Signal Processing 106 (2015) 134–140

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Fast communication

Single-channel speech separation using sequential discriminative dictionary learning ☆

Yangfei Xu, Guangzhao Bao, Xu Xu, Zhongfu Ye *

Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui 230027, People's Republic of China
National Engineering Laboratory for Speech and Language Information Processing, China, Hefei, Anhui 230027, People's Republic of China

ARTICLE INFO

Article history:
Received 22 May 2014
Received in revised form
16 July 2014
Accepted 17 July 2014
Available online 2 August 2014

Keywords:
Single-channel speech separation
Sequential discriminative dictionary learning
Sparse coding

ABSTRACT

A novel sequential discriminative dictionary learning (SDDL) algorithm is presented to suppress the confusion between the separated signals which we denote source confusion. The existing discriminative dictionary learning (DDL) algorithms assume that signals from different speakers have their unique components which makes that the signals can be explained by the corresponding sub-dictionaries. But the signals from different speakers have similar components when divided into speech segments. We take the unique and similar components of different speakers' signals into account, and design a new structured dictionary which contains discriminative and buffer sub-dictionaries. The unique components of different speakers' signals which have better correspondences to the speakers' labels are firstly separated, and the similar components of different speakers' signals are separated in the next layer. An objective function is derived, which guarantees that the unique components of the training sets can be explained by their corresponding discriminative sub-dictionaries and the similar components of the training sets can be explained by the buffer sub-dictionary rather than the cross sub-dictionary. The components distributed in the buffer sub-dictionary are used as training sets in the next layer. Experiments results verify that the proposed algorithm can effectively reduce the source confusion compared to the existing algorithms.

1. Introduction

Single-channel speech separation (SCSS) [1,2] problem, a particularly difficult version of the speech separation problems, occurs when a single-channel recording is available, aiming at recovering the underlying speech signals from a mixed signal. Mathematically, there are infinitely many solutions for the SCSS problem unless we impose some constraints on the sources, such as statistical independence. The existing SCSS algorithms can be categorized into two branches: (1) computational auditory scene analysis (CASA) [3,4] and (2) model-based SCSS [5–9]. The former seeks discriminative features in the observation signals to separate the speech signals while the latter mainly relies on a priori knowledge of sources obtained during a training stage.

Structured signals, like speeches, have approximately sparse representations in suitably chosen dictionaries [5]. The discriminative dictionary learning (DDL) algorithm [5,10] assumes that speech signals from different speakers have their unique components. The column of dictionary called atom is coherent to the signal if the absolute value of the inner product of the atom and the signal is large. After using discriminative dictionary to sparsely code the

☆ This work is supported by the Science and Technology Plan Project of Anhui Province of China (No. 13Z02008-5) and the Youth Innovation Foundation of USTC.

* Corresponding author. Tel.: +86 551 63601314.
E-mail address: yezf@ustc.edu.cn (Z. Ye).

http://dx.doi.org/10.1016/j.sigpro.2014.07.012
0165-1684/

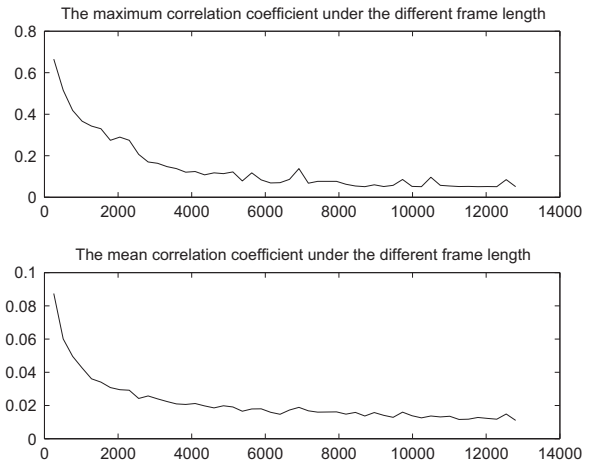

Fig. 1. The correlation between different speakers' speech signals at different lengths of the analysis window.

different structured speech signals, the coding coefficients of the different underlying sources are distributed apart over all dictionary elements.

The speech signal is short-term stationary and should be decomposed into series of short segments [11], referred to as analysis frames, which should be analyzed independently. Fig. 1 shows the correlation between different speakers' speech signals at different lengths of the analysis window. It is apparent that the correlation between the analysis frames of different speakers becomes stronger as the length of analysis window decreases; in other words, the analysis frames of different speakers may have similar components. So, if the DDL algorithm is directly used to separate the mixed signal, the source confusion will be introduced into the separated signals.

