Source Separation for Arbitrary Array Configuration in the Presence of Spatial Aliasing

Masashi Sekikawa
Department of system design engineering, Keio University, 3-14-1, Hiyoshi, Yokohama, Japan
Email: mks12145963@gmail.com

Nozomu Hamada
Department of Electronic Systems Engineering, Malaysia-Japan Institute of Technology (MJIIT), Universi Technologi Malaysia, Kuala Lumpur, Malaysia
Email: hamada@utm.my

Abstract—**This paper proposes a blind source separation method using time-frequency (T-F) decomposition and clustering for estimated propagation direction vectors at T-F slots. The method is applicable to arbitrary array configuration in 3-D space and even in the presence of spatial aliasing. To solve spatial aliasing problem which causes ambiguity of phase, unity norm property of propagation direction vector is employed for solving phase difference ambiguity. With combining our previous direction-of-arrival (DOA) estimation algorithm and clustering in terms of spatial information, efficient separation procedure is achieved by a binary masking in T-F domain. Experimental results demonstrate that the proposed procedure effectively separate three or four speech sources with tetrahedron microphone array with wide sensor spacing where spatial aliasing may occur.**

Index Terms—**microphone array, source separation, time-frequency masking, DOA estimation, spatial aliasing, kernel density estimation**

I. Introduction

Blind source separation (BSS) technique is to estimate individual source signals from their mixtures observed at multiple sensors [1], [2]. The BSS technique for speech is applied to many fields including hands-free remote conference system and robot audition system.

Mathematical model of source mixing process is a convolution operation between source signals and the impulse response from sources to sensors. Frequency domain approach transforms such convolution mixture to an instantaneous mixture, where short-time Fourier transform (STFT) is applied to the sensor observations. Independent component analysis (ICA) and the method based on the sparseness of source signals are two efficient approaches for solving the BSS. The assumption of sparseness of source signals implies that at most one source makes a major contribution to each time-frequency component of STFT representation. This assumption enables to handle the underdetermined problem where the number of sources is larger than that of sensors.

A practical sparseness-based approach is known as time-frequency (T-F) binary mask method, and this paper focuses on this method. Basic idea of binary mask separation is summarized as follows: If the sparseness assumption holds, a histogram of feature vectors, which are obtained from sensor observations, would have the same number of clusters as source's number. Since an individual cluster in the feature space corresponds to an individual source, each source signal can be detected by selecting the time-frequency components contained in each cluster. In an earlier and well-known DUET (Degenerate unmixing estimation technique) approach [3], a pair of sensors is used for defining the geometrical features such as signal level ratio and phase differences, and manual clustering is employed for grouping. Up to now DUET has been generalized in terms of clustering scheme as well as array configuration.

With respect to the clustering algorithm some automated and simplified clustering approaches have been developed such as kernel density estimation [4], maximum likelihood gradient method [5], and *k*-means clustering method [6]. On the other hand, array configuration of sensors is generalized to array with non-linear and non-uniform alignment containing more than three or four sensors. In general we need more than four sensors arranged three-dimensionally for discriminating source direction uniquely.

Previously, the binary mask approach known as MENUET (Multiple sENsor dUET) [7] employs k-means algorithm for a feature utilizing modified level ratio and phase differences between multiple observations at arbitrary three-dimensionally aligned sensor array.

An important issue when treating spatial information such as direction of sources as the feature of time-frequency component is the spatial aliasing problem. As we are concerned with the source directions, the sensor spacing should be no larger than the half of the minimum wavelength of interest in order to hold one-to-one correspondence between the direction and phase difference. Thus conventional approaches [3]-[7] restrict to microphone array with small spatial extent in order to avoid spatial aliasing ambiguity.

