

Single-channel blind source separation based on joint dictionary with common sub-dictionary

Linhui Sun^{1,2} · Cheng Zhao¹ · Min Su¹ · Fu Wang¹

Received: 29 June 2017 / Accepted: 7 October 2017 / Published online: 31 October 2017
© Springer Science+Business Media, LLC 2017

Abstract The cross projection engenders when mixed speech signal is represented over joint dictionary because of the bad distinguishing ability of joint dictionary in single-channel blind source separation (SBSS) using sparse representation theory, which leads to bad separation performance. A new algorithm of constructing joint dictionary with common sub-dictionary is put forward in this paper to this problem. The new dictionary can effectively avoid being projected over another sub-dictionary when a source signal is represented over joint dictionary. In the new algorithm, firstly we learn identify sub-dictionaries using source speech signals corresponding to each speaker. And then we discard similar atoms between two identity sub-dictionaries and construct a common sub-dictionary using these similar atoms. Finally, we combine those three sub-dictionaries together into a joint dictionary. The Euclidean distance among two atoms is used to measure the correlation of them in different identity sub-dictionaries, and similar atoms are searched based on the correlation. In testing stage, each source can be reconstructed with the projection coefficients corresponding to individual sub-dictionary and the common sub-dictionary. Contrast experiments tested in speech database show that the algorithm proposed in this paper performs better, when the Signal-to-Noise Ratio (SNR) is used to measure separation effect. The algorithm set out in this paper has lower time complexity as well.

Keywords Sparse representation · Single-channel blind source separation · Common sub-dictionary · Similar atoms

1 Introduction

Single-channel blind source separation (SBSS) is a process of recovering source signals by only one mixed signal (Roweis 2000), which is an extreme case of underdetermined blind source separation. It is an ill-posed problem in mathematics, being of great theoretical value. In the field of speech signal processing, a robust speech separation is often required at the pre-processing stage prior to the target applications, such as simultaneous interpretation, automatic speech recognition (Xu et al. 2015) and speech coding (Zhen et al. 2016). SBSS has been being widely used in wireless communications, meteorology and Biomedicine (Shapoori et al. 2015). Not only because of the important theoretical but for practical value, SBSS has been becoming a hot point of research in both the academic circles and engineering fields.

It is one of the popular methods to solve SBSS based on sparse representation theory. Many creative researches have been produced in the last few decades. A detailed method in an earlier study to solve SBSS based on sparse representation theory can be found in Bofill and Zibulevsky (2001). At the same time, the way to construct sparse dictionary became the focus of research. In Michal and Elad (2006), a method to design an overcomplete dictionary using K-SVD algorithm was proposed. The key of sparse representation for SBSS is to construct a joint dictionary. More specifically, the large space in which the mixed signal is expressed, is consisted of multiple subspaces, which can represent the signal of each source as complete as possible. The source speech signal can be reconstructed by sparse coefficients and

✉ Linhui Sun
sunlh@njupt.edu.cn

¹ College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China

² Key Lab of Broadband Wireless Communication and Sensor Network Technology, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing, China

the basis corresponding to the subspace (Tan and Liu 2007; Yu et al. 2013).

Generally speaking, joint dictionary is a redundant dictionary. One source signal responds to the identify sub-dictionaries corresponding to additional sources can't be avoided, even for sparse constraints being applied to train dictionaries. We call this phenomenon cross projection (Agrawal et al. 2006), which leads to a decrease in accuracy and efficiency. So it's an urgent demand in SBSS to find a method with high precision, efficiency and robustness. Some methods to learn discriminative dictionary were mentioned in the past time, such as Metaface learning method (Yang et al. 2010). In Bao et al. (2014), Grais and Erdogan (2013) and Lian et al. (2015), a series of methods, such as making joint dictionary discriminative by modifying objective function or adding penalty item, were put forward when learning dictionary. These methods perform perfectly. However, because the objective function is too complex, solving the optimization problem becomes difficult, and time complexity becomes higher. And then the idea that learning dictionary using the differences between each source was proposed in Shapoori et al. (2015), Tang et al. (2015), Rambhatla and Haupt (2014). However, quite a lot of priori information must be known when using these methods. Adding a common sub-dictionary into joint dictionary was presented in Tang et al. (2016). In Tian et al. (2017), the common sub-dictionary to express the common components among each source signals is learned from mixed signal.

In consideration of the advantages and disadvantages for algorithms mentioned above, we come up with a new method to construct joint dictionary with common sub-dictionary by searching similar atoms in all identify sub-dictionaries in this paper. The paper is structured as follows: Sect. 2 illustrates the model of SBSS based on sparse representation. Section 3 presents a new method to construct joint dictionary with a common sub-dictionary, while Sect. 4 gives an account of the experimental processes, simulation results and some crucial factors for algorithm performance. Conclusions and future work are provided in Sect. 5.

2 SBSS based on sparse representation

In this section, we firstly introduce the formulation of SBSS, and then propose the method to deal with this problem by sparse representation theory. At the end of this part, we analyze the cause of cross projection and give the preliminary solution for this problem.

2.1 Problem formulation and notations

The blind source separation (BSS) solves the problem: recovering N underlying speech signals $s_i, i = 1, 2, \dots, N$

from mixed speech signals \mathbf{x} (Yu et al. 2013), which can be called observation signals as well. If matrix being introduced, SBSS can be formulated as (1), where $\mathbf{x} \in R^{m \times N}$ is the matrix of mixed speech signals, $\mathbf{S} \in R^{n \times N}$ is the source speech signals matrix, and \mathbf{A} is the $m \times n$ mixing matrix. The aim of BSS is to reconstruct both \mathbf{A} and \mathbf{s} from \mathbf{x} .

