A Block-Based Blind Source Separation Approach With Equilateral Triangular Microphone Array

Jian Zhang, Zhonghua Fu and Lei Xie
Shaanxi Provincial Key Laboratory of Speech and Image Information Processing
School of Computer Science, Northwestern Polytechnical University, Xi'an, 710129
E-mail: zhangj062464@163.com; mailfzh@nwpu.edu.cn; lxie@nwpu.edu.cn

***Abstract*—In this paper we describe a method for multiple speech sources separation using an equilateral triangular microphone array. Firstly, the azimuths of horizontal plane are divided into many units and the spatial features of some directions observed by the microphone array are modeled precisely. Secondly, the input mixing signals are segmented into blocks, and then the number of active speakers and their directions are estimated in each block. Thirdly, the pre-trained model with the nearest azimuth to each speaker is adapted to obtain a precise model, which is then used for time-frequency binary mask estimation. Finally, we separate every source appeared in each block and concatenate those sounds from same unit to reproduce the whole stream. The experiments are set up in a real meeting room. The results show that our method can separate multiple speech sources correctly with low distortion, and are competitive with the total un-blind separation results.**

***Index Terms*: blind source separation, directions of arrival estimation, time-frequency mask, equilateral triangular microphone array**

I. Introduction

It is known that human has the magical capability of focusing his auditory attention on one talker without being influenced by other interferences. The so-called cocktail party effect remains a challenging problem to machine. The most promising technique to solve this problem might be Blind Source Separation (BSS).

There are enormous works have been reported in the literatures. Generally the BSS approaches can be classed into two types. One is based on Independent Component Analysis, which assumes the underlying sources are independent to each other. It works very well only when the supposed mixture model is correct and generally deals with the over-determined case [1]. However, this is seldom happened in real applications. The other is sparseness-based approach [2,3], which requires that all sources are sparse in Time-Frequency (T-F) domain, and the target is to find a binary T-F mask for each source. We know that speech signal is sparse in T-F domain. It has been reported that once the ideal binary mask is obtained, the speech intelligence can be improved greatly [4].

Multiple speech source separation is important for applications like far-field speech recognition, automatic meeting diarization, etc. The major challenge is that the number of active speakers is unknown and may change with time, but most of previous researches assume that the source number is known [2] or un-changed [3]. In this paperwe propose a new approach to consider this situation. To cope with omnidirectional sound sources, we utilize three microphones to construct an equilateral triangular array.

Firstly, we separate the azimuth of horizontal plane into many units and build Gaussian Mixture Models (GMMs) for some directions. Secondly, the mixing signals are segmented into blocks. Within each block the inter-microphone time difference vectors are transformed into incidence angles. Then the number of active speakers and their locations are estimated using incidence angle histogram. Thirdly, the pre-trained GMM nearest to each active speaker is adapted to obtain each speaker's precise model. Finally, a time-frequency binary mask is used to separate each active speech and the signals from same unit are concatenated to reconstruct the whole stream. We examine this approach in a real meeting room, and the experimental results verify that the mixed speech signals can be separated correctly with low distortion and the performance is competitive with the total un-blind method.

The rest of the paper is organized as follows. Section 2 describes the problem. Section 3 presents our approach. The experiments and the results are shown in Section 4. Section 5 is conclusion.

II. Problem Description

The triangular array is shown in fig. 1, where the three microphones are located at the verdices of equilateral-triangle. Suppose that sources $s_1,\ldots,s_N$ are convolutively mixed and observed by the three sensors as

$$x_i(t) = \sum_{k=1}^{N} \sum_{l} h_{ik}(l) s_k(t-l), \ i = 1,2,3 \qquad (1)$$

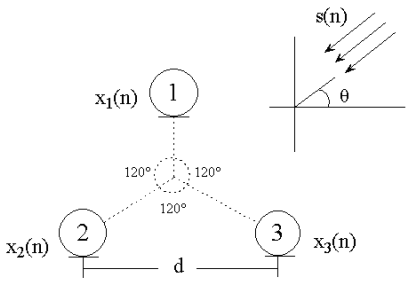

Fig. 1. Triangular microphone array.

