Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 870756, 14 pages
doi:10.1155/2010/870756

*Research Article*

# Estimation of Sound Source Number and Directions under a Multisource Reverberant Environment

**Jwu-Sheng Hu and Chia-Hsin Yang**

*Department of Electrical and Control Engineering, National Chiao-Tung University, Lab 905, Engineering Building No. 5, 1001 Ta Hsueh Road, Hsinchu 300, Taiwan*

Correspondence should be addressed to Chia-Hsin Yang, chyang.ece92g@nctu.edu.tw

Received 3 December 2009; Revised 4 April 2010; Accepted 27 May 2010

Academic Editor: Sven Nordholm



Sound source localization is an important feature in robot audition. This work proposes a sound source number and directions estimation method under a multisource reverberant environment. An eigenstructure-based generalized cross-correlation method is proposed to estimate time delay among microphones. A source is considered as a candidate if the corresponding time delay combination among microphones gives reasonable sound speed estimation. Under reverberation, some candidates might be spurious but their direction estimations are not consistent for consecutive data frames. Therefore, an adaptive K-means++ algorithm is proposed to cluster the accumulated results from the sound speed selection mechanism. Experimental results demonstrate the performance of the proposed algorithm in a real room.

## 1. Introduction

Sound source localization is one of the fundamental features of robot audition for human-robot interaction as well as recognition of the environment. 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) [1–3] was used for robotic applications [4] but it is not robust under multiple sources environment. Improvements on the performance in the multiple sources and reverberant environment have also been discussed [5, 6]. Another approach, proposed by Balan and Rosca [7], 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 [8, 9] combined with spatial smoothing [10] is one of the most popular methods for eliminating the coherence problem and it is also applied to the robot audition [11].

Based on the geometrical relationship among time delay values, Walworth and Mahajan [12] proposed a linear equation formulation for the estimation of the three-dimensional (3D) position of a wave source. Later, Valin et al. [13] gave a simple solution for the linear equation in [12] based on the far-field assumption and developed a novel weighting function method to estimate the time delay. In a real environment, the sound source may move. Valin et al. [14] proposed a localization and tracking of simultaneous moving sound sources method using eight microphones and this method is based on a frequency domain implementation of a steered beamformer along with a particle filter-based tracking algorithm. In addition, Badali et al. [15] investigated the accuracy of different time delay of arrival estimation audio localization implementations in the context of artificial audition for robotic systems.

Yao et al. [16] 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 [12] to estimate the source location and velocity via least square method. Statistical methods [17–19] have also been proposed to solve the DOA problem under complex environment. These methods yield superior performance than conventional DOA method especially when the sound source is not within line-of-sight. However, a training procedure is needed for these methods to obtain the pattern of sound wave arrival. This may not be realistic for the robot applications when the environment is unknown.

The methods above assume that the sound source number is known. But this may not be a realistic assumption because the environment usually contains various kinds of sound sources. Several eigenvalue-based methods have been proposed [20, 21] to estimate the sound source number. However, the eigenvalue distribution is sensitive to noise and reverberation. The work in [22] used the support vector machine (SVM) to classify the distribution with respect to the sound source number. However, it still requires a training stage for a robust result and the binary classification is inadequate when the sound source number is larger than two.

The objective of this work is to estimate the multiple fixed sound source directions without a priori information of the sound source number and the environment. This work utilizes the time delay information and microphone array geometry to estimate the sound source directions [23]. A novel eigenstructure-based GCC (ES-GCC) method to estimate the time delay under a multi-source environment between two microphones is proposed. The theoretical proof of the ES-GCC method is given, and the experimental results show that it is robust in a noisy environment. As a result, the sound source direction and velocity can be obtained by solving the proposed linear equation model using the time delay information. Fundamentally, the sound source number should be known while estimating the sound source directions. Hence, the method which can estimate sound source number and directions simultaneously using the proposed adaptive K-means++ is introduced and all the experiments are conducted in a real environment. This paper is organized as follows. In Section 2, we introduce the novel ES-GCC method for time delay estimation. With the time delay estimation, the sound source direction and speed estimation method is presented in Section 3, where the estimation error is also analyzed. In Section 4, we propose the sound speed selection mechanism and adaptive K-means++ algorithm. Experimental results, presented in Section 5, demonstrate the performance of the proposed algorithm in a real environment. Section 6 concludes the paper.

