Single-channel Speech Separation Using Orthogonal Matching Pursuit

Haiyan Guo
School of Information Science and Engineering, Southeast University, Nanjing, China;
College of Engineering, Nanjing Agricultural University, Nanjing, Jiangsu, China
Email: haiyan.guo@seu.edu.cn

Xiaoxiong Li, Lin Zhou and Zhenyang Wu
School of Information Science and Engineering, Southeast University, Nanjing, China;
Email: {220120726, linzhou, zhenyang}@ seu.edu.cn

***Abstract*—In this paper, we propose a new sparse decomposition based single-channel speech separation method using orthogonal matching pursuit (OMP). The separation is performed using source-individual dictionaries consisting of time-domain training frames as atoms. OMP is used to compute sparse coefficients to estimate sources. We report the separation results of our proposed method and compare them with a separation method based on sparse non-negative matrix factorization (SNMF) which is a classical sparse decomposition based separation method. Experiments show that our proposed method results in higher signal-to-noise ratio (SNR) and signal-to-interference ratio (SIR).**

***Index Terms*—Single-channel speech separation (SCSS), sparse decomposition, orthogonal matching pursuit (OMP), dictionary**

I. 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 [1-2]. Single-channel speech separation (SCSS) discussed in this paper is an extreme case, where only one mixture is known. 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 aims to recover underlying speech sources from a mixture. It is an ill-conditioned problem since the number of mixture is less than the number of sources.

Previous state-of-the-art SCSS approaches can be divided into two groups: source-driven method or method based on computational auditory scene analysis (CASA) [3], and model-driven method [4-6]. CASA-based method tries to achieve human performance in auditory scene analysis (ASA) based on the perceptual organization of sound. Ideal binary time-frequency (T-F) mask has been proposed as the main computational goal of CASA [7]. CASA generally consists of two major stages: segmentation and grouping. In the segmentation stage, the input mixture is decomposed into time-frequency cells dominated by one individual source. In the grouping stage, multiple segments are grouped into simultaneous streams, and subsequently streams are organized into whole streams corresponding to individual sources. The grouping principles in speech organization prominently used are harmonicity 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. It seeks discriminative features in the mixed signal for separation. However, in general, its separation performance is not as good as that of model-based method.

Model-driven method relies heavily on a priori knowledge about the speakers, hence generally outperforms CASA-based method. 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], Gaussian mixture model (GMM) [4] [8] [9-11], hidden Markov model (HMM) [6] and sinusoidal model [12]) or codebooks (e.g. independent component analysis (ICA) basis [5], VQ codebook [10] [12]) 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 [9], maximum likelihood (ML) estimation [5] [10], maximum a posterior (MAP) estimation [5] [9], etc. Though statistical model-driven method has been reported to be effective, its training is rather time consuming and estimation is significantly complex. In [12], every possible combination needs to be

Manuscript received January 27, 2014; revised February 22, 2014; accepted March 7, 2014.
Corresponding author: Haiyan Guo, haiyan.guo@seu.edu.cn.

© 2014 ACADEMY PUBLISHER
doi:10.4304/jsw.9.11.2974-2980

considered during distortion function minimization to find the optimal codebook atoms.

In model-driven SCSS method, sparsity has been proven to be useful for SCSS. In [5], a priori sets of ICA basis filters are learned for SCSS. The associated coefficients of ICA basis functions are made use of as a function of learning algorithm. Based on the observation that only a small number of coefficients of ICA basis functions differ significantly from zero, generalized Gaussian distribution is used. In [13], a sparse-distribute code of spectro-temproral basis functions where basis functions are more than the dimensionality of the space is generated , leading to better separation results than a compact code of basis functions extracted in [14]. On the other hand, due to sparsity, there is less overlap between the sparse decomposition coefficients of different sources, which means different sources are less likely to be simultaneously active in the sparse domain. Obviously, this feature is helpful for separation.