Where $h_{ik}(l)$ represents the impulse response from source k to sensor i.

The mixtures x_i, i=1, 2, 3 are segmented into P blocks x_i^p,p=1,...,P. By using short-time Fourier transform (STFT), the mixing model of each block can be represented in T-F domain as

$$X_i^p(f,t) = \sum_{k=0}^{N-1} H_{ik}(f) \cdot S_k^p(f,t),\ i = 1,2,3 \tag{2}$$

Suppose that, the T-F representations of all source signals are sparse. Then, T-F masking for source separation of each block is performed by

$$\tilde{S}_k^p(f,t) = M_k^p(f,t) \cdot X_i^p(f,t),\ i = 1 \text{ or } 2 \text{ or } 3 \tag{3}$$

$$M_k^p(f,t) = \begin{cases} 1, S_k^p(f,t) \text{ is dominant} \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

We separate every source appeared in each block and concatenate those sounds from same direction to reproduce the whole stream. Accordingly, the tasks are to correctly estimate the number of active speakers and find the binary T-F mask for each speaker within every block.

III. Proposed Method

Our approach includes two major steps. In the first step, we estimate the number and the directions of arrival (DOA) of active sources. We suppose the speech signals with the same DOA belong to the same source.

In the second step, the spatial model for each active source is adapted accordingly and then the separation is performed using the Bayes rule. Finally, the separated source components from all the blocks are concatenated to reconstruct the whole stream.

A. The spatial feature

Generally, the sparseness-based BSS uses spatial feature to represent the location differences of all underlying sources. The common spatial features contain the inter-microphone amplitude difference (IAD) and the inter-microphone phase differences (IPD). In order to avoid phase confusing, the distance between the microphones has a limit [5]. We choose 4 cm as the distance in all experiments in this paper. The small distance makes the IAD neglectable. So we only use the IPD feature. In order to avoid the permutation problem among frequencies, we transfer the IPD feature into time delay. For each T-F point, the IPD can be obtained as

$$IPD(f,t) = \arg\left[\frac{X_j^p(f,t)}{X_i^p(f,t)}\right],\ i,j = 1,2,3 \text{ and } i \neq j \tag{5}$$

The relation between the time delay δ and the phase difference is

$$IPD(f,t) = 2\pi f \delta \tag{6}$$

So, we can obtain the time delay between microphone i and j

$$\delta_{ij} = \frac{1}{2\pi f} \arg[\frac{X_i^p(f,t)}{X_j^p(f,t)}] \tag{7}$$

With three microphones, we compose the three delays to a spatial feature vector as [δ_{13} δ_{21} δ_{32}]. For every T-F point, we can obtain a feature vector denoted as $\psi(\theta)$, where the θ means this point belongs to a source with azimuth θ. The relationship between $\psi(\theta)$ and θ is used to estimate the source direction.

B. To estimate the number and DOAs of active speakers

There are a lot of approaches for source localization [6]. Since our aim is to separate the speech mixtures, we must identify the number of active sources and their directions. We divide the azimuths of horizontal plane into many equivalent units and estimate the histogram.

