Multiple Speech Source Separation with Non-Sparse Components Recovery by Using Dual Similarity Determination

Maoshen JIA\textsuperscript{[a]}, Member, Jundai SUN\dagger, Feng DENG\dagger, and Junyue SUN\ddagger\ddagger, Nonmembers

SUMMARY In this work, a multiple source separation method with joint sparse and non-sparse components recovery is proposed by using dual similarity determination. Specifically, a dual similarity coefficient is designed based on normalized cross-correlation and Jaccard coefficients, and its reasonability is validated via a statistical analysis on a quantitative effective measure. Thereafter, by regarding the sparse components as a guide, the non-sparse components are recovered using the dual similarity coefficient. Eventually, a separated signal is obtained by synthesis of the sparse and non-sparse components. Experimental results demonstrate the separation quality of the proposed method outperforms some existing BSS methods including sparse components separation based methods, independent components analysis based methods and soft threshold based methods.

key words: blind source separation, non-sparse component, similarity coefficient, intelligent systems

1. Introduction

Multiple source separation technique is increasingly important in many computer interaction areas, such as smart home, intelligent robots, teleconference and so on. Considering both network speeds and computational complexity, a more low-delay and low-complexity algorithm is urgently needed in many intelligent systems.

Early, based on the mutual statistical dependence of sources, Independent Components Analysis (ICA)\cite{1} is widely used in Blind Source Separation (BSS). By combining with different algorithms such as principle component analysis, least-squares, non-negative matrix factorization\cite{2}, ICA based methods are applicable to different cases (underdetermined, determined, overdetermined). However, they require a large amount of data recorded in a stationary acoustic condition to provide a reasonable estimate of model parameters. In addition, it imposes a permutation problem due to misalignment of the individual source components\cite{3-5}.

Recently, based on the W-Disjoint Orthogonal (W-DO)\cite{6} property of speech signal, many methods like Sparse Components Analysis (SCA) are applied to solve the BSS issues. By applying the binary time-frequency (TF) masking\cite{7}, they can separate all the sparse components (the set of TF components that satisfy the W-DO property) from mixture signals (i.e., the signals recorded by microphone arrays)\cite{8}. However, the other TF components derived from more than one sources which called non-sparse components are ignored. Moreover, the literature\cite{9} has verified that there are more than one sources active in many TF instants when there are multiple sources occurring simultaneously, i.e., non-sparse components occupy a major proportion.

Due to the absence of non-sparse components, the BSS methods\cite{8} which only focus on the sparse components yield to obtain an imperfect separation quality. To solve this issue, a collaborative BSS method\cite{CBSS} is proposed by using a pair of location informed coincident microphone arrays\cite{10}, the musical and cross-talk distortion is effectively reduced by the combination of the microphone pair and the vector decomposition. But when the number of overlapped sources are three or above, the vector decomposition will get more difficult. Later, as an extension to binary TF masking method, many soft TF masking methods are developed for non-sparse components separation such as Expectation-Maximization (EM) algorithm based method\cite{11}. It achieves a better separation both in anechoic and reverberant environments. However, the success is achieved at the expense of a high computational complexity.

In this paper, a novel BSS scheme is proposed to recover the non-sparse components by using the similarity between the extracted sparse signal and the mixture signal (the extracted sparse signal indicate the signal which is obtained by grouping the sparse components belong to this signal\cite{8}). The dual similarity coefficient is designed based on the normalized cross correlation (NCC) and the Jaccard similarity coefficients. Thereafter, the “contributors” (the “contributors” in this paper indicate the sources whose TF components are overlapped and result in the non-sparse components) and corresponding separation weights of the non-sparse components are determined by the dual similarity coefficient. Eventually, the non-sparse components are separated to each “contributor” by the obtained separation weight. The final separated signals are obtained by combining the sparse and non-sparse components.

The remainder of the paper is organized as follows: Sect. 2 introduces the proposed BSS method. Experimental results are presented in Sect. 3, while conclusions are drawn in Sect. 4.
2. Proposed Method

By regarding the extracted sparse signal as a guidance, a BSS method with non-sparse components recovery is proposed in this work. The illustration of the proposed BSS scheme is shown in Fig. 1. Due to the small size of the soundfield microphone [9] and the complete recording of spatial information, it is chosen as the recording array in this paper, but it should be mentioned that the proposed scheme is not specific to the chosen one. For the input mixture signals (four B-format signals of the soundfield microphone [9]), the Direction of Arrival (DOA) estimation can be obtained by a traditional localization procedure. Then, the sparse components recovery can be achieved by a clustering process of TF bins. To recover the non-sparse components, a dual similarity coefficient is designed by jointly applying the NCC and Jaccard similarity coefficients [12], respectively. The dual similarity coefficient is then utilized to jointly determine the “contributors” and the separation weight of each “contributor”. Eventually, the final TF coefficients of each source will be recovered by the synthesis of sparse and non-sparse components.