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

Single-channel means the number of mixed signal is one. On this case, mixed speech signal $\mathbf{x}(t)$ can be defined as (2).

$$\mathbf{x}(t) = \sum_i a_i s_i(t), 1 \leq t \leq T \tag{2}$$

For convenience of description, mixed signal coming from only two speech sources and being mingled equally-weighted is considered in this paper. SBSS can be formulated as (3) in this case.

$$\mathbf{x} = \mathbf{A}\mathbf{s} = [1 \ 1]^* \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \mathbf{x} = s_1 + s_2 \tag{3}$$

It must be emphasized that the approach presented in this paper can be easily extended to general cases that the number of source speech signals is more than two.

2.2 The solution of SBSS based on sparse representation

The solution of SBSS based on sparse representation contains two stages, training and testing (Michal and Elad 2006). During the training process, a large number of speaker's clean speech signals are used as training sets, and adaptive learning method, such as the K-SVD algorithm, is used to train dictionary. For clarity of statement, we define \mathbf{D}_i as the identity sub-dictionary of the i th speaker, which contains almost all information about this speaker. The way of training identity sub-dictionary can be viewed in Fig. 1. After getting two identity sub-dictionaries, the joint dictionary \mathbf{D} can be formulated as (4).

$$\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2] \tag{4}$$

In the testing phase, every frame of mixed speech signal is sparsely represented over the joint dictionary. Mixed speech signal frame \mathbf{x} and joint dictionary \mathbf{D} are already known here. We can write \mathbf{E} , as $\mathbf{E} = [\mathbf{E}_1 \ \mathbf{E}_2]^T$, where \mathbf{E}_1 and \mathbf{E}_2 are the sparse coefficients vector of training signal

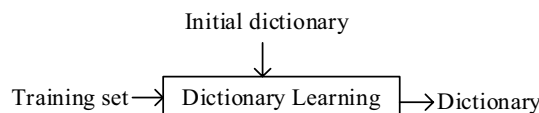


Fig. 1 Adaptive dictionary learning method

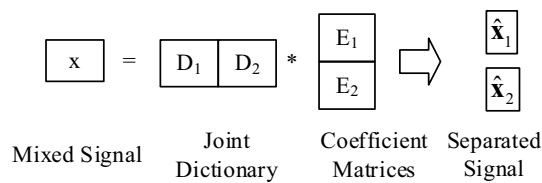


Fig. 2 Detail of the source separation process

frame x over identity sub-dictionary D_1 and D_2 respectively. The process can be formulated as (5).

$$x = D \times E = [D_1, D_2] \times \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} \tag{5}$$

We can get E by many ways such as Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP) and Basis Pursuit (BP) (Yang et al. 2013). We define estimated signals as \hat{x}_1 and \hat{x}_2 , which we can also call reconstruction signal, and then overall SBSS schemes can be shown in Fig. 2.

2.3 Cross projection in SBSS based on sparse representation

The main idea of SBSS based on sparse representation is that signal responds only to corresponding identity sub-dictionary. However, any two source signals always have their own distinctive characteristics as well as some similar components to some extent in fact. When the two source signals have the same type, the same ingredients may become more. The source signals in this paper are all speech signals. The result of this factor is that mixed signal produce larger projections not only on the corresponding identity sub-dictionary, but also on another sub-dictionary. This factor fundamentally leads to great complexity and difficulty of SBSS.

In order to analyze cross projection clearly, we calculate the projection coefficients of one male frame on joint dictionary $D = [D_1, D_2]$, where D_1, D_2 are identity sub-dictionaries corresponding to male and female speakers respectively, whose size is set to be 128×512 . So the size of joint dictionary is 1024. Figure 3a is the waveform of one male frame in time domain. Figure 3b is the projection coefficients on joint dictionary of this frame. The horizontal axis means the atomic number of D and the vertical axis indicates the value of projection coefficients. The part of the transverse axis greater than 512 shows the projection coefficients of this male frame on identity sub-dictionary D_2 in Fig. 3b. We can see from the picture that male signal not only produce projections on its own identity sub-dictionary but also on female identity sub-dictionary.

We can conclude from experiment and analysis above that there exist some analogous ingredients between two source speech signals, leading to bad differentiation of two

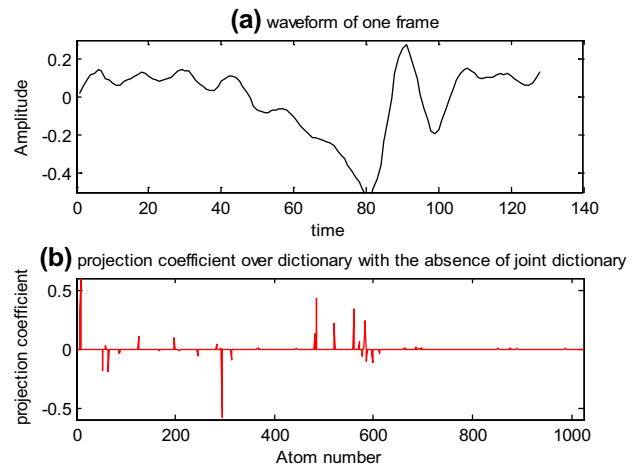


Fig. 3 Projection coefficients on joint dictionary of one male frame

identity sub-dictionaries. When mixed signal is represented over joint dictionary, signal produces large projections not only on the corresponding identity sub-dictionary, but also on another identity sub-dictionary. That is to say, because of the existence of some similar components in source speech signals, the identify sub-dictionary has no distinctive ability learning from source speech signals, which bring about bad performance of separation effect in SBSS.

