The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

# Estimation of Sound Source Number and Directions under a Multi-source Environment

Jwu-Sheng Hu, *Member IEEE*, Chia-Hsing Yang, *Student Member IEEE*, and Cheng-Kang Wang

***Abstract*—Sound source localization is an important feature in robot audition. This work proposes a sound source number and directions estimation method by using the delay information of microphone array. An eigenstructure-based generalized cross correlation method is proposed to estimate time delay between microphones. Upon obtaining the time delay information, the sound source direction and velocity can be estimated by least square method. In multiple sound source case, the time delay combination among microphones is arranged such that the estimated sound speed value falls within an acceptable range. By accumulating the estimation results of sound source direction and using adaptive K-means++ algorithm, the sound source number and directions can be estimated.**

## I. Introduction

AUDITION is one of the important senses for humans, animals, or robot to recognize the environment. The sound information value would be limited if the humans, animals, or robot can not know the position of sound source. Hence, the need for sound source localization has become a fundamental feature for audition. For robot audition, the key advantage is that the robot can use many microphones for hearing system and it's unlike a human-like audition system using only two microphones.

The idea of using multiple microphones to localize sound sources has been developed for a long time. Among various kinds of sound localization methods, generalized cross correlation (GCC) based methods [1]-[3] are one of the major categories discussed for robot localization application [4]. Another approach, proposed by Balan *et al.* [5], explores the eigenstructure of the correlation matrix of the microphone array by separating speech signals and noise signals into two orthogonal subspaces. The direction-of-arrival (DOA) is then estimated by projecting the manifold vectors onto the noise subspace. MUSIC [6] is one of the most popular methods for estimating the sound source direction and it is also applied to the robot audition system [7].

Walworth *et al.* [8] proposed a linear equation formulation for the estimation of the three-dimensional (3-D) position of a wave source based on the time delay values. Valin *et al.* [9] gave a simple solution for the linear equation in [8] based on the far field assumption. They have also developed a novel weighting function method to estimate the time delay.

Yao *et al.* [10] presented an efficient blind beamformer technique to estimate the time delays from the dominant source. This method estimated the relative time delay from the dominant eigenvector computed from the time-averaged sample correlation matrix. They have also formulated a source linear equation similar with [8] to estimate the source location and velocity by using least square method.

The objective of this work is to estimate the multiple sound source directions without the priori information of the sound source number. This work proposes a novel eigenstructure-based GCC (ES-GCC) method to estimate the time delay between two microphones. With the time delay information, the sound source direction and velocity can be obtained by solving the proposed linear equation model. Moreover, without the priori knowledge of sound source number, the method for estimating sound source number and directions using the estimated sound velocity as a filtering criterion and adaptive K-means++ is proposed and evaluated in a real environment.

## II. Time Delay Estimation

Consider an array with $M$ microphone in a noisy environment. The received signal of the $m$-th microphone which contains $D$ sources can be described as:

$$x_m(t) = \sum_{d=1}^{D} a_{md} s_d(t - \tau_{md}) + n_m(t) \tag{1}$$

$a_{md}$ and $\tau_{md}$ are the amplitude and time delay from the $d$-th sound source to the $m$-th microphone and $n_m(t)$ is the non-directional noise. It is assumed that $s_d(t)$ and $n_m(t)$ are mutually uncorrelated and sound source signals are mutually independent. Applying the discrete time Fourier transform (DTFT) to (1), we have:

$$X_m(\omega,k) = \sum_{d=1}^{D} a_{md} S_d(\omega,k) e^{-j\omega\tau_{md}} + N_m(\omega,k) \tag{2}$$
$$\omega = 0,1,\ldots,N_{DTFT}-1$$

where $\omega$ is the frequency band and $k$ is the frame number. Rewrite (2) in matrix form:

$$\boldsymbol{X}(\omega,k) = \mathbf{A}(\omega)\boldsymbol{S}(\omega,k) + \boldsymbol{N}(\omega,k) \tag{3}$$

where

$$\boldsymbol{X}(\omega,k) = \left[X_1(\omega,k), \quad \cdots, \quad X_M(\omega,k)\right]^{\mathrm{T}} \in C^{M\times 1}$$

Manuscript received March 1, 2009. This work was supported in part by the National Science Council of Taiwan, ROC under grant NSC 96-2628-E-009-163-MY3 and the Ministry of Economic Affairs under grant 97-EC-17-A-04-S1-054.