Decomposition based method is a more intuitive way to use sparsity to perform separation [15-18]. It generally works by two steps. First, personalized dictionaries are learned to give sparse representation of training signals. Second, sparse coefficients are obtained by computing sparse decomposition of the mixture on the union of learned dictionaries, and used to perform separation by combining dictionary atoms assigned to each speaker. In decomposition based method, sparsity is exploited to shrink the feasible solution region of the corresponding underdetermined problems, thus further simplifying the search for the optimal solution. Sparse non-negative matrix factorization (SNMF) is a classical sparse decomposition based SCSS method, and has achieved comparable performance [17-18]. Moreover, it has been proved in [19] that the unique sparse representation of a signal in a union of bases can be found by l_0 optimization if the union of bases and the unique sparse representation satisfy certain conditions. Motivated by this result, we expect to perform high quality single-channel speech separation based on sparse decomposition.

Recently, various methods have been proposed for sparse decomposition. The most typical methods are Basis Pursuit (BP) [20] and orthogonal matching pursuit (OMP) [21-22]. The principle of BP is to find a representation of a signal whose l_1 norm is minimal. A BP problem can be equally reformulated as a linear program and solved by linear programming (LP) [20]. OMP proposed in [21-22] is a recursive algorithm to compute sparse decomposition. It is a modification to the MP algorithm and leads to improved convergence. OMP achieves similar performance as BP, but more quickly. It is a greedy algorithm and selects atoms iteratively for signal recovery. At each iteration, an atom most correlated with the residual is chosen and then residual is updated by subtracting off the contribution of the chosen atom.

In this paper, we propose a sparse decomposition based separation method using OMP. A source-specific dictionary is generated as a matrix consisting of time-domain training frames of each speaker as columns termed as atoms. OMP is used to compute sparse coefficients in the union of source-specific dictionaries for the estimation of sources. Experiments show that the proposed separation method is effective since it results in higher signal-to-noise ratio (SNR) and higher signal-to-interference ratio (SIR) than the separation method using SNMF.

The remainder of this paper is structured as follows. In Section II, we introduce the general model of sparse decomposition based SCSS. In Section III, we introduce SCSS algorithm using OMP. The experiment results are reported in section IV. Finally, we conclude and give future perspectives in Section V.

II. Sparse Decomposition Based Scss

Consider the SCSS problem where the mixed signal $y(t)$ is the sum of two individual speech signals: $s_1(t)$ and $s_2(t)$. $y(t)$ is given as

$$y(t) = a_1 s_1(t) + a_2 s_2(t) \tag{1}$$

where $a_i (i = 1,2)$ is the gain of each source fixed over time. Note that the same mixture can be obtained by adjusting the value of a_i or the power of sources. Therefore, the problem in (1) can be simplified as the following equivalent mixing model

$$y(t) = s_1(t) + s_2(t) \tag{2}$$

It can be written in vector as

$$\vec{y} = \vec{s}_1 + \vec{s}_2 \tag{3}$$

where $\vec{y}$ and $\vec{s}_i (i = 1,2)$ denote the mixed signal and the i^{th} individual speech source in time-domain respectively.

Suppose that $\vec{y}$ can be sparsely represented in a known overcomplete dictionary $\mathbf{D}$ which is the concatenation of the source-individual dictionaries $\mathbf{D} = [\mathbf{D}_1 \quad \mathbf{D}_2]$. That is, [17]

$$\vec{y} = \mathbf{D}\vec{\theta} \tag{4}$$

where $\vec{\theta}$ is the sparse code which is the concatenation of the source-individual codes $\vec{\theta}_i (i = 1,2)$, that is, $\vec{\theta} = [\vec{\theta}_1^{\mathrm{T}} \quad \vec{\theta}_2^{\mathrm{T}}]^{\mathrm{T}}$. The sparsest representation of $\vec{y}$ in $\mathbf{D}$ can be found by solving the following problem,

$$\min \left\| \vec{\theta} \right\|_0, s.t. \quad \vec{y} = \mathbf{D}\vec{\theta} \tag{5}$$

where $\|\cdot\|_0$ denotes the l_0 norm of a vector. The problem (5) is equivalent to the following problem,

$$\min f(\vec{\theta}_1, \vec{\theta}_2) = \min \sum_{i=1}^{2} \left\| \vec{\theta}_i \right\|_0, s.t. \quad \vec{y} = \sum_{i=1}^{2} \mathbf{D}_i \vec{\theta}_i \tag{6}$$

