18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

DOA Estimation Based on the Microphone Array for the Time-varying Number of Sound Signals

Hongyan Zhu Kai Guo Yan Lin
Inst.of Integrated Automation
MOE KLINNS Lab
School of Electronic & Information Engineering,
Xi'an Jiaotong University,Xi'an, China 710049
hyzhu@mail.xjtu.edu.cn

Abstract -In this paper, the problem of estimating the directions of arrival (DOAs) for an unknown number of sound sources is addressed using a microphone array. In outdoor environments, the reverberation is not considered, so the pure delay mixture model is typically adopted here. Since the sound signal is usually wide-band and non-stationary, the Short Time Fourier Transform (STFT) is employed. However in practical applications, the missing or appearing of sources frequently occurs, resulting in a time-varying number of targets. To deal with this problem, a new preprocessing method is presented to partition the whole time domain into some small time windows. In each time window, the sample covariance is computed and used to produce the DOA estimate. Moreover, two methods (Eigenvalue-based and Information Theory-based) are used for estimating the number of independent sources. Some selected simulation results are given to demonstrate the estimation performance for the DOAs, and the time accuracy for detecting the change of the number of sources.

Keywords: Direction of Arrival (DOA), Number of Independent Sources, Short Time Fourier Transform, Array Signal Processing

1 Introduction

The estimation of Direction of arrivals (DOAs) for sound sources by using microphone arrays has been an active research topic since the early 1990's [1]. It has been widely used in many application areas, such as video conferencing [2], speech enhancement and speech recognition [3]. The fundamental principle behind the DOA estimation is to capture the phase information present in signals picked up by microphones. The multiple source localization problems can be resolved based on high-resolution subspace techniques, such as the MUSIC [4] and ESPRIT [5] algorithms.

Some methods of estimating DOA is introduced in [6], which is based on the pre-knowledge about the sound's waveform such as gun shots. To reduce the influence of the noise another DOA estimating method based on the Weighted Bispectrum Spatial Correlation Matrix is introduced in [7]. If two or more arrays are used together, then the Kalman Filter or Particle Filter can be used to locate and track the sound source based on the MUSIC algorithm， which is introduced in [8].

But there are also some limits when using MUSIC or ESPRIT algorithm for the DOA estimation. On one hand, the MUSIC and ESPRIT algorithms only work well in narrow-band, stationary signals, such as radar signals [4-5], in which the phase information is simply relative with arriving time lag. But in reality, especially for sound signals, the phase information is not only relative with arriving time lag but also the frequency of the signals, since the sound signal is usually wide-band and non-stationary [9]. On the other hand, estimating the DOAs of multiple sources requires the knowledge of the number of independent sources. However in practical applications, the number of signals is generally not known exactly. Moreover, the sound source may disappear or appear randomly.

In this paper, we develop a new preprocessing method to partition the whole time domain into some small time windows. This proposed method provides a tradeoff between reducing the influence of noise due to a short time window and improving the time accuracy for detecting the change of the number of sources. In each time window, the sample covariance is computed, and the DOAs can be obtained in different frequency points using the ESPRIT algorithm. As for the estimation of the number of sources, the Information Theoretical criteria [10-11] can be adopted. In this paper, performance comparisons are done based on two methods (Eigenvalue-based and Information Theory-based) for estimating the number of independent sources.

2 Problem Formulation

A microphone array with N microphones aligning in a linear form is assumed, and d is the distance between any two adjacent microphones. There are M sound sources placed in one side of the array, with different DOA θ_m for $m = 1,\cdots,M$. In this case, it is assumed that the number of sources is less than the number of microphones, i.e. $M < N$. So we can use the ESPRIT algorithm to estimate the DOAs. Fig.1 shows the microphone-target configuration geometry.

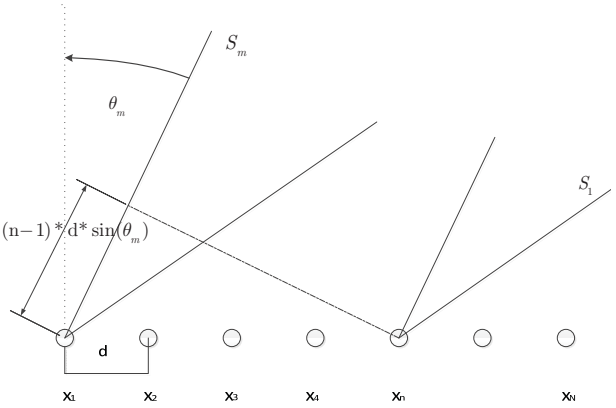

Fig.1. Microphone-target configuration geometry

Following the model in [12], the attenuation coefficient between source m and microphone n is set to 1. It is usually valid in a far field. M source signals are mixed and transmitted to microphone n with additive noise $n_n(t)$, which is zero mean white Gaussian noise. In this way, the mixed signal $x_n(t)$ received by microphone n can be written as