J.S. Hu, C.H.Yang and C.K.Wang are with Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, ROC. (e-mail: {jshu@cn.; chyang.ece92g@; papa.ece91g@}nctu.edu. tw)

$$\boldsymbol{N}(\omega,k)=\left[N_1(\omega,k),\quad \cdots,\quad N_M(\omega,k)\right]^{\mathrm{T}}\in C^{M\times 1}$$
$$\boldsymbol{S}(\omega,k)=\left[S_1(\omega,k),\quad \cdots,\quad S_D(\omega,k)\right]^{\mathrm{T}}\in C^{D\times 1}$$
$$\mathbf{A}(\omega)=\begin{bmatrix} a_{11}e^{-j\omega\tau_{11}} & \cdots & a_{1D}e^{-j\omega\tau_{1d}} \\ \vdots & & \vdots \\ a_{M1}e^{-j\omega\tau_{M1}} & \cdots & a_{MD}e^{-j\omega\tau_{MD}} \end{bmatrix}\in C^{M\times D}$$

The received signal correlation matrix with eigenvalue decomposition can be described as:

$$\begin{aligned}\mathbf{R}_{xx}(\omega)&=\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{X}(\omega,k)\boldsymbol{X}^{\mathrm{H}}(\omega,k)\\&=\sum_{i=1}^{M}\lambda_i(\omega)\boldsymbol{V}_i(\omega)\boldsymbol{V}_i^{\mathrm{H}}(\omega)\end{aligned}\tag{4}$$

where H denotes conjugation transpose; $\lambda_i(\omega)$ and $\boldsymbol{V}_i(\omega)$ are eigenvalues and corresponding eigenvectors with $\lambda_1(\omega)\ge\lambda_2(\omega)\ge\cdots\ge\lambda_M(\omega)$ and $\boldsymbol{V}_1(\omega)$ is the maximum eigenvalue corresponding eigenvector which is defined as:

$$\boldsymbol{V}_1(\omega)=\left[V_{11}(\omega)\quad V_{21}(\omega)\quad\cdots\quad V_{M1}(\omega)\right]^{\mathrm{T}}\in C^{M\times 1}\tag{5}$$

We call eigenvector $\boldsymbol{V}_1(\omega)$ principal component vector since it contains the direction information of the principal sound sources at each frequency. Hence, this work adopts the principal component vector as the microphone received signal for time delay estimation.

Finally, the ES-GCC function between the *i*-th and *j*-th microphone can be represented as:

$$R_{x_ix_j}(\tau)=\sum_{\omega=0}^{N_{DFET}-1}\frac{1}{\left|V_{i1}(\omega)V_{j1}(\omega)\right|}V_{i1}(\omega)V_{j1}(\omega)e^{j\omega\tau}\tag{6}$$

The time delay sample can be estimated by finding the maximum peak of the ES-GCC function:

$$\hat{\tau}^{1}_{x_ix_j}=\arg\max_{\tau}R_{x_ix_j}(\tau)\tag{7}$$

## III. Sound Source Localization and Speed Estimation

### *A. Sound Source Location Estimation Using Least-Square Method*

Once the time delay is obtained, the sound source location can be estimated from the microphone array geometrical calculation. The work in [10] provides a linear equation model for estimating the source localization and propagation speed. Consider sound source location vector $\boldsymbol{r}_s=\left[x_s\quad y_s\quad z_s\right]$, the *i*-th microphone location $\boldsymbol{r}_i=\left[x_i\quad y_i\quad z_i\right]$ and the relative time delays, $t_i-t_j$, between the *i*-th sensor and *j*-th sensor. The relative time delay satisfy

$$t_i-t_1=\frac{\left|\boldsymbol{r}_i-\boldsymbol{r}_s\right|-\left|\boldsymbol{r}_1-\boldsymbol{r}_s\right|}{v}\tag{8}$$

where *v* is the speed of sound. Equation (8) is equivalent to

$$t_i-t_1+\frac{\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|}{v}=\frac{\left|(\boldsymbol{r}_i-\boldsymbol{r}_1)-(\boldsymbol{r}_s-\boldsymbol{r}_1)\right|}{v}\tag{9}$$

Squaring both sides, we have:

$$(t_i-t_1)^2+2(t_i-t_1)\frac{\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|}{v}=\left(\frac{\left|\boldsymbol{r}_i-\boldsymbol{r}_1\right|}{v}\right)^2-\frac{2(\boldsymbol{r}_i-\boldsymbol{r}_1)\cdot(\boldsymbol{r}_s-\boldsymbol{r}_1)}{v^2}\tag{10}$$

With the algebraic manipulation, the equation (10) becomes:

$$-\frac{(\boldsymbol{r}_i-\boldsymbol{r}_1)\cdot(\boldsymbol{r}_s-\boldsymbol{r}_1)}{v\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|}+\frac{\left|\boldsymbol{r}_i-\boldsymbol{r}_1\right|^2}{2v\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|}-\frac{v(t_i-t_1)^2}{2\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|}=(t_i-t_1)\tag{11}$$

Next, define the unit vector of the source direction as,

$$\boldsymbol{w}_s\equiv\left[w_1\quad w_2\quad w_3\right]^{\mathrm{T}}=\frac{\boldsymbol{r}_s-\boldsymbol{r}_1}{v\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|}\tag{12}$$

And two other variables as,

$$w_4=\frac{1}{2v\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|},w_5=\frac{v}{2\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|}\tag{13}$$