One intuitive idea to cope with this issue is that we make dictionary being discriminative by modifying objective function in the process of training identity sub-dictionaries Bao et al. (2014). But an obvious disadvantage of this approach is that constructing objective function become too difficult and solving this optimization problem is a challenging work. Another thought is that we delete similar components from source speech signals and then training identity dictionary by those treated source speech signals. However, it's difficult to find close components between two source speech signals. In view of this factor, an alternative approach is proposed in Tian et al. (2017). In this algorithm, a common sub-dictionary, which is learned from mixed signals, is constructed at first, being used to characterize similar ingredients of source speech signals at first. And then the joint dictionary can be grouped by the two identity sub-dictionary and this common sub-dictionary. After getting the sparse representation of mixed signal on this joint dictionary at last, each source can be reconstructed by using the response corresponding to an identity dictionary and a relatively small percentage of the response on the common sub-dictionary. This method can overcome cross projection effectively. However, some disadvantages still exist. It takes too much time in the process of training common sub-dictionary and separating mixed speech signal for the large size of joint dictionary. Moreover, the same components in identity sub-dictionaries corresponding to each speaker still exist, and only a small

weight is added to the sparse coefficients corresponding to these same components.

We know that the identify sub-dictionary learning from source speech signals has no distinctive ability, because of the existence of some similar components in source speech signals. If these similar components can be deleted, the shortcoming of cross projection can be overcome despite the deficiencies still exist. We guess the close atoms, which we can call close atoms as well, in identify dictionaries are equal to the similar ingredient of source speech signals. We attempt to discard the similar atoms in dictionaries corresponding to two identity sub-dictionaries. However, if we discard a part of atoms, restructured signals will be incomplete, and the fidelity can't be ensured. Supposing that we can establish a new dictionary that is constructed by the combination of those discard atoms. We name the new dictionary as common sub-dictionary. If joint dictionary can be formed by these three sub-dictionaries, we can obtain balance between distinction and fidelity. A new way of constructing a joint dictionary with a common sub-dictionary based on this idea will be brought forward in this paper.

3 SBSS based on joint dictionary with common sub-dictionary

In this section, we firstly introduce the way to construct joint dictionary with common sub-dictionary, including arithmetic statement and parameter selection. And then the procedure to reconstruct source speech signal will be mentioned, which contains parameters selection as well. Overall algorithms are described finally.

3.1 Constructing a joint dictionary with common sub-dictionary

According to the description above, we can construct the new joint dictionary by following way. Search some close atoms between two identity sub-dictionaries firstly. Secondly, fill common sub-dictionary with the linear combination of this a pair of atoms and then discard the pair of atoms from two identity sub-dictionaries respectively. Finally, combine the three sub-dictionaries into a new joint dictionary, including two updated identity sub-dictionaries and a common sub-dictionary.

In order to state the new algorithm unambiguously, we declare some notations firstly. We still use $\mathbf{D}_1, \mathbf{D}_2$ to express identity sub-dictionaries and employ $\mathbf{D}_1^i, \mathbf{D}_2^j$ to indicate i th atom of \mathbf{D}_1 and j th atom of \mathbf{D}_2 respectively. So the Euclidean distance of the two atoms can be formulated as (6).

$$d_{ij} = \|\mathbf{D}_1^i - \mathbf{D}_2^j\|^2 \tag{6}$$

We have reasons to think that the i th atom of \mathbf{D}_1 is similar to the j th atom of \mathbf{D}_2 when $d_{ij} \leq \delta$, where δ is a relatively small constant. We execute the following steps for the sake of finding the most similar atoms from \mathbf{D}_2 for all atoms in \mathbf{D}_1 . Firstly, for the i th atom of \mathbf{D}_1 , we calculate the distance between it and all atoms of \mathbf{D}_2 and put them in a group. And then we find the minimum in this group, noting the minimum as t . If $t \leq \delta$, we believe that the most similar atom to the i th atom of \mathbf{D}_1 has been found. In next step, we set this pair of similar atoms to zero vectors and fill in \mathbf{D}_c with a half of their sum. When traversal of \mathbf{D}_1 is completed, we delete all zero vector in \mathbf{D}_1 and \mathbf{D}_2 , then mark them as $\mathbf{D}'_1, \mathbf{D}'_2$, calling them individual sub-dictionary. The new joint dictionary \mathbf{D}' can be represented as $\mathbf{D}' = [\mathbf{D}'_1, \mathbf{D}'_2, \mathbf{D}_c]$ under these circumstances. If we mark the number of rows of identify sub-dictionary as N , the detailed procedure to construct new joint dictionary can be shown as Table 1.

