# Multiple Sparse Sources Separation Based on Multichannel Frequency Domain Adaptive Filtering

Xiaoyu CHEN, Zhong-hua FU and Lei XIE
Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
Northwestern Polytechnical University, Xi'an
cxy313186@gmail.com, {mailfzh, lxie}@nwpu.edu.cn

***Abstract*—Underdetermined sparse sources separation is a challenge problem especially in adverse environment, where there are often some non-sparse interferences or more than one sparse interferences located closely to the target sources. While in some applications, such as in-car or hands-free environments, references of the interferences ($P \geq 2$) coming from loudspeakers are available. Common sparse source separation approaches have not yet used these reference information, we call them traditional approaches in this paper. We propose a FD-MENUET (Frequency domain aDaptive filtering based Multiple sENsor degenerate Unmixing Estimation Technique) approach, in which we get full use of those reference information to help to separate the target sources. Even if no reference is available, the approach would only degenerate to the traditional approaches. The experimental results show that the proposed approach is more general and could achieves better separation performance than the traditional one.**

## I. Introduction

Underdetermined Sparse Source Separation is more interesting in recent years for its capability of handling the problem that the number of sources $N$ could be bigger than the number of sensors $M$ ($N > M$). For the advantage of being implemented in real time, the binary mask approaches such as Degenerate Unmixing Estimation Technique (DUET) [1] and Multiple Sensor DUET (MENUET) [2] are more attractive in all approaches. It assumes that the sources are efficiently sparse, and assigns the mask with 1 if the energy at a time-frequency (T-F) unit is considered as being from the target, and 0 otherwise. While the ideal binary mask (IBM) approach sets the mask 1 if the target source's energy exceeds other sources and 0 otherwise [3]. However, in practice, there are often interferences that may be non-sparse or locate closely to the target source, and the references are usually available when these interferences come from loudspeakers. In such situation, we could no longer assume that at most one source is dominant at each T-F unit. Then, the traditional approach without using these references would be difficult or even impossible to separate the target sources. So for the problems, our motivation is that using adaptive filtering technique to cancel these interferences' components from the T-F units in frequency domain before features extraction.

Adaptive filtering techniques have been developed from time domain to frequency domain and single channel to multi-channel. Most adaptive algorithms in time domain could be classified into Least Mean Square (LMS) family, Recursive Least Square (RLS) family and Affine Affine Projection (AP) family. In recent years frequency domain adaptive filter (FDAF) become more and more attractive [4–6], because comparing with algorithms in time-domain it converges faster, could be effectively applied in multi-channel (MC) case (MC-FDAF) while requiring poor cross correlation conditions among channels, and update step size of the filter coefficients could be independent to each other at different frequencies [7]. For advantages mentioned above, we apply MC-FDAF into our system.

This paper is organized as follows. In Section II, we introduce the proposed flow model, definitions and problems description. Section III describes the procedures of the proposed approach. Experiments setup and discussions are presented in Section IV. The last section concludes this paper.

## II. Background Description

In the paper, lowercase and uppercase bold font represent vector and matrix quantities, respectively; all vectors are column vector; and $[\cdot]^T$ stands for matrix or vector transposition. Underlined quantities denote DFT-domain variables and $k$ is a discrete time index.

### A. Model and Definition

We consider the integration of all sources and sensors as a MIMO system with $M$ input channels and $Q$ output channels. The received signal from the $q$th sensor at time $k$ is given by

$$x_q(k) = \sum_{m=1}^{M} \mathbf{u}_m^T(k)\mathbf{h}_{mq}(k) = \mathbf{u}^T(k)\mathbf{h}_q(k) \quad (1)$$

where $M = N + P$ and

$$\mathbf{u}_m(k) = [u_m(k), u_m(k-1), \cdots, u_m(k-L+1)]^T \quad (2)$$

$$\mathbf{h}_{mq}(k) = [h_{mq,0}(k), h_{mq,1}(k), \cdots, h_{mq,L-1}(k)]^T \quad (3)$$

are a vector containing the latest $L$ samples captured from the $m$th input channel and the current adaptive filter for the path from source $m$ to sensor $q$, respectively. There are $N$ non-referenced sources and $P$ referenced sources. For convenient, we integrate each of them for all $m$ in one vector

$$\mathbf{h}_q(k) = [\mathbf{h}_{1q}^T(k), \mathbf{h}_{2q}^T(k), ..., \mathbf{h}_{Mq}^T(k)]^T \quad (4)$$