Hence, the linear equation (11) considering all *M* microphones can be written as,

$$\mathbf{A}_g\boldsymbol{w}=\boldsymbol{b}\tag{14}$$

where $\boldsymbol{w}\equiv\left[w_1\quad w_2\quad w_3\quad w_4\quad w_5\right]^{\mathrm{T}}$

$$\mathbf{A}_g=\begin{bmatrix} -(\boldsymbol{r}_2-\boldsymbol{r}_1) & \left|\boldsymbol{r}_2-\boldsymbol{r}_1\right|^2 & -(t_2-t_1)^2 \\ -(\boldsymbol{r}_3-\boldsymbol{r}_1) & \left|\boldsymbol{r}_3-\boldsymbol{r}_1\right|^2 & -(t_3-t_1)^2 \\ \vdots & \vdots & \vdots \\ -(\boldsymbol{r}_M-\boldsymbol{r}_1) & \left|\boldsymbol{r}_M-\boldsymbol{r}_1\right|^2 & -(t_M-t_1)^2 \end{bmatrix},\boldsymbol{b}=\begin{bmatrix} t_2-t_1 \\ t_3-t_1 \\ \vdots \\ t_M-t_1 \end{bmatrix}$$

For more than 5 sensors, the least square solution of equation is given by:

$$\hat{\boldsymbol{w}}=\left[\hat{\boldsymbol{w}}_s^{\mathrm{T}}\quad\hat{w}_4\quad\hat{w}_5\right]^{\mathrm{T}}=\left[\hat{w}_1\quad\hat{w}_2\quad\hat{w}_3\quad\hat{w}_4\quad\hat{w}_5\right]^{\mathrm{T}}=\left(\mathbf{A}_g^{\mathrm{T}}\mathbf{A}_g\right)^{-1}\mathbf{A}_g^{\mathrm{T}}\boldsymbol{b}\tag{15}$$

The source location and speed of sound can be solved as,

$$\tilde{\boldsymbol{r}}_s=\frac{\hat{\boldsymbol{w}}_s}{2\hat{w}_4}+\boldsymbol{r}_1,\tilde{v}=\sqrt{\frac{\hat{w}_5}{\hat{w}_4}}\left(\text{or }\tilde{v}=\frac{1}{\left|\hat{\boldsymbol{w}}_s\right|}\right)\tag{16}$$

### *B. Sound Source Direction Estimation Using Least-Square Method For Far Field Case*

To solve (14), the matrix $\mathbf{A}_g$ must have full rank. However, for matrix $\mathrm{A}_g$, the condition on rank is more complicated and will be ill-conditioned easily. For example, if the microphones are distributed on a spherical surface (i.e. $\boldsymbol{r}_i=\left[R\cos\theta_i\sin\phi_i\quad R\sin\theta_i\sin\phi_i\quad R\cos\phi_i\right]$, $R$ is radius and $\theta_i$ and $\phi_i$ are azimuth and elevation angle respectively.), it can be verified that the fourth column in $\mathrm{A}_g$ is the linear combination of column 1, 2 and 3.

Hence, this section will derive and analyze the linear equation model (11) for far field case. Define $\bar{\boldsymbol{r}}_s$ and $\rho_i$ as,

$$\bar{\boldsymbol{r}}_s=\frac{\boldsymbol{r}_s-\boldsymbol{r}_1}{\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|}\quad\text{and}\quad\rho_i=\frac{\left|\boldsymbol{r}_i-\boldsymbol{r}_1\right|}{\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|}\tag{17}$$

$\bar{\boldsymbol{r}}_s$ represents the unit vector in the source direction and $\rho_i$

means the ratio of the array size to the distance between the array and source, i.e., for far field sources, $\rho_i << 1$. Substituting (17) to (11), we have,

$$-(\boldsymbol{r}_i - \boldsymbol{r}_1)\cdot\frac{\bar{\boldsymbol{r}}_s}{v} + \frac{|\boldsymbol{r}_i - \boldsymbol{r}_1|}{v}\frac{\rho_i}{2} - \frac{1}{v}\frac{v^2(t_i - t_1)^2}{|\boldsymbol{r}_i - \boldsymbol{r}_1|}\frac{\rho_i}{2} = (t_i - t_1) \tag{18}$$

The term $v(t_i - t_1)$ means the distance difference between the sound source to the *i*-th and the first microphones. Let the distance difference be $d_i$, i.e.,

$$d_i = v(t_i - t_1) = |\boldsymbol{r}_s - \boldsymbol{r}_i| - |\boldsymbol{r}_s - \boldsymbol{r}_1| \tag{19}$$

Equation (18) can be re-written as,

$$-\frac{(\boldsymbol{r}_i - \boldsymbol{r}_1)}{v}\cdot\bar{\boldsymbol{r}}_s + f_i\frac{\rho_i}{2} = (t_i - t_1) \tag{20}$$

where

$$f_i = \frac{|\boldsymbol{r}_i - \boldsymbol{r}_1|}{v} - \frac{|d_i|}{v}\frac{|d_i|}{|\boldsymbol{r}_i - \boldsymbol{r}_1|} \tag{21}$$