Manuscript received August 1, 2014; revised November 24, 2014.

doi: 10.12720/ijsps.4.1.32-36

Few time-frequency binary mask approaches have been proposed in the presence of spatial aliasing. By adopting both ICA and binary masking Sawada et al. [8] proposed a grouping procedure based on estimating anechoic propagation model parameters i.e., the time delays of arrival and attenuation from a source to all sensors. Their method solves the problem of spatial aliasing and is applicable to arbitrary array with any number of sensors. As far as the source direction estimation concerned, the other sparseness-based general method by Loesch & Yang [9] avoids the spatial ambiguity by directly comparing phase differences in both observed and model using a distance metric. However, their source separation employs a linear blind beamformer approach.

Based on the above review of previous studies this paper proposes a clustering method performed in the 3-D propagation direction vector, by which both azimuth and elevation angles are provided uniquely. Our recent contribution [10] proposed a method estimating the direction of arrival (DOA) even in the presence of spatial aliasing. This idea is extensively adopted to generate spatial feature of sources, and the feature is also utilized for source separation.

The remainder of this paper is organized as follows. In Section II, our DOA estimation theory in [11] is summarized. The proposed separation algorithm is proposed in Section III. In Section IV, experiments are demonstrated to verify the method. Section V concludes this study.

II. DOA Estimation

This section provides an overview of DOA estimation technique developed in [11]. The method is applicable to arbitrary configuration of sensor array even in the presence of spatial aliasing. [12]

A. Deley of Arrival and Phase Difference

Consider an array with omni-directional M sensors whose location in 3-D space are given by

$$r_m[x_m,y_{m,}\,z_m]^T \text{ where m=1,…M} \tag{1}$$

Here, we assume $\boldsymbol{r_1} = \mathbf{0}$ without loss of generality, and define the matrix $\boldsymbol{R}$ as

$$R{=}[r_2,\ldots r_M]^T \tag{2}$$

Let also assume a sound source locate at a point whose directional unit-length vector, referred to as the *propagation direction vector*, is written by

$$a(\phi,\theta){=}[\sin\theta\cos\phi{,}\sin\theta\sin\phi{,}\cos\theta]^T \tag{3}$$

where $\phi(-\pi\leq\phi<\pi)$ and $\theta(0\leq\theta<\pi)$ denote the azimuth and elevation angles of the source direction respectively. An acoustic source signal with the propagation direction vector $a(\phi,\theta)$ of (3) causes arrival time delay $\tau_m(\phi,\theta),(m=2,\ldots,M)$ between the *m*-th and the reference (*m*=1) sensors which can be represented by

$$\tau_m(\phi,\theta){=}-\frac{r_m^T\,a(\phi,\theta)}{c}{,}m=2,\ldots,M \tag{4}$$

where c is the sound propagating speed. The vector-matrix formulation of (4) can be represented by

$$\tau(\phi,\theta){:=}[\tau_2(\phi,\theta),\ldots\tau_M(\phi,\theta)]^T{=}\frac{R_a(\phi,\theta)}{c} \tag{5}$$

Transformation of the sensor signals into the DFT domain provides the following relationship between the delay of (4) and the phase difference $\varphi_m(\phi,\theta;l)$ between the Fourier components of the observed signals.

$$\varphi_m(\phi,\theta;l)={-}{\vartriangle}\omega l\tau_m(\phi,\theta),\ m=2,\ldots M \tag{6}$$

where $\varphi_m(\phi,\theta;l)$ is unwrapped phase difference at *l*-th frequency bin, $\Delta\omega{=}2\pi f_s/L$ is the angular frequency width between adjacent DFT points, and f_s is the sampling frequency. The vector form of (6) is represented by

$$\begin{aligned}\varphi(\phi,\theta;l)&{=}[\varphi_2(\phi,\theta;l),\ldots\varphi_M(\phi,\theta;l)]^T\\&=-{\vartriangle}\omega l\tau(\phi,\theta)\end{aligned} \tag{7}$$

