Circuits Syst Signal Process (2015) 34:3861–3882
DOI 10.1007/s00034-015-0033-5

Single-channel Speech Separation Using Dictionary-updated Orthogonal Matching Pursuit and Temporal Structure Information

Haiyan Guo[1,2] · **Xiaoxiong Li**[1] · **Lin Zhou**[1] · **Zhenyang Wu**[1]

Received: 20 March 2014 / Revised: 13 March 2015 / Accepted: 14 March 2015 /
Published online: 31 March 2015
© Springer Science+Business Media New York 2015

Abstract In this paper, we propose a two-stage sparse decomposition-based method for single-channel speech separation in time domain. First, we propose a Dictionary-updated orthogonal matching pursuit (DUOMP) algorithm which is used in both separation stages. In the proposed DUOMP algorithm, all atoms of each source-specific dictionary are updated by subtracting off the current approximation of each source to the original atoms. It is proved that the DUOMP algorithm can limit the separated sources within a region where they are uncorrelated in statistical sense more quickly. Then, we propose an adaptive dictionary generation method followed by a frame labeling method to perform a second-stage separation on the mixed frames having certain temporal structure. Experiments show that the proposed method outperforms a separation method using sparse non-negative matrix factorization (SNMF), a separation method using OMP and a source-filter-based method using pitch information in overall. Additionally, what affects the performance of the proposed method is also shown.

Keywords Single-channel speech separation (SCSS) · Sparse decomposition · Orthogonal matching pursuit (OMP) · Dictionary

✉ Haiyan Guo
haiyan.guo@seu.edu.cn; guohaiyan198311@163.com

Xiaoxiong Li
220120726@seu.edu.cn

Lin Zhou
linzhou@seu.edu.cn

Zhenyang Wu
zhenyang@seu.edu.cn

[1] School of Information Science and Engineering, Southeast University, Nanjing 210096, China

[2] College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

1 Introduction

In a natural environment, several speech signals are usually mixed. Speech separation aims to estimate such individual speech sources from their mixture. It has several obvious applications, e.g., in hearing aids or as a preprocessor to offer robustness in speech recognition, speaker recognition and speech coding. Single-channel speech separation (SCSS) discussed in this paper is an extreme case, where only one mixture is observed. It is considered as the most difficult case since no information of mixing matrix can be used. However, the human auditory system has impressive ability to solve this problem, that is, even using an ear, we can still isolate each individual speech when multi-talkers speak at the same time.

SCSS is an ill-posed problem, aiming to recover underlying speech sources from an observed mixture. Previous state-of-the-art SCSS approaches can be divided into two groups: source-driven or computational auditory scene analysis (CASA)-based method [11,12,36,43,44], and model-driven method [1,13,30,32,45]. CASA-based method tries to achieve human performance in auditory scene analysis (ASA) based on the perceptual organization of speech. Ideal binary time–frequency (T–F) mask has been proposed as the main computational goal of CASA [41]. The grouping principles prominently used in speech organization are harmonicity and periodicity of voiced speech, temporal continuity, onset and offset synchrony, common amplitude modulation, etc. CASA-based method does not rely heavily on priori knowledge of sources. However, in general, its separation performance is not as good as that of model-driven method.

Model-driven method generally outperforms CASA-based method, since it utilizes priori information of speakers. From a separation viewpoint, model-driven method can be divided into two classes: statistical model-driven method and decomposition-based method. Statistical model-driven method is based on statistical models (e.g., vector quantization (VQ) [8,27,33], Gaussian mixture model (GMM) [1,8,17,25,26,30], hidden Markov model (HMM) [10,31,42] and sinusoidal model [15,19,20,24,38]) or codebooks [e.g., independent component analysis (ICA) bases [13,14,16] and VQ codebooks [19–21,26]] trained for individual speakers. It tries to solve out model parameters or find codebook atoms which can generate mixture optimally to estimate sources by statistical methods, e.g., minimum mean square error (MMSE) estimation [25,29,30], maximum-likelihood (ML) estimation [13,15,26] and maximum a posterior (MAP) estimation [13–15,25]. Though statistical model-driven method has been reported to be effective, its training is rather time-consuming and estimation is significantly complex. In [32], 8192 states are required to fit each HMM to carefully model each transition state. In [19,20], every possible combination needs to be considered during distortion function minimization to find the optimal codebook atoms.

In statistical model-driven SCSS method, sparsity has been proven to be useful for SCSS. For example, generalized Gaussian distribution is used based on the observation that only a small number of coefficients of ICA basis functions differ significantly from zero [13]; a sparse-distributed code of basis functions is generated in [30], leading to better separation results than a compact code of basis functions in [28].

Sparsity has been also used in decomposition-based SCSS method [2,18,23,34,35,37,40], which is called as sparse decomposition-based SCSS method in this paper. Sparse decomposition-based method performs separation by mapping a mixture feature onto the union of learned source dictionaries and then computing the parts which fall in each dictionary by sparse decomposition. Each source dictionary is learned prior to give sparse representation of corresponding speaker's training speech features. Generally, sources are only sparse in their own specific dictionaries; therefore, they have less overlap in the source-specific dictionary union. Obviously, this feature is very helpful for separation. SCSS using sparse non-negative matrix factorization (SNMF) is a classical sparse decomposition-based method and has achieved comparable performance [2,35,40].

In sparse decomposition-based SCSS method, sparse coefficients of dictionary atoms need to be computed by sparse decomposition. Various methods have been proposed recently, of which the most typical methods are basis pursuit (BP) [3,4,6,9] and orthogonal matching pursuit (OMP) [7,22,39]. OMP can achieve similar performance to BP with the major advantages of its speed and ease for implementation. It is a greedy algorithm in which an atom most strongly correlated with the residual is chosen and its contribution is subtracted off to update the residual at each iteration. In OMP, the dictionary is fixed, while residuals are updated iteratively. To make a better match between chosen atoms and updated residuals, we propose a dictionary-updated OMP (DUOMP) algorithm in which dictionary is also updated at each iteration in this paper. We also prove its benefit for separation theoretically.

In addition, sparse decomposition-based method generally use the same trained source-specific dictionaries for separation of all mixed frames. However, it is not reasonable since mixed frames very different in temporal or frequency structure are not distinguished. Therefore, we propose to generate adaptive source dictionaries based on temporal structure information to perform a second-stage separation on mixed frames of certain temporal structure in this paper. Such frames are labeled out by a proposed frame labeling method mainly using pitch period and frame label results.

To evaluate the proposed method, we access the performance in various ways in Grid Corpus [5]. First, we compare DUOMP to OMP intuitively by showing their selected atoms for separation of the same mixed frame. Second, pitch period tracking and frame labeling results obtained are reported. Third, the proposed method is compared with a method using SNMF [35], a method using OMP and a source-filter-based method using pitch information [38] in terms of SNR, SDR, SIR and SAR. It is observed that the proposed method achieves better separation results in overall.

The remainder of this paper is structured as follows. In Sect. 2, we introduce the general model for sparse decomposition-based SCSS. In Sect. 3, we propose a two-stage sparse decomposition-based SCSS algorithm using DUOMP and temporal structure information. First, a DUOMP algorithm is proposed to compute sparse decomposition. Second, a frame labeling procedure is introduced to label mixed frames of certain temporal structure. Third, an adaptive dictionary generation method is presented for a second-stage separation of labeled mixed frames. The experiment results are reported and discussed in Sect. 4. Finally, we conclude and give future perspectives in Sect. 5.

2 Sparse Decomposition-Based Approach

Consider the standard linear instantaneous mixing model where a mixed signal $y(t)$ at time t is the linear combination of K speaker sources at the same time, that is,

$$\vec{y} = \sum_{k=1}^{K} \vec{s}_k \tag{1}$$

where $\vec{y} = [\, y(1),\ y(2),\ \ldots\ y(N)\,]^{\mathrm{T}}$ is a vector of size N denoting the single mixture, and $\vec{s}_k = [\, s_k(1),\ s_k(2),\ \ldots\ s_k(N)\,]^{\mathrm{T}}$ is a vector of the same size representing the kth source.