In this paper, we propose a novel sequential discriminative dictionary learning (SDDL) algorithm to separate the mixed signal in SCSS system. Besides the last layer, in each layer the dictionary contains two discriminative sub-dictionaries and one buffer sub-dictionary. The main motivation of this paper is that the unique components of the speech signals will be firstly separated, because the unique components of the speech signals have better correspondences to the speaker labels which can be explained by the corresponding discriminative sub-dictionaries. The similar components of the speech signals which are distributed in the buffer sub-dictionary will be separated in the next layer, if not, the similar components of the speech signals will be distributed in their cross discriminative sub-dictionary in the current layer. Then, the separated signals have better correspondences to the speaker labels compared to the DDL algorithms, so the source confusion can be suppressed. In the last layer the sequential dictionary only have the discriminative sub-dictionaries which ensure that the overall reconstruction error of the signals will be small; that is to say, the source distortion will not be introduced and meantime the source confusion will be suppressed. It should be noted that speech signals have better sparsity in the time-frequency (TF) domain compared to the time domain [12]. The phase angles of the signals in TF domain were usually ignored and only the information of signals' magnitude spectra were used [5,6]. Hence, the proposed algorithm is performed in TF domain.

2. Learn a sequential discriminative dictionary and speech separation

2.1. Signal modeling and the conventional DDL algorithm

Straightforwardly, SCSS can be treated as a problem with one equation and two unknown variables where we have

$$x(t) = s_1(t) + s_2(t), \quad 1 \le t \le T. \tag{1}$$

After applying a short time Fourier transform (STFT) on both sides of (1) and only using the magnitude [5,6], it can be expressed in the TF domain as follows:

$$|X(t,f)| \approx |S_1(t,f)| + |S_2(t,f)| \tag{2}$$

where t and f are the time and frequency index; $|S_1(t,f)|$ and $|S_2(t,f)|$ are the unknown STFT magnitudes of the first and second sources in the mixed signal. The magnitude spectra can be written in a matrix form as follows:

$$\mathbf{X} \approx \mathbf{S}_1 + \mathbf{S}_2. \tag{3}$$

In the training stage, the clean speech signals' magnitude spectra are used as training sets to learn the discriminative sub-dictionary $\mathbf{D}_i$, $i = 1, 2$ and they are concatenated into a structured dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2]$.

In the separation stage, sparse coding is exploited in the source separation module: the mixed speech signal $\mathbf{X}$ is sparsely represented over the structured dictionary.

$$\mathbf{X} = \mathbf{D} \times \mathbf{C} = [\mathbf{D}_1, \mathbf{D}_2] \times \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix} \tag{4}$$

where $\mathbf{C}$ denotes the matrix of the sparse coding coefficients. The conventional DDL based SCSS algorithms expect that the underlying signals can be approximately represented by their corresponding sub-dictionaries alone when the mixed signal is sparsely represented over the structured dictionary. So the mixed signal is separated by corresponding sub-dictionaries and coding coefficients as follows:

$$\hat{\mathbf{S}}_i = \mathbf{D}_i \times \mathbf{C}_i \tag{5}$$

where $\hat{\mathbf{S}}_i$ denotes the separated magnitude spectra of the ith source.

2.2. Sequential discriminative dictionary learning

According to the descriptions in Section 1, the conventional DDL algorithms ignore the similar components of analysis frames from different speakers, which introduce the source confusion. To solve the problem, a sequential discriminative dictionary based on layered learning, containing two discriminative sub-dictionaries and a buffer sub-dictionary, is designed. And the similar components of the training sets should be distributed over the buffer sub-dictionary rather than the cross discriminative sub-dictionary in each layer. If the energy of the components that corresponding to the buffer sub-dictionary is below a certain threshold, then, the

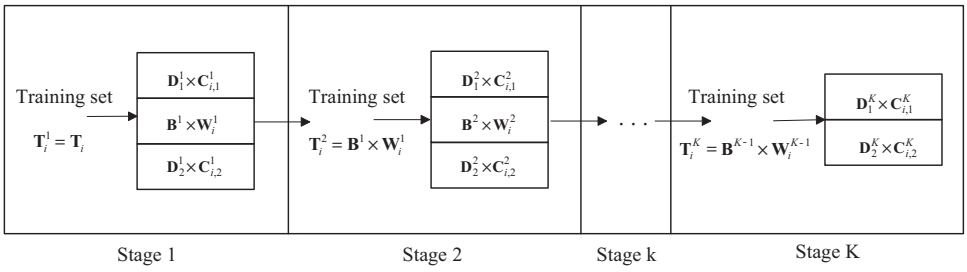

Fig. 2. Sequential discriminative dictionary learning scheme.

layered learning procedure stops and the dictionary should only have two discriminative sub-dictionaries. In this way, the discriminative sub-dictionaries have better correspondences to the speaker labels in each layer. The proposed sequential discriminative dictionary learning scheme is shown in Fig. 2.