It is straightforward to see that $f_i \geq 0$ since

$$d_i \leq |\boldsymbol{r}_i - \boldsymbol{r}_1| \tag{22}$$

Also, $f_i$ achieves it maximum of $|\boldsymbol{r}_i - \boldsymbol{r}_1|/v$ when $d_i = 0$ (i.e., when the source is located along the line passing through the mid point of and perpendicular to the segment connecting the *i*-th and the first microphone). This also means that $f_i$ has the order of magnitude less than or equal to the vector $(\boldsymbol{r}_i - \boldsymbol{r}_1)/v$. Therefore, from (20), it is clear that for far field sources ($\rho_i << 1$), the delay relation approaches,

$$-\frac{(\boldsymbol{r}_i - \boldsymbol{r}_1)}{v}\cdot\bar{\boldsymbol{r}}_s = (t_i - t_1) \tag{23}$$

Thus, the left hand side of (20) consists of the far field term and near field influence of the delay relation. We define $\rho_i$ as the field distance ratio and $f_i$ the near field influence factor for their roles in the source localization using microphone array. Equation (23) can also be derived from a plane wave assumption. Consider a single incident plane wave and a pair of microphones as shown in Fig. 1 and the relative time delay between two microphones can be described as:

$$\frac{|\boldsymbol{r}_i - \boldsymbol{r}_1|\cos(\theta_i)}{v} = t_1 - t_i \tag{24}$$

The parameters $\cos(\theta_i)$ can be represented as:

$$\cos(\theta_i) = \frac{(\boldsymbol{r}_i - \boldsymbol{r}_1)}{|\boldsymbol{r}_i - \boldsymbol{r}_1|}\cdot\frac{(\boldsymbol{r}_s - \boldsymbol{r}_1)}{|\boldsymbol{r}_s - \boldsymbol{r}_1|} \tag{25}$$

Equation (23) can be derived by substituting (25) into (24).

For far field sources ($\rho_i << 1$), the over determinant linear equation system (14) becomes (from (23)),

$$\mathbf{A}_s\boldsymbol{w}_s = \boldsymbol{b} \tag{26}$$

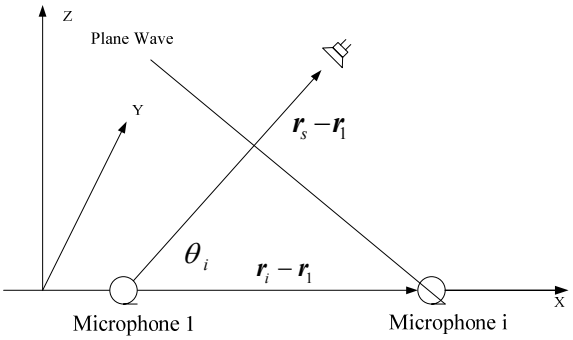


Fig. 1. Geometry model of plane wave and two microphones

where

$$\mathbf{A}_s = \begin{bmatrix} -(\boldsymbol{r}_2 - \boldsymbol{r}_1) \\ -(\boldsymbol{r}_3 - \boldsymbol{r}_1) \\ \vdots \\ -(\boldsymbol{r}_M - \boldsymbol{r}_1) \end{bmatrix}$$

Notice that it is unlikely to obtain the source location but the unit vector of the source direction ($\boldsymbol{w}_s$). And the speed of sound is obtained by:

$$\hat{v} = \frac{1}{|\hat{\boldsymbol{w}}_s|} = \frac{1}{\left|(\mathbf{A}_s^{\mathrm{T}}\mathbf{A}_s)^{-1}\mathbf{A}_s^{\mathrm{T}}\boldsymbol{b}\right|} \tag{27}$$

Then, the sound source direction for far field case can be given by:

$$\hat{\boldsymbol{r}}_s = \frac{\hat{\boldsymbol{w}}_s}{|\hat{\boldsymbol{w}}_s|} = \frac{(\mathbf{A}_s^{\mathrm{T}}\mathbf{A}_s)^{-1}\mathbf{A}_s^{\mathrm{T}}\boldsymbol{b}}{\left|(\mathbf{A}_s^{\mathrm{T}}\mathbf{A}_s)^{-1}\mathbf{A}_s^{\mathrm{T}}\boldsymbol{b}\right|} \tag{28}$$

### *C. Estimation Error Analysis*

Equation (26) is an approximation by considering plane wave only. It will give errors both in the source direction and the speed of sound. The error in the speed of sound is more interesting as it can reveal the relative distance information of sources to the microphone array. It can be shown that the closer the sound source, the larger the estimate of the speed. To see this, consider the original close form relation of equation (20) by moving the second term on the left hand side to the right.

$$-\frac{(\boldsymbol{r}_i - \boldsymbol{r}_1)}{v}\cdot\bar{\boldsymbol{r}}_s = (t_i - t_1) - f_i\frac{\rho_i}{2} \tag{29}$$

Without loss of generality, assume that $t_i > t_1$. Since both $\rho_i$ and $f_i$ are non-negative, equation (29) shows if the far field assumption is utilized (see (23)), the delay shall be decreased to match the real situation. However, when solving (23), there is no modification of the value $t_i - t_1$. Therefore, one possibility to match the case of augmented delay is to decrease the speed of sound. Another possibility is to change the direction of the source vector $\bar{\boldsymbol{r}}_s$. However, for an array spans the 3-D space, the possibility of adjusting the source direction for all sensor pairs is small since the least square method is applied. For example, changing the direction may

work for sensor pair (1, *i*) but has adverse effect on sensor pair (1, *j*) if $(\boldsymbol{r}_i - \boldsymbol{r}_1)$ and $(\boldsymbol{r}_j - \boldsymbol{r}_1)$ are perpendicular to each other.