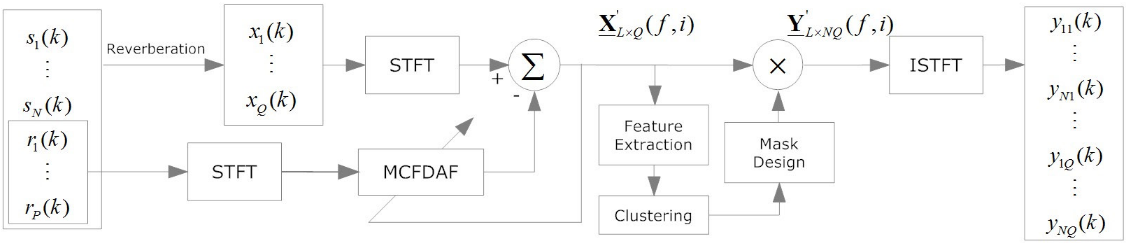


Fig. 1. The proposed model and flow procedures.

$$\begin{aligned}\mathbf{u}(k) &= \left[\mathbf{u}_1^T(k), \mathbf{u}_2^T(k), \cdots, \mathbf{u}_M^T(k)\right]^T \\ &= \left[\mathbf{s}_1^T(k), \cdots, \mathbf{s}_N^T(k), \mathbf{r}_1^T(k), \cdots, \mathbf{r}_P^T(k)\right]^T\end{aligned} \tag{5}$$

where $\mathbf{s}_n^T(k)$ and $\mathbf{r}_p^T(k)$ are the $n$th non-referenced source and the $p$th referenced source, respectively. We mainly concern about the situation that non-referenced sources number $N$ is greater than sensors number $Q$ ($N > Q$), namely underdetermined situation. There is no restriction to the number of reference sources ($P \geq 0$), but we focus on the situation that the number of reference sources $P \geq 2$. The proposed approach aims to separate out each source $y_{nq}(k)$ from every mixed signal $x_q(k)$, and make good use of all valuable information to obtain more pure $y_{nq}(k)$.

### B. Problems Description

*1) Problem of traditional approach:* According to the sparse theory [1, 2], we know each source sparsely distributes in the time-frequency domain, and it is a small probability event that multiple sources appear in the same T-F unit of the mixture. Accordingly, it can be approximated that the T-F unit likely belongs to the source whose energy is major here

$$\underline{\mathbf{x}}_{L,q}(f,i) \approx \underline{\mathbf{u}}_{L,m}^T(f,i)\underline{\mathbf{h}}_{L,mq}(f,i) \tag{6}$$

where $i$ is a block time index, $L$ is DFT length. However, in practice, some sources are not sparse enough or even non-sparse, not only the prominent-energy source but also other considerable-energy sources exist in the same T-F unit. Due to the greatly increased probability of multiple sources' energy overlap in one T-F unit, we can not approximate it as (6).

*2) Problem of sensor pair:* Interaural Level Difference (ILD) and Interaural Time Difference (ITD) [8] are two important and popular features to estimate the binary mask. When the locations of source and sensor pair follow

$$|SF_1| - |SF_2| = \pm 2a \ (0 \leq |2a| \leq |F_1F_2|) \tag{7}$$

where $a$ is a constant scalar, and $|SF_1|$, $|SF_2|$, $|F_1F_2|$ and $|2a|$ represent the length from source $S$ to sensor $F_1$ and sensor $F_2$, the length between $F_1$ and $F_2$, the distance difference between $S$ to $F_1$ and $F_2$, respectively. The ITD features will be exactly the same, then for omnidirectional sensor pair, it would be difficult to separate the sources. Particularly, when $a = 0$, even directional sensor pair could not separate them.

*3) Problem of multiple sources nearby:* In practice, there are often some sources located closely to each other, noticed that the positions of these sources and sensor pair would not satisfy (7). No matter for omnidirectional or directional sensor pair, the ILDs and ITDs of these sources are extremely similar, especially in reverberate environment. It is difficult to separate these sources efficiently. In Fig.2, there are 5 sparse sources in (a) and (c), in which 2 sources are close to each other. Results shows that there are only 4 peaks, the 2 sources are mixed into one peak. For detail discussion, see IV-B.