2.1 Sparse Components Separation and Proportion Exploring among Multiple Sources

Under the sparse assumption [10], it is clear that one source will have the approximate DOA estimates at different TF instants. So the sparse components of sources can be recovered by grouping the TF components using these DOAs.

In detail, the recorded signals are transformed to B-format which consists of one omnidirectional signal ($S_w$) and three figure-of-eight directional (Cartesian bi-directional, so-called for its shape) signal ($S_x, S_y, S_z$), i.e., input mixture signals.

When there are $Q$ sources simultaneously occurring, for $S_i(n, k)$ (one of the $Q$ sources, $i \in [1, Q]$), a sparse components signal of $S_i(n, k)$, denoted by $\hat{S}_i(n, k)$, can be obtained by extracting the TF components from $S_w$ [8]:

$$\hat{S}_i(n, k) = \begin{cases} S_w(n, k), & \text{if } \hat{\mu}(n, k) \in [\mu_i - \Delta \mu, \mu_i + \Delta \mu] \\ 0, & \text{otherwise} \end{cases}$$

(1)

where $n$ and $k$ are the time and frequency index, respectively. $\hat{\mu}(n, k)$ denotes the DOA estimate [9] at TF instant $(n, k)$, $\mu_i$ represents the estimated DOA of source $i$ and $\Delta \mu$ is a range threshold to get a better tolerance.

Aiming to separate the non-sparse components by using the sparse components as a guidance, a necessary percentage of the TF components must be recovered. That is, the separated sparse components need to occupy an adequate proportion. In order to investigate the proportion of the sparse components among multiple speech sources, a measurement of the sparse components ratio is designed in this subsection.

A Local-zone Sparse Components Proportion (LSCP) is defined as the occupied ratio of the sparse components among the total TF samples in a TF zone $\Omega$ (a series of frequency-adjacent TF components $(n, k)$), i.e.,

$$LSCP(\Omega) = \frac{||\hat{S}_i(\Omega)||_0}{||S_i(\Omega)||_0}$$

(2)

where the bold-type letter means the signal vector that contains all the TF samples in $\Omega$ and $|| \cdot ||_0$ counts the number of non-zero components in its argument. It is evident that a higher value will be obtained when there are more sparse components separated.

In order to examine the average LSCP among multiple speech sources, a statistic analysis is taken. A total of 40 sentences (the sampling frequency is 16kHz) from the NTT speech database were used for testing. Each sentence was divided into a group with the other $Q - 1 (Q \geq 2)$ sentences in the time domain resulting in $Q$ simultaneously occurring speech conditions. A room of $6.25 \times 4.75 \times 2.5 \text{ m}^3$ is simulated by using Roomsim [13]. The soundfield microphone was placed in the center of the room to generate the mixture signals and the power of sound sources from different directions was equal in each simulation. The statistic results
about the average LSCP versus different source counts are shown in Fig. 2. The TF zone contains 64 TF samples and similar conclusion can also be drawn by other different values.

It can be observed that the percent degrades with the source count raising, but fortunately, there are about 40% TF instants recovered, i.e., the extracted sparse components occupies a proportion about 40% even when the source number is 5 or 6. Above all, the statistical results imply that the proposal of firstly completing source separation among multiple sources based on sparse components is reasonable. Furthermore, the separated sparse components can also be utilized for the non-sparse components recovery.

2.2 Construction of Dual Similarity Coefficient

From the Monte Carlo method [14], we know that the higher similarity between the signal A and the mixture signal B, the greater the likelihood that mixture signal B is from signal A, where a mixture signal B is the recorded signal oriented from multiple speech sources while signal A represents one of the pure speech signal from them. Thus, similarity between the extracted sparse signal and the mixture signal in a certain TF zone is adopted to determine the related sources of the non-sparse components. Besides, the similarity coefficient is assumed to imply the contribution of each source to the non-sparse components. That is to say higher similarity indicates a greater contribution.