In the kth layer, we define the sequential discriminative dictionary as $\mathbf{D}^k = [\mathbf{D}_1^k, \mathbf{B}^k, \mathbf{D}_2^k]$, where $\mathbf{D}_i^k$ denotes the discriminative sub-dictionary; $\mathbf{B}^k$ denotes the buffer sub-dictionary. The training set of the ith source is $\mathbf{T}_i^k$. The coding coefficient matrix of $\mathbf{T}_i^k$ over $\mathbf{D}^k$ is $\mathbf{E}_i^k$, where $\mathbf{E}_i^k = [(\mathbf{C}_{i,1}^k)^T, (\mathbf{W}_i^k)^T, (\mathbf{C}_{i,2}^k)^T]^T$, superscript T denotes the transpose of a vector or a matrix; $\mathbf{C}_{i,i}^k$ denotes the coding coefficient matrix corresponding to the discriminative sub-dictionary $\mathbf{D}_i^k$; $\mathbf{W}_i^k$ denotes the coding coefficient matrix corresponding to the buffer sub-dictionary $\mathbf{B}^k$; $\mathbf{C}_{i,j}^k, i \neq j$ denotes the cross coefficients matrix corresponding to the cross discriminative sub-dictionary $\mathbf{D}_j^k$.

The training set $\mathbf{T}_i^k$ is the ith signal distributed in the buffer sub-dictionary in the $(k-1)$th layer,

$$\begin{cases} \mathbf{T}_i^k = \mathbf{B}^{k-1} \times \mathbf{W}_i^{k-1} & \text{if} k \neq 1 \\ \mathbf{T}_i^k = \mathbf{T}_i & \text{if} k = 1 \end{cases} \tag{6}$$

where $\mathbf{T}_i$ denotes the clean speech signals' magnitude spectra of the ith speaker. In order to learn a dictionary that satisfies the requirements, an objective function is designed as follows:

$$\mathbf{D}^k = \arg\min_{\mathbf{D}^k} J_k \tag{7}$$

$$J_k = \sum_{i=1}^{2} ||\mathbf{T}_i^k - \mathbf{D}^k \times \mathbf{E}_i^k||_F^2 + \alpha \sum_{i=1}^{2} ||\mathbf{T}_i^k - \mathbf{D}_i^k \times \mathbf{C}_{i,i}^k||_F^2 + \beta \sum_{i,j=1, i \neq j}^{2} ||\mathbf{D}_j^k \times \mathbf{C}_{i,j}^k||_F^2 \tag{8}$$

where $||\bullet||_F^2$ denotes the Frobenius norm. The first term of (8) indicates the reconstruction error of $\mathbf{T}_i^k$ and it should be as small as possible to ensure that the source distortion would be small; the third term indicates the cross error of $\mathbf{T}_i^k$ over the cross discriminative sub-dictionary $\mathbf{D}_j^k$ and it should be as small as possible to ensure that the source confusion would be small. The Eq. (8) indicate that the similar components of the training sets should be distributed in the buffer sub-dictionary $\mathbf{B}^k$ rather than the cross discriminative sub-dictionaries; on this basis, the second term indicates that the training set $\mathbf{T}_i^k$ could be as much as possible explained by the corresponding discriminative sub-dictionary. It is not required that the training sets $\mathbf{T}_i^k$ can be totally reconstructed by the corresponding discriminative sub-dictionary in the current layer, but the reconstruction error of $\mathbf{T}_i^k$ over the dictionary $\mathbf{D}^k$ should be as small as possible, so the weight of the second term in (8) should be smaller than the first term. In each layer, the source confusion and the reconstruction error should be minimized simultaneously, so the weight of the third term of (8) should be equal to the first term.

Dictionary learning is realized by decomposing a data matrix $\mathbf{Y} \in \mathbf{R}^{M \times N}$ into a dictionary matrix $\mathbf{D} \in \mathbf{R}^{N \times L}$ ($L \gg N$) and a coding matrix $\mathbf{E} \in \mathbf{R}^{L \times N}$ [13], by solving the following optimization problem:

$$\min_{\mathbf{D},\mathbf{E}} ||\mathbf{Y} - \mathbf{D} \times \mathbf{E}||_2^2, \tag{9}$$

subject to a sparsity constraint on $\mathbf{E}$ and a unit constraint on $\mathbf{D}$. However, matrix factorization is a difficult problem, since the joint optimization of $\mathbf{E}$ and $\mathbf{D}$ is non-convex. Iterative solvers may yield locally optimal solutions, by alternating between the coding update and the dictionary update. Without loss of generality, we will take the kth layer as an example.

