, the interference and the artifacts respectively.

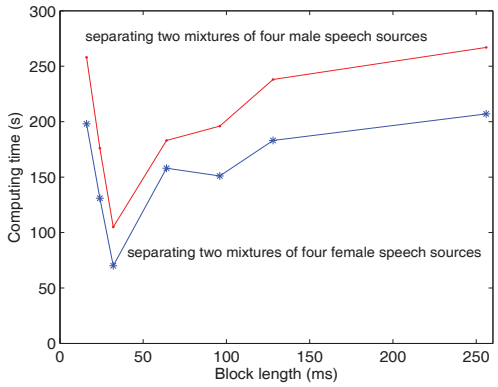


Fig. 3. The effect of different block length L on the computational efficiency of the proposed algorithm.

First, we perform an experiment to evaluate the effect of the block size L on the computational efficiency and separation performance of the proposed algorithm. In this experiment, we generated two mixture signals by mixing together

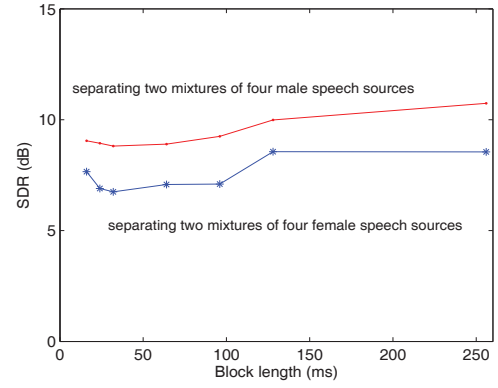


Fig. 4. The effect of different block length L on the separation performance measured by SDR.

four speech sources using the following mixing matrix.

$$\mathbf{A} = \begin{pmatrix} 0.3420 & 0.6428 & 0.7934 & 0.9239 \\ 0.9397 & 0.7660 & 0.6088 & 0.3827 \end{pmatrix} \quad (12)$$

For the case of $L = 512$ (i.e. 32ms), the average of $\hat{\mathbf{A}}$ obtained by the clustering algorithm for all blocks is shown in (13). We see that $\hat{\mathbf{A}}$ is reasonably close to $\mathbf{A}$ except the permutation ambiguity.

$$\hat{\mathbf{A}} = \begin{pmatrix} 0.6275 & 0.9184 & 0.3620 & 0.7896 \\ 0.7754 & 0.3850 & 0.9270 & 0.6102 \end{pmatrix} \quad (13)$$

The computing time required for running the proposed algorithm varies with L, as shown in Figure 3, where the results for both the male and female groups of speech signals are plotted. From this figure, it can be observed that the algorithm is most efficient when the block size is equal to 32ms. It only takes 70s and 105s for running the algorithm on both groups of speech signals. In constrast, the average computing time required by the algorithm for the same two groups of sources without blocking is 1208s and 1355s respectively. In this case, the block-based algorithm is approximately 12 times faster than the algorithm without blocking. The separation performance measured by the SDR is shown in Figure 4. This figure suggests that using different L, there are only slight changes in the separation performance.

We also created 20 pairs of mixtures by using the mixing matrices whose elements were drawn from uniformly distributed random variables. The proposed algorithm was applied to each pair of these mixtures. The averaged separation performance of these experiments is given in Table 1. In comparison with other thirteen algorithms involved in the campaign (see the results reported in [4]), the proposed algorithm provides competitive separation performance. To show this, we plot the results measured by the SDR in Figure 5 for each source tested in the experiments, where our proposed algorithm is compared with the two algorithms with best performance, i.e. algortithm 1 and 3 in the campaign [4]. The proposed algorithm offers a higher SDR in six out of the eight tested sources.

	SDR	ISR	SIR	SAR
Male speech 1	13.26	16.77	14.17	17.38
Male speech 2	5.89	7.82	6.08	8.01
Male speech 3	4.83	5.67	5.08	9.32
Male speech 4	9.87	17.36	12.67	13.49
Female speech 1	10.40	19.98	12.52	14.12
Female speech 2	4.42	6.86	5.58	7.71
Female speech 3	3.92	5.74	4.95	7.15
Female speech 4	8.64	18.62	10.40	12.79

Table 1. SDR, ISR, SIR and SAR (in dB) measured for each signal within the two groups of speech sources.

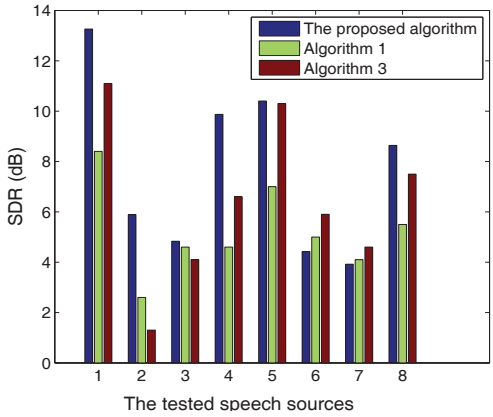


Fig. 5. SDR comparison between the proposed algorithm and two recent algorithms (algorithm 1 and 3 in the campaign [4]), with sources 1-4 corresponding to the four male speech signals 1-4 in Table 1, and 5-8 corresponding to the four female speech signals 1-4).

4. CONCLUSIONS AND FUTURE WORK

We have presented a multi-stage method for underdetermined blind speech separation using block-based compressed sensing incorporating binary mask. Numerical experiments have shown the improved separation performance and computational efficiency by the proposed method, as compared with recent underdetermined BSS approaches. Extension of the proposed approach to the separation of convolutive speech mixtures is 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] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Comput.*, vol. 13, pp. 863–882, 2001.

[3] T. Xu and W. Wang, "A compressed sensing approach for underdetermined blind audio source separation with sparse representation," in *Proc. IEEE Int. Workshop on Statistical Signal Processing*, 2009.

[4] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation*, 2007, pp. 552–559.

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

[6] 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.

[7] T. Jan, W. Wang, and D. L. Wang, "A multistage approach for blind separation of convolutive speech mixtures," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1713–1716.