It is generally known that the NCC coefficient is often utilized to measure the similarity between two signals. However, for super-sparse signal—meaning that the signals with many zero-value elements, the similarity determined by the NCC coefficient might prefer the signals with higher energy. That is, a limited similar values with a high energy between two signal may probably reach a high similarity if it is calculated based on NCC function only. Unfortunately, the recovered sparse components signal is just the super-sparse signal defined above, especially when the source count is more than four. It is naturally caused by the absence of the non-sparse components. Thus, detecting “contributors” by using a single threshold determination may lose the other “contributors” with low energy and find “pseudo-contributors” sometimes as well. In this subsection, this phenomenon is investigated and solved.

The configuration of the investigation is illustrated in Fig. 3. A total of 40 sentences from the NTT speech database were used for testing. Each speech signal is transmitted to TF domain via short-time Fourier transform (STFT). Then, the frequency coefficients are categorized into three groups in accordance with the energy from high to low in turn, i.e., high energy, medium energy and low energy datasets. In the experiment, $K(i = 1, 2, 3)$ sample points were randomly selected elements from the dataset, to generate three vectors. Following, each vector is processed via a sparse matrix technique, i.e., insert zero-values randomly to obtain three new vectors with the same dimension denoted by $V_1^{K 	imes K}, V_2^{K 	imes K}, V_3^{K 	imes K}$, respectively. Here, $K = K_1 + K_2 + K_3$, and $K = 64$ in the following experiments. Later, the three vectors will be mixed to generate an “imitation” mixture $V$. To measure the performance of the single similarity, we calculate the NCC coefficient between $V_i$ and $V$ by:

$$r_{V_i,V} = \frac{|V_i^T V|}{||V_i|| \cdot ||V||}, i = 1, 2, 3$$

where $T$ is the transpose operation, $r_{V_i,V} \in [0, 1]$ and a closer value to 1 means a higher similarity between the two vectors. By adjusting the proportion of three sample points, the statistical relation between NCC coefficient and the proportion is shown in Fig. 4, where the proportion represent the ratio $K_i/K$.

It can be seen that the greater the proportion of a component in the mixture, the higher the NCC coefficient is,
which also implies that calculating the similarity by using NCC coefficient has certain rationality. However, another significant trend can also be founded that in the case of the same percentage, the calculated NCC coefficient of the high energy component is larger than that of the low energy component, which shows that the NCC coefficient is affected by the energy and can not accurately respond to the similarity.

In other words, the same similarity coefficient can be obtained even when the proportion of a component in the mixture is small if this component has a high energy.

To solve the drawback of the single similarity coefficient, Jaccard similarity coefficient which is often used for comparing similarity and distance of the data set is combined with NCC coefficient to form a dual similarity coefficient as follow:

\[ r_{d,V,V}^i = \alpha r_{V,V} + \beta \tilde{J}_{V,V} \]

where \( r_{d,V,V} \) denotes the proposed dual similarity coefficient, \( \alpha \) and \( \beta \) are the weight parameters \((\alpha + \beta = 1)\) and \( \tilde{J}_{V,V} \) denotes the Jaccard similarity coefficient between \( V_i \) and \( V \) which is defined as follow:

\[ \tilde{J}_{V,V} = \frac{P(V_i \cap V)}{P(V_i \cup V)} \]

where \( P(\cdot) \) represent the probability function. By applying the Jaccard similarity coefficient to the similarity determination, the final results will be balanced because the number of efficient values is also taken into account. Under the same condition with Fig. 4, the statistical relation between dual similarity coefficient and the proportion is shown in Fig. 5, where \( \alpha = \beta = 0.5 \) (informal testing found this value generally led to satisfactory results but future work can explore the optimization of this value). It can be seen that for three kinds of signals, their curves are very close. That implies the dual similarity can better reflect the relationship between the proportion and the similarity coefficient with regardless of the influence of signal energy. Therefore, a dual similarity determination based non-sparse components recovery method is proposed in following.

2.3 Non-Sparse Components Recovery by Using Dual Similarity Determination

The proposed non-sparse components recovery algorithm is performed on a frame-by-frame basis in TF domain. The NCC coefficient in \( \Omega \) between the extracted sparse signal \( \hat{S}_i \) and the mixture signal \( S_w \) is defined by:

\[ r_{S_i,S_w}(\Omega) = \frac{||\hat{S}_i(\Omega)^T S_w(\Omega)||}{||\hat{S}_i(\Omega)|| \cdot ||S_w(\Omega)||} \quad i \in [1, Q] \cap \mathbb{N} \]  

where \( r_{S_i,S_w}(\Omega) \in [0, 1] \) and a closer value to 1 means a higher similarity between the two vectors. In the zone \( \Omega \), the TF mask vector \( W_i \) of source \( i \) is designed by:

\[ W_i(l) = \begin{cases} 0, & \text{if } \hat{S}_i(l) = 0 \\ 1, & \text{otherwise} \end{cases} \quad l \in [1, L(\Omega)] \cap \mathbb{N} \]

where \( L(\cdot) \) is the width of the zone (i.e., cardinality of the zone) and the TF mask vector \( W_{S_w} \) can be obtained by similar form of (7) by processing the B-format signal \( S_w \).