## 2. Time Delay Estimation

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

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

where $a_{md}(t)$ is the transfer function from the $d$th sound source to the $m$th microphone assumed to be time-invariant over the observation period and $\otimes$ represents the convolution operation. $s_d(t)$ and $n_m(t)$ are the $d$th sound source and the nondirectional noise, respectively. It is assumed that $s_d(t)$ and $n_m(t)$ are mutually uncorrelated and sound source signals are mutually independent. Applying the short-time Fourier transform (STFT) to (1), we have

$$X_m(\omega, k) = \sum_{d=1}^{D} A_{md}(\omega) S_d(\omega, k) + N_m(\omega, k), \quad \omega = 0, 1, \ldots, N_{\text{STFT}} - 1, \tag{2}$$

where $\omega$ is the frequency band, $k$ is the frame number, and $N_{\text{STFT}}$ is the STFT point. $A_{md}(\omega)$, $X_m(\omega, k)$, $S_d(\omega, k)$, and $N_m(\omega, k)$ are the STFT of the respective signals. Rewrite (2) in matrix form:

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

where

$$\begin{aligned} \mathbf{X}(\omega, k) &= \left[X_1(\omega, k), \ \ldots, \ X_M(\omega, k)\right]^{\mathrm{T}} \in C^{M\times 1}, \\ \mathbf{N}(\omega, k) &= \left[N_1(\omega, k), \ \ldots, \ N_M(\omega, k)\right]^{\mathrm{T}} \in C^{M\times 1}, \\ \mathbf{S}(\omega, k) &= \left[S_1(\omega, k), \ \ldots, \ S_D(\omega, k)\right]^{\mathrm{T}} \in C^{D\times 1}, \\ \mathbf{A}(\omega) &= \begin{bmatrix} A_{11}(\omega) & \cdots & A_{1D}(\omega) \\ \vdots & & \vdots \\ A_{M1}(\omega) & \cdots & A_{MD}(\omega) \end{bmatrix} \in C^{M\times D}. \end{aligned} \tag{4}$$

Suppose the noises are spatially white, and the noise correlation matrix is diagonal matrix $\sigma_n^2\mathbf{I}$. Therefore, the received signal correlation matrix using $K$ frames with eigenvalue decomposition (EVD) can be described as

$$\begin{aligned} \mathbf{R}_{xx}(\omega) &= \frac{1}{\mathrm{K}}\sum_{k=1}^{K}\mathbf{X}(\omega, k)\mathbf{X}^{\mathrm{H}}(\omega, k) = \mathbf{A}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{A}^{\mathrm{H}}(\omega) + \sigma_n^2\mathbf{I} \\ &= \sum_{i=1}^{M}\lambda_i(\omega)\mathbf{V}_i(\omega)\mathbf{V}_i^{\mathrm{H}}(\omega), \end{aligned} \tag{5}$$

where H denotes conjugation transpose; $\mathbf{R}_{ss}(\omega) = (1/\mathrm{K}) \sum_{k=1}^{K} \mathbf{S}(\omega, k)\mathbf{S}^{\mathrm{H}}(\omega, k)$; $\lambda_i(\omega)$ and $\mathbf{V}_i(\omega)$ are eigenvalues and corresponding eigenvectors with $\lambda_1(\omega) \geq \lambda_2(\omega) \geq \cdots \geq \lambda_M(\omega)$. The signal-only correlation matrix $\mathbf{A}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{A}^{\mathrm{H}}(\omega)$ can be expressed as (6) using the property $\sigma_n^2\mathbf{I} = \sum_{m=1}^{M}\sigma_n^2\mathbf{V}_m(\omega)\mathbf{V}_m^{\mathrm{H}}(\omega)$ (the proof of this property is given in the appendix):

$$\mathbf{A}_s(\omega)\mathbf{R}_{ss}(\omega)\mathbf{A}_s^{\mathrm{H}}(\omega) = \sum_{m=1}^{M}\left(\lambda_m(\omega) - \sigma_n^2\right)\mathbf{V}_m(\omega)\mathbf{V}_m^{\mathrm{H}}(\omega). \tag{6}$$