© 2014 ACADEMY PUBLISHER

The solution of problem (5) is denoted as $\hat{\vec{\theta}}$, which is the concatenation of estimated source-individual codes $\hat{\vec{\theta}}_i (i=1,2)$, the solution of problem (6), that is, $\hat{\vec{\theta}} = \left[\hat{\vec{\theta}}_1^{\mathrm{T}} \quad \hat{\vec{\theta}}_2^{\mathrm{T}} \right]^{\mathrm{T}}$. If the source-individual dictionaries are diverse enough, we can separate $\vec{y}$ into its individual sources $\hat{\vec{s}}_i$ as [17]

$$\hat{\vec{s}}_i = \mathbf{D}_i \hat{\vec{\theta}}_i \tag{7}$$

As a consequence of above, there are two connected tasks to be solved in sparse decomposition based SCSS: learning source-individual dictionaries and computing sparse decomposition in (5). In this paper, we do not focus on dictionary learning and generate $\mathbf{D}_i (i=1,2)$ as the matrix consisting of the i^{th} speaker's training frames as columns called atoms. The proposed separation method using that dictionary is proved effective since it leads to a higher SNR and SIR in our experiments than the separation method based on SNMF which is a classical sparse decomposition based method. (The performance of separation using $\mathbf{D}_i$ generated as unsupervised clustering of training frames has also been tested, and it is much lower in SNR and SIR.)

For the computation of sparse decomposition, two classical approaches are used extensively: BP and OMP. The principle of BP is to find the optimal decomposition coefficients having the smallest l_1 norm. That is, one solves the problem [20]

$$\min \left\| \vec{\theta} \right\|_1, s.t. \quad \vec{y} = \mathbf{D}\vec{\theta} \tag{8}$$

where $\|.\|_1$ denotes the l_1 norm of a vector. In [23], it has been proven that, the solution to the problem (5) can be approximated by the solution of the problem (8). Since problem (5) is NP-hard and difficult to solve, one turns to solve the problem (6) to obtain an approximate solution to (5). The solution of BP problem (6) can be obtained by solving an equivalent linear program as [20]

$$\min \vec{c}^{\mathrm{T}} \vec{z}, s.t. \quad \mathbf{A}\vec{z} = \vec{b} \tag{9}$$

where

$$\mathbf{A} \Leftrightarrow (\mathbf{D}, \quad -\mathbf{D}) ; \ \vec{b} \Leftrightarrow \vec{y} ; \ \vec{c} \Leftrightarrow (1; \quad 1) ; \ \vec{z} \Leftrightarrow (\vec{u}; \quad \vec{v}) ;$$

$$\vec{\theta} \Leftrightarrow \vec{u} - \vec{v} .$$

OMP is a sparse approximation algorithm which is not based on optimization. It is a recursive algorithm to compute coefficients of atoms in a dictionary which is nonorthogonal and possibly overcomplete. It is a modification to MP by maintaining orthogonality of residual at each iteration, thus leads to improved convergence. The major advantage of OMP is its speed and ease of implementation. It achieves comparable performance as BP. The recovery of $\vec{y}$ based on sparse decomposition with OMP is given as follows [21-22].

Algorithm (Signal recovery with OMP)

INPUT:
Dictionary $\mathbf{D}$
Mixed signal $\vec{y}$
The sparsity level m of $\vec{y}$
OUTPUT:
An estimate $\hat{\vec{y}}$ of mixed signal
A set $\mathbf{D}^m$ containing chosen atoms to approximate $\vec{y}$
A sparse approximation $\hat{\vec{\theta}}^m$ of $\vec{y}$
A residual $\vec{r}^{\,m} = \vec{y} - \mathbf{D}^m \hat{\vec{\theta}}^m$

PROCEDURE:

1) Initialize the residual $\vec{r}^{\,0} = \vec{y}$, the matrices of chosen atoms $\mathbf{D}^0 = \varnothing$, the iteration counter $t = 0$.

2) Find the index that solves the optimization problem [21-22]

$$\lambda^t = \arg \quad \max_{j=1,\ldots,K} \left| \left\langle \vec{r}^{\,t}, \vec{d}_k \right\rangle \right| \tag{10}$$

where $\vec{d}_k$ is the k^{th} atom in $\mathbf{D}$, and K is the number of atoms in $\mathbf{D}$. Select $\vec{d}_{\lambda_t}$ as the newly selected atom.

3) Argument the matrix of chosen atoms $\mathbf{D}^t = \left[\mathbf{D}^{t-1} \quad \vec{d}_{\lambda^t} \right]$.