We define a distance matrix $T = (d_{ij})$ here in order to save all Euclidean distances between any two atoms in two identity sub-dictionaries. From the definition of T , we know that the T_{ij} is the distance between the i th atom of \mathbf{D}_1 and the j th atom of \mathbf{D}_2 . So the size of T is $N \times N$, where N is the number of atoms in sub-identify dictionary. Because there are the large number of elements in matrix T , we are obliged to calculation the statistical distribution of T . The distribution of values of Euclidean distance between any two atoms is shown in Fig. 4. The

Table 1 The procedure to construct joint dictionary with common sub-dictionary

y
Task: Construct joint dictionary with common sub-dictionary
Input: identify sub-dictionaries $\mathbf{D}_1, \mathbf{D}_2$, threshold δ
Output: Joint dictionary \mathbf{D}
for $i = 1 : N$
calculate the distance between \mathbf{D}_1^i and each atom of \mathbf{D}_2 , save them in the array $\mathbf{Dis}(i :)$, i is the serial number of \mathbf{D}_1
find the smallest distance, use function in Matlab $[x, y] = \min(\mathbf{Dis}(i :))$, x is the minimum number, y is the position of the element.
if $(x < \delta)$
put $0.5 * (\mathbf{D}_1(:, i) + \mathbf{D}_2(:, y))$ as a new atom into \mathbf{D}_c , set the $\mathbf{D}_1(:, i)$ $\mathbf{D}_2(:, y)$ to zero vector
End if
End for
for $j = 1 : N$
if (\mathbf{D}'_1 isn't zero vector) put \mathbf{D}'_1 as a new atom into \mathbf{D}'_1
if (\mathbf{D}'_2 isn't zero vector) put \mathbf{D}'_2 as a new atom into \mathbf{D}'_2
End if
End for
$\mathbf{D}' = [\mathbf{D}'_1, \mathbf{D}'_2, \mathbf{D}_c]$

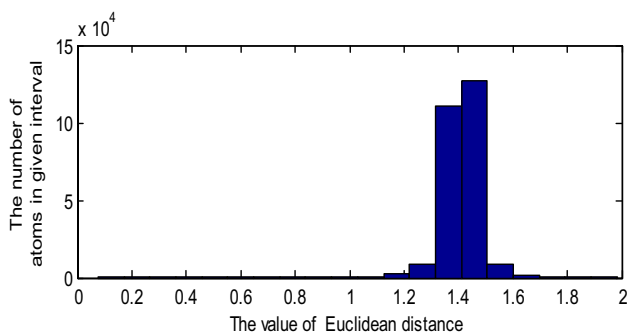


Fig. 4 The distribution of Euclidean distance between any two atoms

horizontal axis of this picture means the value of Euclidean distance and the vertical axis indicates the number of atoms in given distance interval.

As is pictured in Fig. 4, the majority values of Euclidean distance between any two atoms range from 1.3 to 1.5. Only a few of them are smaller than 1.3 or bigger than 1.5. Approximate range of δ can be found from this chart. In order to get accurate value of threshold, we calculate the SNR when δ changes. Experimental results are shown in Table 2. It should be pointed that the results are the average of multiple experiments.

To find out the relationship between threshold and SNR clearly, we draw Fig. 5 using the data from Table 1. We can see from the curves that SNR is the smallest when $\delta=0$ and SNR rises obviously with the increasing of δ . When the value of δ falls between 0.8 and 1.2, SNR is quite high and stay steady. When $\delta > 1.2$, SNR declines sharply with the increasing of δ . What’s more, the same distribution of SNR is observed using the female signal. Average SNR is the average of SNR-male and SNR-female. From experimentations above, we can conclude that when $0.8 \leq \delta \leq 1.2$, the size of \mathbf{D}_c is reasonable, the separation effect remain idea.

When the optional threshold is obtained, we can construct joint dictionary using the algorithm stated in Table 1, and then mixed signal can be sparsely represented on the joint dictionary. At this moment, BP algorithm can be utilized to get sparse coefficients. Finally we can get estimation of each source speech signal and use SNR to measure separation effect, as is mentioned in Vincent et al. (2006).

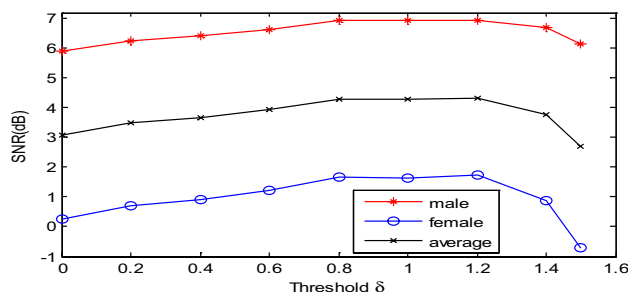


Fig. 5 The relationship curves between the threshold and SNR

3.2 Reconstructing source speech

As is shown in (5), the way to solve SBSS can be translated into solve underdetermined equation $\mathbf{D} \times \mathbf{E} = \mathbf{x}$, where \mathbf{D} is joint dictionary and \mathbf{x} is mixed signal. When using the joint dictionary with common sub-dictionary, the equation can be modified to $\mathbf{D}' \times \mathbf{E}' = \mathbf{x}$. In this equation, $\mathbf{D}' = [\mathbf{D}'_1, \mathbf{D}'_2, \mathbf{D}_c]$, where $\mathbf{D}'_1, \mathbf{D}'_2$ and \mathbf{D}_c are two individual sub-dictionaries and common sub-dictionary respectively. \mathbf{x} is mixed signal which is algebraic sum of two source speech signal. $\mathbf{E}' = [\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_c]$, where $\mathbf{E}_1, \mathbf{E}_2$ and \mathbf{E}_c are sparse coefficients of \mathbf{x} over $\mathbf{D}'_1, \mathbf{D}'_2$ and \mathbf{D}_c respectively. In mathematics, we can get \mathbf{E}' by translating the equation into an optimization problem as is shown in (7).

$$\begin{aligned} \min_E \|\mathbf{x} - \mathbf{D}\mathbf{E}'\|_2^2 \\ \text{s.t. } \|\mathbf{E}'\|_0 \leq K \end{aligned} \tag{7}$$