The eigenvalues and eigenvectors are divided into two groups. The first group, consisting of $D$ eigenvectors ($\mathbf{V}_1(\omega)$

to $\mathbf{V}_D(\omega)$) is referred to as signal eigenvectors and spans the signal subspace. The second group, consisting of $M$-$D$ eigenvectors ($\mathbf{V}_{D+1}(\omega)$ to $\mathbf{V}_M(\omega)$) is referred to as noise eigenvectors and spans the noise subspace. The MUSIC algorithm [8, 9] uses the orthogonal property of the signal and noise subspaces to estimate the signal directions and it mainly uses the eigenvectors that lie in the noise subspace. Rather than using the noise subspace information, this paper considers the eigenvectors that lie in the signal subspace for time delay estimation (TDE) to minimize the influence of noise. The idea that employs the eigenvectors in the signal subspace can also be referred as the Blackman-Tukey frequency estimation method [24]. In the signal eigenvectors, $\mathbf{V}_1(\omega)$ is the eigenvector associated with the maximum eigenvalue:

$$\mathbf{V}_1(\omega) = \begin{bmatrix} V_{11}(\omega) & V_{21}(\omega) & \cdots & V_{M1}(\omega) \end{bmatrix}^{\mathrm{T}} \in C^{M\times 1}. \quad (7)$$

This paper chooses the eigenvector $\mathbf{V}_1(\omega)$ for TDE because it lies in the signal subspace and it contributes most to construct the signal-only correlation matrix. We call the eigenvector $\mathbf{V}_1(\omega)$ first principal component vector since it contains the information of the speech sound sources and is robust to the noise. It is different from the conventional GCC methods where a number of weighting functions are adjusted for different applications. In essence, this paper replaces the microphone-received signal $\mathbf{X}(\omega, k)$ with $\mathbf{V}_1(\omega)$ for TDE since $\mathbf{V}_1(\omega)$ can be considered as the approximation of $\mathbf{A}(\omega)\mathbf{S}(\omega, k)$. A detailed explanation is given in the appendix. Hence, the ES-GCC function between the $i$th and $j$th microphone can be represented as

$$R_{x_i x_j}(\tau) = \sum_{\omega=0}^{N_{\mathrm{STFT}}-1} \frac{1}{\left| V_{i1}(\omega) V_{j1}(\omega) \right|} V_{i1}(\omega) V_{j1}(\omega) e^{j\omega\tau}. \quad (8)$$

The weighting function in (8) follows the idea of GCC-PHAT [2] and the reason is that studies [3, 25] showed it is more immune to reverberation time than other cross-correlation-based methods but sensitive to noise. By replacing the original signals with the principal component vectors, the robustness to noise can be enhanced. As a result, the time delay sample can be estimated by finding the maximum peak of the ES-GCC function as

$$\hat{\tau}^1_{x_i x_j} = \arg\max_{\tau} R_{x_i x_j}(\tau). \quad (9)$$

## 3. Sound Source Localization and Speed Estimation

*3.1. Sound Source Location Estimation Using Least-Square Method.* The sound source location can be estimated from geometrical calculation of the time delays among the microphone array elements. The work in [16] provides a linear equation model for estimating the source localization and propagation speed. The following derivations explain the idea. Consider sound source location vector $\mathbf{r}_s = [x_s \quad y_s \quad z_s]$, the $i$th microphone location $\mathbf{r}_i = [x_i \quad y_i \quad z_i]$, and the relative time delays, $t_i - t_1$, between the $i$th microphone and the first microphone. The relative time delay satisfies

$$t_i - t_1 = \frac{|\mathbf{r}_i - \mathbf{r}_s| - |\mathbf{r}_1 - \mathbf{r}_s|}{v}, \quad (10)$$

where $t_i$ is the time delay from the sound source to the $i$th microphone and $v$ is the speed of sound. Equation (10) is equivalent to

$$t_i - t_1 + \frac{|\mathbf{r}_s - \mathbf{r}_1|}{v} = \frac{|(\mathbf{r}_i - \mathbf{r}_1) - (\mathbf{r}_s - \mathbf{r}_1)|}{v}. \quad (11)$$