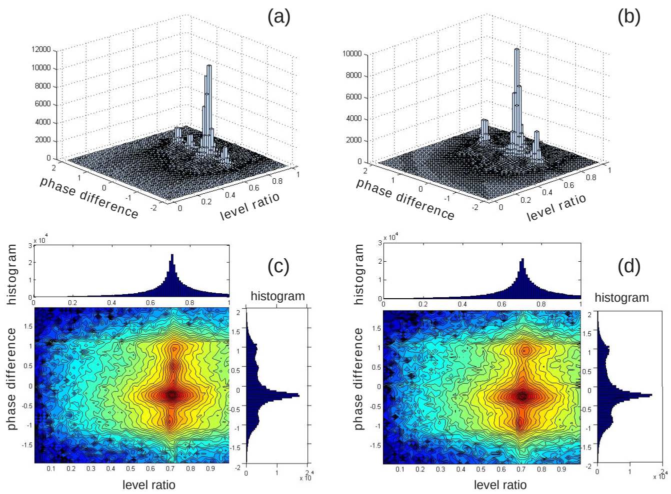


Fig. 2. Example contours and histograms of traditional and proposed approaches in the condition two sources are close to each other with $T60 = 60ms$. We chose one level ratio and the phase difference of each features to plot. (a) and (c) are traditional results; (b) and (d) are proposed results.

## III. Proposed Method

In order to improve the separation performance, especially in the challenging situations mentioned above, we propose to make full use of the possible references. The main procedures of our proposed approach are present in Fig.1.

### A. Adaptive filtering Procedures

Before feature extraction, we process adaptive filtering first whenever only referenced sources are present. Here we can use *double talk detector* (DTD) [9, 10] as a controller to determine when the adaptive filter should work. And we must take different DFT transforms before and after every adaptive filtering iteration. In adaptive filtering procedure, $2L$th DFT is employed, for referenced source $r_p$

$$\underline{\mathbf{R}}_{2L\times 2LP}(i) = \left[\underline{\mathbf{R}}_{2L,1}(i), \underline{\mathbf{R}}_{2L,2}(i), ..., \underline{\mathbf{R}}_{2L,P}(i)\right] \tag{8}$$

$$\underline{\mathbf{R}}_{2L,p}(i) = diag\left\{\mathbf{F}_{2L\times 2L}\begin{bmatrix} r_p(iL-L) \\ \vdots \\ r_p(iL+L-1) \end{bmatrix}\right\} \tag{9}$$

where $\mathbf{F}_{2L\times 2L}$ is a $2L \times 2L$ DFT matrix with elements $e^{-j2\pi vl/2L}$, where $v, l = 0, ..., 2L-1$; for $q$th sensor received signal $x_q$

$$\underline{\mathbf{X}}_{2L\times Q}(i) = \mathbf{F}_{2L\times 2L}\begin{bmatrix} \mathbf{0}_{L\times Q} \\ \mathbf{x}_1(i), ..., \mathbf{x}_q(i), ..., \mathbf{x}_Q(i) \end{bmatrix} \tag{10}$$

$$\mathbf{x}_q(i) = [x_q(iL), x_q(iL+1), ..., x_q(iL+L-1)] \tag{11}$$

and for frequency domain adaptive filter

$$\underline{\mathbf{H}}_{2LP\times Q}(i) = \left[\underline{\mathbf{h}}_{2LP,1}(i), ..., \underline{\mathbf{h}}_{2LP,q}(i), ...\underline{\mathbf{h}}_{2LP,Q}(i)\right] \tag{12}$$

$$\underline{\mathbf{h}}_{2LP,q}(i) = \mathbf{F}_{2L\times 2L}\begin{bmatrix} \mathbf{h}_{1q}(i), ..., \mathbf{h}_{pq}(i), ...\mathbf{h}_{Pq}(i) \\ \mathbf{0}_{L\times P} \end{bmatrix} \tag{13}$$

The MC-FDAF is used [5, 7] to update the adaptive filter until non-referenced sources are present.