2.2.1. Coding update stage

In the coding update stage, the dictionary is fixed and the coding coefficient matrix is updated by solving the following optimization problem.

$$\min ||\mathbf{E}_i^{k,n}||_1 \quad \text{subject to} \quad \mathbf{T}_i^k = \mathbf{D}^{k,n-1} \times \mathbf{E}_i^{k,n} \tag{10}$$

where $||\mathbf{E}_i^{k,n}||_1$ denotes the sparsity constraint term [14]; $\mathbf{D}^{k,n-1}$ denotes the dictionary of the kth layer in the $(n-1)$th iteration. If $n=1$, $\mathbf{D}^{k,0}$ is the initial dictionary which contains three sub-dictionaries $\mathbf{D}_1^{k,0}$, $\mathbf{D}_2^{k,0}$ and $\mathbf{B}^{k,0}$. The initial sub-dictionaries $\mathbf{D}_i^{k,0} \in \mathbf{R}^{N \times L}$ are trained by the K-SVD algorithm [15], and the initial buffer sub-dictionary $\mathbf{B}^{k,0} \in \mathbf{R}^{N \times L}$ should be coherent to signals from different speakers simultaneously. The algorithm to obtain the initial dictionary is described in Table 1. The components which are corresponding to the initial buffer sub-dictionary are formulated as follows:

$$\mathbf{T}_{i(b)}^k = \mathbf{B}^{k,0} \times \mathbf{W}_i^k. \tag{11}$$

If $||\mathbf{T}_{i(b)}^k||_2^2 \leq \varepsilon$, in the kth layer, the dictionary $\mathbf{D}^k$ only contains two discriminative sub-dictionaries, and $K = k$; otherwise, the dictionary $\mathbf{D}^k$ contains two discriminative sub-dictionaries and one buffer sub-dictionary; where ε denotes threshold; K denotes the number of layer of the structured dictionary. In this stage the basis pursuit (BP) [16] algorithm is used to solve the coding update problem in (10).

Table 1
Algorithm 1: The algorithm to obtain the initial dictionary.

Input: The training set $\mathbf{T}_1$ and $\mathbf{T}_2$
Output: The initial dictionary $\mathbf{D}^0=[\mathbf{D}_1^0,\mathbf{B}^0,\mathbf{D}_2^0]$.
Step 1: For the training set $\mathbf{T}_1$ and $\mathbf{T}_2$, use K-SVD algorithm to obtain the initial sub-dictionary $\mathbf{D}_1^0$ and $\mathbf{D}_2^0$
Step 2: Calculate the mean for each row of $(\mathbf{D}_1^0)^T\times\mathbf{T}_2$ and $(\mathbf{D}_2^0)^T\times\mathbf{T}_1$, denote the means as $\mathbf{m}_1$ and $\mathbf{m}_2$.
Step 3: Obtain the initial buffer sub-dictionary.
Find the $L/2$ largest value elements of the $\mathbf{m}_1$ and $\mathbf{m}_2$, and the index is $\mathbf{i}_1$ and $\mathbf{i}_2$ respectively.
$\mathbf{B}^0=[\mathbf{D}_1^0(:,\mathbf{i}_1),\mathbf{D}_2^0(:,\mathbf{i}_2)]$ where $\mathbf{D}_1^0(:,\mathbf{i}_1)$ denotes the $\mathbf{i}_1$th column of $\mathbf{D}_1^0$, $\mathbf{D}_2^0(:,\mathbf{i}_2)$ denotes the $\mathbf{i}_2$th column of $\mathbf{D}_2^0$
End

2.2.2. Dictionary update stage

In the dictionary update stage, the coding matrix is fixed and the dictionary is updated by solving the following optimization problem.

$$\mathbf{D}^{k,n}=\min_{\mathbf{D}^{k,n}} J_k, \tag{12}$$

Furthermore, we introduce selecting matrix $\mathbf{Q}_i\in\mathbf{R}^{L\times L}$ $(i=1,2)$ to optimize the sub-dictionaries jointly, where $\mathbf{Q}_1=\begin{bmatrix}\mathbf{I}&\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{0}&\mathbf{0}\end{bmatrix}$, $\mathbf{Q}_2=\begin{bmatrix}\mathbf{0}&\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{0}&\mathbf{I}\end{bmatrix}$; while $k=K$, $\mathbf{Q}_1=\begin{bmatrix}\mathbf{I}&\mathbf{0}\\\mathbf{0}&\mathbf{0}\end{bmatrix}$, $\mathbf{Q}_2=\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{0}&\mathbf{I}\end{bmatrix}$; $\mathbf{0}$ denotes zero matrix whose entries are all zeroes; $\mathbf{I}$ denotes the identity matrix. Then J_k can be reformatted as follows:

$$\begin{aligned}J_k=&\sum_{i=1}^{2}||\mathbf{T}_i^k-\mathbf{D}^{k,n}\times\mathbf{E}_i^{k,n}||_F^2+\alpha\sum_{i=1}^{2}||\mathbf{T}_i^k-\mathbf{D}^{k,n}\times\mathbf{Q}_i\\&\times\mathbf{E}_i^{k,n}||_F^2+\beta\sum_{i,j=1,i\neq j}^{2}||\mathbf{D}^{k,n}\times\mathbf{Q}_j\times\mathbf{E}_i^{k,n}||_F^2.\end{aligned} \tag{13}$$