Squaring both sides, we have

$$\begin{aligned}(t_i - t_1)^2 + 2(t_i - t_1)\frac{|\mathbf{r}_s - \mathbf{r}_1|}{v} &= \left(\frac{|\mathbf{r}_i - \mathbf{r}_1|}{v}\right)^2 \\ &- \frac{2(\mathbf{r}_i - \mathbf{r}_1)\cdot(\mathbf{r}_s - \mathbf{r}_1)}{v^2}.\end{aligned} \quad (12)$$

By some algebraic manipulations, (12) becomes

$$-\frac{(\mathbf{r}_i - \mathbf{r}_1)\cdot(\mathbf{r}_s - \mathbf{r}_1)}{v|\mathbf{r}_s - \mathbf{r}_1|} + \frac{|\mathbf{r}_i - \mathbf{r}_1|^2}{2v|\mathbf{r}_s - \mathbf{r}_1|} - \frac{v(t_i - t_1)^2}{2|\mathbf{r}_s - \mathbf{r}_1|} = (t_i - t_1). \quad (13)$$

Next, define the normalized sound source position vector as,

$$\mathbf{w}_{\mathbf{s}} \equiv [w_1 \quad w_2 \quad w_3]^{\mathrm{T}} = \frac{\mathbf{r}_{\mathbf{s}} - \mathbf{r}_1}{v|\mathbf{r}_{\mathbf{s}} - \mathbf{r}_1|}. \quad (14)$$

And define two other variables as

$$w_4 = \frac{1}{2v|\mathbf{r}_s - \mathbf{r}_1|}, \qquad w_5 = \frac{v}{2|\mathbf{r}_s - \mathbf{r}_1|}. \quad (15)$$

The linear equation (13) considering all $M$ microphones can be written as

$$\mathbf{A}_g \mathbf{w} = \mathbf{b}, \quad (16)$$

where $\mathbf{w} = [\mathbf{w}_{\mathbf{s}}^{\mathrm{T}} \quad w_4 \quad w_5]^{\mathrm{T}} = [w_1 w_2 w_3 w_4 w_5]^{\mathrm{T}}$,

$$\mathbf{A}_g = \begin{bmatrix} -(\mathbf{r}_2 - \mathbf{r}_1) & |\mathbf{r}_2 - \mathbf{r}_1|^2 & -(t_2 - t_1)^2 \\ -(\mathbf{r}_3 - \mathbf{r}_1) & |\mathbf{r}_3 - \mathbf{r}_1|^2 & -(t_3 - t_1)^2 \\ \vdots & \vdots & \vdots \\ -(\mathbf{r}_M - \mathbf{r}_1) & |\mathbf{r}_M - \mathbf{r}_1|^2 & -(t_M - t_1)^2 \end{bmatrix},$$
$$\mathbf{b} = \begin{bmatrix} t_2 - t_1 \\ t_3 - t_1 \\ \vdots \\ t_M - t_1 \end{bmatrix}. \quad (17)$$

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

$$\begin{aligned}\hat{\mathbf{w}} &= \begin{bmatrix} \hat{\mathbf{w}}_{\mathbf{s}}^{\mathrm{T}} & \hat{w}_4 & \hat{w}_5 \end{bmatrix}^{\mathrm{T}} \\ &= \begin{bmatrix} \hat{w}_1 & \hat{w}_2 & \hat{w}_3 & \hat{w}_4 & \hat{w}_5 \end{bmatrix}^{\mathrm{T}} \\ &= \left(\mathbf{A}_g^{\mathrm{T}} \mathbf{A}_g\right)^{-1} \mathbf{A}_g^{\mathrm{T}} \mathbf{b}.\end{aligned} \quad (18)$$