$$x_n(y) = \sum_{m=1}^{M} s_m(y\text{-}\tau_{nm}) + n_n(y) \tag{1}$$

where y is discrete time index, assume that $x_n(y)$ have L discrete points.

Then, the STFT is used to $x_n(y)$ due to the linearity of the STFT. So, we have:

$$X_n(p,k) = \sum_{m=1}^{M} S_m(p,k) e^{-j2\pi f_s \frac{k-1}{K}\tau_{nm}} + N_n(p,k) \tag{2}$$

where p is the window index, k represents the discrete frequency index, $k = 1,2...K$, K is the window length. To simplify the notation, we replace the discrete frequency index k by true frequency $f = \frac{k-1}{K} f_s$, $f = 0,\cdots, f_s/2$ due to the symmetry of the DFT.

$$X_n(p,k) = \sum_{l=1}^{K-1} x_n(p,l) e^{-j2\pi k \frac{l}{K}} \tag{3}$$

$$S_m(p,k) = \sum_{l=1}^{K-1} s_m(p,l) e^{-j2\pi k \frac{l}{K}} \tag{4}$$

$$N_n(p,k) = \sum_{l=1}^{K-1} n_n(p,l) e^{-j2\pi k \frac{l}{K}} \tag{5}$$

Considering microphone n at each frequency f, the $X_n(1,f), X_n(2,f), \cdots, X_n(P,f)$ is still a time sequence signal, where P is the number of windows. It is assumed that the noise at different microphone is uncorrelated, and the noise and the signals are uncorrelated.

To simplify the notation, the equations can be written as follows:

$$\mathbf{X}(p,f) = \mathbf{A}(f)\mathbf{S}(p,f) + \mathbf{N}(p,f) \tag{6}$$

where $\mathbf{X}(p,f) = [\mathrm{X}_1(p,f), \cdots, \mathrm{X}_N(p,f)]^{\mathrm{T}}$,

$\mathbf{A}(f) = [\mathbf{a}_1(\theta_1, f), \cdots, \mathbf{a}_M(\theta_M, f)]$,

Each column of $\mathbf{A}(f)$ is

$\mathbf{a}_m(\theta_m, f) = [e^{-j2\pi f*0*d*\sin(\theta_m)/c}, e^{-j2\pi f*1*d*\sin(\theta_m)/c}, \cdots, e^{-j2\pi f*(N-1)*d*\sin(\theta_m)/c}]^{\mathrm{T}}$ so $\mathbf{A}(f)$ is a $N \times M$ mixing matrix. The source signal is $\mathbf{S}(p,k) = [S_1(p,f), \cdots, S_N(p,f)]^{\mathrm{T}}$, and the noise is $\mathbf{N}(p,k) = [N_1(p,f), \cdots, N_N(p,f)]^{\mathrm{T}}$.

The main objective is to estimate the number of sound sources and the corresponding DOAs θ_{m}. After that, the mixing matrix $\tilde{\mathbf{A}}(f)$ can be estimated. The frequency-domain Blind Source Separation (BSS) problem [13] could be done by using de-mixing matrix $\tilde{\mathbf{W}}(f) = \tilde{\mathbf{A}}(f)^{-1}$ .The permutation problem [14] can be solved by the estimated DOAs. Finally the source signals can be recovered by inverse STFT.

3 Estimating DOA of Variable Number of Sound Sources

To implement the STFT, the received signal is divided into P windows, and each window has K values. The window width K is a critical parameter. Even if the sound signal is not stationary, the short time period of the signal can be approximately regarded as stationary. However, a short window width is not good for the frequency resolutions, whereas a long window width is not good for the time resolutions. What we want to do is to present a preprocessing method to partition the time window efficiently.

Before using the ESPRIT algorithm, it is required to determine the number of the sources in every time window, for the sound source may appear or disappear randomly. Then, time-frequency domain mixed signals are normalized, so that the normalized signals have zero mean and unit variance.