It is very difficult to find the global minimizer of J_k. Here, the limited-memory BFGS algorithm (L-BFGS) [17] can be used to solve the large scale optimization problem described in (12). The gradient of J_k is as follows:

$$\begin{aligned}\frac{\partial J_k}{\partial\mathbf{D}^{k,n}}=&2\sum_{i=1}^{2}(\mathbf{D}^{k,n}\times\mathbf{E}_i^{k,n}\times(\mathbf{E}_i^{k,n})^T-\mathbf{T}_i^k\times(\mathbf{E}_i^{k,n})^T)\\&+2\alpha\sum_{i=1}^{2}(\mathbf{D}^{k,n}\times\mathbf{Q}_i\times\mathbf{E}_i^{k,n}\times(\mathbf{Q}_i\times\mathbf{E}_i^{k,n})^T-\mathbf{T}_i^k\\&\times(\mathbf{Q}_i\times\mathbf{E}_i^{k,n})^T)+2\beta\sum_{i,j=1,2,i\neq j}^{2}\mathbf{D}^{k,n}\\&\times\mathbf{Q}_j\times\mathbf{E}_i^{k,n}\times(\mathbf{Q}_j\times\mathbf{E}_i^{k,n})^T\end{aligned} \tag{14}$$

which is required in L-BFGS algorithm.

2.3. Speech separation

In the separation stage, the mixed signal $\mathbf{X}$ is sparsely represented by the first layer dictionary $\mathbf{D}^1$ and the similar component which distributed in the buffer sub-dictionary $\mathbf{B}^1$ will be separated in the next layer. In the kth $(k>1)$ layer, the unseparated mixed signal $\mathbf{X}^k$ is defined as follows:

$$\mathbf{X}^k=\mathbf{B}^{k-1}\times\mathbf{W}_S^{k-1} \tag{15}$$

and sparsely represented over the structured dictionary $\mathbf{D}^k$, where $\mathbf{W}_S^{k-1}$ denotes the coefficients matrix corresponding to the buffer sub-dictionary $\mathbf{B}^{k-1}$.

$$\mathbf{X}^k=\begin{cases}\mathbf{D}^k\times\mathbf{C}_S^k=[\mathbf{D}_1^k,\mathbf{B}^k,\mathbf{D}_2^k]\times\begin{bmatrix}\mathbf{C}_{S_1}^k\\\mathbf{W}_S^k\\\mathbf{C}_{S_2}^k\end{bmatrix}, & k=1,\ldots,K-1\\\mathbf{D}^K\times\mathbf{C}_S^K=[\mathbf{D}_1^K,\mathbf{D}_2^K]\times\begin{bmatrix}\mathbf{C}_{S_1}^K\\\mathbf{C}_{S_2}^K\end{bmatrix}, & k=K\end{cases} \tag{16}$$

where the $\mathbf{C}_S^k$ denotes the sparse coefficients matrix of $\mathbf{X}^k$ over $\mathbf{D}^k$; $\mathbf{C}_{S_1}^k$, $\mathbf{C}_{S_2}^k$ denote the sparse coefficients matrix corresponding to the discriminative sub-dictionaries $\mathbf{D}_1^k$ and $\mathbf{D}_2^k$, respectively. The component of the ith source in the kth layer is estimated as follows:

$$\hat{\mathbf{S}}_i^k=\mathbf{D}_i^k\times\mathbf{C}_{S_i}^k. \tag{17}$$

The sum of the components in all layers is taken as the final estimate of the ith source, and the separated underlying sources will have better correspondence to their corresponding discriminative sub-dictionaries.

$$\hat{\mathbf{S}}_i=\sum_{k=1}^{K}\hat{\mathbf{S}}_i^k=\sum_{k=1}^{K}\mathbf{D}_i^k\times\mathbf{C}_{S_i}^k \tag{18}$$

Supervised dictionary learning algorithms involve the common problem that is the mismatch between training set and test set. In order to suppress the influence of the mismatching, the dictionary self-learning process [18], namely adaptive separation, is used. The overall algorithm is described in Table 2. The overall structure of the proposed algorithm is shown in Fig. 3.