For the triangular array shown in fig.1, there are three pairs of microphones and each of them has an azimuth difference of 120. They receive the speech signal $s(n)$ propagating from the direction θ with elevation angle φ. We ignore the φ firstly. The relationship between the azimuth of a single speech source θ and the three delays are described by the following equations.

$$\delta_{13} = \frac{d}{c}\cos(\theta - \frac{2}{3}\pi) \tag{8}$$

$$\delta_{21} = \frac{d}{c}\cos(\theta + \frac{2}{3}\pi) \tag{9}$$

$$\delta_{32} = \frac{d}{c}\cos(\theta) \tag{10}$$

Where d is the distance of two microphones, c is the sound velocity. We define a transformation matrix T. to make the x axis of the new coordinate corresponds to the $\psi(0)$, and make y axis corresponds to the $\psi(\pi)$ [7].

$$T = [\ e1 \quad e2\ ]^T \tag{11}$$

$$e1 = \psi(0) = [\ \tfrac{d}{c}\cos(-\tfrac{2}{3}\pi) \quad \tfrac{d}{c}\cos(\tfrac{2}{3}\pi) \quad \tfrac{d}{c}\ ]^T \tag{12}$$

$$e2 = \psi(\frac{\pi}{2}) = [\ \tfrac{d}{c}\cos(-\tfrac{1}{6}\pi) \quad \tfrac{d}{c}\cos(\tfrac{7}{6}\pi) \quad 0\ ]^T \tag{13}$$

Then we transform $\psi(\theta)$ as follows

$$T \cdot \psi(\theta) = \frac{x}{y} = \frac{3d^2}{2c^2}\begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \tag{14}$$

Then the direction θ can obtained by

$$\theta(f,t) = \begin{cases} \operatorname{arc\,cot}(\frac{x}{y}),\ if(y > 0) \\ \operatorname{arc\,cot}(\frac{x}{y}) + \pi,\ else \end{cases} \tag{15}$$

The elevation angle will not affect the result because the each component of the delay vector is multiplied by the same factor $\cos(\varphi)$. We divide the horizontal plane 360° into K equivalent units, so every unit is $360°/K$. Then the total energy of every unit is accumulated to build the histogram as

$$w(k) = sum(A(f,t)|\frac{360(k-1)}{K} < \theta(f,t) \leqslant \frac{360 \cdot k}{K}) \tag{16}$$

$$A(f,t) = \sqrt{\sum_{i=1}^{2} |X_i^p(f,t)|^2},\ k = 1,....,K \tag{17}$$

The value of K is determined according to the requirement of the application. The larger K value means the higher accuracy of DOA estimation. But, if K is too large, the error of the source number estimation will increase. In our experiments the K is 36.

To eliminate the interferences caused by room reflections, we need a threshold to select the distinct peaks in the histogram. The tiny peaks might be caused by the fake sound image or the insufficient data of one source available in the current block. This threshold depends on the value of K and block size, the larger K or block size the larger threshold. In our experiment the threshold is set empirically. The number of peaks that are larger than the threshold can be considered as the number of active speakers N in this block. Accordingly, the azimuth corresponding to each peak is taken as the DOA of that source.

C. The Directional GMMs

After the estimation of the number and DOAs of active speakers, the directional model for each active source must be built before separation. To avoid the data insufficient problem, we use model adaption technique. Specifically, we firstly train some directional GMMs off-line. Note that the number of GMMs does not have to match that of active speakers. But if sufficient data are available beforehand, more directional GMMs will lead to more precise target directional models.

Now for each active speaker with known azimuth in current block, the pre-trained GMM with the nearest azimuth is selected. Suppose the number of active speaker in current block is N, the selected GMMs are denoted as $\{\lambda_{base-1}, \ldots, \lambda_{base-N}\}$. Note that these basic models do not have to be different.

The adaption data is an important issue. Since the signals observed by the triangular microphones usually contain not only the direct-path signals that are attenuated and delayed replicas of the sources, but also the multi-path reflected signals. Additionally, speech signal is not ideal sparse. Hence, sometimes different sources may contribute to the same T-F point. To eliminate these disturbances, it is necessary to find the reliable T-F points that only one source is dominant and we can use the corresponding features to adaptive the basic source models. Generally, for each usable peak in the DOA histogram, the T-F point of which the direction is closer is more reliable [8]. Specifically, the T-F point of which the direction θ satisfies the following option is selected for the k-th source

$$\{(f,t)|(k-1)\times\frac{360}{K}<\theta(f,t)\leqslant k\times\frac{360}{K}\} \tag{18}$$