3.1 Preprocessing

The covariance matrix of $\mathbf{X}(p,f)$ is

$$\mathbf{R}_{\mathbf{xx}}(p,f) = \mathbf{X}(p,f)\mathbf{X}^{\mathrm{H}}(p,f) \tag{7}$$

where $(\bullet)^{\mathrm{H}}$ is the conjugate transpose operator. In reality, the eigenvalue decomposition is not directly used to matrix

$\mathbf{R}_{xx}(p,f)$, because the $\mathbf{R}_{xx}(p,f)$ is often singularity when given a certain time and frequency. So, in general, the time-domain mean of $\mathbf{R}_{xx}(p,f)$ is used jointly, i.e. $\mathbf{R}_{xx}(f)=\frac{1}{P}\sum_{p=1}^{P}\mathbf{X}(p,f)\mathbf{X}^{H}(p,f)$. However, doing this ignores the signal difference due to the randomly appearing or missing of sound sources. In fact, it is difficult for us to determine when the sources appear or disappear. So, we aim to find a tradeoff between reducing the influence of noise due to a short time window and improving the time accuracy for detecting the change of the number of sources.

The time-domain mean of $\mathbf{R}_{xx}(p,f)$ is still needed here. However, we only sum up those time windows that overlap with each other, but not the whole time-domain.

Define the time mean of $\mathbf{R}_{xx}(p,f)$ in a short time period as

$$\bar{\mathbf{R}}_{xx}(q,f)=\frac{K-O}{K}\sum_{p=(q-1)*K+1}^{(q-1)*K+K/(K-O)}\mathbf{R}_{xx}(p,f) \qquad (8)$$

where $q=[1,2....Q]$ is a new window index, Q is the minimum integer larger than L/K, O is the overlap points in STFT. In fact, $\bar{\mathbf{R}}_{xx}(q,f)$ includes not only all the data between time $(q-1)K+1$ to $(q-1)K+K$, but also the next K points. That means $\bar{\mathbf{R}}_{xx}(q,f)$ represent a $2K$ points information, and the time-domain accuracy is $2K/f_s$. For example, the frequency-domain window width is $K=8$ points, and the overlap of two adjacent time windows is $O=7$ points. We only sum up $K/[K-O]=8$ windows from $p=1$ to $p=8$ when computing the time-domain mean. Fig.2 shows this basic idea.

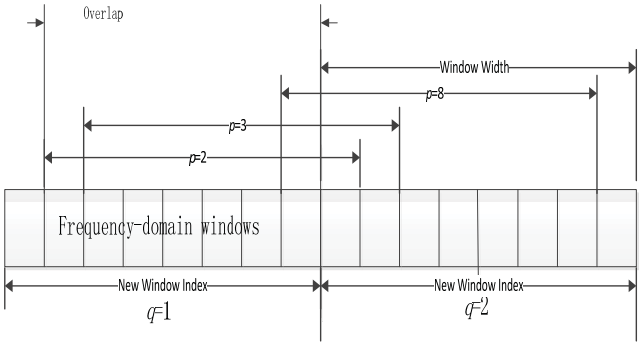

Fig.2. Example for $\bar{\mathbf{R}}_{xx}(q,f)$

3.2 Estimating the number of sources

$\bar{\mathbf{R}}_{xx}(q,f)$ carries the information at frequency f in time window $(q-1)K+1$ to $(q-1)K+K$. In this time window, the eigenvalue decomposition as for $\bar{\mathbf{R}}_{xx}(q,f)$ used to estimate the number of sources.

$$\bar{\mathbf{R}}_{xx}(q,f)\mathbf{V}(q,f)=\mathbf{V}(q,f)\mathbf{\Lambda}(q,f) \qquad (9)$$

where $\mathbf{V}(q,f)$ is the matrix formed by the corresponding eigenvectors, $\mathbf{V}(q,f)=[\mathbf{v}_1(q,f)\ \mathbf{v}_2(q,f)\ \cdots\ \mathbf{v}_N(q,f)]$, and $\mathbf{\Lambda}(q,f)$ is the diagonal matrix formed by eigenvalues in the descending order,

$$\mathbf{\Lambda}(q,f)=\mathrm{diag}\{\lambda_1(q,f),\lambda_2(q,f)\cdots\lambda_N(q,f)\},$$
$$\lambda_i(q,f)\le\lambda_j(q,f),\text{for } i\ge j$$

Two ways to estimate the number of sources are adopted here.