$$\begin{aligned} \underline{\mathbf{E}}_{2L\times Q}(i) =& \underline{\mathbf{X}}_{2L\times Q}(i) \\ &- \mathbf{G}^{01}_{2L\times 2L}\underline{\mathbf{R}}_{2L\times 2LP}(i)\underline{\mathbf{H}}_{2LP\times Q}(i-1) \end{aligned} \tag{14}$$

$$\begin{aligned} \underline{\mathbf{H}}_{2LP\times Q}(i) =& \underline{\mathbf{H}}_{2LP\times Q}(i-1) \\ &+ \mathbf{G}^{10}_{2LP\times 2LP}\mathbf{K}(i)\underline{\mathbf{E}}_{2L\times Q}(i) \end{aligned} \tag{15}$$

where $\mathbf{G}^{01}_{2L\times 2L}$, $\mathbf{G}^{10}_{2LP\times 2LP}$ and the kalman gain $\mathbf{K}(i)$ are derived in [5, 11]. We can also calculate the kalman gain by using [12]. Then all sources are present and it is only needed to process $L$th DFT with $\mathbf{F}_{L\times L}$ similar to $\mathbf{F}_{2L\times 2L}$, adding a frame window is optional here. The interferences with references can be eliminated from the mixture by

$$\underline{\mathbf{X}}'_{L\times Q}(i) = \underline{\mathbf{X}}_{L\times Q}(i) - \sum_{p=1}^{P}\underline{\mathbf{r}}_{L,p}(i)\cdot\underline{\mathbf{H}}_{L,p}(i) \tag{16}$$

where

$$\underline{\mathbf{X}}_{L\times Q}(i) = \left[\underline{\mathbf{x}}_{L,1}(i), ..., \underline{\mathbf{x}}_{L,q}(i), ..., \underline{\mathbf{x}}_{L,Q}(i)\right] \tag{17}$$

$$\underline{\mathbf{H}}_{L,p}(i) = \left[\underline{\mathbf{h}}_{L,p1}(i), ..., \underline{\mathbf{h}}_{L,pq}(i), ..., \underline{\mathbf{h}}_{L,pQ}(i)\right] \tag{18}$$

and $\underline{\mathbf{h}}_{L,pq}(i)$ is $L$th DFT transform of the current filter for the path from source $p$ to sensor $q$. After the adaptive elimination, then we could make separation on $\underline{\mathbf{X}}'_{L\times Q}(i)$.

## B. Separation Procedures

*Setp 1. Clustering:* Because an individual cluster in the histogram corresponds to an individual source [1], we can separate each source by selecting the observation signal at T-F units in each cluster with a binary mask [2]. Suppose we know the number of non-referenced sparse sources is $N$, then by using the popular clustering methods such as *K-means* or *GMM* to cluster the features $\mathbf{\Theta}(f,i)$ extracted from $\underline{\mathbf{X}}'_{L\times Q}(i)$, and we can obtain $N$ current clustered centers $\mathbf{C}_n(i)$, where $1 \le n \le N$. Kinds of features are discussed in [2].

*Step 2. Pattern recognization:* By extracting new feature $\Theta_{q_1q_2}(f,i)$ from $\underline{\mathbf{x}}'_{L,q_1}(i)$ and $\underline{\mathbf{x}}'_{L,q_2}(i)$, we can use distance or probability as the criterion to determine which clustered center, generated in *step 1*, the new feature belongs to. Then we could obtain the binary mask accordingly by

$$M_{n,q_1q_2}(f,i) = \begin{cases} 1 & \Theta_{q_1q_2}(f,i) \in \mathbf{C}_{n,q_1q_2}(i) \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

By applying the binary mask to $\underline{\mathbf{x}}'_{L,q_1}$ and $\underline{\mathbf{x}}'_{L,q_2}$, we can obtain the $n$th source's components from sensor $q_1$ and $q_2$, respectively

$$\begin{aligned} \underline{\mathbf{y}}'_{n,q_1}(f,i) = M_{n,q_1q_2}(f,i)\cdot\underline{\mathbf{x}}'_{L,q_1}(f,i) \\ \underline{\mathbf{y}}'_{n,q_2}(f,i) = M_{n,q_1q_2}(f,i)\cdot\underline{\mathbf{x}}'_{L,q_2}(f,i) \end{aligned} \tag{20}$$

We can finally obtain the separated signal $\mathbf{y}_{n,q}(i)$ from the $q$th sensor's mixed-signal $x_q(k)$ by using *overlap-add* method [2, 13],

$$\mathbf{y}_{n,q}(i) = \frac{1}{\lambda}\sum_{\beta=0}^{\lambda-1}\mathbf{y}'^{\beta}_{n,q}(i-\beta) \tag{21}$$

where $\lambda = L/S$ is overlap factor, $S$ is frame shift length, $\mathbf{y}_{n,q}(i)$ is a vector with length $S$,

$$\begin{aligned} \mathbf{y}'^{\beta}_{n,q}(j) =& [y'_{n,q}((j+\beta)S+1), y'_{n,q}((j+\beta)S+2), \\ & ..., y'_{n,q}((j+\beta+1)S)]^T \end{aligned} \tag{22}$$

where $y'_{n,q}(v)$ is the $v$th element of $\mathbf{y}'_{n,q}(j)$, and $\mathbf{y}'_{n,q}(j)$ is the inverse DFT transform of $\underline{\mathbf{y}}'_{n,q}(j)$.