## IV. Sound Source Number and Directions Estimation

Due to the singular problem of the matrix $\mathbf{A}_g$ described in section III-B, this paper assumes the distance form source to the array is much larger than the array aperture and equation (26) is used to solve the sound source direction estimation problem. If the sound source number is known, the sound source directions can be estimated by putting each sound source corresponding time delay vector $\boldsymbol{b}$ into equation (28) and how to obtain each sound source time delay vector $\boldsymbol{b}$ is the first problem. However, if the sound source number is unknown, the sound source directions estimation will become more complicated since there are several combinations to form the timed delay vector.

This section describes how to estimate sound sources number and directions simultaneously using the proposed method in section II and III-B. For multiple sound sources environment, the GCC function should have multiple peaks [11]. Because we assume there is no priori knowledge of the sound source number, the time delay sample for each microphone pair which meets the constraint below will be selected as the time delay sample candidates:

$$\begin{aligned} &R_{x_i x_1}(\hat{\tau}^{n_i}_{x_i x_1}) > \alpha \cdot R_{x_i x_1}(\hat{\tau}^{1}_{x_i x_1}) \\ &n_i = 2,3,\cdots,n_i^{\max} \quad i = 2,3,\cdots,M \end{aligned} \tag{30}$$

where $\alpha$ is a specified saclar and $\hat{\tau}^{n_i}_{x_i x_1}$ is the time delay sample corresponding to the $n_i$-th largest peak in ES-GCC function $R_{x_i x_1}$. If $R_{x_i x_1}$ has no time delay sample can meet the constraint above, then the $n_i^{\max}$ will be set to one. Hence, there are $n_2^{\max} \times n_3^{\max} \times \cdots \times n_M^{\max}$ possible combinations to form the time delay vector $\boldsymbol{b}_u$ and there should be $D$ correct combinations in those possible combinations. The next problem is how to choose the accurate combinations and determine the sound source number.

This paper proposes an acceptable sound velocity based method to arrange the combination and we assume the accurate combination should correspond to an acceptable sound velocity using equation (27). Therefore, each possible combination will be taken into equation (27) ($\mathbf{A}_s$ is known), and if the estimated sound velocity falls within an acceptable range, then this combination will be considered as the accurate combination; that is, we consider the combination as correct one when

$$\begin{aligned} &\left| \frac{1}{\left\| \left(\mathbf{A_s}^{\mathrm{T}}\mathbf{A_s}\right)^{-1}\mathbf{A_s}^{\mathrm{T}}\boldsymbol{b}_u \right\|} - \bar{v} \right| < \varepsilon \\ &u = 1,2,3,\cdots,n_2^{\max} \times n_3^{\max} \times \cdots \times n_M^{\max} \end{aligned} \tag{31}$$

where $\bar{v} = 34300$ is the sound velocity in *cm/sec* and $\varepsilon$ is a specified threshold. Assume there are $\tilde{D}$ combinations $(\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, \cdots, \tilde{\boldsymbol{b}}_{\tilde{D}})$ which correspond with the equation (31) and hence the $\tilde{D}$ sound sources direction can be obtained by:

$$\begin{aligned} &\tilde{\boldsymbol{r}}_{s_u} = \begin{bmatrix} x_{s_u} & y_{s_u} & z_{s_u} \end{bmatrix} = \frac{\left(\mathbf{A_s}^{\mathrm{T}}\mathbf{A_s}\right)^{-1}\mathbf{A_s}^{\mathrm{T}}\tilde{\boldsymbol{b}}_u}{\left\| \left(\mathbf{A_s}^{\mathrm{T}}\mathbf{A_s}\right)^{-1}\mathbf{A_s}^{\mathrm{T}}\tilde{\boldsymbol{b}}_u \right\|} \\ &\theta_u = \tan^{-1}\left( y_{s_u} \Big/ x_{s_u} \right) \qquad \phi_u = \tan^{-1}\left( z_{s_u} \Big/ \sqrt{x_{s_u}^2 + y_{s_u}^2} \right) \\ &u = 1,2,3,\cdots,\tilde{\mathrm{D}} \end{aligned} \tag{32}$$

where $\theta_u$ and $\phi_u$ are azimuth and elevation angle for the sound source respectively.