If K is too large or the block size is too small, the option can be relaxes as follows

$$\{(f,t)|(k-2)\times\frac{360}{K}<\theta(f,t)\leqslant (k+1)\times\frac{360}{K}\} \tag{19}$$

With these reliable T-F points, the corresponding feature vectors are used for adaption. The adaption is implemented by only $2\sim 4$ expectation maximum (EM) operations. Then we obtain the precise target directional model for each active speaker, denoted as $\{\lambda_{tar-1}, \ldots, \lambda_{tar-N}\}$.

D. The Separation and the Reconstruction

After we obtain every speaker's model, we can merge the other models together to make a background model λ_{bk} for every source.

$$\lambda_{bk-i}=\bigcup_{j=1,j\neq i}^{N}\lambda_{tar-j} \tag{20}$$

When both the background model and the target model are available, the binary mask for each source can be estimated using Bayes rule. The binary mask is simply estimated based on the likelihood comparison. For each T-F point, the spatial feature of the mixed signal is denoted as $O(f,t)$. The binary mask of the source i can be obtained using

$$M_i^p(f,t)=\begin{cases}1,\ p(\lambda_{bk-i}|O(f,t))<p(\lambda_{tar-i}|O(f,t))\\ 0,\ \text{otherwise}\end{cases} \tag{21}$$

Then the separation is performed use (2). The speech signal of the source i can be reconstructed using Inverse short time Fourier transform (ISTFT) and overlap-adding method.

The stream concatenation is simple here. We suppose the speech signals with the same unit belong to the same source. Hence the whole speech stream $s_j, j=1,\ldots,N$ are reconstructed by concatenating together all the segments of that source $s_j^p, p=1,\ldots,P$.

IV. Experiments and Discussions

A. Experiments setup

The experiments are performed in a real meeting room and the deployment of the microphones and loudspeakers is shown in Fig.2. A high-fidelity loudspeaker is placed respectively at sixteen points 1.8m around of the triangular microphone array. The audio signals played through the loudspeaker are some clean speech. These audio signals are played with the same average sound level and then mixed in computer.

- Loudspeaker (120cm height, high-fidelity)
- Microphone (120cm height, omni-directional)
- Room Size: 6.5m(W)*4.6m(L)*3.3m(H)
- Reverberation time: 160ms
- Ground Noise: 25dB

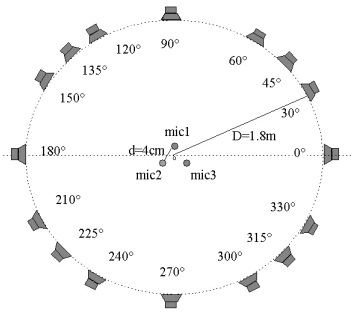

Fig. 2. Experiment setup.

We use the interference reduction ratio (IRR) and log-likelihood ratio (LLR)[9] to measure the interference reduction

and the speech distortion respectively.

$$IRR = \frac{\sum_{f,t} |Y_i(f,t) \cdot (1 - M(f,t))|^2}{\sum_{f,t} |Y_i(f,t)|^2} \cdot 100\% \quad (22)$$

$$LLR = E\left\{ \log\left(\frac{\alpha_p R_c \alpha_p^T}{\alpha_c R_c \alpha_c^T} \right) \right\} \quad (23)$$

Where $Y_i(f,t)$ is the pure interference; $M(f,t)$ is the binary mask obtained using (18); α_p is the LPC vector of the separated speech; α_c is that of the original clean speech and R_c is the autocorrelation matrix of the original clean speech signal; $E\{\}$ is the expectation function.