Suppose that $\vec{y}$ can be sparsely represented in a known overcomplete dictionary $\mathbf{D}$ which is the concatenation of K source-individual dictionaries, $\mathbf{D} = \left[\, \mathbf{D}_1\ \mathbf{D}_2\ \cdots\ \mathbf{D}_K \,\right]$,

$$\vec{y} = \mathbf{D}\vec{\theta} \tag{2}$$

where $\vec{\theta}$ is the complete code matrix which is the concatenation of the source-individual codes, $\vec{\theta} = \left[\, \vec{\theta}_1^{\mathrm{T}}\ \vec{\theta}_2^{\mathrm{T}}\ \cdots\ \vec{\theta}_K^{\mathrm{T}} \,\right]^{\mathrm{T}}$. The sparsest representation $\vec{\theta}$ of $\vec{y}$ in $\mathbf{D}$, denoted as $\hat{\vec{\theta}}$, can be found by solving the following problem,

$$\min \left\| \vec{\theta} \right\|_1 ,\ \text{s.t. } \vec{y} = \mathbf{D}\vec{\theta} \tag{3}$$

where $\|.\|_1$ denotes the l_1 norm of a vector. If the source dictionaries are diverse enough, it is possible to separate $\vec{y}$ into its individual sources $\hat{\vec{s}}_k$ as

$$\hat{\vec{s}}_k = \mathbf{D}_k \hat{\vec{\theta}}_k \tag{4}$$

where $\hat{\vec{\theta}}_k$ is a part of $\hat{\vec{\theta}}$, denoting the estimation of source-individual code $\vec{\theta}_k$.

As a consequence of above, there are two connected tasks to be solved in sparse decomposition-based SCSS: source dictionary learning and sparse decomposition computation. For the first task, we simply generate $\mathbf{D}_k$ as a matrix consisting of the kth speaker's time-domain training frames as columns called atoms in the first separation stage and focus on generating adaptive dictionary $\mathbf{\Psi}_k$ by selecting atoms of certain temporal structure from $\mathbf{D}_k$ in the second separation stage. As shown in our simulations, it is effective to use such time-domain source dictionaries for separation since better overall results are achieved as compared to the separation method using SNMF. It is worth mentioning that $\mathbf{D}_k$ generated as unsupervised clustering of training frames has also been tested, but results in much lower SNR. The greatest contribution for the first task is that we propose an adaptive source dictionary generation method in the second separation stage. The method is based on pitch period and frame label information, leading to improvement of separation on the mixed frames having certain temporal structure.

For the second task, we propose a DUOMP algorithm in which all atoms of a source dictionary are updated at each iteration so as to match residual better in temporal struc-

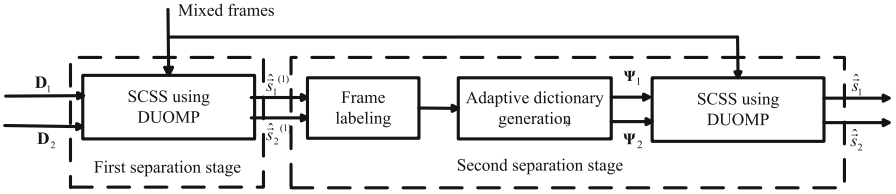

Fig. 1 Block diagram of the proposed sparse decomposition-based speech separation method using DUOMP and temporal structure information ($\hat{\vec{s}}_k^{(1)}$ is the source estimated in the first separation stage, and $\mathbf{\Psi}_k$ is the adaptive source dictionary generated using temporal structure information)

ture. We prove that DUOMP can achieve uncorrelated sources after fewer iterations in statistical sense as compared to OMP theoretically. The proposed DUOMP is used in both separation stages, leading to better separation performance than OMP as observed in our simulations.

3 Proposed Separation Method

We will now proceed to describe the proposed two-stage sparse decomposition-based separation approach using DUOMP and temporal structure information. In this paper, we focus on separating speech mixture composed of two speakers, i.e., $K = 2$ and $k \in \{1, 2\}$. Figure 1 shows the block diagram of the proposed separation approach.

As shown in Fig.1, the system is composed of the following blocks: the first-stage separation using DUOMP algorithm, frame labeling, adaptive dictionary generation and the second-stage separation using DUOMP algorithm. All mixed frames are separated in the first separation stage, while only labeled mixed frames are separated again in the second separation stage. Separation is performed frame after frame, and then, speech is synthesized by overlap-adding.

3.1 First-stage Separation Using DUOMP

In this subsection, we propose a DUOMP algorithm to compute sparse decomposition for separation. We update all the atoms of each source dictionary by subtracting off the current approximation of the corresponding source iteratively. In this way, atoms are expected to match more with updated residuals in temporal structure; thus, improved separation results should be achieved. Now we will first present the proposed DUOMP algorithm for SCSS and then explain its benefit theoretically.

The first-stage separation algorithm using DUOMP is presented as follows.

Now we will proceed to explain the benefit of the proposed DUOMP algorithm for SCSS.

Theorem 1 *On the assumption that correlation coefficients between atoms in* $\mathbf{D}_1^0$ *and* $\mathbf{D}_2^0$ *have a distribution with mean 0, we have estimation of* $\vec{s}_1$ *and* $\vec{s}_2$ *at iteration* $t = p + q$ *achieved using DUOMP, denoted as* $\hat{\vec{s}}_1^t$ *and* $\hat{\vec{s}}_2^t$*, satisfying*

Algorithm 1 (first-stage separation using DUOMP)

INPUT:

Two time-domain source dictionaries $\mathbf{D}_1$ and $\mathbf{D}_2$

Mixed signal $\vec{y}$

OUTPUT (suppose convergence satisfies after m iterations) :

Two estimated sources $\hat{\vec{s}}_1^{(1)}$ and $\hat{\vec{s}}_2^{(1)}$

Two matrices $\mathbf{I}_1^m$ and $\mathbf{I}_2^m$ consisting of chosen atoms from $\mathbf{D}_1$ and $\mathbf{D}_2$ respectively

Two approximations $\hat{\vec{\theta}}_1^m$ and $\hat{\vec{\theta}}_2^m$ of sources $\vec{s}_1$ and $\vec{s}_2$ respectively

A residual $\vec{r}^m = \vec{y} - \left[\begin{array}{cc}\mathbf{I}_1^m & \mathbf{I}_2^m\end{array}\right]\left[\begin{array}{cc}\hat{\vec{\theta}}_1^{m\mathrm{T}} & \hat{\vec{\theta}}_2^{m\mathrm{T}}\end{array}\right]^{\mathrm{T}}$

PROCEDURE:

(1) Initialize the residual $\vec{r}^0 = \vec{y}$, the source dictionaries $\mathbf{D}_1^0 = \mathbf{D}_1, \mathbf{D}_2^0 = \mathbf{D}_2$ and their union $\mathbf{D}^0 = \left[\begin{array}{cc}\mathbf{D}_1^0 & \mathbf{D}_2^0\end{array}\right]$, the matrices consisted of chosen atoms $\mathbf{I}_1^0 = \varnothing, \mathbf{I}_2^0 = \varnothing$, and the iteration counter $t = 0$.

(2) Find the index that solves the optimization problem [22,39]

$$\lambda^t = \arg\max_{j=1,\ldots,M} \left|\left\langle \vec{r}^t, \vec{d}_j^t \right\rangle\right| \qquad (5)$$

where $\vec{d}_j^t$ is the jth atom of $\mathbf{D}^t$ which is the concatenation of source dictionaries, $\mathbf{D}^t = \left[\begin{array}{cc}\mathbf{D}_1^t & \mathbf{D}_2^t\end{array}\right]$, and M is the number of atoms in $\mathbf{D}^t$.

(3) Merge the newly selected atom $\vec{d}_{\lambda^t}$ with the previous matrices of chosen atoms:

$$\mathbf{I}_1^t = \begin{cases} \left[\begin{array}{cc}\mathbf{I}_1^{t-1} & \vec{d}_{\lambda^t}\end{array}\right] & \lambda^t \le M_1 \\ \mathbf{I}_1^{t-1} & M_1 \le \lambda^t \le M \end{cases} \qquad (6)$$

$$\mathbf{I}_2^t = \begin{cases} \mathbf{I}_2^{t-1} & \lambda^t \le M_1 \\ \left[\begin{array}{cc}\mathbf{I}_2^{t-1} & \vec{d}_{\lambda^t}\end{array}\right] & M_1 \le \lambda^t \le M \end{cases} \qquad (7)$$

where M_1 and M_2 denote the numbers of atoms in $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively.