4) Solve a least squares problem to obtain a new approximation of $\vec{y}$ supported in $\mathbf{D}^t$ [21-22],

$$\vec{\theta}^t = \arg\min_{\vec{\theta}} \left\| \vec{y} - \mathbf{D}^t \vec{\theta} \right\|_2 \tag{11}$$

5) Calculate residual as [21-22],

$$\vec{r}^{\,t+1} = \vec{y} - \hat{\vec{\theta}}^t \mathbf{D}^t \tag{12}$$

6) Increment t, and return to step 2) if $t < m$.

The mixed signal is estimated by $\hat{\vec{y}} = \mathbf{D}^m \hat{\vec{\theta}}^m$.

As stated above, OMP picks atoms to recover the original signal in a greedy fashion. At each iteration, an atom that is most strongly correlated with the remaining part of the original signal called residual is chosen from

© 2014 ACADEMY PUBLISHER

the dictionary, and its contribution to the residual is subtracted off for the next iteration.

III. Proposed Separation Method

We will now proceed to describe the proposed new sparse decomposition based separation method using OMP. Fig.1 shows the block diagram of the proposed separation algorithm.

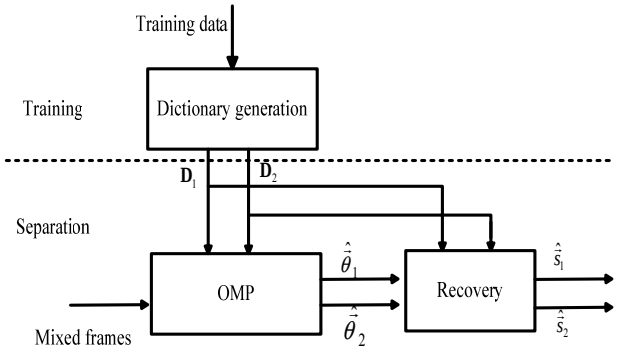


Fig.1. Block algorithm of proposed speech separation method using OMP.

As stated in Fig.1, the proposed algorithm works in two stages: training and separation. In the training stage, two source-individual dictionaries consisting of training frames of individual speakers as atoms are generated. In the separation stage, separation is performed frame after frame and then speech is synthesized by overlap-adding. Each mixed frame is separated based on sparse decomposition using OMP algorithm proposed.

In the following, we give the separation algorithm using OMP.

Algorithm (Separation using OMP)

INPUT:

Source-individual dictionaries $\mathbf{D}_1, \mathbf{D}_2$

Mixed frame $\vec{y}$

Stopping criterion

OUTPUT: (suppose procedure stops after m iterations):

Two estimates $\hat{\vec{s}}_1, \hat{\vec{s}}_2$ for original individual source frames $\vec{s}_1, \vec{s}_2$

Two sets $\mathbf{D}_1^m, \mathbf{D}_2^m$ containing chosen atoms to approximate $\vec{s}_1, \vec{s}_2$

Two decomposition coefficients $\hat{\vec{\theta}}_1^m, \hat{\vec{\theta}}_2^m$ supported in $\mathbf{D}_1^m, \mathbf{D}_2^m$ to approximate $\vec{s}_1, \vec{s}_2$

A residual $\vec{r}^m = \vec{y} - \left[\mathbf{D}_1^m \quad \mathbf{D}_2^m\right]\left[\hat{\vec{\theta}}_1^{m\mathrm{T}} \quad \hat{\vec{\theta}}_2^{m\mathrm{T}}\right]^{\mathrm{T}}$

PROCEDURE:

1) Initialize the residual $\vec{r}^0 = \vec{y}$, the matrices of chosen atoms $\mathbf{D}_1^0 = \varnothing, \mathbf{D}_2^0 = \varnothing$, the iteration counter $t = 0$.

2) Find the index that solves the optimization problem (10).

3) Merge the newly selected atom with the previous matrices of chosen atoms,

$$\mathbf{D}_1^t = \begin{cases} \left[\mathbf{D}_1^{t-1} \quad \vec{d}_{\lambda^t}\right], & \text{if} \quad \lambda^t \le K_1 \\ \mathbf{D}_1^{t-1}, & \text{others} \end{cases} \tag{13}$$

$$\mathbf{D}_2^t = \begin{cases} \left[\mathbf{D}_2^{t-1} \quad \vec{d}_{\lambda_t}\right], & \text{if} \quad K_1 \le \lambda^t \le K_1 + K_2 \\ \mathbf{D}_{t-1}^2, & \text{others} \end{cases} \tag{14}$$

where K_i is the number of atoms in $\mathbf{D}_i$, satisfying $K = \sum_{i=1}^{2} K_i$.