For the robustness consideration, the final sound source number and directions will be determined over Q-times results of equation (32). Let define all the accumulated estimation angle results over Q-times estimation as:

$$\begin{aligned} &\tilde{\boldsymbol{\theta}} = \begin{bmatrix} \tilde{\theta}_1 & \tilde{\theta}_2 & \cdots & \tilde{\theta}_R \end{bmatrix} \\ &\tilde{\boldsymbol{\varphi}} = \begin{bmatrix} \tilde{\phi}_1 & \tilde{\phi}_2 & \cdots & \tilde{\phi}_R \end{bmatrix} \\ &R = Q \times \left( \tilde{D}_1 + \tilde{D}_2 + \cdots + \tilde{D}_Q \right) \end{aligned} \tag{33}$$

where $\tilde{D}_q$ represents the combination number which meets the equation (31) constraint at the *q*-th testing.

So far, we have a set of *R* data points. Then, we cluster these accumulated results and this paper proposes an adaptive K-means ++ algorithm for clustering which is based on the K-means [12] and K-means++ [13] algorithms. We wish to choose $\hat{D}$ cluster center so as to minimize the potential function:

$$\begin{aligned} &\min \sum_{\hat{d}=1}^{\hat{D}} \sum_{\boldsymbol{\sigma}_r \in C_{\hat{d}}} (\boldsymbol{\sigma}_r - \boldsymbol{\mu}_{\hat{d}})^2 \\ &\boldsymbol{\sigma}_r = \begin{bmatrix} \tilde{\theta}_r & \tilde{\phi}_r \end{bmatrix} \\ &r = 1,2,3,\cdots,R \end{aligned} \tag{34}$$

where there are $\hat{D}$ cluster $C_{\hat{d}}$ and $\boldsymbol{\mu}_{\hat{d}}$ is the center of all the points $\boldsymbol{\sigma}_r \in C_{\hat{d}}$. Because the sound source number is unknown, we would set the cluster number $\hat{D}$ to be one and initial center $\boldsymbol{\mu}_1$ to be the median of $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\varphi}}$ as the initial condition to execute K-means. When the K-means algorithm converges, the constraint below is checked:

$$\begin{aligned} &E\left[ \left| \boldsymbol{\sigma}_r - \boldsymbol{\mu}_{\hat{d}} \right|^2 \right] < \delta \\ &\boldsymbol{\sigma}_r \in C_{\hat{d}} \ , \quad \hat{d} = 1,2,\cdots,\hat{D} \end{aligned} \tag{35}$$

where $E(\cdot)$ is the expectation operation and $\delta$ is a specified threshold. Equation (35) is used to check the variance of each cluster when the K-means algorithm converges. If one of the variance of each cluster is not less than a threshold, then we will make $\hat{D}$ plus one, find the other initial center using K-means++ [13] and execute K-means algorithm again. Otherwise, the final sound source number is $\hat{D}$ and the sound source directions are:

$$\left[\hat{\theta}_{\hat{d}} \quad \hat{\phi}_{\hat{d}}\right] = \boldsymbol{\mu}_{\hat{d}} \qquad (36)$$
$$\hat{d} = 1,2,\cdots,\hat{D}$$

The adaptive K-means ++ algorithm flowchart for estimating sound sources number and direction is shown in Fig. 2.

In summary, the steps of the proposed sound source number and directions estimation method are:

Step1. Calculate ES-GCC function $R_{x_i x_1}(\tau)$ and pick peaks satisfied with equation (30) from $R_{x_i x_1}(\tau)$ for each microphone pair.

Step2. Select $\tilde{D}$ time delay vector from equation (31) and estimate the corresponding sound source direction using equation (32).

Step3. Repeat Step 1 to 2 Q times and accumulate the results.

Step4. Cluster the accumulated results using adaptive K-means++ algorithm and the final cluster number and centers are sound source number and directions respectively.

## V. Experimental Results

This section provides the experimental results to evaluate the capability of the proposed algorithm. Two experiments, ES-GCC performance and sound source number and directions estimation, are tested in this section. The experiments are performed in a real room approximately of the size 10.5 m × 7.2 m and height of 3.6 m. The self-designed 8 channel digital microphone array platform is installed on the mobile robot for the experiment shown in Fig. 3 and the microphone position is marked with the circle symbol. The room temperature is approximately 20℃ and the sampling rate is 16k Hz. The experimental condition is shown in Fig. 4 and the distance from each sound source to the origin is 2.4 m. In Fig.4, the distance from the origin to Mic.1~4 and to Mic.5~8 is 0.22 m and 0.14m respectively. Seven microphone pairs listed in Table I are selected for the experiments. The air conditioner which is 4 m from the first microphone is turned on during this experiment (Noise 1 in Fig. 4). The parameters of α, ε and δ are set to be 0.9, 5000, and 16.