(1) Eigenvalue-based

It is well known that, the larger eigenvalues correspond to the signals, and the smaller eigenvalues correspond to the noise. A simple way is given to find the number of larger eigenvalues .

By calculating

$$e_1=\lambda_1(q,f)/\lambda_2(q,f)$$
$$e_2=\lambda_2(q,f)/\lambda_3(q,f)$$
$$\cdots$$
$$e_{N\text{-}1}=\lambda_{N-1}(q,f)/\lambda_N(q,f)$$

The number $\hat{M}$ of sources can be determined by finding out the maximum,

$$\hat{M}=\arg\max_m(e_1,e_2,\cdots e_{N-1}) \qquad (10)$$

(2) Information Theory-based

Another way to find the number of sources is the Information Theoretical Criteria. The Akaike Information Criteria (AIC) is a mature criterion usually used in estimating the model order. The AIC is defined by

$$AIC=-2\log f(\mathrm{X}\mid\hat{\Theta})+2\varphi \qquad (11)$$

where φ is the number of independent parameter, $\hat{\Theta}$ is the estimated parameter set, and $f(\mathrm{X}\mid\hat{\Theta})$ is the likelihood function.

For given q,f, the AIC criterion is computed by

$$AIC(m)=-2\frac{K}{K-O}(N-m)\log\rho(m)+2m(2N-m) \qquad (12)$$

where

$$\rho(m)=\frac{(\lambda_{m+1}\lambda_{m+2}\cdots\lambda_N)^{\frac{1}{N-m}}}{\frac{1}{N-m}(\lambda_{m+1}+\lambda_{m+2}+\cdots+\lambda_N)}$$

So, the number of sources can be estimated by

$$\hat{M}=\arg\min_m(AIC(m)),m=1,2...N-1 \qquad (13)$$

3.3 ESPRIT algorithm

when given a certain parameter f and q, with the estimated number $\hat{M}$ of sources, the ESPRIT algorithm can be used to find out the DOAs $\hat{\theta}_m(q,f)$ .The details are as follows.

(1) Select out $\hat{M}$ eigenvector $\mathbf{v}_1(q,f),\cdots,\mathbf{v}_{\hat{M}}(q,f)$ corresponding to the $\hat{M}$ largest eigenvalues;

(2) Form matrixs: $\mathbf{G}_1(q,f)=[\mathbf{v}_1(q,f),\cdots,\mathbf{v}_{\hat{M}-1}(q,f)]$ and $\mathbf{G}_2(q,f)=[\mathbf{v}_2(q,f),\cdots,\mathbf{v}_{\hat{M}}(q,f)]$;

(3) Calculate the matrix: $\boldsymbol{\mu}(q,f)=(\mathbf{G}_1^{\mathrm{H}}(q,f)\mathbf{G}_1(q,f))^{-1}\mathbf{G}_1^{\mathrm{H}}(q,f)\mathbf{G}_2(q,f)$

(4) Find out the $\hat{M}$ eigenvalues of matrix $\boldsymbol{\mu}(q,f)$ $\left\{\lambda_{\mu 1}(q,f),\cdots,\lambda_{\mu m}(q,f),\cdots,\lambda_{\mu\hat{M}}(q,f)\right\}$

(5) Calculate the DOAs: $\hat{\theta}_m(q,f)=\arcsin\{\mathrm{Im}\{\ln(\lambda_{\mu m}(q,f))c/(2\pi fd)\}\},\mathrm{m}=1,2,\cdots\hat{M}$

where the $\mathrm{Im}(\bullet)$ show the image part of a complex number.

From this, it can be seen that, in a certain time window q, there are lot of DOA estimates $\hat{\theta}_m(q,f)$ in different frequency points. Although these estimates correspond with the same source, they may differ severely in some frequency points. Such a situation is mainly due to noise, calculating errors and the inherent defect of the ESPRIT algorithm. So, some measures must be taken to achieve the final DOA in time window q.

3.4 Final DOA estimation

Note that the estimation performance of DOAs is sensible to the frequency in the ESPRIT algorithm. Too low or high frequency may result in an inaccurate DOA estimates. Given d, the maximum frequency that the array can work normally is $c/2d$. If the frequency is higher than $c/2d$, the array will meet the spatial aliasing problem. For the lower frequency point, the wavelength is too long. When the phase only changes a little bit, the array fails to capture the changes precisely.