3. Experiment result

The performance of SDDL is analyzed through experiments in this section. Firstly, the data and the performance metrics used in the experiments are considered. Secondly, the proposed SDDL, NMF [9], CMF [8], supervised PLCA [7] and single-layer DDL [10] based SCSS algorithms are simulated simultaneously for comparison to verify the effectiveness of the proposed algorithm.

Table 2
The overall algorithm of SDDL for SCSS.

Input: The mixed signal $x(t)$, the training sets $\mathbf{T}_1$ and $\mathbf{T}_2$, the iteration number of SDDL: N, the maximum layer of SDDL: K.
Output: The separated speech signals.

Step 1: Learn the sequential discriminative dictionary:
for $k=1:K$
Use algorithm 1 to obtain the initial dictionary $\mathbf{D}^{k,0}$.
Use BP algorithm to calculate the coefficient of $\mathbf{T}_i^k$ over the initial dictionary $\mathbf{D}^{k,0}$ by solving (11), if $||\mathbf{T}_{i(b)}^k||_2^2 \le \varepsilon$, $\mathbf{D}^k=[\mathbf{D}_1^k,\mathbf{D}_2^k], K=k$; else $\mathbf{D}^k=[\mathbf{D}_1^k,\mathbf{D}^k,\mathbf{D}_2^k]$.
for $n=1:N$
Coding update stage:
Use BP algorithm to calculate the coefficient of $\mathbf{T}_i^k$ over the dictionary $\mathbf{D}^{k,(n-1)}$ by solving (11).
Dictionary update stage:
Use L-BFGS algorithm to obtain the sequential discriminative dictionary $\mathbf{D}^{k,n}$ by solving (12).
end for
$\mathbf{D}^k \leftarrow \mathbf{D}^{k,N}$.
end for
Step 2: Speech separation:
Use STFT to obtain the magnitude spectra $\mathbf{X}$ of the mixed signal $x(t)$.
Obtain the separated magnitude spectra $\hat{\mathbf{S}}_1$ and $\hat{\mathbf{S}}_2$ by using (15)–(18).
Use the dictionary self-learning algorithm post processing the separated magnitude spectra $\hat{\mathbf{S}}_1$ and $\hat{\mathbf{S}}_2$.
Use inverse short time Fourier transform (ISTFT) and the mixed signals' phase angles to recover the separated signal in time domain.
End

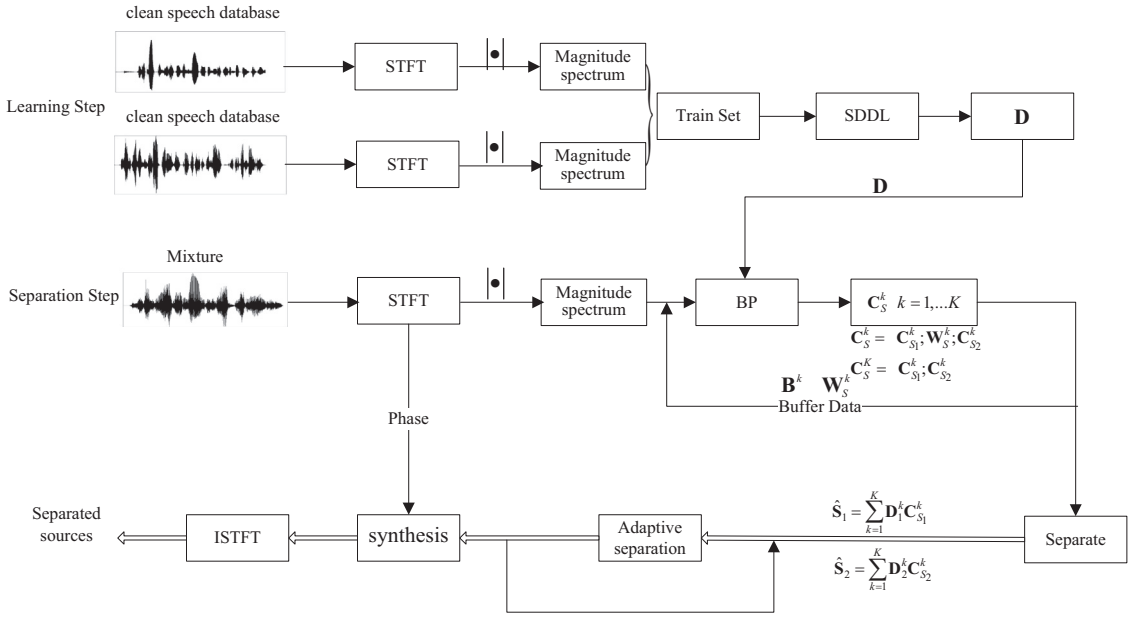

Fig. 3. The overall structure of the proposed algorithm.