### *A. Eigenstructure-based GCC performance evaluation*

In this experiment, the sound source is female speech in English and the signal to noise ratio (SNR) is 14.58 dB. The proposed ES-GCC method is compared with the conventional GCC using PHAT weighting (PHAT-GCC) method [2]. In both methods, each frame is 512 samples long and the measurement times are 100 in each microphone pair. The correct time delay samples of this experiment for each microphone pair are (-4,-9,-14,-20,-16,-11,-5). Fig. 5 and Fig. 6 show the estimated time delay samples of PHAT-GCC and ES-GCC respectively. As can be seen, the estimated time delay samples of the ES-FCC method almost maintain a constant value along the measure times for each condition. However, the estimated time delay samples of the PHAT-GCC method fluctuate along the measurement times. This is because the principal component vector in ES-GCC method can characterize the time delay relation of the sound source better than PHAT-GCC method.

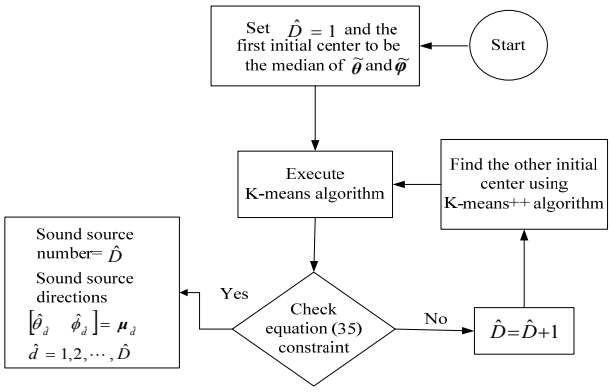


Fig. 2. The flowchart of adaptive K-means++ algorithm

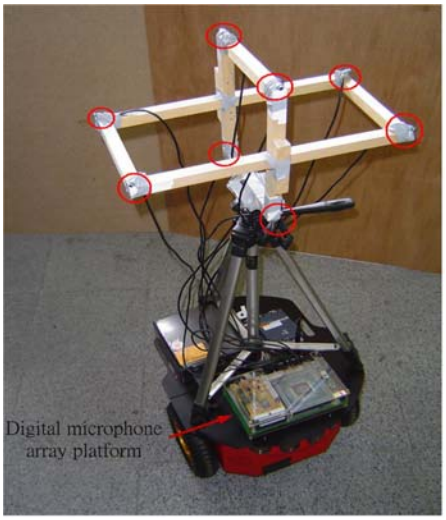


Fig. 3. Digital microphone array mounted on the robot

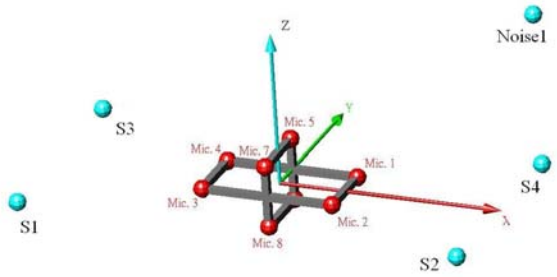


Fig. 4. Arrangement of microphone array and sound sources

Table I
THE CONDITION NUMBERS
CORRESPOND TO THE MICROPHONE PAIRS

| Condition | Microphone pairs | Condition | Microphone pairs |
|---|---|---|---|
| C1 | (1,2) | C5 | (1,6) |
| C2 | (1,3) | C6 | (1,7) |
| C3 | (1,4) | C7 | (1,7) |
| C4 | (1,5) | | |

### *B. Single source and dual sources direction estimation*

Three cases, single, dual, and four sound sources are tested in this experiment to evaluate the performance of the proposed sound source number and directions estimation system. The sound sources are all speech in English.

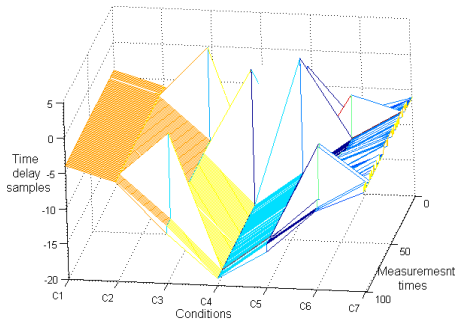


Fig. 5. Estimated time delay samples (PHAT-GCC)

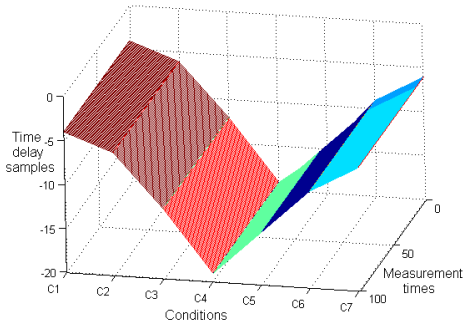


Fig. 6. Estimated time delay samples (ES-GCC)

TABLE II
EXPERIMENTAL RESULTS OF SINGLE SOURCE ESTIMATION WITH NOISE 1

| Experimental Conditions | | | Experimental Results (Mean) | |
|---|---|---|---|---|
| Source number | SNR (dB) | CorrectAngle ($\theta$ $\phi$) | Estimated Angle | Estimated number |
| 1 | 14.78 | (-153°, -6°) | (-152.74°, -6.15°) | 1 |
| 1 | 15.51 | (-27°, -3°) | (-28.61°, -3.05°) | 1 |
| 1 | 16.35 | (207°, 0°) | (207.89°, 0°) | 1 |
| 1 | 16.21 | (27°, 6°) | (25.2°, 6.07°) | 1 |