4) Solve the least squares problem (11) to obtain a new decomposition coefficient to approximate $\vec{y}$ supported in $\mathbf{D}^t$ [21-22], the concentration of $\mathbf{D}_1^t$ and $\mathbf{D}_2^t$, $\mathbf{D}^t = \left[\mathbf{D}_1^t \quad \mathbf{D}_2^t\right]$. The solution of (11) is given by $\hat{\vec{\theta}}^t = \left(\mathbf{D}^t\left(\mathbf{D}^t\right)^{\mathrm{T}}\right)^{-1}\mathbf{D}^t\vec{y}$ [21-22]. We denote $\hat{\vec{\theta}}_1^t$ and $\hat{\vec{\theta}}_2^t$ as the parts of $\hat{\vec{\theta}}^t$ belonging to the support $\mathbf{D}_1^t$ and $\mathbf{D}_2^t$ respectively.

5) Update residual as (12).

6) Increment t, and return to step 2) until satisfying $\left\|\vec{r}^t\right\|_2 \le \delta_e$ or $\max_{j=1,\ldots,K}\left|\left\langle \vec{r}^t, \vec{d}_k \right\rangle\right| \le \delta_c$ where δ_e and δ_c are chosen thresholds.

The separated speech source is estimated by $\hat{\vec{s}}_i = \hat{\vec{\theta}}_i^m \mathbf{D}_i^m (i = 1,2)$.

IV. Experiments

As a proof of concept, we evaluate the proposed separation algorithm using the Grid corpus provided for SCSS by Cooke et. al [24] and compare its performance with SNMF based SCSS method which is a classical sparse decomposition based method [17-18]. 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 kHz to 8 kHz and a

© 2014 ACADEMY PUBLISHER

hamming window of duration 32 ms with a frame-shift of 16 ms was used.

To evaluate the separation performance, average of signal-to-noise ratio (SNR) and average of source-to-interferences ratio (SIR) are used. SNR is defined as

$$\mathrm{SNR}_i = 10 \times \lg\left(\frac{\vec{x}_i \vec{x}_i^{\mathrm{T}}}{(\vec{x}_i - \hat{\vec{x}}_i)(\vec{x}_i - \hat{\vec{x}}_i)^{\mathrm{T}}}\right) \quad (15)$$

where $\vec{x}_i$ and $\hat{\vec{x}}_i$ are original and estimated source speech signals respectively, and SNR_i is the SNR of estimated $\hat{\vec{x}}_i$. SIR is defined as in [25].

In our experiments we set $\delta_e = 10^{-10}$ and $\delta_c = 10^{-5}$.

We compare the separation results of the proposed method using OMP with that of the separation method using SNMF in SNR and SIR respectively. The average results of 600 mixtures tested are shown in TABLE I and II. 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 [36].

TABLE I
PERFORMANCE OF SEPARATION USING SNMF AND PROPOSED SEPARATION METHOD USING OMP IN SNR (DB) ON 600 MIXTURES

	SNMF	Proposed method
F/F	4.7/4.5	4.8/4.7
F/M	5.5/5.3	5.7/5.8
M/M	3.3/3.9	3.7/4.4

TABLE II
PERFORMANCE OF SEPARATION USING SNMF, TRADITIONAL OMP AND OMP IN SIR (DB) ON 600 MIXTURES

	SNMF	Proposed method
F/F	5.0/6.3	8.7/11.4
F/M	9.3/8.7	12.7/11.7
M/M	3.3/4.9	8.9/8.9

As shown in TABLE I and II, the proposed separation method outperforms the method using SNMF in both SNR and SDR. The average SNR result of the proposed method is 0.15dB, 0.35dB and 0.45dB higher than that of the method using SNMF for female/female, female/male and male/male mixture respectively. The average SIR result of the proposed method is 4.4dB, 3.2dB and 4.8dB higher than that of the method using SNMF for female/female, female/male and male/male mixture respectively.