BP algorithm is utilized to solve this optimization problem in this paper. When getting \mathbf{E}' , estimated source speech signal can be calculated by (8).

$$\left. \begin{aligned} \hat{\mathbf{x}}_1 &= \mathbf{D}_1 \mathbf{E}'_1 + \alpha \mathbf{D}_c \mathbf{E}'_c \\ \hat{\mathbf{x}}_2 &= \mathbf{D}_2 \mathbf{E}'_2 + \beta \mathbf{D}_c \mathbf{E}'_c \end{aligned} \right\} \tag{8}$$

We know that the weight coefficients α and β in (8) are of great influence on the value of reconstructed signals, resulting in different SNRs between of the estimated signals. In order to get appropriate weight coefficients, we investigate the impact of coefficients α and β on system performance. Figure 6 shows the curve of male-SNR changing with the weight α . Some conclusion can be brought that as the weight

Table 2 SNR when δ changes (dB)

Threshold δ	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.5
SNR-male	5.9214	6.2644	6.4152	6.6418	6.9390	6.8480	6.2424	6.6852	6.1321
SNR-female	0.2508	0.7024	0.9237	1.2363	1.6682	1.5388	1.1281	0.8820	-0.7126
Average SNR	3.0861	3.4834	3.6844	3.9390	4.3036	4.1934	3.6853	3.7836	2.7098

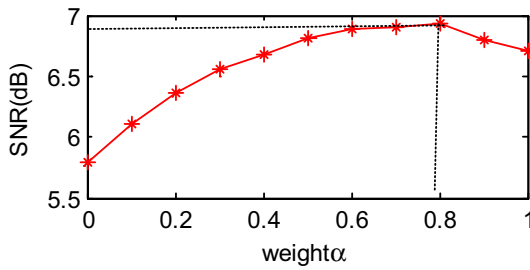


Fig. 6 Weight α for male-SNR

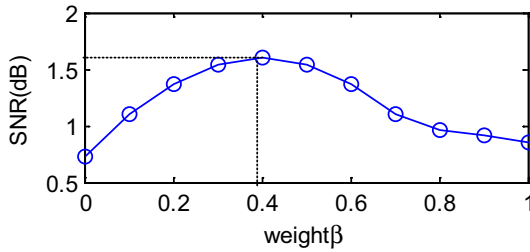


Fig. 7 Weight β for female-SNR

Table 3 Constructing joint dictionary with common sub-dictionary

Task: Learning identify sub-dictionary, using them to construct common sub-dictionary and joint dictionary
Input: Clean speech signals of corresponding speaker, initial dictionary, the number of iterations K and threshold δ
Output: Joint dictionary \mathbf{D}'

Step 1: Learn the dictionary $\mathbf{D}_1, \mathbf{D}_2$ by K-SVD algorithm using clean speech signals
Step 2: Construct common sub-dictionary \mathbf{D}_c and individual sub-dictionaries $\mathbf{D}'_1, \mathbf{D}'_2$ by algorithm shown in Table 1.
Step 3: Set joint dictionary $\mathbf{D}' = [\mathbf{D}'_1, \mathbf{D}'_2, \mathbf{D}_c]$

Table 4 Separating mixed signal using BP algorithm based joint dictionary

Task: Separate mixed signal using BP algorithm based on joint dictionary.
Input: Clear speech signals, mixed speech signal \mathbf{x} , weight coefficients α and β , joint dictionary \mathbf{D}'
Output: SNR between source speech signals $\mathbf{x}_1, \mathbf{x}_2$ and the reconstructed speech signals $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2$

Step 1: Use BP algorithm to sparsely code every frame of \mathbf{x} over \mathbf{D}' and obtain the separated speech signal of the current frame by (8)
Step 2: Synthesize all the frames into complete speech signals and calculate SNR

α increasing, separation effect of male improves firstly, and then starts to drop when the weight of α is too large. The best performance is achieved when α is 0.8. The kinked line for female-SNR is shown in Fig. 7. The trend of female-SNR is the same as male-SNR, rising firstly and then falling. When β is set to 0.4, the best performance is obtained.

3.3 The overall algorithm

The entire experiment contains two stages, training and testing. The purpose of training stage is to learn identify sub-dictionary and construct joint dictionary with common sub-dictionary. The aim of testing stage is to solve sparse coefficients of

mixed speech signal over joint dictionary. The detailed steps of the two stages are shown in Tables 3 and 4 respectively.

4 Simulation experiment and results analysis

Simulation experiment and result analysis are described in this section. Experimental environment is introduced at first. And then, the validity of our algorithm is confirmed. What's more, the comparisons of performance between our proposed algorithms and some others are presented. Then, algorithm complexity is analyzed. At last, the effects of some important factors on algorithm performance are considered.

To ensure the persuasiveness of experiment, all speech signals used in our paper come from Chinese Speech library constructed by the Institute of Automation, Chinese Academy of Sciences (CASIA). We choose one female and one male speaker and there are 265 sentences for each speaker. Moreover, we choose 200 sentences of each speaker as training set and the 65 sentences as testing set for each speaker. K-SVD algorithm is used to learn identify sub-dictionary, in which the initial dictionary is Discrete Cosine Transform (DCT) dictionary and the number of iterations is 80. Simulation experiments

are carried out on Matlab 2013. The performance of the separation is measured by Signal-to-Noise Ratio (SNR), mentioned in Vincent et al. (2006), which can be formulated by (9).

$$SNR = 10 \log_{10} \left(\frac{\|\mathbf{x}\|_2^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2} \right) \tag{9}$$