In [12], the estimates $\hat{\theta}_m(q,f)$ in different frequency points are used to approximate the probability density function of the real DOAs in a certain time window q. Then, the final DOA $\tilde{\theta}_m(q)$ can be obtained by using the maximum likelihood estimation (MLE) .

The Parzen-windows is a non-parametric method to estimate the probability distribution by

$$P(\theta_m(q))=\frac{1}{Fh}\sum_f Ker(\frac{\theta_m(q)-\hat{\theta}_m(q,f)}{h})$$

where F is the samples. In our problem, it is the number of frequency points. h is a smoothing parameter called bandwidth. $Ker(\bullet)$ is the kernel. Here, we adopt Gaussian kernel. The algorithm in [15] provided a way to choose the optimal bandwidth h .The final DOAs is

$$\tilde{\theta}_m(q)=\arg\max_{\theta_m(q)}P(\theta_m(q))$$

4 Simulation results

In this section, some selected simulation results are given to illustrate the estimation performance for the DOAs, and the time accuracy for detecting the change of the number of sources.

4.1 Parameter settings

Two typical scenarios are considered here.

(1) Scenario 1

Two sources appears all the time. This stage is used to show the performance for the DOA estimation of wide-band signals, and illustrate the effect by using the Parzen window to estimate the probability density function and then determine the final DOA based on the MLE.

(2)Scenario 2

Two sources appears all the time. The third source appears in some time after the beginning and stays for a period of time, and then disappears. This stage is designed to detect the change of the number of sources, and determine the appearing and missing time for the third source. The design parameters are shown in Table.1.

TABLE 1 PARAMETER SETTING

Parameters	Values
Source categories	S1：English speech; S2：Drum; S3：Sound of water flow
Source1 DOA	Scenario 1: 45 degree Scenario 2: 60 degree
Source2 DOA	Scenario 1: -60 degree Scenario 2: 30 degree
Source3 DOA	Scenario 2: -60 degree
Source1,2 length	80000 points ;3.628s
Source3 length	30000 points;1.36s
Source3 appear disappear time	Appear at 1.36s; Disappear at 2.72s
Number of microphones	Scenario 1: 4 Scenario 2: 5
Array spacing *d*	0.005m
Sample rate *fs*	22050Hz
Window Width	512 points
Overlap	Scenario 1:256 points Scenario 2:496 points
FFT window	Hamming
Monte Carlo	100

4.2 Experiment results

(1) Scenario **1**

Fig.3 shows the results of the DOA estimate in different frequencies for scenario 1.

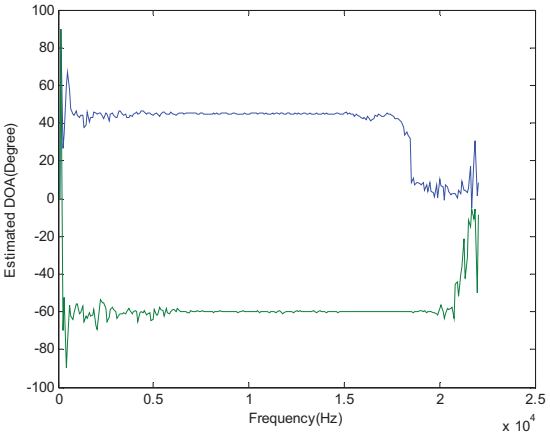

Fig.3. The DOA estimate in different frequency, (one source is at 45 degree and another at -60 degree)

From Fig.3, it can be seen that two DOAs for two sources can be determined well for most frequency values. However, the estimation performance is poor in parts of low frequency and high frequency. The final DOA estimate will be obtained in the sense of MLE based on the estimated probability density function. Here, only the estimated probability density for source 2 is plotted in Fig.4.

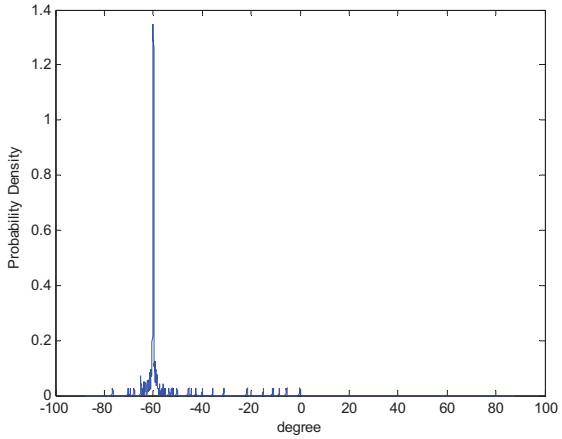

Fig.4. The final DOA estimate for source 2 (DOA: -60 degree, SNR=30dB).