We also report the separation results of the sentences which are shown in TABLE III.

TABLE III
LABELS OF SPEAKERS AND FILE NAMES USED FOR TESTING

F1	speaker 18	"lwix2s" "sbil4a" "prah4s"
F2	speaker 20	"lwwy2a" "sbil2a" "prbu5p"
M1	speaker 1	"pbbv6n" "sbwozn" "prwkzp"
M2	speaker 2	"lwwm2a" "sgai7p" "priv3n"

For each speech pair, the speech sources are added directly to form a mixture, resulting 54 mixtures including 36 male-female mixtures, 9 male-male mixtures and 9 female-female mixtures.

Fig.2. shows the first 16 atoms chosen using OMP to separate a mixed frame. In this example, 70 and 51 atoms are selected to estimate two sources based on OMP.

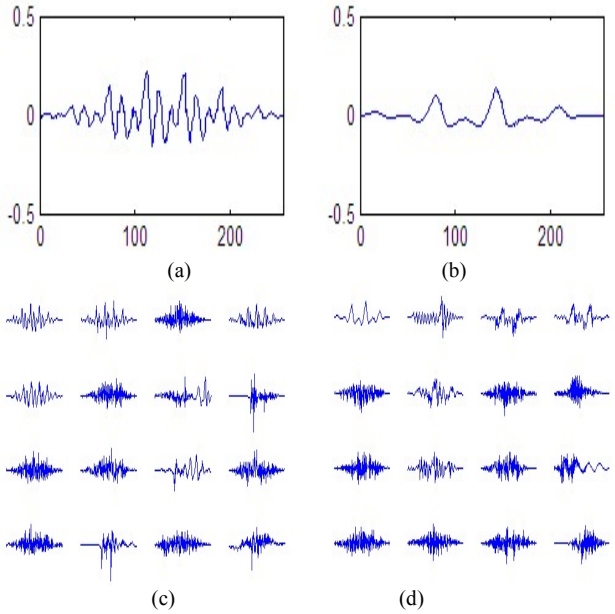


Fig.2. Waveforms of sources and selected atoms.(a-b) source frames 1,2; (c-d) the first 16 atoms selected to estimate source frames 1,2 using traditional OMP;

We compare the average results of separation using OMP with the separation method using SNMF in SNR and SIR respectively. The results are shown in TABLE IV and V.

From TABLE IV and V, it is also observed that the proposed method outperforms the separation method using SNMF in both SNR and SIR. The proposed method achieves 0.2dB higher SNR and 2.55dB higher SIR results respectively than the method using SNMF for female/female mixtures. It achieves o.15 dB higher SNR and 1.35dB higher SIR result than the method using SNMF for female/male mixtures. It achieves o.3 dB higher SNR and 3.9dB higher SIR result than the method using SNMF for male/male mixtures.

© 2014 ACADEMY PUBLISHER

TABLE IV
PERFORMANCE OF SEPARATION USING SNMF AND PROPOSED SEPARATION METHOD USING OMP IN SNR (DB) ON 54 MIXTURES

	SNMF	Proposed method
F/F	4.4/4.1	4.2/4.3
F/M	5.5/5.7	5.7/5.8
M/M	3.2/3.3	3.4/3.7

TABLE V
PERFORMANCE OF SEPARATION USING SNMF, TRADITIONAL OMP AND OMP IN SIR (DB) ON 54 MIXTURES

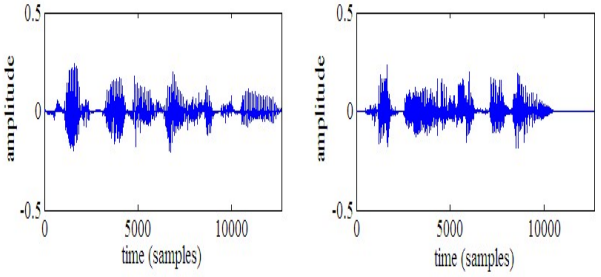
	SNMF	traditional OMP
F/F	6.8/3.5	6.0/7.8
F/M	8.3/9.6	9.3/11.3
M/M	1.3/3.6	5.1/7.6

Finally, Fig.2. illustrates the waveforms of the original sources and the separated results for the mixture of a female and male speech.