The estimated sound source location and speed of sound can be obtained as

$$\tilde{\mathbf{r}}_s = \frac{\hat{w}_s}{2\hat{w}_4} + \mathbf{r}_1, \qquad \tilde{v} = \sqrt{\frac{\hat{w}_5}{\hat{w}_4}} \quad \left(\text{or } \tilde{v} = \frac{1}{|\hat{\mathbf{w}}_s|}\right). \tag{19}$$

*3.2. Sound Source Direction Estimation Using Least-Square Method for Far-Field Case.* To solve (16), the matrix $\mathbf{A}_g$ must be full rank. However, for matrix $\mathbf{A}_g$, the condition on rank is more complicated and can be ill-conditioned easily. For example, if the microphones are distributed on a spherical surface (i.e., $\mathbf{r}_i = [R_m \cos\theta_i \sin\phi_i \;\; R_m \sin\theta_i \sin\phi_i \;\; R_m \cos\phi_i]$, $R_m$ is radius, and $\theta_i$ and $\phi_i$ are azimuth and elevation angle resp.), it can be verified that the fourth column in $\mathbf{A}_g$ is the linear combination of column 1, 2, and 3. Secondly, if the aperture of the array is small compared with the source distance (far-field), the distance estimation is also sensitive to noise. In the following, a detailed analysis of (13) is presented which leads to a formulation for the far-field case. Define $\bar{\mathbf{r}}_s$ and $\rho_i$ as,

$$\bar{\mathbf{r}}_s = \frac{\mathbf{r}_s - \mathbf{r}_1}{|\mathbf{r}_s - \mathbf{r}_1|}, \qquad \rho_i = \frac{|\mathbf{r}_i - \mathbf{r}_1|}{|\mathbf{r}_s - \mathbf{r}_1|}. \tag{20}$$

$\bar{\mathbf{r}}_s$ 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, that is, for far-field sources, $\rho_i \ll 1$. Substituting (20) to (13), we have,

$$-(\mathbf{r}_i - \mathbf{r}_1) \cdot \frac{\bar{\mathbf{r}}_s}{v} + \frac{|\mathbf{r}_i - \mathbf{r}_1|}{v}\frac{\rho_i}{2} - \frac{1}{v}\frac{v^2(t_i - t_1)^2}{|\mathbf{r}_i - \mathbf{r}_1|}\frac{\rho_i}{2} = (t_i - t_1). \tag{21}$$

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 $d_i$, that is,

$$d_i = v(t_i - t_1) = |\mathbf{r}_s - \mathbf{r}_i| - |\mathbf{r}_s - \mathbf{r}_1|. \tag{22}$$

Equation (21) can be rewritten as

$$-\frac{(\mathbf{r}_i - \mathbf{r}_1)}{v} \cdot \bar{\mathbf{r}}_s + f_i\frac{\rho_i}{2} = (t_i - t_1), \tag{23}$$

where

$$f_i = \frac{|\mathbf{r}_i - \mathbf{r}_1|}{v} - \frac{|d_i|}{v}\frac{|d_i|}{|\mathbf{r}_i - \mathbf{r}_1|}. \tag{24}$$

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

$$d_i \leq |\mathbf{r}_i - \mathbf{r}_1|. \tag{25}$$

Also, $f_i$ achieves its maximum value of $|\mathbf{r}_i - \mathbf{r}_1|/v$ when $d_i = 0$ (i.e., when the source is located along the line passing through the midpoint 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 magnitude of vector$(\mathbf{r}_i - \mathbf{r}_1)/v$.

From (23), it is clear that for far-field sources ($\rho_i \ll 1$), the delay relation approaches

$$-(\mathbf{r}_i - \mathbf{r}_1) \cdot \mathbf{w}_s = (t_i - t_1). \tag{26}$$

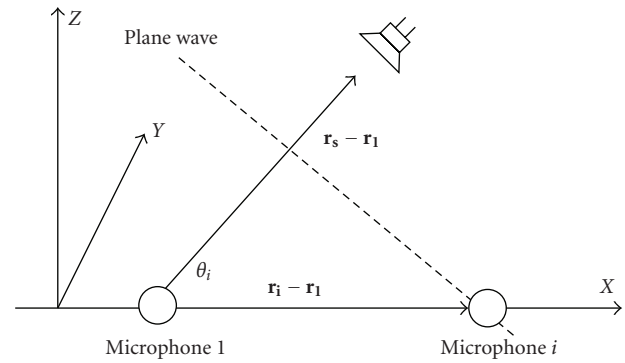


Figure 1: Geometry model of plane wave and two microphones.