(4) Solve a least-squares problem to obtain a new approximation of $\vec{y}$ supported in $\mathbf{D}^t$:

$$\vec{\theta}^t = \arg\min_{\vec{\theta}} \left\| \vec{y} - \mathbf{I}^t \vec{\theta} \right\|_2 \qquad (8)$$

[22,39] where $\mathbf{I}^t$ is the concatenation of $\mathbf{I}_1^t$ and $\mathbf{I}_2^t$, $\mathbf{I}^t = \left[\begin{array}{cc}\mathbf{I}_1^t & \mathbf{I}_2^t\end{array}\right]$. The solution of (8) is given by $\hat{\vec{\theta}}^t = \left(\mathbf{I}^t \left(\mathbf{I}^t\right)^{\mathrm{T}}\right)^{-1} \mathbf{I}^t \vec{y}$ [7].

(5) Update all the atoms in $\mathbf{D}^t$:

$$\vec{d}_j^t = \begin{cases} \vec{d}_j^0 - \hat{\vec{\theta}}_1^t \mathbf{I}_1^t & 1 \le j \le M_1 \\ \vec{d}_j^0 - \hat{\vec{\theta}}_2^t \mathbf{I}_2^t & M_1 + 1 \le j \le M \end{cases} \qquad (9)$$

where $\hat{\vec{\theta}}_1^t$ and $\hat{\vec{\theta}}_2^t$ denote the parts of $\hat{\vec{\theta}}^t$ supported in $\mathbf{D}_1^t$ and $\mathbf{D}_2^t$, respectively, satisfying $\hat{\vec{\theta}}^t = \left[\begin{array}{cc}\hat{\vec{\theta}}_1^{t\mathrm{T}} & \hat{\vec{\theta}}_2^{t\mathrm{T}}\end{array}\right]^{\mathrm{T}}$.

(6) Calculate the new residual [9,22,39],

$$\vec{r}^t = \vec{y} - \hat{\vec{\theta}}^t \mathbf{I}^t \qquad (10)$$

Increment t, and return to step 2 until satisfying $\left\|\vec{r}^t\right\|_2 \le \delta_e$ or $\max_{j=1,\ldots,M} \left|\left\langle \vec{r}^t, \vec{d}_j^t \right\rangle\right| \le \delta_c$ where δ_e and δ_c are chosen thresholds and $\|.\|_2$ denotes the l_2 norm of a vector.

(7) Estimate the speech source in the first-stage separation as $\hat{\vec{s}}_k^{(1)} = \mathbf{D}_k^m \hat{\vec{\theta}}_k^m$.

$$E\left(\left\langle \hat{\vec{s}}_1^t, \hat{\vec{s}}_2^t \right\rangle\right) \approx 0$$

when $pq > Q$ where Q is a large number, p and q denote the number of selected atoms to estimate $\vec{s}_1$ and $\vec{s}_2$ at iteration t, respectively, and $\left\langle \hat{\vec{s}}_1^t, \hat{\vec{s}}_2^t \right\rangle$ denotes correlation between $\hat{\vec{s}}_1^t$ and $\hat{\vec{s}}_2^t$.

Proof Suppose that at iteration $t = p + q$, the k_qth atom in $\mathbf{D}_2^t$ denoted as $\vec{d}_{k_q}^t$ is selected, we have the estimated sources $\hat{\vec{s}}_1^t, \hat{\vec{s}}_2^t$ at iteration t satisfying

$$\hat{\vec{s}}_1^t = \mathbf{I}_1^t \hat{\vec{\theta}}_1^t = \mathbf{I}_1^{t-1}(\hat{\vec{\theta}}_1^{t-1} + \vec{u}) \tag{11}$$

$$\hat{\vec{s}}_2^t = \mathbf{I}_2^t \hat{\vec{\theta}}_2^t = [\,\mathbf{I}_2^{t-1}\ \vec{d}_{q_2}^t\,]\left[\,(\hat{\vec{\theta}}_2^{t-1} + \vec{v})^{\mathrm{T}}\ \hat{\theta}_2^t(q)\,\right]^{\mathrm{T}} \tag{12}$$

where $\vec{\mu}$ and $\vec{v}$ are vectors of the same sizes as $\hat{\vec{\theta}}_1^{t-1}$ and $\hat{\vec{\theta}}_2^{t-1}$, respectively. Since $\hat{\vec{\theta}}^t$ is obtained by solving (8), according to experience, when p and q are larger, we have

$$\|\vec{\mu}\|_0 \le A_1, \quad \|\vec{\mu}\|_1 \le \sigma_1 \tag{13}$$

$$\|\vec{v}\|_0 \le A_2, \quad \|\vec{v}\|_1 \le \sigma_2 \tag{14}$$

where A_1 and A_2 are small integers, and σ_1 and σ_2 are small values which can be ignored. Therefore, by combing (11–14), we have

$$\left\langle \hat{\vec{s}}_1^t, \hat{\vec{s}}_2^t \right\rangle \approx \left\langle \hat{\vec{s}}_1^{t-1}, \hat{\vec{s}}_2^{t-1} \right\rangle + \hat{\theta}_2^t(q) \left\langle \hat{\vec{s}}_1^{t-1}, \vec{d}_{k_q}^t \right\rangle \tag{15}$$

From (11), we have

$$\hat{\vec{s}}_1^{t-1} = \sum_{m=1}^{p-1} c_m \vec{d}_{j_m}^0 \tag{16}$$

where c_m is a variable scalar dependent on $\hat{\vec{\theta}}_1^1, \hat{\vec{\theta}}_1^2, \ldots, \hat{\vec{\theta}}_1^{p+q-1}$ and $\vec{d}_{j_m}^0$ denotes the j_mth atom in $\mathbf{D}_2^0$. From (9) and (12), we have

$$\begin{aligned} \vec{d}_{k_q}^t &= \frac{\vec{d}_{k_q}^0 - \hat{\vec{s}}_2^{t-1}}{\left\| \vec{d}_{k_q}^0 - \hat{\vec{s}}_2^{t-1} \right\|_2} = \frac{1}{\left\| \vec{d}_{k_q}^0 - \hat{\vec{s}}_2^{t-1} \right\|_2} (\vec{d}_{k_q}^0 - \mathbf{I}_2^{t-1}(\hat{\vec{\theta}}_2^{t-1} + \vec{v})) \\ &= \frac{1}{\left\| \vec{d}_{k_q}^0 - \hat{\vec{s}}_2^{t-1} \right\|_2} \sum_{r=1}^{q} b_r \vec{d}_{k_r}^0 \end{aligned} \tag{17}$$

where b_r is variable scalar dependent on $\hat{\vec{\theta}}_2^1, \hat{\vec{\theta}}_2^2, \ldots, \hat{\vec{\theta}}_2^{p+q-1}$ and $\vec{d}_{k_r}^0$ denotes the k_rth atom in $\mathbf{D}_2^0$. Combining (16) and (17), we have

$$\left\langle \hat{\vec{s}}_1^{t-1}, \vec{d}_{k_q}^t \right\rangle = \sum_{m=1}^{p-1} \sum_{r=1}^{q} c_m b_r \left\langle \vec{d}_{j_m}^0, \vec{d}_{k_r}^0 \right\rangle \tag{18}$$

Since c_m and b_r are bounded and $\left\langle \vec{d}_j^{\,0}, \vec{d}_k^{\,0} \right\rangle$ has a distribution with mean 0, according to the central limit theorem, when pq is a large number, $\left\langle \hat{\vec{s}}_1^{\,t-1}, \vec{d}_{k_q}^{\,q} \right\rangle$ has approximately normal distribution with mean 0, that is,

$$E\left(\left\langle \hat{\vec{s}}_1^{\,t-1}, \vec{d}_{k_q}^{\,t} \right\rangle\right) = 0 \quad \text{when } pq > Q \tag{19}$$

where Q is a large number. Similarly, we have

$$E\left(\left\langle \hat{\vec{s}}_1^{\,t-1}, \hat{\vec{s}}_2^{\,t-1} \right\rangle\right) = 0 \quad \text{when } pq > Q \tag{20}$$

Combing (15), (19) and (20), we have

$$E\left(\left\langle \hat{\vec{s}}_1^{\,t}, \hat{\vec{s}}_2^{\,t} \right\rangle\right) = 0 \quad \text{when } pq > Q \tag{21}$$

(21) can be also concluded in a similar way on the suppose that the j_pth atom in $\mathbf{D}_1^t$ denoted as $\vec{d}_{j_p}^{\,t}$ is selected at iteration t. Theorem 1 is proved. □

By a similar analysis, we can easily find that $E(\langle \hat{\vec{s}}_1^{\,t}, \hat{\vec{s}}_2^{\,t} \rangle) \approx 0$ holds when $p > Q$ and $q > Q$ by using OMP for SCSS. The comparison means that separated sources by using DUOMP tend to be limited within a region where they are uncorrelated more quickly in statistical sense than those using OMP. It is obviously helpful for SCSS since sources are generally independent of each other.