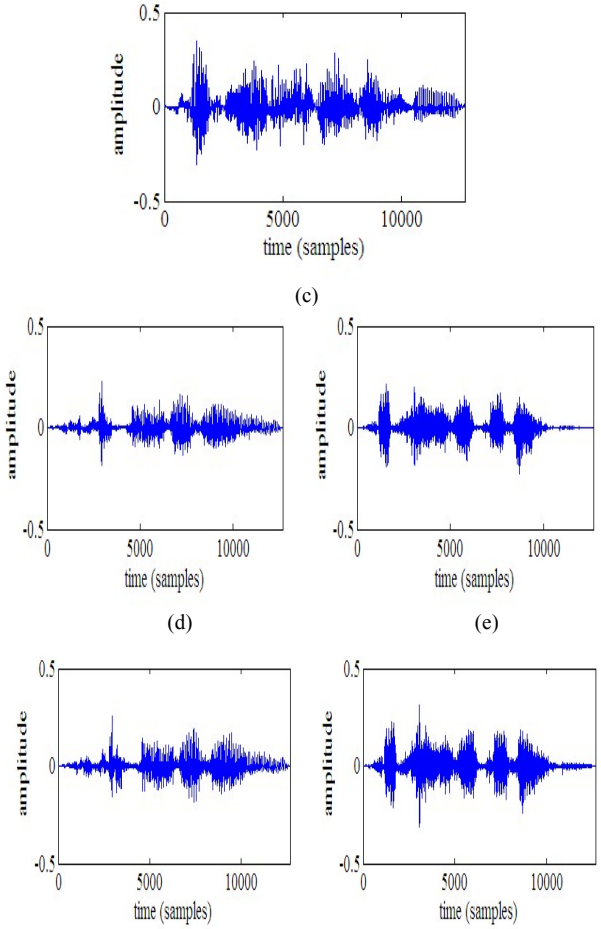
V. CONCLUSION AND FUTURE WORK

In this paper, we presented a new sparse decomposition based SCSS method using OMP. In the proposed method, we generate source dictionaries as matrices consisting of time-domain training frames as atoms and use OMP for the computation of sparse coefficients in the union of source dictionaries. We compared our separation results to the results of separation using SNMF, and showed that our proposed method achieved higher SNR and SIR.

In the proposed separation method, matrices consisting of training frames as atoms are used as source-individual dictionaries directly. In the future, we plan to unite dictionary learning and the presented separation work to improve separation speed. Moreover, we plan to discuss whether we can improve the OMP algorithm for separation by considering mutual independence between sources.



(a) (b)



(c) (d) (e) (f) (g)

Fig.2. Separated speech waveforms of a female and a male speech. (a-b)original sources; (c) mixed signal ; (d-e) separated sources estimated using SNMF based method; (f-g) separated sources estimated using OMP.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (No. 61302152, No. 61201345 and No. 61271240), the Beijing Key Laboratory of Advanced Information Science and Network Technology. (No. XDXX1308) and the Engineering Colledge of Nanjing Agricultural University Introduced Talent Start-up Fund Project (No. rcqd11-02).

REFERENCES

[1] J. Wen 1, S. Zhang, and J. Yang, "A fast algorithm for undedetermined mixing matrix identification based on mixture of guassian (MoG) sources model," *Journal of Software*, vol. 9, no. 1, pp. 184-189, 2014.

[2] B. Yu, H.F Li, C.Y. Fang, "Speech Emotion Recognition based on Optimized Support Vector Machine," *Journal of Software*, vol. 7, no. 12, pp. 2726-2733, 2012.

[3] Y. Shao, S. Srinivasan, Z.Z Jin and D.L. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," Computer Speech and Language, vol. 24, pp. 77-93, 2010.

[4] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust.Soc. Amer.*, vol. 126, pp. 1486–1494, 2009.