Given two objects, \( W_i \) and \( W_{S_w} \), each of them is with the same binary attributes, i.e., 0 or 1. The total number of each combination of the attributes for both \( W_i \) and \( W_{S_w} \) are specified as: a) \( Mat_{11} \) represents the total number of attributes where \( W_i \) and \( W_{S_w} \) both have a value of 1; b) \( Mat_{01} \) represents the total number of attributes where the attribute of \( W_i \) is 0 while the attribute of \( W_{S_w} \) is 1; c) \( Mat_{10} \) represents the total number of attributes where the attribute of \( W_i \) is 1 and the attribute of \( W_{S_w} \) is 0; d) \( Mat_{00} \) represents the total number of attributes where \( W_i \) and \( W_{S_w} \) both have a value of 0. Above all, it can be concluded that:

\[ Mat_{11} + Mat_{01} + Mat_{10} + Mat_{00} = L(\Omega) \]  

Based on the the Jaccard similarity theory, the other similarity coefficient:

\[ J_{\hat{S}_i,S_w}(\Omega) = \frac{Mat_{11}}{L(\Omega) - Mat_{00}} \]
Then, the dual similarity is defined by:

$$r_{S_i}^d(\Omega) = \alpha r_{S_i,S_w}(\Omega) + \beta 3_{S_i,S_w}(\Omega)$$

(10)

It implies that the higher similarity, the greater the contribution of the corresponding sound source might make to the non-sparse component of the zone. Coinciding with the Monte Carlo method in theory, all the “contributors” of the non-sparse components in the mixture signal over the zone \(\Omega\) can be approximately determined by the dual similarity coefficient as follow:

$$O(\Omega) = \{i | r_{S_i}^d(\Omega) > \eta, i \in [1, Q] \cap \Omega\}$$

(11)

Similar to the assumption that the probability implies the contribution of the corresponding sound source might make to the non-sparse component of the zone. Thus, the proposed dual similarity determination based separation weight is given by:

$$\gamma_S(\Omega) = \frac{\alpha r_{S_i,S_w}(\Omega) + \beta 3_{S_i,S_w}(\Omega)}{\sum_{i \in O(\Omega)} \alpha r_{S_i,S_w}(\Omega) + \sum_{j \in O(\Omega)} \beta 3_{S_j,S_w}(\Omega)}, i \in O(\Omega)$$

(12)

So considering both sparse components and non-sparse components of the sources, the final separated signals of “contributor” \(i\) at each TF bin can be obtained by:

$$S_i'(n,k) = \begin{cases} \hat{S}_i(n,k), & \text{if } \gamma_S(\Omega)S_w(n,k) \neq 0, (n,k) \in \Omega \\ \gamma_S(\Omega)S_w(n,k), & \text{Otherwise} \end{cases}$$

(13)

The similarity determination based separation principle is then conducted for all the TF components and the final separated signals will be obtained by using the inverse TF transformation.

3. Experimental Results

Aiming to evaluate the separation quality of the proposed approach, both simulated and real environments were considered. The experimental parameters are shown in Table 1 and it should be noted that the simulated room is generated by Roomsim[13]. In both simulated and real rooms, the soundfield microphone was placed in the center of the room and the power of sound sources from different directions was equal in each simulation. Note that both reference methods and the proposed method were proceeded by using the localization algorithm in literature [9].

Both objective and subjective evaluation tests were adopted to assess the algorithm. In detail, the objective evaluation was proceeded by three measurements, i.e., Perceptual Evaluation of Speech Quality (PESQ)[15], Signal-to-Distortion Ratio (SDR) [16], Signal-to-Interference Ratio (SIR) [16], which implied the quality of the separation algorithm in different aspects. While in subjective evaluation, a MUSHRA [17] test was conducted to obtain the subjective evaluation of the speech signals separated by different BSS methods.