3.2 Frame Labeling

By analyzing the sources estimated in the first separation stage, we present a frame labeling approach for the second-stage separation of mixed frames having certain temporal structure mainly including quasi periodicity and sample value concentration. In this paper, a frame is termed concentrated if the values of its most samples are mainly positive or negative. Figure 2 shows some examples of normalized concentrated frames. It can be seen that a concentrated frame may be unvoiced or transition, and has certain temporal structure.

Three mixed types are considered here, which are voiced/voiced (V/V), voiced/concentrated (V/C) and concentrated/voiced (C/V). In V/V frames, both $\vec{s}_1$ and $\vec{s}_2$ are voiced. In V/C frames, only $\vec{s}_1$ is voiced and $\vec{s}_2$ is concentrated. In C/V frames, only $\vec{s}_2$ is voiced and $\vec{s}_1$ is concentrated. V/V, V/C and C/V frames are labeled by using the following measures:

- pitch period p_k, the pitch period of $\hat{\vec{s}}_k^{\,(1)}$ in samples.
- pitch period difference $|p_1 - p_2|$, the difference between the pitch periods of two estimated sources $\hat{\vec{s}}_1^{\,(1)}$ and $\hat{\vec{s}}_2^{\,(1)}$ in samples.

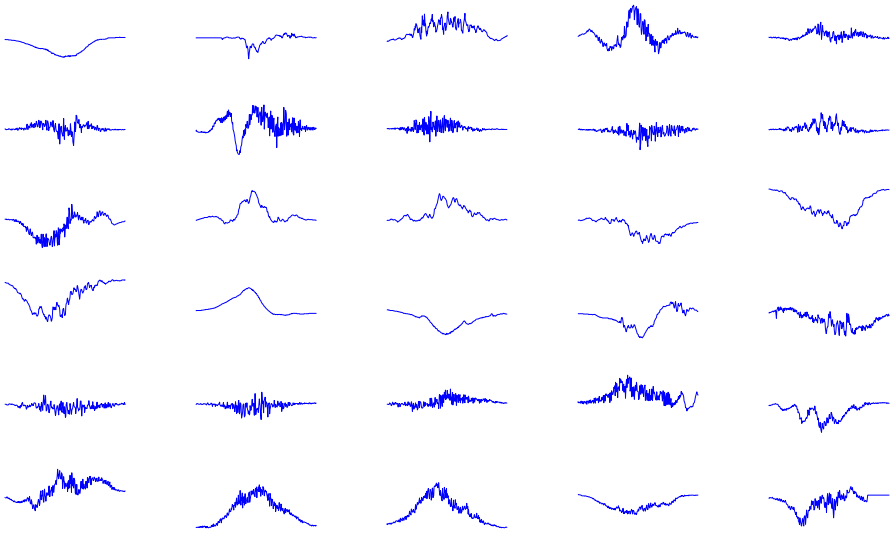

Fig. 2 Concentrated frame examples

– sample concentration ratio (SCR) r_k, defined as

$$r_k = \frac{\left|\sum_{n=1}^{N} \hat{\hat{s}}_k^{(1)}(n)\right|}{\sum_{n=1}^{N} \left|\hat{\hat{s}}_k^{(1)}(n)\right|} \tag{22}$$

where $\hat{\hat{s}}_k^{(1)}(n)$ denotes the value of the nth sample in $\hat{\hat{s}}_k^{(1)}$. Concentrated frames have much higher SCRs than the other frames.

– zero crossing number (ZCN) z_k, the zero crossing number of $\hat{\hat{s}}_k^{(1)}$ in samples. Voiced and concentrated frames generally have larger ZCNs than the other frames.

The proposed frame labeling method works as follows:

(1) All mixed frames are labeled as V/V when neither p_1 nor p_2 is not zero, $|p_1 - p_2|$ is bigger than a chosen threshold ε_p, and both z_1 and z_2 are lower than a chosen threshold ε_z. For the frames labeled as V/V, clear their labels if neither of their nearest neighbor frames is labeled as V/V. For unlabeled frames, label them V/V if their nearest neighbor frames are both labeled V/V.
(2) Calculate $\bar{r}_k$ and $\bar{z}_k$, the average of r_k and z_k of labeled V/V frames in step (1), respectively.
(3) For the frames not labeled as V/V in step (1), label them as follows when neither p_1 nor p_2 is not zero and $|p_1 - p_2|$ is smaller than ε_p:

All unlabeled frames are labeled V/V when for both estimated sources, r_k is lower than $\alpha \bar{r}_k$ where α is a chosen scalar invariant, z_k is lower than $\beta \bar{z}_k$ where β is a chosen scalar invariant, and p_k is continuous. p_k is considered continuous in this paper when there exists three consecutive frames including the present frame satisfying that the

differences between the pitch periods of the nearest neighbor frames are both smaller than a chosen threshold ε_d.

All the frames not labeled V/V above are considered V/C when r_1 is lower than $\alpha\bar{r}_1$, z_1 is lower than $\beta\bar{z}_1$ and p_1 is continuous. Then, set p_2 to 0. C/V frames are labeled and processed in a similar way.

All the remaining unlabeled frames are considered V/C (or C/V) when there exists a frame labeled V/C (or C/V) among the three consecutive frames and set $p_2 = 0$ (or $p_1 = 0$). Then, the rest of unlabeled frames are labeled V/C (or C/V) when both of their previous and next frames are labeled V/C (or C/V) and set $p_2 = 0$ (or $p_1 = 0$).

(4) For the frames not labeled in step (1) or (3), the present frame is considered a concentrated frame included when r_1 or r_2 is higher than a chosen threshold ε_r. Label the present frame as a V/C frame when satisfying that $p_1 \neq 0$ and the pitch difference between the present and the next (or previous) frame is lower than ε_d. Label the present frame C/V in a similar way.

(5) For the frames not labeled in step (1), (3) or (4), the present frame is considered voiced frame included, denoted as V/X or X/V temporarily, when satisfying that r_k is lower than $\alpha\bar{r}_k$, $p_k \neq 0$, and the pitch difference between the present and the next (or previous) frame is lower than ε_d for the same k. Label the V/X or X/V frames V/V when $r_{k'}$ is lower than $\alpha\bar{r}_{k'}$ and $z_{k'}$ is lower than $\beta\bar{z}_{k'}$ for $k' = 1, 2, k' \neq k$. Label the V/X frames V/C when r_2 is lower than $\alpha\bar{r}_2$ or z_2 is lower than $\beta\bar{z}_2$. Label the X/V frames C/V when r_1 is lower than $\alpha\bar{r}_1$ or z_1 is lower than $\beta\bar{z}_1$.

3.3 Adaptive Source Dictionary Generation Using Temporal Structure Information

In this subsection, we propose an adaptive dictionary generation method to perform a second-stage separation on labeled V/V, V/C and C/V frames. The proposed method utilizes pitch period and frame labeling results obtained in the subsection above to incorporate priori temporal structure information into dictionary generation. In this way, mixed frames showing great variety are distinguished; thus, improved separation performance can be expected. To be distinguished from source dictionaries $\mathbf{D}_k$ used in the first separation stage, source dictionaries generated here are denoted as $\mathbf{\Psi}_k$.