4.1 Validity verification

As is described in Sect. 2, we know that cross projection can be avoided effectively by making use of common sub-dictionary. For strengthening the persuasiveness, the

distribution of projection coefficients of one male frame signal over joint dictionary constructed by algorithm proposed in this paper is shown. In contrast, the distribution of projection coefficients of one male frame signal over joint dictionary combining with the two identify sub-dictionaries is demonstrated too. More specifically, Fig. 8a is the waveform of one male frame in time domain. Figure 8b is the projection coefficients of this frame on joint dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2]$, where $\mathbf{D}_1, \mathbf{D}_2$ are identity sub-dictionaries corresponding to male and female speakers respectively. The size of identity sub-dictionary is set to be 128×512 , so the size of joint dictionary is 1024. The part of the transverse axis greater than 512 shows the projection coefficients of this male frame on identity sub-dictionary \mathbf{D}_2 . We can see from the picture that male signal not only produce projections on its own identity sub-dictionary but also on female identity sub-dictionary. Figure 8c is the projection coefficients of this frame on joint dictionary $\mathbf{D}' = [\mathbf{D}'_1, \mathbf{D}'_2, \mathbf{D}_c]$, where $\mathbf{D}'_1, \mathbf{D}'_2, \mathbf{D}_c$ are two individual sub-dictionaries and common sub-dictionary respectively. The common sub-dictionary consists of 116 atoms, so the number of atoms in \mathbf{D}'_1 or \mathbf{D}'_2 is 396 and the size of \mathbf{D}' is 908. We can see from the picture that male signal not only produce projection on its own individual sub-dictionaries \mathbf{D}'_1 but also on common sub-dictionary \mathbf{D}_c .

Comparing Fig. 8b, with Fig. 8c, we can perorate that cross projection disappear when using joint dictionary with common sub-dictionary by the algorithm presented in this paper. How about separation effect of a whole speech signal? In order to validate the algorithm in this paper convectively. A series of schematics of signal synthesis and decomposition when the size of the initial dictionary is 128×512 ,

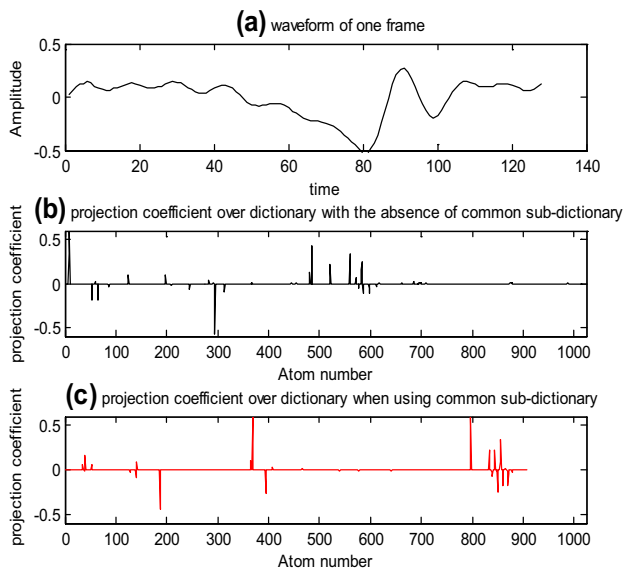


Fig. 8 Waveform in time domain and projection coefficients over dictionary

threshold δ is 0.8, α is 0.8 and β is 0.4 are shown in Fig. 9. It is worth noting that the test signals are chosen stochastically.

Contrasting Fig. 8a, with Fig. 8d as well as Fig. 8b with Fig. 8e, we can find that the source speech signal has similar outline to estimated speech signal, proving that using algorithm proposed in this paper can realize SBSS effectively.

4.2 Contrast experiments and complexity analysis

Some contrast experiments of the proposed algorithm in the paper and other algorithms will be shown. One of the schemes is to construct joint dictionary by combining the two identify sub-dictionary together and it is mentioned in Yu et al. (2011). Another plan mentioned in Tian et al. (2017) is to construct joint dictionary by combining all identify sub-dictionaries with common sub-dictionary learning from mixed signal. All algorithms mentioned above are tested in the same speech database. The separation results measured by SNR of estimated speech signal for these methods are vividly presented in Table 5 and time consumption will be analyzed.

It can be seen from Table 5 that our approach can obtain a better result. Specifically speaking, our proposed method improves SNR by more than 1 dB for male speech signal, about 1.5 dB for female speech signal, compared to method in Yu et al. (2011). Our method improves SNR by about 0.9 dB for male speech signal, 1.5 dB for female speech signal, compared to method in Tian et al. (2017). In addition, because of large amplitude for male speech signal, the

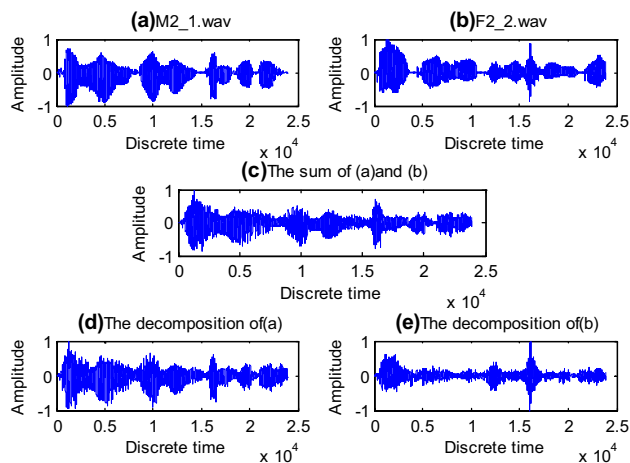


Fig. 9 Schematics of signal synthesis and decomposition

Table 5 Performance comparison between different methods (dB)

SNR	Male	Female	Average
Yu et al. (2011)	5.9214	0.2508	3.0861
Tian et al. (2017)	6.5275	0.8509	3.6892
Proposed algorithm	7.03	1.7436	4.3868

separation effect of male is better than female. We can conclude that when one of the source speech signals is stronger than another signal, the weaker signal is hard to be separated, which results in a poor average result.