Fig.4. shows that a satisfied DOA estimation can be achieved by introducing the MLE. Fig.5. shows the RMSE of estimated DOAs versus SNR. It is easily observed that the RMSE decreases as the SNR grows.

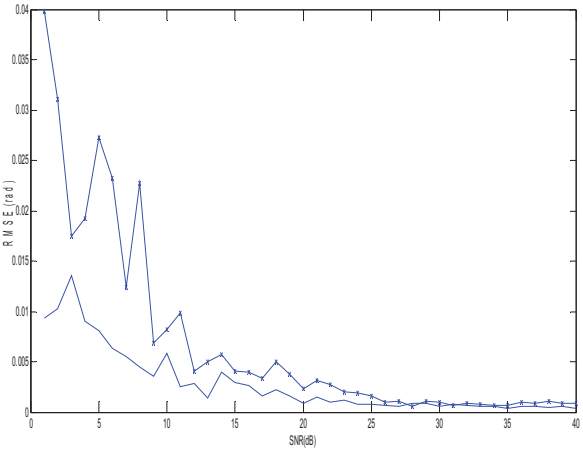

Fig.5. RMSE for the DOA estimation

(2) Scenario **2**

Figs. 6-7 depict the probability densities for a certain DOA (x-axis), and time (y-axis). The darker the points are, the higher the probability densities are.

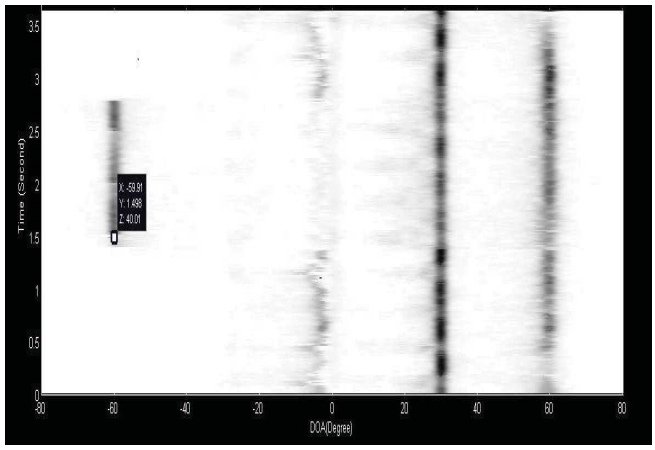

Fig.6. The appearing of the third source at time 1.498s,DOA -59.91 degree using AIC

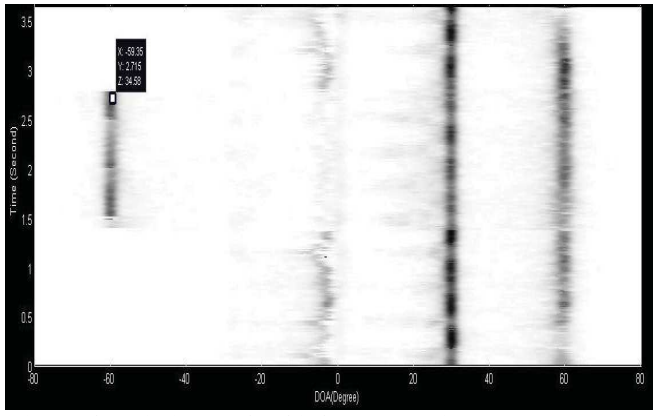

Fig.7. The missing of the third source at time 2.715s DOA -59.35 degree using AIC

The Information Theory-based method to estimate the source number can discover efficiently the appearance and disappearance of the third source (-60 degree). The estimated appearing time is1.498s (the true time is 1.36s), and the estimated missing time is 2.715s (the true time is 2.72). The reason for the deviation lies in two aspects. One is that the time accuracy for $\bar{\mathbf{R}}_{xx}(q,f)$ is $2K / f_s$ points (about 0.045s) ; The other one is that the signal's amplitude may be so small at the beginning that the algorithm considered it as noise.