As presented in Theorem 1, DUOMP is appropriate for SCSS on the assumption that correlation coefficients between atoms in two source dictionaries have a distribution with mean 0. Therefore, to try to satisfy the assumption for labeled V/V frames, we generate an adaptive dictionary for the estimation of one source by adding the limitation on pitch period while keeping the other source dictionary unchanged. In this way, correlation coefficients are most likely to satisfy the assumption since they vary greatly. It is worth noting that we have observed that it indeed results in poor separation performance by generating adaptive dictionaries for both sources in our experiments.

For a labeled V/V frame, an adaptive dictionary $\mathbf{\Psi}_k$ is generated for the source with simpler temporal structure denoted as $\vec{s}_k$. $\vec{s}_k$ is considered simpler here when fewer atoms are selected for its estimation in the first-stage separation, that is, $\mathbf{I}_k^m$ consists of fewer atoms. $\mathbf{\Psi}_k$ is generated as a matrix consisting of atoms chosen in $\mathbf{D}_k$ based

on pitch periods as columns. The difference between the pitch period of each atom in $\mathbf{\Psi}_k$ and p_k is smaller than a chosen threshold ε'_p.

For a labeled V/C frame, $\mathbf{\Psi}_2$ is generated as a matrix consisting of atoms selected from $\mathbf{D}_2$ whose pitch periods are zero, SCRs are greater than a chosen threshold, ε'_r, and ZCNs are smaller than a chosen threshold. $\mathbf{\Psi}_1$ is generated in the same way as that for the estimation of the voiced source having simpler temporal structure above. Though the assumption in Theorem 1 may be not satisfied strictly since atoms in $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ both show certain temporal structure, DUOMP is still expected to be appropriate due to that the atoms in $\mathbf{\Psi}_2$ still show great randomness. Indeed, as shown in our simulation, it is helpful for separation by generating adaptive dictionaries for V/C frames in this way.

For a labeled C/V frame, adaptive source dictionaries are generated in a similar way.

4 Experiment

As a proof of concept, we evaluate the performance of the proposed SCSS method and compare it with the method using SNMF [35] and the method using OMP. We also report the separation performance of our proposed method on the mixtures available online in a source-filter-based method using pitch information [38]. To evaluate the separation performance, average of signal-to-noise ratio (SNR), average of source-to-distortion ratio (SDR), average of source-to-interferences ratio (SIR) and sources-to-artifacts ratio (SAR) are used. The SNR of the kth separated sentence $\hat{\vec{r}}_k$ is defined as

$$\mathrm{SNR} = 10\lg\left[\frac{(\vec{r}_k)^{\mathrm{T}}\vec{r}_k}{(\vec{r}_k-\hat{\vec{r}}_k)^{\mathrm{T}}(\vec{r}_k-\hat{\vec{r}}_k)}\right] \tag{23}$$

where $\vec{r}_k$ is the original sentence of kth speaker and $\hat{\vec{r}}_k$ is the respective separated sentence.

To evaluate the proposed separation algorithm, we used the Grid Corpus provided for SCSS by Cooke et. al [5]. We selected four speakers including two female (speakers 18 and 20) and two male speakers (speakers 1 and 2) from the database and denoted them as F1, F2, M1 and M2 in sequel. For each speaker, half of the sentences in the database were used for training and ten other sentences are selected randomly for testing. Speech sources are added directly at 0 dB SNR for each speech pair to have 400 female–male mixtures, 100 female–female mixtures and 100 male–male mixtures. The original sampling frequency was decreased from 25 to 8 kHz, and a hamming window of duration 32 ms with a frame shift of 16 ms was used.

In this section, we first give an example of selected atoms for SCSS using the proposed DUOMP and OMP algorithm, respectively. Then, we report the results of the proposed pitch tracking and frame labeling method. Thirdly, we compare the separation performance of the proposed method with that of the method using SNMF and OMP, and report our separation SNR results on the six mixtures available online in [38]. Finally, we discuss what affects the separation performance of the proposed method and address our future work.

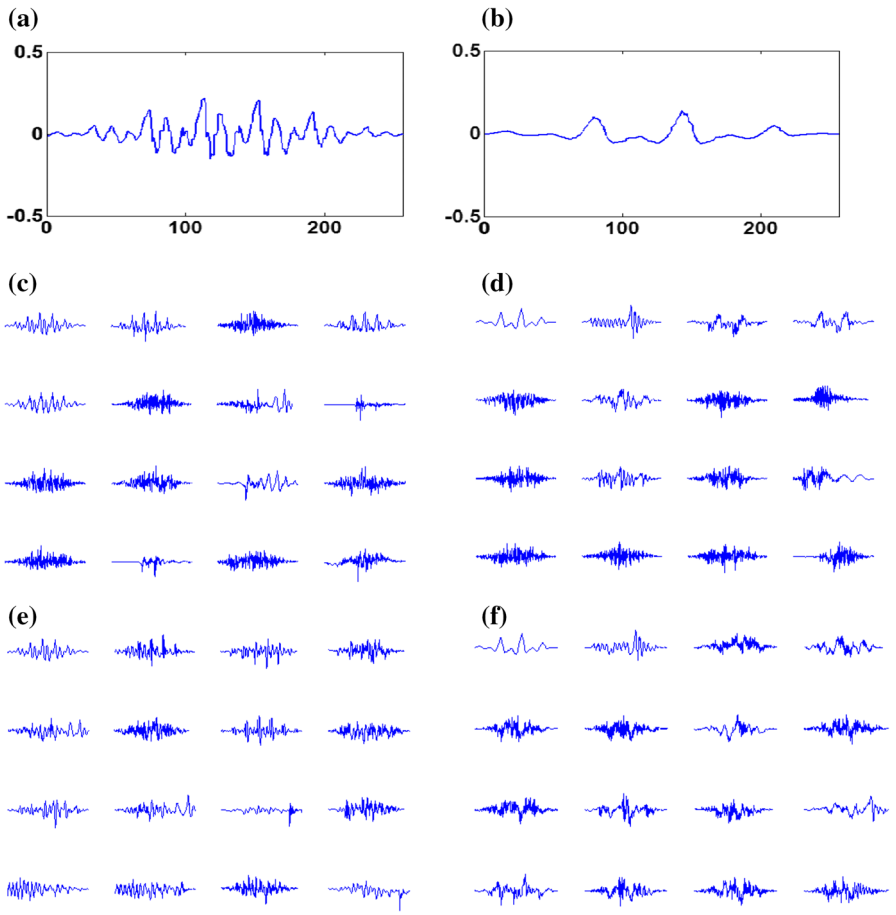

Fig. 3 Waveforms of sources and selected atoms. **a**, **b** Sources 1,2; **c**, **d** the first 16 atoms selected to estimate sources using OMP; **e**, **f** the first 16 atoms selected to estimate sources using DUOMP

4.1 Selected Atoms in DUOMP

Figure 3 shows the first 16 atoms chosen using OMP and DUOMP, respectively, for the separation of a mixed frame. In this example, 70 and 51 atoms are selected for the estimation of two sources using OMP, while 32 and 89 atoms are selected using DUOMP. In our simulations, we set $\delta_e = 10^{-10}$, $\delta_c = 10^{-5}$.

4.2 Pitch Tracking and Frame Labeling Results

In this subsection, we report the results of pitch period tracking and frame labeling, which are used to find out the mixed frames with temporal structure to perform a second-stage separation. The relationship between their performance and separation results will be discussed in the following discussion subsection. In our simulations,

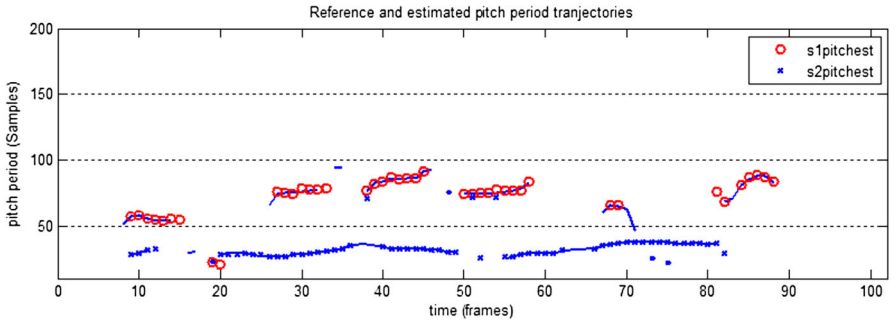

Fig. 4 Pitch period tracking results on test mixture of a male (M1) and a female (F1) speakers ("pbbv6n" and "sbil4a"), together with truth pitch period trajectories (*black solid lines*)

Table 1 Pitch period tracking results in terms of η_{p_k} (%)

	η_{p_1}	η_{p_2}
F/F	75.9	73.1
F/M	74.0	81.1
M/M	62.1	68.9

we set $\varepsilon_p = 2$, $\varepsilon_r = 0.6$, $\varepsilon_z = 80$, $\alpha = 3$, $\beta = 1.2$, $\varepsilon_d = 3$. The parameters are experimentally determined and can lead to better SCSS performance than other parameters.