A common problem in using binary mask is that there are inevitable errors in estimating the masks, this would lead to musical noise. The problem can be prevented by using spectral smoothing technique [14].

# IV. Experiments

In this section, we show the performance of the proposed approach and compare it with the traditional [2] and the IBM approaches.

## A. Experiment Environment

The experiments are carried out in a typical-sized room with $T60 = 200ms$ and there are one sensor pair. We obtain the sensor received signals by convolving the source signals and the related measured impulse responses. The inter-distance of the sensor pair is $4cm$ and sampling frequency is 8KHz. The sources include chinese speech and white noise. Instead of

DTD, we update the adaptive filter first until non-referenced sources are present.

The frame size of STFT is $L = 512$, the frame shift length is $S = 128$ and the chosen feature [2] is

$$\Theta_{q_1 q_2}(f,i) = \left[ \frac{|\mathbf{x}_{q_1}(f,i)|}{A_{q_1 q_2}(f,i)}, \frac{|\mathbf{x}_{q_2}(f,i)|}{A_{q_1 q_2}(f,i)}, \frac{1}{\alpha_{q_1 q_2} f} \arg \left[ \frac{\mathbf{x}_{q_2}(f,i)}{\mathbf{x}_{q_1}(f,i)} \right] \right]$$

where $A_{q_1 q_2}(f,i) = \sqrt{|\mathbf{x}_{q_1}(f,i)|^2 + |\mathbf{x}_{q_2}(f,i)|^2}$, $\alpha_{q_1 q_2} = 4c^{-1} d_{q_1 q_2}$ is the normalization factor, $c = 343\ m/s$ is the speed of sound propagating in the air and $d_{q_1 q_2}$ is the distance between the $q_1$th sensor and the $q_2$th sensor. We cluster the extracted features by using *k-means* method. The adaptive filter length is $L$, the update rate is $\omega = (1 - 1/(3L))^L$ and the frame overlap factor is $\lambda = L/S$.

We performed 4 group experiments, see Fig.3. There are 5 sources with 3 non-referenced ones and 2 referenced ones in every group. In the first group, two sparse sources are both satisfied (7) with $a = 0$; in the second group, two sparse sources are located closely; in the third group, there is a non-sparse source with reference; the last group is the situation that all sparse sources can be separated by traditional approach. We compared the separation performances of the approaches using Signal to Interference Ratio (SIR) and Signal to Distortion Ratio (SDR) as the evaluation measures.

$$\begin{aligned} SIR_n =& 10\log_{10} \frac{\sum_k |y^n_{nq}(k)|^2}{\sum_k \left| \sum_{j \neq n} y^j_{nq}(k) \right|^2} \\ & - 10\log_{10} \frac{\sum_k |x^n_q(k)|^2}{\sum_k \left| \sum_{j \neq n} x^j_q(k) \right|^2} \\ SDR_n =& 10\log_{10} \frac{\sum_k |x^n_q(k)|^2}{\sum_k |x^n_q(k) - \beta y^n_{nq}(k - D)|^2} \end{aligned}$$

where $y^v_{nq}(k)$ and $x^v_q(k)$, $(1 \leq v \leq N)$, are the components that belonged to $s_v(k)$, mixed in $y_{nq}(k)$ and $x_q(k)$ respectively; $\beta$ and $D$ are used to compensate the amplitude attenuation and the time delay between $y^v_{nq}(k)$ and $x^v_q(k)$. $\beta = 1$ and $D = 0$ in our statistics.

### B. Results And Discussion

Note that we only concern the 3 non-referenced sources and the results are shown in Fig.4 and Fig.5 in which *(A)-(D)* are relative to *(A)-(D)* in Fig.3.

Fig.4 is generated in the situation (D) of Fig.3. We can see that the proposed results are more pure than the traditional's. The intuitive comparison in the situation (B) is shown in Fig.2, in which (a) and (c) are results of traditional approach, there are 5 sources are present but only 4 peaks are distinct; (b) and (d) are generated by the proposed approach, in the 5 sources, the 3 sources are respect to the 3 distinct peaks respectively.