B. Estimated Phase Difference and Uncertainty

As in [8] let define $X_m(k,l)$ as the L-point STFT (Short Time Fourier Transform) of $x_m(t)$ where *k* is time frame index, *l* is frequency bin index. The phase difference inherently has uncertainty by an amount of the integer multiplying 2π, the estimate of delay (4) using the relationship (6) can be written as follow.

$$\hat{\tau}_m(k,l){=}-\frac{1}{{\vartriangle}\omega l}\{ARG[\frac{X_m(k,l)}{X_1(k,l)}]+2\pi\rho_m(l)\} \tag{8}$$

where $p_m(l)$ represents an unknown integer depending on *l*, and ARG[Y] means the principal value of complex number Y.

$$p(l){:=}[p_2(l),p_3(l),\ldots p_M(l)]^T \tag{9}$$

Now, defining a vector with integer elements yields the following vector form representation of the estimated delays (8) which contains unknown vector $p(l)$.

$$\hat{\tau}_p(L)(k,l){=}[\hat{\tau}_2(k,l),\hat{\tau}_3(k,l),\ldots\hat{\tau}_M(k,l)] \tag{10}$$

The issue addressing here is to select one appropriate vector $\hat{\rho}(l)$ at each *l* and obtain an estimate of $\hat{a}_{\rho(L)}(k,l)$ by solving the equation (5).

C. DOA Estimation in the Spatial Aliasing Case

As discussed in [11] the finite extent of unknown integer $p_m(l)$, that is $|p_m(l)|{<}p_m(l)$, is determined by the length of the location vector r_m of (1). The idea of selecting an appropriate $p(l)$ in [11] from possible combination of $|p_m(l)|$'s is the fact that any propagating vector should have unit norm, namely $a(\phi,\theta)=1$.

Assume N sources with different directions and for each (*k*,*l*)-cell, apply the following steps A~C to obtain an estimate propagation vector $\hat{a}_{p(L)}(k,l)$.

Step A: For all elements of $p(l)$, compute the estimate $\hat{\tau}_{p(L)}(k,l)$ by using (8) and (10).

Step B: Solve the equation to obtain $\hat{a}_{p(l)}(k,l)$ It is noted that under the conditions $M\geq$ 4 and rank***R***=3 Eq.(11) may yield unique solution. The generalized inverse of ***R*** and the Gram-Schmidt orthogonalization [10] are two known methods.

$$-\frac{R\hat{a}_{p(l)}(k,l)}{c}=\hat{\tau}_{p(L)}(k,l) \tag{11}$$

Step C: Obtain the following $\hat{p}(l)$ as the most probable unknown integer vector.

$$\hat{p}(l)=\mathrm{Arg\,Min}(\left\|a_{\mathrm{p(l)}}(k,l)\right\|\text{-}1) \tag{12}$$

For a set of $\hat{a}_{\hat{p}(l)}(k,l)$ with available (*k*, *l*), we may have N clusters, and *n*-th cluster gives a propagation direction vector a_n. Several methods for determining a_n (n=1~N) have been proposed, such as the histogram peak search, centroid search by *k*-means algorithm, and the peak search of kernel density estimation. [3]-[6]

III. Source Separation via T-F Masking

The task of separation process based on time-frequency masking is to determine the most dominant source at each time-frequency slot (*k*, *l*) in STFT domain. As shown in the previous section, estimated propagation direction vector $\hat{a}_{\hat{p}(l)}(k,l)$ at each (*k*, *l*) is used to attribute a spatial feature of T-F slot. In this context with multiple source case, the separation process is performed by classifying $\hat{a}_{\hat{p}(l)}(k,l)$ of all T-F slots (*k*, *l*) into N classes. After that, the *n*-th class consists of mixtures of T-F component where the *n*-th source is dominant. As in [8], we use the following notation for representing a case that an estimated $\hat{a}_{\hat{p}(l)}(k,l)$ belongs to the *n*-th class.

$$\mathrm{C\ (k,\ l) = n} \tag{13}$$

The proposed separation process consists of updating both (ϕ,θ) and *n* by replacing $\tilde{p}(l)$ and $\tilde{n}$ thorough the following minimization process.