Fig.8 show the results by using Eigenvalue-based method to estimate the number of sources. We can see that the probability density for the signal located at 60 degree is less concentrated than that in Fig.6. It reveals that the Information Theory-based method outperforms the Eigenvalue-based method when estimating the number of sources.

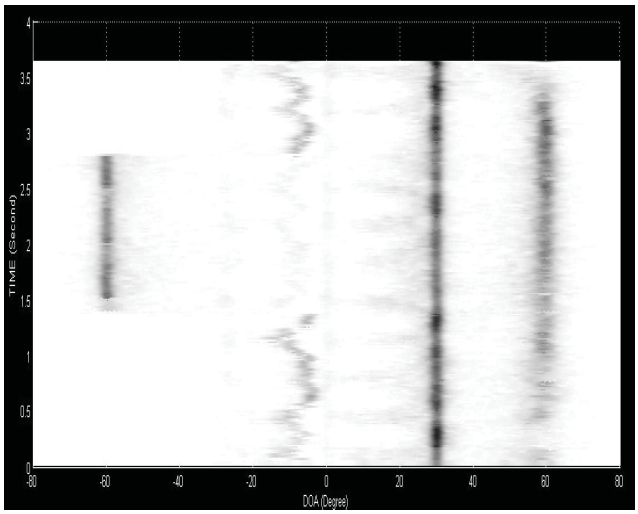

Fig.8. The example of using eigenvalue-based method

5 Conclusions and Future Work

This paper proposed an algorithm to estimate the wide-band signals' DOAs based on the microphone array when the number of sound signals changes over time. The algorithm is robust to noise and to the non-stationary signals. But the proposed approach has its own problems. The DOA estimation based on the ESPRIT algorithm does not behave well for the low and high frequency components. In addition, as the number of microphones and sound sources grow, the computation burden may increase dramatically. Future work includes how to reduce the computational load, and the investigation for multiple moving sound sources.

References

[1] M. Brandstein and H. Silverman, "A practical methodology for speech localization with microphone arrays," *Technical Report*, Brown University, November 13, 1996.

[2] S. Birchfield and D. Gillmor, "Acoustic source direction by hemisphere sampling," *Proc. of ICASSP2001*, pp.3053–3056, May 2001.

[3] M. S. Brandstein and S. M. Griebel, "Nonlinear, model based microphone array speech enhancement," *Acoustic Signal Processing for Telecommunications*, Kluwer Academic Publishers, 2000.

[4] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation*, IEEE Transactions on, vol. 34, no. 3, pp.276–280, 1986.

[5] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *IEEE Trans. ASSP*, vol. 37, no. 7, pp. 984-995, July 1989.

[6] Borzino, A.M.C.R.; Apolinario, J.A.; de Campos, M.L.R., "Estimating direction of arrival of long range gunshot signals," Telecommunications Symposium (ITS), 2014 International , vol., no., pp.1,5, 17-20 Aug. 2014

[7] Xue, W.; Liu, W.; Liang, S., "Noise Robust Direction of Arrival Estimation for Speech Source with Weighted Bispectrum Spatial Correlation Matrix," Selected Topics in Signal Processing, IEEE Journal of , vol.PP, no.99, pp.1,1

[8] Hung-Kuang Hao; Hang-Ming Liang; Yi-Wen Liu, "Particle methods for real-time sound source localization based on the Multiple Signal Classification algorithm,"Intelligent Green Building and Smart Grid (IGBSG), 2014 International Conference on, vol., no., pp.1,5, 23-25 April 2014

[9] J. Dmochowski, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," *Signal Processing*, IEEE Transactions on, vol. 57, no. 4, pp. 1383–1395, 2009.

[10] H. Akaike, "A new look at the statistical model identification," *Automatic Control*, IEEE Transactions on, vol. 19, no. 6, pp. 716–723, 1974.

[11] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[12] Sun L, Cheng Q, "Real-time microphone array processing for sound source separation and localization", *Information Sciences and Systems (CISS)*, 2013 47th Annual Conference on, 2013:1 - 6.

[13] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw.*, vol. 13, no. 4-5, pp. 411–430, 2000.

[14] Andrzej Cichocki,Juha Karhunen and Wlodzimierz Hasprzak, "Neural networks for blind separation with unknown number of sources," *Neurocomputing*, 1999.24: 55~93

[15] Wong K, Zhang Q, Reilly J P, "On Information theoretic criteria for determining the number of signal in high resolution array processing," IEEE, Trans.ASSP-38, Nov.1990, p1959-1970