Figure 4 depicts the pitch period tracking result of a female–male mixture. For each mixture, two pitch period trajectories in samples are estimated by analyzing the sources estimated in the first separation stage using autocorrelation method. The ground truth pitch period trajectories are obtained directly from original speech sources.

To evaluate the overall performance of pitch period tracking, we use a correctness measure defined as,

$$\eta_{p_k} = \frac{\left|\left\{n_f | -\varepsilon'_p \le \vec{q}_k(n_f) - \hat{\vec{q}}_k(n_f) \le \varepsilon'_p\right\}\right|}{N_f} \tag{24}$$

where $\vec{q}_k$ and $\hat{\vec{q}}_k$ denote the true and estimated pitch period trajectory of speaker k in samples, respectively, n_f denotes frame index, $\left\{n_f | -\varepsilon'_p \le \vec{q}_k(n_f) - \hat{\vec{q}}_k(n_f) \le \varepsilon'_p\right\}$ denotes an index set consisting of frames satisfying $-\varepsilon'_p \le \vec{q}_k(n_f) - \hat{\vec{q}}_k(n_f) \le \varepsilon'_p$, $|\{\cdot\}|$ denotes the cardinality of a set, and N_f denotes the number of total frames evaluated. In our experiment, we set $\varepsilon'_p = 1$. Tables 1 and 2 summarize the averaged pitch period tracking and frame labeling results of 600 mixtures tested.

As shown in Table 1, the proposed pitch period tracking method performs best for female–male pairs and worst for male–male pairs for the reason that the first-stage separation performs best on female–male mixtures and worst on male–male mixtures. From Table 2, it can be seen that our proposed method performs best for V/V frames and much worse for V/C and C/V frames. The reason is that the method relies heavily on the estimated waveforms. Concentrated frames show much more randomness than

Table 2 Frame labeling results (%)

Reference frame label	Estimated frame label											
	V/V			*V* / *C*			C/V			Others		
	F/F	F/M	M/M	F/F	F/M	M/M	F/F	F/M	M/M	F/F	F/M	M/M
V/V	**83.6**	**86.3**	**52.3**	5.3	0.9	4.1	3.9	2.9	2.3	7.2	9.9	41.3
V/C	13.8	40.5	21.0	**62.5**	**22.2**	**27.7**	4.7	2.4	0.8	19.0	34.9	50.5
C/V	36.8	19.3	25.6	1.5	1.9	3.6	**26.5**	**42.5**	**19.6**	35.2	36.3	51.2
others	19.0	18.0	11.8	8.7	2.8	5.3	5.3	6.1	3.2	**67.0**	**73.1**	**79.7**

The results in bold represents the ratio of the frames which are labeled out correctly

Table 3 Results of separation using SNMF, OMP and DUOMP (first-stage, second-stage and first-stage plus second-stage separation) in SNR (dB)

Method	F/F	F/M	M/M
SNMF	4.7/4.5	5.5/5.3	3.3/3.9
OMP	4.8/4.7	5.7/5.8	3.7/4.4
DUOMP (first-stage separation)	5.1/5.0	6.2/6.3	4.0/4.5
DUOMP (second-stage separation)	5.3/5.1	6.4/6.6	4.1/4.6
DUOMP (first-stage plus second-stage separation)	5.7/5.5	6.7/6.8	4.3/4.8

Table 4 Results of separation using SNMF, OMP and DUOMP (first-stage, second-stage and first-stage plus second-stage separation) in SDR (dB)

Method	F/F	F/M	M/M
SNMF	4.2/5.0	6.2/6.0	1.9/3.1
OMP	3.8/4.1	5.4/5.4	2.2/3.3
DUOMP (first-stage separation)	4.3/4.7	6.2/6.2	2.5/3.4
DUOMP (second-stage separation)	4.4/4.8	6.4/6.3	2.6/3.6
DUOMP (first-stage plus second-stage separation)	5.0/5.3	6.8/6.7	3.0/3.8

voiced frames in time domain; thus, it results in much more difference between original and estimated waveforms.

4.3 Separation Results

We compare the separation results of the proposed first-stage, second-stage and first-stage plus second-stage separation using DUOMP with that of the separation method using SNMF and the method using OMP in SNR, SDR, SIR and SAR, respectively. The average results of 600 mixtures tested are shown in Tables 3, 4, 5 and 6. The same training sentences are used to generate each SNMF source dictionary with the sparsity $\lambda = 0.1$ and the size of 560 as in [35]. The dictionaries used in the OMP method are the same as those used in the DUOMP method in the first-stage separation. First-stage plus second-stage separation method selects the separated frames of higher SNR from the first-stage and second-stage separation results as its separation results.

As shown in Table 3, first-stage separation using DUOMP outperforms separation using OMP in SNR by 0.3 dB in female–female mixtures, 0.5 dB in female–male mixtures and 0.2 dB in male–male mixtures, respectively. As compared to the method using SNMF, it achieves 0.45 dB higher SNR in female–female mixtures, 0.85 dB higher SNR in female–male mixtures and 0.65 dB higher SNR in male–male mixtures, respectively. Second-stage separation using DUOMP outperforms first-stage separation using DUOMP by 0.15 dB in female–female mixtures, 0.25 dB in female–male mixtures and 0.1 dB in male–male mixtures, respectively. Although second-stage separation leads to higher frame SNRs for most of the labeled frames than first-

Table 5 Results of separation using SNMF, OMP and DUOMP (first-stage, second-stage and first-stage plus second-stage separation) in SIR (dB)

Method	F/F	F/M	M/M
SNMF	5.0/6.3	9.3/8.7	3.3/4.9
OMP	8.7/11.4	12.7/11.7	8.9/8.9
DUOMP (first-stage separation)	9.5/12.2	13.5/12.9	8.8/9.8
DUOMP (second-stage separation)	9.3/12.7	14.3/12.9	8.7/10.0
DUOMP (first-stage plus second-stage separation)	10.2/13.3	14.7/13.7	9.4/10.5

Table 6 Results of separation using SNMF, OMP and DUOMP (first-stage, second-stage and first-stage plus second-stage separation) in SAR (dB)

Method	F/F	F/M	M/M
SNMF	10.4/10.1	10.1/10.0	10.2/10.6
OMP	6.2/5.3	6.6/6.9	3.9/5.2
DUOMP (first-stage separation)	6.5/5.7	7.3/7.5	4.3/4.9
DUOMP (second-stage separation)	6.7/5.9	7.4/7.7	4.4/5.0
DUOMP (first-stage plus second-stage separation)	7.0/6.3	7.8/7.9	4.7/5.3

stage separation, it leads to lower frame SNRs for rare labeled frames. Therefore, the improvement is small. First-stage plus second-stage separation can achieve better SNR results by selecting separated frames of higher frame SNRs from first-stage and second-stage separation as its separated frames. As shown in Table 3, it achieves 0.5 dB higher SNR in female–female mixtures, 0.5 dB higher SNR in female–male mixtures and 0.3 dB higher SNR in male–male mixtures, respectively, than the first-stage separation. However, it is not practical due to that frames of higher frame SNRs cannot be selected since speech sources are not known as a priori. Still, we can consider incorporating a perceptual evaluation of speech quality (PESQ) 563 system to select separated frames which can lead to higher mean opinion score (MOS) to improve our proposed method in our future work.