TABLE III
EXPERIMENTAL RESULTS OF DUAL SOURCES ESTIMATION WITH NOISE 1

| Experimental Conditions | | | Experimental Results (Mean) | |
|---|---|---|---|---|
| Source number | SNR (dB) | CorrectAngle ($\theta$ $\phi$) | Estimated Angle | Estimated number |
| 2 | 13.38 | (-153°, -6°)<br>(-27°, -3°) | (-152.74°, -6.15°)<br>(-28.61°, -3.05°) | 2 |
| 2 | 13.53 | (-27°, -3°)<br>(207°, 0°) | (-28.61°, -3.05°)<br>(207.25°, 0°) | 2 |
| 2 | 14.24 | (207°, 0°)<br>(27°, 6°) | (206.56°, -2.01°)<br>(26.57°, 6.07°) | 2 |

TABLE IV
EXPERIMENTAL RESULTS OF FOUR SOURCES ESTIMATION WITH NOISE 1

| Experimental Conditions | | | Experimental Results (Mean) | |
|---|---|---|---|---|
| Source number | SNR (dB) | Correct Angle ($\theta$ $\phi$) | Estimated Angle | Estimated number |
| 4 | 15.38 | (-0.25°,1.2 °)<br>(88.45°, 1°)<br>(-92.6°, 0.83°)<br>(178.65°, 1.7°) | (1.35°, 0.32°)<br>(90.13°,0.45°)<br>(-90.36°, 0.36°)<br>(179.65°,0.12 °) | 4 |

The experimental results of single, dual and four sources cases are shown in Table II, III and IV. The trial number in each condition is 300. Two-dimensional angles, 360° azimuth θ and 180° elevation $\phi$ are considered to be estimated in this experiment. As shown in Table II, III and IV, the statistical estimation results of the mean error of the proposed sound source direction estimation system is less than 3° which is accurate enough for other robot system application (like human face tracking system). Especially, in multiple sound sources case, we find that the wrong combination of time delay vector $\boldsymbol{b}_u$ will cause the estimated sound speed to range between 9000 and 15000 or more than 50000. Hence, the estimated sound velocity is reliable to arrange the combinations of time delay vector.

## VI. CONCLUSION

This work provides the sound source number and directions estimation algorithm. The multiple source time delay vector combination problem can be solved by the proposed acceptable sound velocity based method. By accumulating the estimated sound source angle, the sound source number and directions can be obtained by the proposed adaptive K-means++ algorithm. The proposed algorithm is evaluated in a real environment.

## REFERENCES


[1] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform (SCOT)," Naval Underwater Systems Center, New London Lab., New London, CT, Tech. Memo TC-159-72, Aug. 1972.

[2] C. H. Knapp, and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustic, Speech, Signal Processing*, vol. 24, pp. 320-327, Aug 1976.

[3] M. S. Brandstein, and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," *IEEE conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 375-378, April 1997.

[4] Q. H. Wang, T. Ivanov, and P. Aarabi, "Acoustic robot navigation using distributed microphone arrays," *Information Fusion*, vol. 5, pp.131-140, June 2004

[5] R. V. Balan, and J. Rosca, "Apparatus and method for estimating the direction of Arrival of a source signal using a microphone array," European Patent, No US2004013275, 2004.

[6] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation* vol. 34, pp. 276-280. Mar 1986.

[7] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoto, " Robust Speech Interface Based on Audio and Video Information Fusion for Humanoid HRP-2," *Proc. IEEE/RSJ Conf. on Intelligent Robots and Systems*, pp.2402-2410, 2004.

[8] M. Walworth, and A. Mahajan, "3D position sensing using the difference in the time-of-flights from a wave source to various receivers," *Proc. IEEE Conf. on Advanced Robotics*, pp.611-616, 1997.

[9] J. M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," *Proc. IEEE/RSJ Conf. on Intelligent Robots and Systems*, pp. 1228-1233, 2003.

[10] K. Yao, R. E. Hudson, C. W. Reed, D. Chen, and F. Lorenzelli, "Blind beamforming on a randomly distributed sensor array system," *IEEE J. Select. Areas Commun.*, vol. 16, pp.1555-1567, Oct. 1998.

[11] D. Bechler, and K. Kroschel, "Considering the second peak in the GCC function for multi-Source TDOA estimation with a microphone array," *International Workshop on Acoustic Echo and Noise Control,* pp.315-318, Sept. 2003.

[12] J. A. Hartigan, and M. A. Wong, "A k-means clustering algorithm," *In Applied Statistics*, pp.100-108, 1979.

[13] D. Arthur, and S. Vassilvitskii, "K-means++ the Advantages of careful seeding," *2007 Symposium on Discrete Algorithms* (SODA).