Thus, the left hand side of (23) consists of the far-field term and near field influence of the delay relation. We define $\rho_i$ as *the field distance ratio* and $f_i$ as *the near field influence factor* for their roles in the sound source localization using microphone array. Equation (26) can also be derived from a plane wave assumption. Consider a single incident plane wave and a pair of microphones as shown in Figure 1 and the relative time delay between two microphones can be described as:

$$\frac{|\mathbf{r}_i - \mathbf{r}_1|\cos(\theta_i)}{v} = t_1 - t_i. \tag{27}$$

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

$$\cos(\theta_i) = \frac{(\mathbf{r}_i - \mathbf{r}_1)}{|\mathbf{r}_i - \mathbf{r}_1|} \cdot \frac{(\mathbf{r}_s - \mathbf{r}_1)}{|\mathbf{r}_s - \mathbf{r}_1|}. \tag{28}$$

Equation (26) can be derived by substituting (28) into (27).

For far-field sources ($\rho_i \ll 1$), the overdetermined linear equation system (16) becomes (from (26))

$$\mathbf{A}_f \mathbf{w}_s = \mathbf{b}, \tag{29}$$

where

$$\mathbf{A}_f = \begin{bmatrix} -(\mathbf{r}_2 - \mathbf{r}_1) \\ -(\mathbf{r}_3 - \mathbf{r}_1) \\ \vdots \\ -(\mathbf{r}_M - \mathbf{r}_1) \end{bmatrix}. \tag{30}$$

The unit vector of the source direction ($\mathbf{w}_s$) can be estimated using the least square method similar with (18). And the speed of sound is obtained by

$$\hat{v} = \frac{1}{|\hat{\mathbf{w}}_s|} = \frac{1}{\left|\left(\mathbf{A}_f^{\mathrm{T}}\mathbf{A}_f\right)^{-1}\mathbf{A}_f^{\mathrm{T}}\mathbf{b}\right|}. \tag{31}$$

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

$$\hat{\mathbf{r}}_s = \frac{\hat{\mathbf{w}}_s}{|\hat{\mathbf{w}}_s|} = \frac{\left(\mathbf{A}_f^{\mathrm{T}}\mathbf{A}_f\right)^{-1}\mathbf{A}_f^{\mathrm{T}}\mathbf{b}}{\left|\left(\mathbf{A}_f^{\mathrm{T}}\mathbf{A}_f\right)^{-1}\mathbf{A}_f^{\mathrm{T}}\mathbf{b}\right|}. \tag{32}$$

*3.3. Estimation Error Analysis.* Equation (29) 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 (23) by moving the second term on the left-hand side to the right:

$$-\frac{(\mathbf{r}_i - \mathbf{r}_1)}{v} \cdot \bar{\mathbf{r}}_s = (t_i - t_1) - f_i \frac{\rho_i}{2}. \tag{33}$$

Without loss of generality, assume that $t_i > t_1$. Since both $\rho_i$ and $f_i$ are nonnegative, (33) shows that if the far-field assumption is utilized (see (26)), the delay shall be decreased to match the real situation. However, when solving (26), 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 $\bar{\mathbf{r}}_s$. However, for an array spans the 3D 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 $(\mathbf{r}_i - \mathbf{r}_1)$ and $(\mathbf{r}_j - \mathbf{r}_1)$ are perpendicular to each other. A simple simulation for estimation error is illustrated for the microphone locations depicted in Figure 7. We assume that there is no time delay estimation error and the sound velocity is 34300 cm/sec. The sound source location is moved on the direction vector (0.3256, 0.9455, 0) to make sure that $t_i > t_1$. The estimated sound source direction and velocity are obtained by using (31) and (32). Figure 2 shows the relation between direction estimation error and the factor $1/\rho_2$. The direction estimation error is defined as the difference between real angel and estimated angle. As it can be seen, the estimation error becomes smaller and converges to a small value when $1/\rho_2$ is increased. In particular, the estimation error would not change dramatically when $1/\rho_2$ is larger than 5 ($|\mathbf{r}_s - \mathbf{r}_1|$ is larger than five times of $|\mathbf{r}_2 - \mathbf{r}_1|$). Figure 3 shows the relation between estimated velocity and $1/\rho_2$. The estimated velocity converges to 34300 when $1/\rho_2$ is increased and this is consistent with the analysis at the beginning of this section.