For comparison, one of the most efficient sparse component separation (SCS) method was selected as the reference method [8] in the first subsection to indicate the effect of the non-sparse component recovery by using the proposed method. In the second subsection, five outstanding existing methods were chosen for a further evaluation.

3.1 Objective Evaluation between the Proposed Method and SCS

Table 1: Experimental parameters

| Parameter                  | Value                        |
|----------------------------|------------------------------|
| simulated room size        | \(6.25 \times 4.75 \times 2.5 \text{ m}^3\) |
| real room size             | \(6.25 \times 4.75 \times 2.5 \text{ m}^3, S N R = 20 \text{dB}, R T 60 = 0.5 \text{s}\) |
| \(L(\Omega)\)              | 64                           |
| \(\rho\)                   | 4                            |
| \(\alpha\)                 | 0.5                          |
| \(\beta\)                  | 0.5                          |
| angle between sources      | 40°, 50°, 60°                |

The PESQ, SDR, SIR results in simulated anechoic and real room are shown in Fig. 6. Condition Mixture (W) is the mixed signal which is used for indicating the worst quality. Condition SCS is the reference method [8] and it is chosen since it has reached a better comprehensive performance in PESQ, SDR and SIR when there were more sources than the microphones, i.e., better quality while with less crosstalk and distortion. Condition Pro-BSS is the proposed algorithm in this paper.

It can be concluded that the proposed method achieves a better performance of the extracted sources than the method which only focuses on sparse components in PESQ, SDR, and SIR. Especially, when there are more multiple simultaneously occurring sources than microphones, the proposed method can achieve evidently better separation quality. In addition, the performance of all conditions degrades slightly in real room which may caused by the reverberations and noise.

3.2 Objective and Subjective Evaluation of Different BSS Methods

To further evaluate separation quality of the proposed method, we compared the performance of the proposed method with existing efficient separation methods in this subsection. Considering the practical acoustic environment always contains reverberations and noise, thus the data recorded in real environment was selected as testing data and the source count is three to compare the ICA based methods. We conducted a comparison with five other outstanding existing approaches: (a) Spatio-Temporal ICA [18] applied using a single soundfield microphone (S-ICA); (b) Spatio-Temporal ICA applied using two soundfield microphones (D-ICA); (c) source DOA-based BSS using a single soundfield microphone (S-BSS) [8]; (d) DOA-based col-
collaborative BSS (C-BSS)\cite{10} using two soundfield microphones; (e) EM algorithm based soft threshold separation method (SS-EM) using a soundfield microphone. The proposed method using a single soundfield microphone (Pro-BSS). Specific spatial configuration of this evaluation see as the literature\cite{10}. The PESQ results of different BSS methods are shown in Fig. 7 and it can be concluded that the proposed method reaches the best performance compared to the other methods. Condition SS-EM also reaches a similar PESQ result when the angle between sources=60° while the computational complexity of it can not be ignored. In addition, the SDR and SIR results are omitted since they have similar phenomenon as PESQ measurement according to the results in Sect. 3.1.

Besides, a MUSHRA test (containing 16 listeners) was conducted to obtain the subjective evaluation of the speech object separated by different BSS methods. The results are shown in Fig. 8.

It can be observed that significant improvement on the separation quality is achieved by applying the proposed scheme. The MUSHRA score for the proposed method is over than 80 which implies an “Excellent” quality; the second best score is achieved by SS-EM, however, the computational complexity with EM algorithm is very high; the third best score is approximately rated as “Good”, it should be noted that we just use one soundfield microphone while C-BSS adopts a pair.
4. Conclusion

This paper proposed a multiple speech source separation method with jointly sparse and non-sparse components recovery based on the dual similarity determination. Experimental results show that the proposed method can achieve superior separation quality compared with the reference BSS methods which only focus on sparse components. In addition, the comparison is also conducted with other BSS approaches, and according to both objective and subjective evaluation, the proposed method achieved a superior perceptual quality of separated sources than others.

Acknowledgments

This work has been supported by China Postdoctoral Science Foundation (2017M610731), the Project supported by Beijing Postdoctoral Research Foundation and “Ri xin” Training Programme Foundation for the Talents by Beijing University of Technology.