3.1. Data and performance metrics

All the speech signals used in our experiments come from TIMIT [19] with the sampling rate of 16 kHz. Speech samples are arbitrarily taken from 4 females and 4 males including 10 sentences for each speaker. Moreover, 2 sentences are chosen as test set and the left 8 sentences are chosen as training set for each speaker.

Three kinds of metrics: signal to distortion ratio (SDR) [20], signal to interference ratio (SIR) [20] and perceptual evaluation of speech quality (PESQ) [21] are used to reflect the overall source separation quality. The higher SDR and SIR indicate that the separated speeches are less distortion and less confusion. The higher PESQ indicates that the separated speeches have better speech quality.

Some issues in the experiments are described as follows. The number of training samples is fixed as 2000. The length of analysis window for STFT is 512; because the conjugate symmetry properties of STFT, the size of initial sub-dictionary is 257×512. To simplify the problem, we assuming that each atom of the dictionary has the same effect on signal distortion, because $\mathbf{D}$ is three times the size of $\mathbf{D}_i$, in order to ensure that the cross error has the same weight of the reconstruction error, the parameter $\beta=3$; because the weight of the second term of (8) should be smaller than the first term, we choose $\alpha=0.5$ and the

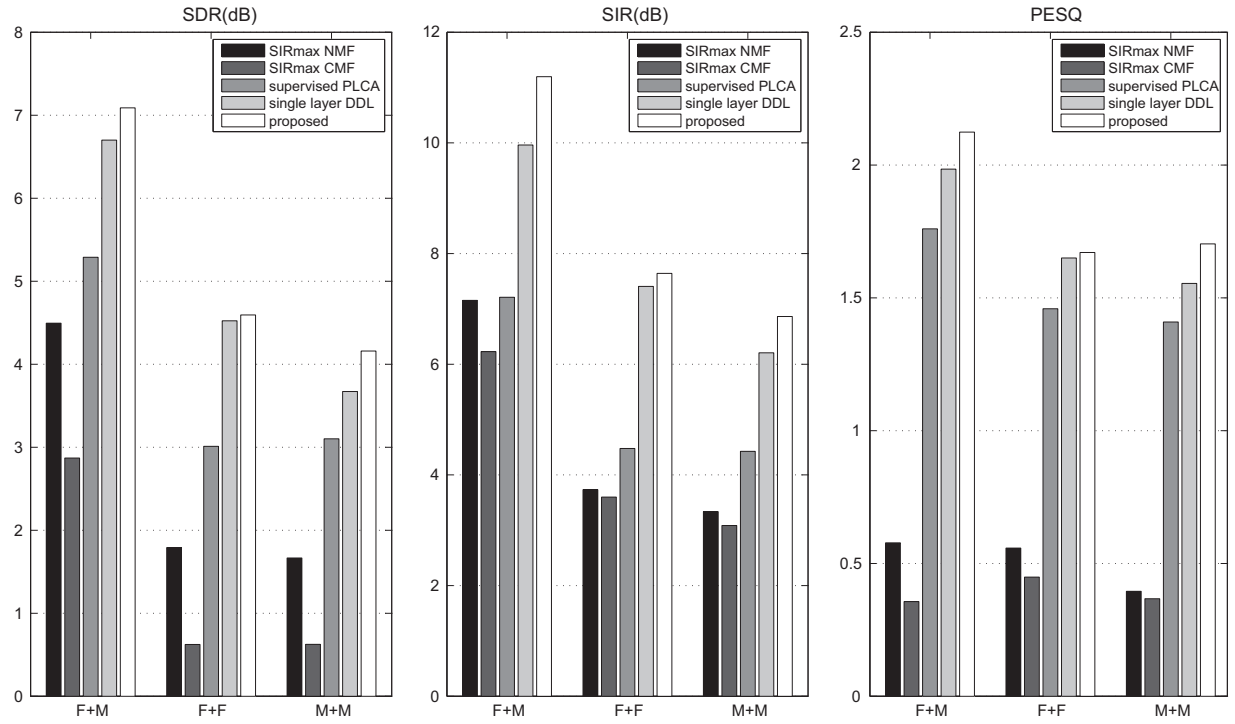

Fig. 4. Performance comparison among NMF, CMF, supervised PLCA, single layer DDL and SDDL based SCSS algorithms for different genders of speakers.

parameters α and β may be not optimal. The threshold ε is chosen as $0.05 \times \min(||\mathbf{T}_1||_2^2, ||\mathbf{T}_2||_2^2)$.

3.2. The performance comparison

The baseline DDL algorithm in the experiment is performed in TF domain and the size of the discriminative dictionary is 257×1536. The results of NMF and CMF based SCSS are obtained by the toolbox [22] provided by Brian King, using the 'sirmax' configuration. The final results are the average values for 4 trials of different mixtures, i.e., female and male, female and female, male and male.