In situation (A), the features of the two sources are exactly the same, the traditional approach can not separate these two sources. The proposed approach cancel the confused referenced source and separate other sources well, see Fig.5.

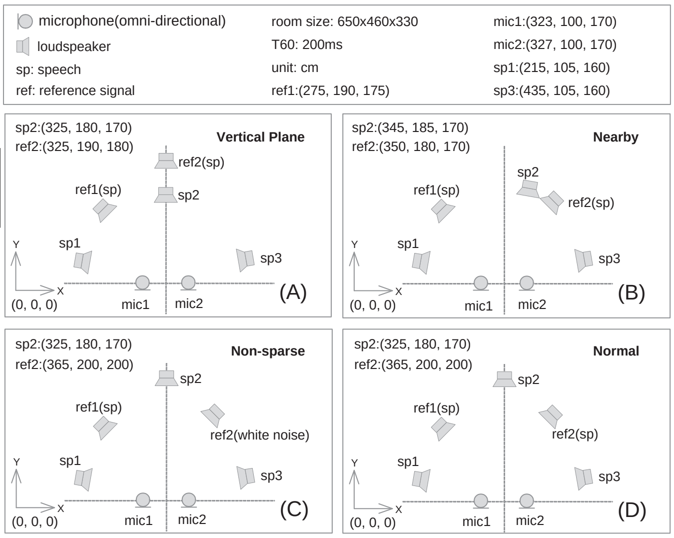


Fig. 3. Experiment environments setup.

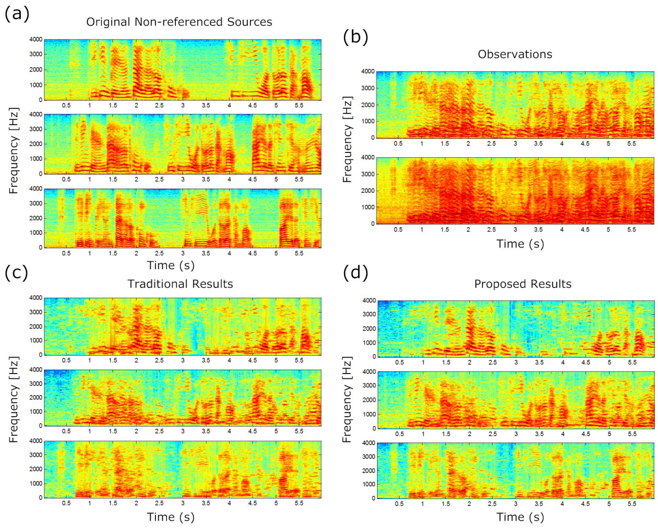


Fig. 4. Spectrum of non-referenced sources resulted from traditional and proposed approach in the condition Fig.3 (D). (a): the sensor received signal's spectrum of each non-referenced sources; (b): the spectrum of the 5 sources' mixtures; (c): traditional approach results; (d): proposed approach results.

Results in (B) is similar to (A) and the features of the two nearby sources are similar. In traditional results, these two separated sources interfere each other, in worst case, this situation even cause to large deviation of the other 3 cluster's regular center. In (C), there is a non-sparse source, white noise with -2dB, in the mixture. The traditional approach could still work with small energy of the non-sparse source, but the separation results are poor. While heavy energy will lead to impossible separation. The proposed approach eliminates the components of the non-sparse source and obtains a good separation result. (D) is a normal situation that the traditional approach generally deal with. All sources are sparse and spatial features are different from each other. The result shows

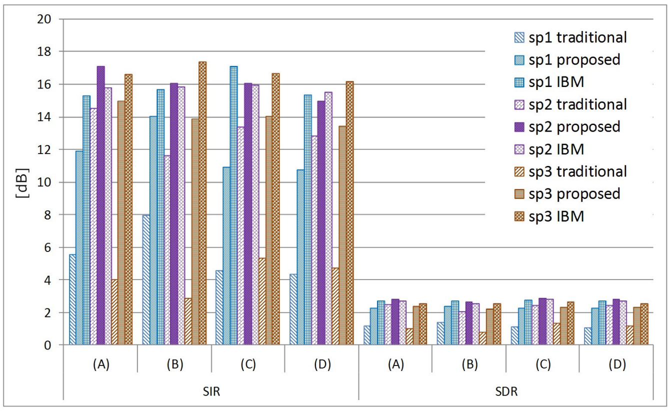


Fig. 5. Average SIR improvement and SDR of non-referenced sources in using traditional, proposed and IBM approaches in each conditions (A)-(D) relative to (A)-(D) in Fig.3.

that the performance of the proposed approach is also better than the traditional one in this situation.