© 2014 ACADEMY PUBLISHER

[5] G.J. Jang and T.W. Lee, "A maximum likelihood approach to single-channel source separation", *Journal of Machine Learning Research*, 4, pp. 1365-1392, 2003.
[6] M. Stark, M. Wohlmayr and F. Pernkopf, "Source–Filter-Based Single-Channel Speech Separation Using Pitch Information," *IEEE Transcations on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 242–255, Feb. 2011.
[7] D.Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines*, Kluwer Academic, Norwell MA, pp. 181-197, 2005.
[8] M.G. Christensen, "Metrics for vector quantization-based parametric speech enhancement and separation," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp.3062-3065, 2013.
[9] M.H. Radfar and R.M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Transcations on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.
[10] M.H. Radfar, R.M. Dansereau and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation", *EURASIP Journal on Audio, Speech, and Music Processing*, 2007(1).
[11] R.J. Weiss andD.P.W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," Computer Speech and Language, vol. 24, pp. 16-29, 2010.
[12] P. Mowlaee, M.G. Christensen, S.H. Jensen, "New results on single-channel speech separation using sinusoidal modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.19, no. 5, pp. 1265-1277, 2011.
[13] M.V.S Shashanka, B. Raj and P. Smaragdis, "Sparse overcomplete decomposition for single channel speaker separation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. II-641-II-644, 2007.
[14] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 17-20, 2005.
[15] M. Moussallam, G. Richard and L. Daudet, "Audio source separation informed by redundancy with greedy multiscale decompositions," *IEEE European Signal Processing Conference (EUSIPCO)*, pp. 2644-2648, 2012.
[16] B.A. Pearlmutter and R.K. Olsson, "Linear program differentiation for single-channel speech separation," *IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pp. 421-426, 2006.
[17] M.N. Schmidt and R.K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," *International Conference on Spoken Language Processing (INTERSPEECH)*, 2006.
[18] B.L. Zhu, W.L., R.J. Li and X.Y. Xue, "Multi-Stage Non-Negative Matrix Factorization for Monaural Singing Voice separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no.10, pp. 2096-2107, 2013.
[19] Rémi Gribonval and Morten Nielsen, "Sparse Representations in Unions of Bases," IEEE Transactions on Information Theory, vol.49, no.12, pp: 3320-3325, 2003.
[20] Y. Li, A. Cichocki, and S.-I. Amari, "Analysis of sparse representation and blind source separation," *Neural Comput.*, vol. 16, no. 6, pp. 1193–1234, 2004.
[21] Y.C. Pati, R. Rezaiifar and P.S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," *1993 Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pp. 40-44, 1993.
[22] J.A. Tropp and A.C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no.12, pp. 4655-4666, 2007.
[23] Y.Q. Li, A. Cichocki, S. Amari, S.L. Xie and C. Guan, "Equivalence Probability and Sparsity of Two Sparse Solutions in Sparse Representation," IEEE *Transactions on Neural Networks*, vol. 19, no. 12, pp. 2009-2021, Dec 2008.
[24] M.P. Cooke, J. Barker, S.P. Cunningham, and X. Shao, "An audiovisual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.
[25] E. Vincent, C. Fevotte and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

Haiyan Guo received her Ph.D. degree in signal and information processing from Nanjing University of Posts & Telecommunication (NUPT), Nanjing, China in 2011.

She is a lecturer in the college of engineering, Nanjing Agricultural University. She is currently working as a postdoctoral researcher in the school of information science and engineering, Southeast University. Her research interests mainly include speech signal processing, sparse decomposition and compressed sensing.

Xiaoxiong Li received his B.S. degree in signal and information processing from Hohai University, Nanjing, China in 2012. He is currently pursuing the M.S. degree in signal and information processing in the school of information science and engineering, Southeast University.

His research interests mainly include speech signal processing.

Lin Zhou received her Ph.D. degree in signal and information processing from Southeast University, Nanjing, China in 2005.

She is currently an associate professor of signal and information processing in the school of information science and engineering, Southeast University. Her research interests mainly include speech signal processing and spatial hearing.

Zhenyang Wu received his M.S. degree in electrical engineering from Southeast University, Nanjing, China in 1982.

He is currently a professor of signal and information processing in the school of information science and engineering, Southeast University. During 2000–2008, he was the vice dean of the School of Information Science and Engineering, Southeast University. His research interests include speech and audio processing, image and multimedia processing, and sensor array signal processing. He has published more than 100 international journal and conference paper. From 1990s, he has been an invited reviewer of several famous scientific journals. He is a number of IEEE.

Prof. Wu served as a member of technical program committee for the 13th IEEE International Symposium on Consumer Electronics (ISCE2009). In 2008, he received the National Award for Distinguished Teacher.

© 2014 ACADEMY PUBLISHER