References

[1] P. Comon, “Independent component analysis, a new concept?,” Signal Process., vol.36, no.3, pp.287–314, April 1994.
[2] A. Hyvärinen, J. Karhunen, and E. Oja, “Introduction,” Independent Component Analysis, pp.1–475, John Wiley & Sons, Inc., 2002.
[3] F. Nesta and M. Fakhry, “Unsupervised spatial dictionary learning for sparse underdetermined multichannel source separation,” 2013 IEEE International Conference on Acoustics, Speech Signal Process., pp.86–90, May 2013.
[4] L. Wang, H. Ding, and F. Yin, “A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures,” IEEE Trans. Audio, Speech, Language Process., vol.19, no.3, pp.549–557, March 2011.
[5] B. Schölkopf, J. Platt, and T. Hofmann, An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments, pp.953–960, MIT Press, 2007.
[6] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” IEEE Trans. Signal Process., vol.52, no.7, pp.1830–1847, July 2004.
[7] M.I. Mandel, R.J. Weiss, and D.P.W. Ellis, “Model-based expectation-maximization source separation and localization,” IEEE Trans. Audio, Speech, Language Process., vol.18, no.2, pp.382–394, Feb. 2010.
[8] M. Shujau, C.H. Ritz, and I.S. Burnett, “Separation of speech sources using an acoustic vector sensor,” 2011 IEEE 13th International Workshop on Multimedia Signal Process. (MMSP), pp.1–6, Oct. 2011.
[9] M. Jia, J. Sun, F. Deng, and J. Sun, “Single source bins detection-based localisation scheme for multiple speech sources,” Electronics Letters, vol.53, pp.430–432(2), March 2017.
[10] X. Zheng, C. Ritz, and J. Xi, “Collaborative blind source separation using location informed spatial microphones,” IEEE Signal Process. Lett., vol.20, no.1, pp.83–86, Jan. 2013.
[11] X. Chen, W. Wang, Y. Wang, X. Zhong, and A. Alineghi, “Reverberant speech separation with probabilistic time-frequency masking for b-format recordings,” Speech Communication, vol.68, pp.41–54, 2015.
[12] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, “Using of jaccard coefficient for keywords similarity,” Lecture Notes in Engineering & Computer Science, vol.2202, no.1, 2013.
[13] D. Campbell, K. Palomaki, and G. Brown, “A matlab simulation of “shoebox” room acoustics for use in research and teaching,” Computing and Information Systems, vol.9, no.3, p.48, 2005.
[14] J.M. Hammersley and D.C. Handscomb, Monte Carlo Methods, Springer Netherlands, 1964.
[15] “Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs,” ITU-T Rec P.862.2, 2005.
[16] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” IEEE Trans. Audio, Speech, Language Process., vol.14, no.4, pp.1462–1469, July 2006.
[17] “Method for the subjective assessment of intermediate quality levels of coding systems,” ITU Rec BS.1534, 1997.
[18] S.C. Douglas, M. Gupta, H. Sawada, and S. Makino, “Spatio-temporal fastica algorithms for the blind separation of convolutive mixtures,” IEEE Trans. IEEE Trans. Audio, Speech, Language Process., vol.15, no.5, pp.1511–1520, 2007.

Maoshen Jia received the B.E. degree in electronic information engineering in 2005 from Hebei University, China. He received the Ph.D. degree in electronic science and technology in 2010 from Beijing University of Technology, China. Recently, he is an associate professor of Faculty of Information Technology in Beijing University of Technology. His current research interests include spatial audio signal processing, multichannel audio signal processing and audio coding. He is a senior member of IEEE, a senior member of The Chinese Institute of Electronics, a member of Audio Engineering Society, a member of IEICE, and a TC member of the APSIPA Speech, Language and Audio technical committee.

Jundai Sun received the B.E. degree in Science and Technology of Electronic Information in 2015 from Agricultural University of Hebei, China. He is currently working towards the M.S. degree in Information and Communication Engineering at Beijing University of Technology, China. His current research interests include source localization and separation methods for environmental audio scene analysis, speech and audio signal processing. He is a student member of IEEE, a student member of Audio Engineering Society, a student member of China Computer Federation.

Feng Deng received the M.S. and Ph.D. degrees in electronic science and technology from the Beijing University of Technology, China, in 2012 and 2015, respectively. During August 2013 to August 2014, he was a Visiting Researcher at the Communications and Signal Processing Group, Victoria University of Wellington, New Zealand. His research interests include single and multichannel speech enhancement. He is currently working as a Senior Research Engineer at Dolby Laboratories (Beijing).
Junyue Sun received the B.E. degree of Electronic Information Engineering in 2004 from Northeast Petroleum University, China. And then he received the M.S. degree of Physical electronics in 2008 from Dalian University of Technology, China. He is currently working in Dialog Semiconductor Company, and the research field includes high speed analog digital converter, audio amplifier, audio codec, etc.