$$(\tilde{p}(l),C(k,l)=\tilde{n}$$
$$:=Min_{\hat{p}(l),n=1\sim N}\left\|a_{\hat{p}(l)}(k,l)-a_n\right\| \tag{14}$$

The clustering or separation process is given by $(k,l)=\tilde{n}$. This classification result directly generates a T-F mask for separating $\tilde{n}$ -th source signal by the following masking procedure.

$$Y_{\tilde{n}}(k,l)=Y_{\tilde{n}}(k,l)X_1(k,l):=\left\{\begin{array}{ll} X_1(k,l), & if\ C(k,l)=\tilde{n} \\ 0, & otherwise \end{array}\right\} \tag{15}$$

Finally, application of inverse STFT to $Y_{\tilde{n}}(k,l)$ yields time-domain separated signal $Y_{\tilde{n}}(t)$ which is an approximation of $\tilde{n}$ -th source signal observed at the reference sensor.

IV. Experiments

DOA estimation and separation experiments using tetrahedron microphone array, as shown in Fig. 1, with the following condition are conducted.

A. Experimental Setup

Sampling frequency:	8000Hz
Microphone distance:	8cm
Window shape:	Hamming
Window length:	512points
Room (Width, Depth):	(18m, 15m)
Reverberation Time:	1200ms

It is noted that this setting of 8-cm spacing for 8kHz sampling may cause special aliasing.

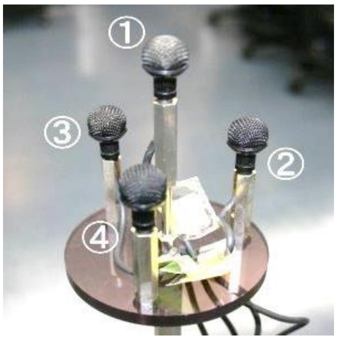


Figure 1. Tetrahedron microphone array

B. DOA Estimation and Clustering

The azimuth and elevation angles (degree) of four sources are given as follows.

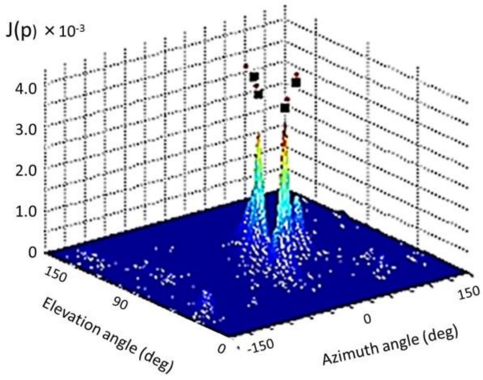
Source1: $(\phi_1,\theta_1)=(0,90)$,
Source2: $(\phi_2,\theta_2)=(0,120)$),
Source3: $(\phi_3,\theta_3)=(30,60)$,
Source4: $(\phi_4,\theta_4)=(60,90)$)

The results of DOA estimation and clustering by the proposed method are shown in Fig. 2. Fig. 2 (a) shows a profile of kernel density estimation using $\hat{a}_{\hat{p}(l)}(k,l)$ on the (ϕ,θ) plane. Four prominent peaks appeared in the profile correspond to individual source’s directions. The estimated individual DOAs are given as follows.

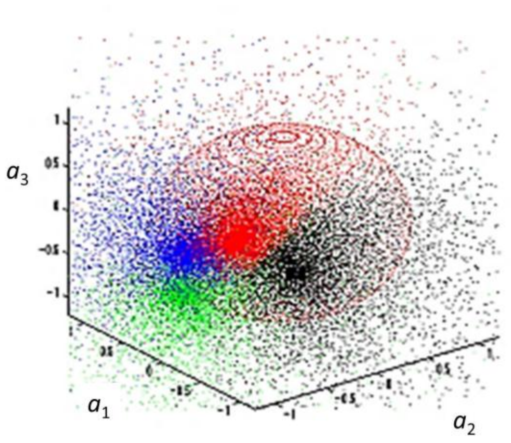
$$(\hat{\phi}_1,\hat{\theta}_1)=(0,92)\,,\ (\hat{\phi}_2,\hat{\theta}_2)=(-2{,}199)$$

$$(\hat{\phi}_3,\hat{\theta}_3)=(32,66)\,,\ (\hat{\phi}_4,\hat{\theta}_4)=(59,90)$$