Compared to the single layer DDL algorithm, the SDDL algorithm firstly separate the unique components of the signals and the similar components of the signals which will introduce source confusion are separated in the next layer. From Fig. 4 one can find that compared to the single layer DDL algorithm the SDDL algorithm can obtain 0.1 improvement of PESQ, 0.5 dB improvement of SDR and 1 dB improvement of SIR for different gender of speakers, which gives evidence that the SDDL algorithm can suppress the speech confusion without introducing source distortion. When compared to supervised PLCA and single layer DDL algorithms, the improvement of the metric SDR for the gender combination F+F is very small, but the metrics SIR and PESQ have significant improvement. In other words, the proposed algorithm can suppress the speech confusion and improve the speech quality. Overall, the experiment results verify the effectiveness of the proposed algorithm compared to the existing algorithms.

4. Conclusion

An algorithm termed SDDL has been proposed to solve the SCSS problem. The sequential discriminative dictionary with the discriminative and buffer sub-dictionaries is learned. The proper objective function is derived to ensure that the unique and the similar components of the signals can be explained by the corresponding discriminative sub-dictionary and the buffer sub-dictionary and in the last layer the dictionary only contains two discriminative sub-dictionaries, so the source confusion can be suppressed without introducing the source distortion. Experiment results verify the advantage of the proposed algorithm compared to the tested SCSS algorithms.

References

[1] N. Tengtrairat, W.L. Woo, Extension of DUET to single-channel mixing model and separability analysis, Signal Process. 96 (2014) 261–265.
[2] S.T. Roweis, One microphone source separation, Adv. Neural Inf. Process. Syst. 13 (2001) 793–799.
[3] Y. Wang, K. Han, D.L. Wang, Exploring monaural features for classification-based speech segregation, IEEE Trans. Audio Speech Lang. Process. 21 (2) (2013) 270–279.
[4] P. Li, Y. Guan, B. Xu, W. Liu, Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech, IEEE Trans. Audio Speech Lang. Process. 14 (6) (2006) 2014–2023.
[5] E.M. Grais, H. Erdogan, Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation, in: Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH), August 2013.
[6] Y. Litvin, I. Cohen, D. Chazan, Monaural speech/music source separation using discrete energy separation algorithm, Signal Process. 90 (12) (2010) 3147–3163.

[7] P. Smaragdis, B. Raj, M. Shashghanka, Supervised and semi-supervised separation of sounds from single-channel mixtures, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2007, pp. 414–421.
[8] B.J. King, L. Atlas, Single-channel source separation using complex matrix factorization, IEEE Trans. Audio Speech Lang. Process. 19 (8) (2011) 2591–2597.
[9] M. Schmidt, R. Olsson, Single-channel speech separation using sparse non-negative matrix factorization, in: Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH), September 2006.
[10] G. Bao, Y. Xu, Z. Ye, Learning a discriminative dictionary for single-channel speech separation, IEEE Trans. Audio Speech Lang. Process. 22 (7) (2014) 1130–1138.
[11] X. Huang, A. Acero, H.W. Hon, Spoken Language Processing, Prentice-Hall, Inc., New York, NJ, USA, 2001.
[12] Y. Li, S.I. Amari, A. Cichocki, D.W.C. Ho, Underdetermined blind source separation based on sparse representation, IEEE Trans. Signal Process. 54 (2) (2006) 423–437.
[13] C.D. Sigg, T. Dikk, M. Joachim, J.M. Buhmann, Speech enhancement with sparse coding in learned dictionaries, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010, pp. 4758–4761.
[14] N. Hurley, S. Rickard, Comparing measures of sparsity, IEEE Trans. Inf. Theory 55 (10) (2009) 4723–4741.
[15] M. Aharon, M. Elad, A. Bruckstein, K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[16] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1) (1998) 33–61.
[17] D.C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, Math. Program. 45 (1989) 503–528.
[18] X. Wei, G. Bao, Z. Ye, X. Xu, Compressed sensing based underdetermined blind source separation with unsupervised sparse dictionary self-learning, in: Signal Processing, Communication and Computing (ICSPCC), August 2013.
[19] V. Zue, S. Seneff, J. Glass, Speech database development at MIT: TIMIT and beyond, Speech Commun. 9 (4) (1990) 351–356.
[20] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE Trans. Audio Speech Lang. Process. 14 (4) (2006) 1462–1469.
[21] ITU-T P.835, Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, ITU-T Recommendation P.835, 2003.
[22] ⟨https://sites.google.com/a/uw.edu/isdl/projects/cmf-toolbox⟩.