The results of the IBM approach are present as well. Note that the IBM results are implemented without adaptive filtering and all 5 sources are separated. However, in the proposed approach, only the 3 non-referenced sources are needed to be separated. Hence the proposed results are sometimes better than those of the IBM.

The comparison differences of average SIR and average SDR in situation (A)-(D) among traditional, proposed and IBM approaches are listed in Table I. It also shows the proposed approach is better than the traditional one.

TABLE I
DIFFERENCE OF AVERAGE SIR AND SDR AMONG THE THREE APPROACHES

| | SIR[dB] | | | SDR[dB] | | |
|---|---|---|---|---|---|---|
| Speech | IBM-P[a] | IBM-T[b] | P-T | IBM-P | IBM-T | P-T |
| sp1 | 3.95 | 10.23 | 6.28 | 0.40 | 1.52 | 1.12 |
| sp2 | -0.28 | 2.68 | 2.96 | -0.08 | 0.34 | 0.43 |
| sp3 | 2.63 | 12.46 | 9.83 | 0.26 | 1.47 | 1.22 |

[a] Proposed approach
[b] Traditional approach

## V. CONCLUSIONS

We proposed a novel FD-MENUET (Frequency domain aDaptive filtering based MENUET) approach. It extends the applicable situations of the traditional approach as follows: 1) sparse sources are lived with referenced non-sparse sources; 2) some referenced interferes are close to target sources or follow the relation (7) with the sensor pair; 3) the situation that traditional approach could achieve and some sources have reference signals. We derived a general MIMO solution, and when there is none referenced sources, the proposed approach will degenerate to the traditional one. Due to the finite length of the adaptive filters, the separation performance is still subject to the condition of reverberation.

There are many issues remain such as source number estimation, precise DTD and reverberation problem, which will be further studied in our future work.

## ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (60901077), Aeronautical Science Foundation of China (20095553028) and Doctoral Program of Higher Education in China (20096102120044).

## REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[3] Y. Li and D. Wang, "On the optimality of ideal binary timefrequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009.

[4] J. Benesty and D. R. Morgan, "Frequency-domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation," *Proc of ICASSP 2000*, vol. 2, pp. II789—-II792 vol.2, 2000.

[5] H. Buchner, J. Benesty, and W. Kellermann, "Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication," *Signal Processing*, vol. 85, no. 3, pp. 549–570, 2005.

[6] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans Signal Processing*, vol. 51, no. 1, pp. 11–24, 2003.

[7] W. Herbordt, H. Buchner, S. Nakamura, and W. Kellermann, "Multichannel Bin-Wise Robust Frequency-Domain Adaptive Filtering and Its Application to Adaptive Beamforming," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1340–1351, 2007.

[8] J. F. Culling, M. Hawley, and R. Litovsky, "The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources," *Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 1057–1065, 2004.

[9] T. Gänsler and J. Benesty, "The fast normalized cross-correlation double-talk detector," *Signal Processing*, vol. 86, no. 6, pp. 1124–1139, 2006.

[10] H. Buchner, J. Benesty, T. Gänsler, and W. Kellermann, "Robust Extended Multidelay Filter and Double-Talk Detector for Acoustic Echo Cancellation," *IEEE Transactions On Audio Speech And Language Processing*, vol. 14, no. 5, pp. 1633–1644, 2006.

[11] H. Buchner and W. Kellermann, "Improved Kalman gain computation for multichannel frequency-domain adaptive filtering and application to acoustic echo cancellation," *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 2, no. 1, pp. 1909–1912, 2002.

[12] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, pp. 1140–1156, June 2006.

[13] P. Sommen and J. Jayasinghe, "On frequency domain adaptive filters using the overlap-add method," *Circuits and Systems,IEEE International Symposium on*, vol. vol.1, pp. 27 –30, 1988.

[14] T. Jan, W. Wang, and D. Wang, "A multistage approach to blind separation of convolutive speech mixtures," *Speech Communication*, vol. 53, pp. 524–539, Apr. 2011.