On the other hand, what's the difference of time consumption between these schemes at the same time? We know that the whole algorithm contains two stages, training and testing. Qualitative analysis of time complexity is given from detailed steps to construct joint dictionary. In training phase, we use K-SVD algorithm mainly, which contains OMP arithmetic. Main time consumption is depended on the size of the identify sub-dictionary, iteration times and complexity of objective function. Let's suppose that the iteration times are equal and initial dictionary are same for these ways. Two sub-dictionaries must be learned in Yu et al. (2011), and three sub-dictionaries in Tian et al. (2017). Two sub-dictionaries must be learned and cyclic traversal is executed of method proposed in this paper. The order of time consumption in training phase from less to more is Yu et al. (2011), our method, Tian et al. (2017). In testing phase, the BP algorithm is used, so the size of joint dictionary for those methods being of great effect on time complexity. The atom number of identify sub-dictionary is defined as N here. We know that atom number is $2N$ in Yu et al. (2011), less $2N$ of our method, $3N$ in Tian et al. (2017). The order of time consumption in testing stage from less to more is our method, Yu et al. (2011), Tian et al. (2017).

However, it is worth pointing out that the training phase is completed offline for practical application, the timeliness requirement is not high relatively. Our algorithm has advantages in separation performance and time complexity, compare with two other methods.

4.3 Effects of some factors

There are many factors influencing the performance of our proposed algorithms, such as the number of atoms in common sub-dictionary, the size of initial dictionary, the number of training samples, the degree of convergence and iteration number of K-SVD algorithm. In this part, we will discuss the effects of some primary factors.

The relationship between SNR and threshold has been discussed and the best threshold has been found in Sect. 3. On the one hand, we know that the size of common sub-dictionary can be controlled by adjusting threshold. It's the scale of common sub-dictionary that affects separation effect and it's only a means to control the number of similar atoms by adjusting threshold. How many atoms are thought to be similar exactly is the key to this problem? On the other hand, when the size of initial dictionary varies, the identify sub-dictionary also changes. How many atoms should be put into common sub-dictionary at the moment and what's the impact of initial dictionary size on separation effect?

We define some notations firstly. With the proposed algorithm being used, the joint dictionary is formulated as $\mathbf{D}' = [\mathbf{D}'_1, \mathbf{D}'_2, \mathbf{D}'_c]$ and the size of it can be set as m . When not using common sub-dictionary, the joint dictionary \mathbf{D} is equal to $[\mathbf{D}_1, \mathbf{D}_2]$, whose size is the sum of \mathbf{D}_1 and \mathbf{D}_2 . We noted the size of \mathbf{D} as n . We define $\eta = \frac{m}{n}$, which we called compression ratio, for measuring the size change of joint dictionary when using the method proposed in this paper. Suppose the number of identify sub-dictionary and common sub-dictionary are a and b respectively. So $m = 2(a - b) + b = 2a - b$, $n = 2a$, and then η can be described as (10).

$$\eta = \frac{2a - b}{2a} \quad (10)$$

when $b = 0$, $\eta = 1$, the column number of common dictionary is zero, that is, the common dictionary is out of usage. When $\eta = 0.5$, the maximum number of columns for common sub-dictionary can be equal to the identify sub-dictionary \mathbf{D}_1 or \mathbf{D}_2 . Thus, the range of η is between 0.5 and 1. In order to verify the effects of compression ratio η on SNR, some experiments will be done follow the ideas below.

We choose the dictionary size to be 128×256 , 128×512 , 128×768 , 128×896 respectively. Under this condition, we adjust the threshold constantly, and record the compression ratio and SNR in each test. The trend that separation effect SNR varies with compression ratio η when the size of initial dictionary changes is shown in Fig. 10.

Figure 10 can be analyzed from two perspectives. On the one hand, we can observe each individual curve. When η approaches to 0.5, that is, when almost all atoms are thought to be similar, the SNR is low. SNR increases slowly with η growing until $\eta \leq 0.75$. SNR decreases sharply with η increasing if $\eta \geq 0.85$, only seldom atoms being regarded as similar atoms at this time. SNR remains stable when η lies between 0.75 and 0.85. We can conclude from the curves

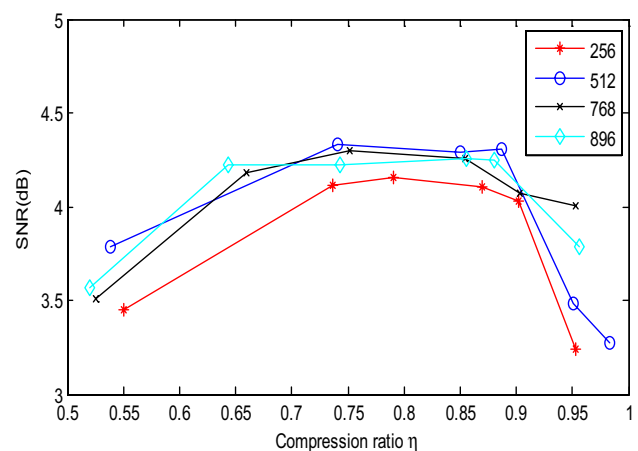


Fig. 10 The trend that separation effect SNR vary with compression ratio η

that separation effect is closely related to the size of common sub-dictionary. On the other hand, some conclusion can be made by comparing the four curves. The performance is relatively poor, when the initial dictionary is 128×512 . When the column of initial dictionary is 512,768 and 896, separation effect isn't quite different. That is to say, the algorithm can perform well regardless of the size of initial dictionary, as long as we can find those similar atoms. However, as is described in Sect. 4.2, time consumption rises fast with the increasing of the size of initial dictionary.