Fig. 2 (b) shows a distribution of estimation $\hat{a}_{\hat{p}(l)}(k,l)$ for all (*k*, *l*)-slots. In the figure, each different color clustered set (each class has different color) corresponds to an individual source. Fig. 2 (c) shows the evaluation function J for the same example. J is the introduced index defined and utilized for DOA estimation in [9]. From the resulted evaluation index it is so hard to detect four distinctive peaks at accurate positions.

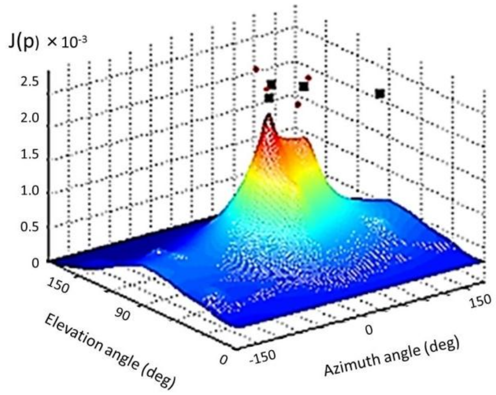


(a) Estimated density result by proposed method
Four prominent peaks at accurate positions are appeared



(b) Clustering Result: Four clustering points are shown;

Source1: Blue, Source2: Green, Source3: Red, Source4: Black



(c) Evaluation function J by the method [9]
Three prominent peaks are appeared at correct positions, but one peak is misestimated)

Figure 2. DOA and clustering results

C. Separation Performance

To verify the effectiveness of the proposed clustering, we conducted experiments for the following two cases and compare it with the results by Loesch & Yang [9].

Case 1: Three-sources

$\{(\phi_{i,}\theta_i)\}_{i=1,2,3}$={(0,90),(120,60),(240,120)}

Case 2: Four -sources

$\{(\phi_{i,}\theta_i)\}_{i=1,2,3}$={(0,90),(120,60),(240,120)}

Here we evaluated the separation performance in terms of W-disjoint orthogonality [3]. The index formulation is given by,

$$WDO_M := \frac{\left\|M(k,l)S_D(k,l)\right\|^2 - \left\|M(k,l)S_I(k,l)\right\|^2}{\left\|S_D(k,l)\right\|^2} \qquad (16)$$

where $S_D(k,l)$ and $S_I(k,l)$ are the STFTs of the original and its separated signals respectively and *M*(*k*,*l*) is a binary mask. The results of WDO_M are shown in Table I.

TABLE I. SEPARATION PERFORMANCE WDO$_M$

METHOD	CASE 1	CASE 2
Loesch & Yang[8]	0.72	0.66
Proposed	0.78	0.69

The separation performance difference between the proposed and that of [9] is small. However, the advantage of the proposed method exists in its computational efficiency. Average computation times of the proposed and the conventional method [9] for separation with aliasing condition are 1 and 20 seconds respectively.

V. CONCLUSION

Source separation based on sparseness of speech signals is proposed. The method is applicable to arbitrary array configuration for multiple sources and even in spatial aliasing. At first, our recently established DOA estimation method is adopted to clustering spatial feature vectors. Then the ambiguity of phase is solved by using the unit norm property of propagating vector. Experimental results proved that the proposed method is an effective and fast clustering method for spatial aliasing case to source separation. The extension of this approach to source tracking method [13], [14] for spatial aliasing case will be future issues.