For comparison, the un-blind results and the ideal binary mask results are also shown. Un-blind means the location of every source is known, and its model, i.e. the GMM, is trained well separately in advance. The ideal binary mask (IBM) is the upper bound of the separation performance. It is created using knowledge of the signals before they were mixed.

$$M_{ideal}(i) = |S_i(f,t)| > |Y_i(f,t)| \ , i = 1, ..., N \quad (24)$$

Where $S_i(f,t)$ is the STFT of the desired signal.

We mix six sources, and their angles are $\{30°, 90°, 150°, 210°, 270°, 330°\}$. We also select eight directional models as the basic models. Their angles are $\{0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°\}$. The block length is 5 seconds, and within each block, 2$\sim$6 speech sources as random selected and mixed. Totally, we simulate 120 blocks with 10 minutes of mixed data. The 360° of horizontal plane is divided to 36 units, so every unit is 10°.

Fig.3 and Fig.4 show an example of one block. There are four active sources and their angles are 30°, 150°, 210° and 330°, respectively. Fig.3 shows the original sources and fig.4 shows the mixed observations.

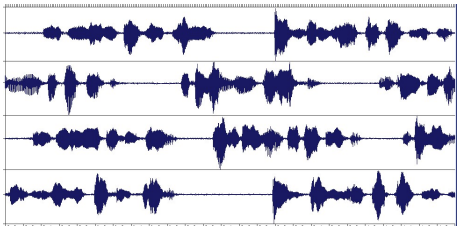

Fig. 3. original signals.

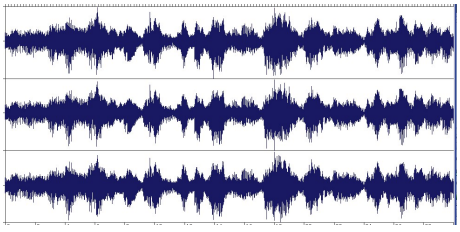

Fig. 4. mixed signals.

The corresponding DOA histogram is show in fig.5. It is very clear four peaks are distinct on the correct azimuth angles.

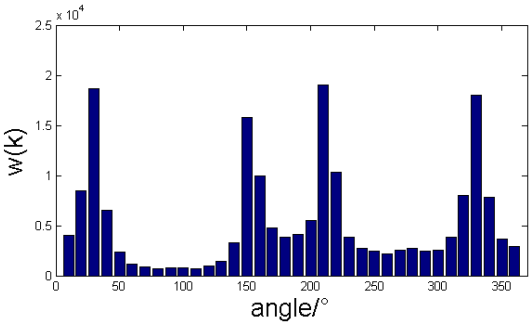

Fig. 5. result of DOA.

The separation results of our method are shown in fig.6. The results of un-blind separation are show in fig.7 and the IBM results are shown in fig.8.

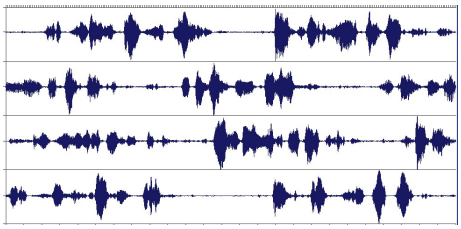

Fig. 6. Separation results of our approach. IRR {0.8580, 0.9159, 0.9594, 0.9636}; LLR {0.5714,0.6255, 0.6230, 0.5303}

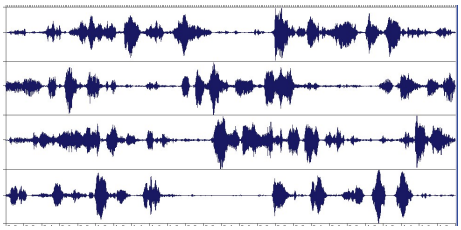

Fig. 7. Separation results of un-blind approach. NRR-U {0.9395, 0.9250, 0.9141, 0.9655}; LLR-U {0.5689, 0.5773, 0.5740, 0.5046}