From Tables 4, 5 and 6, it can be included that compared to separation using SNMF, second-stage separation using DUOMP achieves 0.3 dB higher SDR, 4.8 dB higher SIR and 3.3 dB lower SAR in average. Thus, the proposed method outperforms separation using SNMF in overall. It can be easily seen that DUOMP still outperforms OMP for SCSS in SDR, SIR and SAR. As shown in Tables 4, 5 and 6, First-stage separation achieves 0.15 dB higher SDR, 0.8 dB higher SIR and 0.5 dB higher SAR in average than separation using OMP. Moreover, compared to first-stage separation, second-stage separation achieves slightly higher SDR, SIR and SAR results, while first-stage plus second-stage separation achieves relatively much higher results.

Finally, we compare the separation performance of our proposed method on the six mixtures available online reported in a source-filter-based method using pitch information [38] in SNR. The mixtures include four female–male, a female–female and a male–male speech pairs. Table 7 shows the comparison results. The results of

Table 7 Comparing the SNR results of the proposed method with the method using SNMF and source-filter-based method

Mixture #	Method using SNMF		Source-filter-based method		Proposed method			
					DUOMP (second-stage separation)		DUOMP (first-stage plus second-stage separation)	
	Source 1	Source 2	Source 1	Source 2	Source 1	Source 2	Source 1	Source 2
1	6.1	6.5	6.7	6.7	6.4	7.2	6.6	7.3
2	5.9	6.0	5.0	8.0	6.9	6.3	7.0	6.4
3	6.1	6.4	7.4	7.7	6.4	6.5	6.6	6.7
4	5.2	5.5	5.2	6.5	5.7	5.7	5.7	5.9
5	3.4	4.1	6.4	6.7	5.2	6.7	5.3	6.8
6	3.1	3.1	0.8	1.8	3.1	3.1	3.3	3.3
Ave	5.0	5.3	5.3	6.1	5.6	5.9	5.8	6.5

Table 8 Results of (a–g) in SNR (dB)

Method	F/F	F/M	M/M
(a)	5.2/4.9	6.4/6.5	4.0/4.5
(b)	5.3/4.9	6.4/6.4	4.0/4.5
(c)	5.5/5.3	6.5/6.7	4.3/4.7
(d)	5.5/5.3	6.4/6.5	4.1/4.6
(e)	6.1/5.8	6.6/6.9	4.6/5.0
(f)	5.8/5.7	6.9/6.9	4.3/4.8
(g)	6.3/6.3	7.2/7.3	5.1/5.5

Table 9 Results of (a–g) in SDR (dB)

Method	F/F	F/M	M/M
(a)	4.4/4.7	6.3/6.3	2.5/3.4
(b)	4.3/4.8	6.3/6.2	2.5/3.4
(c)	4.7/5.2	6.6/6.6	3.0/3.7
(d)	4.7/5.0	6.3/6.3	2.7/3.5
(e)	5.5/5.9	6.7/6.7	3.5/4.2
(f)	5.2/5.6	6.9/7.0	3.0/3.9
(g)	6.0/6.5	7.4/7.4	4.3/5.0

separation using SNMF are also given. It is shown that our proposed method performs better than source-filter-based method using pitch information in SNR in overall.

4.4 Discussion and Future Work

In the proposed method, a second-stage separation is presented based on adaptive dictionaries and performed on labeled V/V, V/C and C/V frames. In this subsection, we experimentally discuss how pitch period tracking, frame labeling and dictionary generation impact the proposed second-stage separation performance and have a consideration of our potential future work.

Tables 8, 9, 10 and 11 shows the results of : (a) second-stage separation only on labeled V/V frames; (b) second-stage separation only on labeled V/C and C/V frames; (c) second-stage separation only on true V/V frames using true pitch periods; (d) second-stage separation only on true V/C and C/V frames using true pitch periods; (e) second-stage separation on true V/V, V/C and C/V frames using true pitch periods; (f) second-stage separation on labeled V/V frames using optimal dictionaries; (g) second-stage separation on true V/V frames using true pitch periods and optimal dictionaries. True V/C frames are defined as frames satisfying $p_1 \neq 0, p_2 = 0, r_2 \geq \varepsilon_r'$, and true C/V frames are defined in a similar way. Optimal dictionaries are defined as the dictionaries leading to the highest frame SNRs which are selected from adaptive dictionaries generated based on the temporal structure of source 1, 2 and dictionaries the same as those used in the first-stage separation. The results are averaged on the tested 600 mixtures.

Table 10 Results of (a–g) in SIR (dB)

Method	F/F	F/M	M/M
(a)	9.4/12.4	13.8/12.9	8.6/9.9
(b)	9.8/12.5	13.6/12.9	8.8/9.9
(c)	10.2/12.5	14.5/13.3	9.3/10.5
(d)	10.0/12.7	13.7/13.1	9.1/10.0
(e)	11.0/14.5	15.1/13.4	9.7/10.8
(f)	11.0/12.9	14.4/14.3	9.5/10.5
(g)	11.6/14.6	15.4/17.8	11.0/12.2

Table 11 Results of (a–g) in SAR (dB)

Method	F/F	F/M	M/M
(a)	6.7/5.8	7.4/7.6	4.4/4.9
(b)	6.6/5.9	7.3/7.7	4.3/5.0
(c)	6.7/6.1	7.7/7.9	4.7/5.2
(d)	6.7/6.1	7.4/7.6	4.4/5.1
(e)	7.5/6.7	7.7/8.0	5.2/5.5
(f)	7.0/6.7	8.1/8.2	4.7/5.4
(g)	7.6/7.3	8.4/9.6	5.7/6.2

From (a–b) in Tables 8, 9, 10 and 11, it can be seen that frame labeling improves separation performance since the results of (a) and (b) are both higher than those of first-stage separation and lower than those of second-stage separation. Moreover, by comparing (c) to (a), (d) to (b) and (e) to the second-stage separation results, we find out that, by improving the accuracy of pitch period tracking and frame labeling, we can achieve at most 0.2 dB higher SNR, 0.3 dB higher SDR, 0.3 dB higher SIR and 0.1 dB higher SAR for the separation of V/V frames, 0.1 dB higher SNR, 0.1 dB higher SDR and 0.1 dB higher SIR for the separation of V/C and C/V frames, 0.4 dB higher SNR, 0.5 dB higher SDR, 0.2 dB higher SIR and 0.1 dB higher SAR in overall.

More importantly, it is observed that the proposed method can be improved more by optimal dictionaries used for the separation of V/V mixtures. Comparing (f) to (a), we can see that 0.3 dB higher SNR, 0.7 dB higher SDR, 1.0 dB higher SIR and 0.6 dB higher SAR can be achieved. Although optimal dictionaries leading to the highest frame SNRs cannot be selected since sources are not known, we can still expect to use a PESQ 563 system as a feedback to choose the dictionaries leading to highest MOS score as optimal dictionaries.

Comparing (g) to (a), we can see that by the combination of improving the accuracy of pitch period tracking and using optimal dictionaries, it can lead to 1.1 dB higher SNR, 1.3 dB higher SDR, 2.9 dB higher SIR and 1.4 dB higher SDR. Thus, we will consider improving pitch period tracking and using optimal dictionaries to improve our proposed method in our future work.

In addition, although DUOMP outperforms OMP in SCSS as shown in Tables 3, 4, 5 and 6, the complexity of the algorithm is higher. The reason is that all atoms are updated at each iteration. As a potential future work, we expect to reduce the algorithm complexity by updating part of atoms and study on which atoms to be updated.

5 Conclusion and Future Work

In this paper, we presented a two-stage sparse decomposition-based SCSS method. A DUOMP algorithm has been proposed to compute sparse decomposition, and an adaptive dictionary generation method has been presented for a second-stage separation of mixed frames having certain temporal structure. In our proposed DUOMP algorithm, all atoms of each source dictionaries are updated by subtracting off the present approximation to the source at each iteration, leading to separated sources which are limited within a region in which they are uncorrelated more quickly in a statistical sense than OMP. Adaptive dictionaries are generated based on pitch period and frame label information to distinguish mixed frames having different temporal structure. By comparison to other separation methods, it was observed that our proposed method achieved better separation performance in SNR, SDR, SIR and SAR.

We have discussed what affects the performance of the proposed separation method and considered selecting optimal source dictionaries and improving the pitch period tracking and frame labeling accuracy as our potential work. In addition, we will consider reducing the complexity of our proposed method by studying on updating part of atoms and incorporating dictionary learning into the presented separation work.