5 Conclusion

In order to dispose of cross projection when mixed speech signal is represented over joint dictionary in SBSS, a construction method of common sub-dictionary is put forward. We come up with a novel algorithm of constructing joint dictionary with common sub-dictionary in this paper. We discard similar atoms in identify sub-dictionary and combine these similar atoms into common sub-dictionary at first. And then joint dictionary can be structured by those sub-dictionaries. The correlation between each atoms of each identify sub-dictionary is utilized to search similar atoms. The relevant experimental datum show that comparing to conventional methods, algorithm proposed in this paper performs more outstanding, when it comes to the ability of discrimination and fidelity as well as time complexity. These factors, including the size of identify sub-dictionary, the atom numbers of common sub-dictionary are of great impact on the performance of our algorithm.

The performance of the proposed algorithm has been analyzed in the presence of only two sources of speech signals. However, this method can be extended to the case when the number of source speech signal is bigger than two from the principle of algorithm described in this paper. We discuss the case that all the identify sub-dictionaries have the same size, so the case that each identify sub-dictionary has different size is a future research direction. What's more, because of the difference between male and female voice signals, the separation effect of female is not ideal. How to design algorithm separating all signals better is an issue worthy of study.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No.61501251), the Natural Science Foundation of Jiangsu Province (BK20140891) and the Scientific Research Foundation of Nanjing University of Posts and Telecommunications (NY214038).

References

- Agrawal, A., Raskar, R., & Chellappa, R. (2006). Edge suppression by gradient field transformation using cross-projection tensors computer vision and pattern recognition, 2006 IEEE Computer Society Conference on. IEEE, 2301–2308.
- Bao, G., Xu, Y., & Ye, Z. (2014). Learning a discriminative dictionary for single-channel speech separation. *IEEE/ACM Transactions on Audio Speech & Language Processing*, 22(7), pp. 1130–1138.
- Bofill, P., & Zibulevsky, M. (2001). Underdetermined blind source separation using sparse representations. *Signal Processing*, 81(11), 2353–2362.
- Grais, E., Erdogan, H. (2013). *Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation* (pp. 808–812). France: INTERSPEECH, Lyon.
- Lian, Q., Shi, G., & Chen, S. (2015). Research progress of dictionary learning model, algorithm and its application. *Journal of Automation*, 41(2), 240–260.
- Michal, A., Elad, M. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11), 4311–4322.
- Rambhatla, S., Haupt, J. (2014). Semi-blind source separation via sparse representations and online dictionary learning. In *2013 Asilomar conference on signals, systems and computers* (pp. 1687–1691). IEEE.
- Roweis, S. (2000). One microphone source separation, NIPS, pp. 793–799.
- Shapoori, S., Sanei, S., & Wang, W. (2015). Blind source separation of medial temporal discharges via partial dictionary learning. In *2015 IEEE 25th International workshop on machine learning for signal processing (MLSP)*, Boston, MA, pp. 1–5.
- Tan, H., & Liu, H. L. (2007). On recoverability of blind source separation based on sparse representation. *Journal of Guangdong University of Technology*, 2008(02), 44–46.
- Tang, S., Guo, H., Zhou, N., Huang, L., & Zhan, T. (2016). Coupled dictionary learning on common feature space for medical image super resolution, 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, pp. 574–578.
- Tang, Y., Chen, Y., & Xu, N., et al. (2015). Speech reconstruction via sparse representation using harmonic regularization. IEEE: International Conference on Wireless Communications and Signal Processing, pp. 1–4.
- Tian, Y., Wang, X., & Zhou, Y. (2017). A new algorithm for single channel blind source separation based on sparse representation. *Journal of Electronics and Information*, 39(6), 1371–1378.
- Vincent, E., Gibbonval, R., & Fevotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio Speech and Language Processing*, 14(4), 1462–1469.
- Xu, L., Yang, Z., & Shao, X. (2015). Dictionary design in subspace model for speaker identification. *International Journal of Speech Technology*, 18(2), 177–186.
- Yang, M., Zhang, L., Yang, J., & Zhang, D. (2010). Metaface learning for sparse representation based face recognition. IEEE International Conference on Image Processing, 1601–1604.
- Yang, Z., Yang, Z., & Sun, L. (2013). A review of orthogonal matching pursuit algorithms for signal compression reconstruction. *Signal Processing*, 29(4), 486–496.
- Yu, F., Xi, J., & Zhao, L., et al. (2011). Analysis of sparse component underdetermined blind source separation based on CS and K-SVD. *Journal of Southeast University*, 41(6), 1127–1131.
- Yu, X., Hu, D., & Xu, J. (2013). Blind Source Separation: Theory and Applications. *Journal of the Acoustical Society of America*, 105(2), 1101–1102.
- Zhen, L., Peng, D., & Yi, Z., et al. (2016). Underdetermined blind source separation using sparse coding. *IEEE Transactions on Neural Networks and Learning Systems*, 99, 1–7.