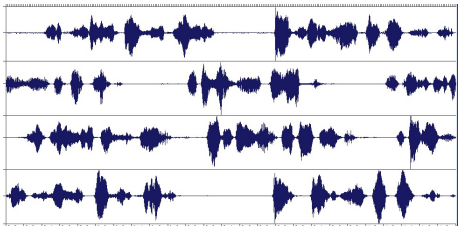

Fig. 8. Separation results of IBM. IRR-I {0.9749, 0.9584, 0.9397, 0.9740}; LLR-I {0.3950, 0.3967, 0.4216, 0.4522}

The quantitative results are listed in Table I. Note that the higher the IRR is, the more interference is eliminated, and the lower the LLR is, the smaller distortion the speech has.

TABLE I
EXPERIMENT RESULTS

source	1	2	3	4	5	6
LLR	0.5901	0.7059	0.5734	0.5680	0.5298	0.6085
LLR-U	0.5659	0.7115	0.5605	0.5508	0.5352	0.6514
LLR-I	0.4483	0.4622	0.3813	0.4514	0.4449	0.4511
IRR	0.9044	0.9325	0.8975	0.9474	0.9351	0.9459
IRR-U	0.9045	0.9389	0.9072	0.8779	0.9355	0.9594
IRR-I	0.9666	0.9553	0.9684	0.9679	0.9581	0.9686
T-len.(min)	6.17	6.42	6.42	6.92	6.33	5.67

T-Len means the total length of speech. According to the experimental results, the IBM achieves the best performance, and our approach has the competitive performance with the total un-blind separation approach, and both are close to the IBM performance. Almost 90% interferences are eliminated. The results show that our approach can separate the mixtures correctly with low speech distortion without knowing the number and DOAs of the active sources.

V. Conclusion

This paper introduces a block-based approach for the underdetermined blind source separation using binary T-F masks. The main idea is to model each sound source. Firstly, we estimated the source number and their directions. And select reliable T-F points for each source. Secondly, we load model for every source and use these reliable T-F points to adapt the model. Than we calculate binary mask according these models. The last step, the separated source components from all the blocks are concatenated to reconstruct the whole signal. The experimental results show the effectiveness.

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). This work was also supported By graduate starting seed fund of Northwestern Polytechnical University (Z2011143).

References

[1] Ricardo Vigario And Erkki Oja, "BSS and ICA in Neuroinformatics: From Current Practices to Open Challenges. IEEE Reviews in Biomedical Engineering," vol 1: pp. 50-61, 2008.

[2] C. Hummersone, R. Mason and T. Brookes, "Ideal Binary Mask Ratio: A Novel Metric for Assessing Binary-Mask-Based Sound Source Separation Algorithms." IEEE Transactions on Audio, Speech, and Language Processing, in press.

[3] S. Araki, H. Sawada, R. Mukai and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors, "*Signal Processing*, vol. 87(8): pp. 1833-1847, 2007.

[4] S. Araki, T. Nakatani, H. Sawada and S. Makino, "Blind Sparse Source Separation For Unknown Number of Sources Using Gaussian Mixture Model Fitting With Dirichlet Prior," *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009 (ICASSP 2009)*. 2009.

[5] Z.H. Fu, L. Xie and D.M. Jiang, "Dual-microphone noise reduction based on semi-blind DUET," *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2010.

[6] M. Swartling, B. Säberg, N. Grbic, "Source localization for multiple speech sources using low complexity non-parametric source separation and clustering,"*Signal Processing*, vol. 91 (8) , pp.1781-1788, 2011.

[7] M. Yoshida, D. Ning and N. Hamada, "Blind Speech Separation by Integrating Three Pairs of Phase Differences of Equilateral Triangular Microphone Array, " *2010 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2010)*, 2010.

[8] Y. Lv and S.T. Li, "Underdetermined Blind Source Separation of Anechoic Speech Mixtures," *9th International Conference on Signal Processing, 2008 (ICSP 2008)*, 2008.

[9] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16(1), pp. 229-238, 2008.