REFERENCES

[1] S. Makino, J. Chen, and Y. Huang, *Blind Speech Separation*, Springer, 2007.

[2] N. Hamada and N. Ding, "Source separation and DOA estimation for underdetermined auditory scene," *Soundscape Semiotics–Localization and Categorization*, H. Glotin, ed. ch. 1, 2014.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via. time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830-1847, July 2004.

[4] N. Roman, *et al.* "Speech segregation based on sound localization," *J. Acoust. Soc. Am*. vol. 114, no. 4, 2003.

[5] S. Ricard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," in Proc. ICA 2001, 2001, pp. 651-656.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiely Interscience, 2000, ch. 4.

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

[8] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 5, July 2007.

[9] B. Loesch and B. Yang, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," LVA/ICA 2010, St. Malo, Frankreich, September 2010.

[10] K. Fujimoto, N. Ding, and N. Hamada, "Multiple sources' direction finding by using reliable component on phase difference manifold and kernel density estimator," in *Proc. IEEE ICASSP '12*, Kyoto, Mar. 2012.

[11] M. Sekikawa and N. Hamada, "DOA estimation of multiple source using arbitrary microphone array configuration in the presence of spatial aliasing," in *Proc. IEEE, ISPACS2014*, Kuching, Malaysia, Dec. 2014.

[12] J. Dmochowskim, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," *IEEE Trans. SP*, vol. 57, no. 4, April 2009.

[13] D. Su, M. Sekikawa, K. Nakazawa, and N. Hamada, "Novel scheme of real-time direction finding and tracking of multiple speakers by robot-embedded systems," *1st Int. Con. on Robot Intelligence Tech. & Appl. (RiTA 2012)*, Gwangju, Republic of Korea, Dec. 2012.

[14] A. Kijima, Y. Hioka, and N. Hamada, "Tracking of multiple moving sound sources using particle filter for arbitrary microphone array configurations," in *Proc. 2012 IEEE ISPACS*, New Taipei city, Taiwan, November 4-7, 2012.

Masashi Sekikawa was born in Tokyo, Japan, in 1988. He received the Bachelor's of Engineering degree (System Design Engineering) and the Master's of Engineering from Keio University, Yokohama, Japan, in 2011 and 2013, respectively. During the Master program, he worked on microphone array system for blind separation and direction finding of multiple sound sources. From April 2013, he has been the network solution division of NTT DATA Corporation.

Nozomu Hamada was born in Tokushima, Japan in 1947. He received B.S., M.S. and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan. His research interests include circuit theory, stability theory of dynamical systems, digital signal processing, and image processing. His recent research field is the realization of a human interface system using microphone arrays. The main topics in this study are the acquisition and localization of audio signals from spatially distributed sound sources and the separation of multiple speech signals. In 1974, he joined the Department of Electrical Engineering in Keio University. From 1993 to 2013, he was a Professor in the System Design Engineering, Faculty of Science and Technology, Keio University. He is now a professor in the Electronic System Engineering at Malaysia-Japan International Institute of Technology (MJIIT), University Technology Malaysia (UTM). He is also a professor emeritus of Keio University. He was a visiting researcher at the Australian National University in 1982, an adjunct professor of Xi'an Jiaotong University and Xi'an Jiaotong University City College during 2006-2009, and a visiting scholar of EMARO program of Warsaw University of Technology in 2010. He is the author of "Linear Circuits, Systems and Signal Processing (Chapter 5), Marcell Dekker Inc., 1990, the co-author of "Two-Dimensional Signal and Image Processing" ,SICE1996, and the author of the book "Signal Processing" (Ohm-sha Pub. Inc. 2012). He served as a chair of the IEEE signal Processing Society, Japan Chapter (2004). Prof. Dr. Nozomu Hamada received the 2012 Best Paper Award from RISP. He is a fellow of IEICE, and a member of IEEE. He currently serves the chief editor of research paper in Journal of Signal Processing.