Acknowledgments This work is supported by National Natural Science Foundation of China (No. 61302152, Nos. 61201345, 61271335 and 61271240), the Beijing Key Laboratory of Advanced Information Science and Network Technology (No. XDXX1308), the Major Science Research Project of Jiangsu Provincial Education Department (13KJA510003) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

1. L. Benaroya, F. Bimbot, R. Gribonval, Audio source separation with a single sensor. IEEE Trans. Audio Speech **14**(1), 191–199 (2006)
2. L. Benaroya, L.M. Donagh, F. Bimbot, R. Gribonval, Non negative sparse representation for wiener based source separation with a single sensor. ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. Proc. **6**, 613–616 (2003). doi:10.1109/ICASSP.2003.1201756
3. S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1998)
4. S.S. Chen, D.L. Donoho, Atomic decomposition by basis pursuit. SIAM Rev. **43**(1), 129–159 (2001)
5. M.P. Cooke, J. Barker, S.P. Cunningham, X. Shao, An audiovisual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**(5), 2421–2424 (2006)
6. G.B. Dantzig, *Linear Programming and Extensions* (Princeton University Press, Princeton, 1963)
7. D.L. Donoho, Y. Tsaig, I. Drori, J.L. Starck, Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. IEEE Trans. Inform. Theory **58**(2), 1094–1121 (2012)
8. D.P.W. Ellis, R.J. Weiss, Model-based monaural source separation using a vector-quantized phase-vocoder representation. ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. Proc. **5**, 957–960 (2006)
9. P.E. Gill, W. Murray, M.H. Wright, *Numerical Linear Algebra and Optimization* (Addison-Wesley, Redwood City, 1991)
10. J.R. Hershey, S.J. Rennie, P.A. Olsen, T.T. Kristjansson, Superhuman multi-talker speech recognition: a graphical modeling approach. Comput. Speech Lang. **24**(1), 45–66 (2010)
11. G. Hu, D.L. Wang, Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Netw. **15**(5), 1135–1150 (2004)

12. G. Hu, D.L. Wang, Auditory segmentation based on onset and offset analysis. IEEE Trans. Audio Speech **15**(2), 396–405 (2007)
13. G.J. Jang, T.W. Lee, A maximum likelihood approach to single-channel source separation. J. Mach. Learn. Res. **4**(7–8), 1365–1392 (2003)
14. G.J. Jang, T.W. Lee, A probabilistic approach to single channel source separation. in *16th Annual Neural Information Processing Systems Conference* (2003)
15. G.J. Jang, T.W. Lee, Y.H. Oh, Single-channel signal separation using time-domain basis functions. IEEE Signal Process. Lett. **10**, 168–171 (2003)
16. G.J. Jang, T.W. Lee, Y.H. Oh, A subspace approach to single channel signal separation using maximum likelihood weighting filters. IEEE Int. Conf. Acoust. Speech Signal Process. **5**, 45–48 (2003). doi:10.1109/ICASSP.2003.1199864
17. H. Katmeoka, T. Nishimoto, S. Sagayama, Separation of harmonic structures based on tied Gaussian mixture model and information criterion for concurrent sounds. IEEE Int. Conf. Acoust. Speech Signal Process. **4**, 297–300 (2004)
18. M. Moussallam, G. Richard, L. Daudet, Audio source separation informed by redundancy with greedy multiscale decompositions. in *European Signal Processing Conference* (2012), pp. 2644–2648
19. P. Mowlaee, M.G. Christensen, S.H. Jensen, Improved single-channel speech separation using sinusoidal modeling. ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. Proc. 21–24 (2010). doi:10.1109/ICASSP.2010.5496263
20. P. Mowlaee, M.G. Christensen, S.H. Jensen, New results on single-channel speech separation using sinusoidal modeling. IEEE Trans. Audio Speech **19**(5), 1265–1277 (2011)
21. P. Mowlaee, A. Sayadiyan, H. Sheikhzadeh, Evaluating single-channel speech separation performance in transform-domain. Sci. C J. Zhejiang Univ. **11**(3), 160–174 (2010)
22. Y.C. Pati, R. Rezaiifar, P.S. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. in *Conference Record of Asilomar Conference Signals Systems Computers* (1993), pp 40–44
23. B.A. Pearlmutter, R.K. Olsson, Linear program differentiation for single-channel speech separation. in *Proceedings of IEEE Signal Processing Society Workshop. Machine Learning Signal Processing MLSP 2006*. pp. 421–426 (2006). doi:10.1109/MLSP.2006.275587
24. T.F. Quatieri, R.G. Danisewicz, An approach to co-channel talker interference suppression using a sinusoidal model for speech. IEEE Trans. Audio Speech **38**(1), 56–69 (1990)
25. M.H. Radfar, R.M. Dansereau, Single-channel speech separation using soft mask filtering. IEEE Trans. Audio Speech **15**(8), 2299–2310 (2007)
26. M.H. Radfar, R.M. Dansereau, A. Sayadiyan, A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation. EURASIP J. Audio Speech Music Process **2007**, 084186 (2007). doi:10.1155/2007/84186
27. M.H. Radfar, R.M. Dansereau, A. Sayadiyan, Monaural speech segregation based on fusion of source-driven with model-driven techniques. Speech Commun. **49**(6), 464–476 (2007)
28. B. Raj, P. Smaragdis, Latent variable decomposition of spectrograms for single channel speaker separation. in *IEEE ASSP Workshop Applications Signal Processing to Audio Acoustics*, pp. 17–20, doi:10.1109/ASPAA.2005.1540157
29. A.M. Reddy, B. Raj, A minimum mean squared error estimator for single channel speaker separation. in *INTERSPEECH- 2004*, pp. 2445–2448 (2004)
30. A.M. Reddy, B. Raj, Soft mask methods for single-channel speaker separation. IEEE Trans. Audio Speech **15**(6), 1766–1776 (2007)
31. M.J. Reyes-Gomez, D.P.W. Ellis, N. Jojic, Multiband audio modeling for single-channel acoustic source separation. ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. Proc. **5**, 641–644 (2004). doi:10.1109/ICASSP.2004.1327192
32. S.T. Roweis, One microphone source separation. Adv. Neural. In. **13**, 793–799 (2000)
33. S.T. Roweis, Factorial models and refiltering for speech separation and denoising. in *EUROSPEECH* (2003), pp. 1009–1012
34. M.N. Schmidt, R.K. Olsson, Linear regression on sparse features for single-channel speech separation. in *IEEE ASSP Workshop Applications of Signal Processing to Audio Acoustics*, pp. 26–29 (2007). doi:10.1109/ASPAA.2007.4393010
35. M.N. Schmidt, R.K. Olsson, Single-channel speech separation using sparse non-negative matrix factorization. in *INTERSPEECH* 2006

36. Y. Shao, S. Srinivasan, Z. Jin, D. Wang, A computational auditory scene analysis system for speech segregation and robust speech recognition. Comput. Speech Lang. **24**(1), 77–93 (2010)
37. M.V.S. Shashanka, B. Raj, P. Smaragdis, Sparse overcomplete decomposition for single channel speaker separation. IEEE Trans. Audio Speech **2**, 641–644 (2007)
38. M. Stark, M. Wohlmayr, F. Pernkopf, Source-filter-based single-channel speech separation using pitch information. IEEE Trans. Audio Speech **19**(2), 242–255 (2011)
39. J.A. Tropp, A.C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans. Inform. Theory **53**(12), 4655–4666 (2007)
40. T. Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech **15**(3), 1066–1074 (2007)
41. T. Virtanen, Speech recognition using factorial hidden Markov models for separation in the feature space. in *INTERSPEECH* 2006, pp. 89–92 (2006)
42. D.L. Wang, On ideal binary mask as the computational goal of auditory scene analysis, in *Speech Separation by Humans and Machines*, ed. by D.L. Wang (Kluwer Academic, Norwell, 2005), pp. 181–197
43. D.L. Wang, G.J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, NY, 2006)
44. D.L. Wang, G.J. Brown, Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Netw. **10**, 684–697 (1999)
45. R.J. Weiss, D.P.W. Ellis, Speech separation using speaker-adapted eigenvoice speech models. Comput. Speech Lang. **24**(1), 16–29 (2010)