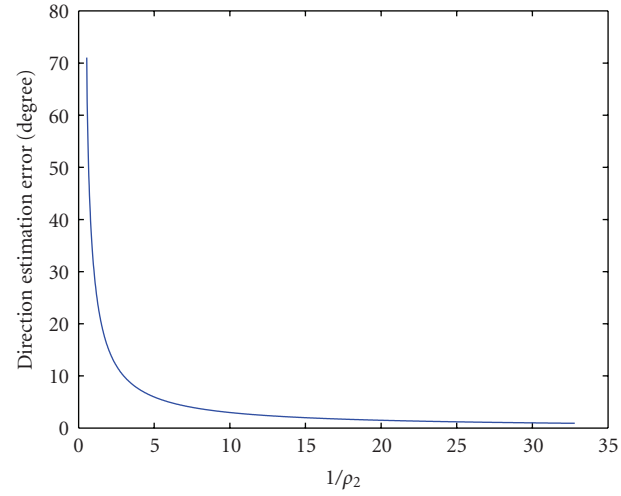


Figure 2: Direction estimation error versus $1/\rho_2$.

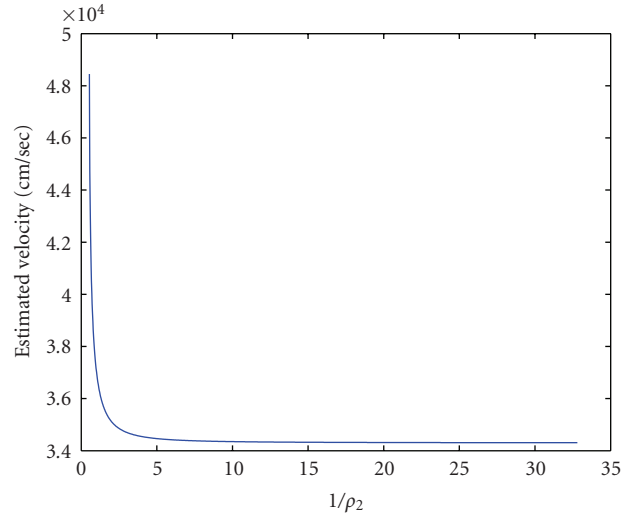


Figure 3: Estimated velocity versus $1/\rho_2$.

## 4. Sound Source Number and Directions Estimation

This paper assumes that the distance from source to the array is much larger than the array aperture, and (29) is used to solve the sound source direction estimation problem. If the number of sound sources is known, the sound source directions can be estimated by putting time delay vector $\mathbf{b}$ of corresponding sound source into (32). 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 vectors. This section describes how to estimate the sound sources number and directions simultaneously using the proposed method in Sections 2 and 3.2. A two-step algorithm is proposed to estimate the source number. First, the combinations of delays are filtered by the estimated sound velocity which does not fall within a reasonable range of the true one. But in a reverberant environment, it is still possible to have a phantom source that results in reasonable sound speed estimation. This paper assumes that the power level of phantom source is much weaker than that of the true source. Therefore, only a true source can exhibit a consistent estimation of direction on consecutive frames of signals because the weighting function of ES-GCC also has certain robustness to reverberation. The second step of source number estimation is to cluster the accumulated results from the first step using clustering technique and the reverberation can be considered as the outlier for the clustering technique. The well-known clustering method, K-means, is sensitive to initial conditions and is not robust to outliers. In addition, the cluster number should be known in

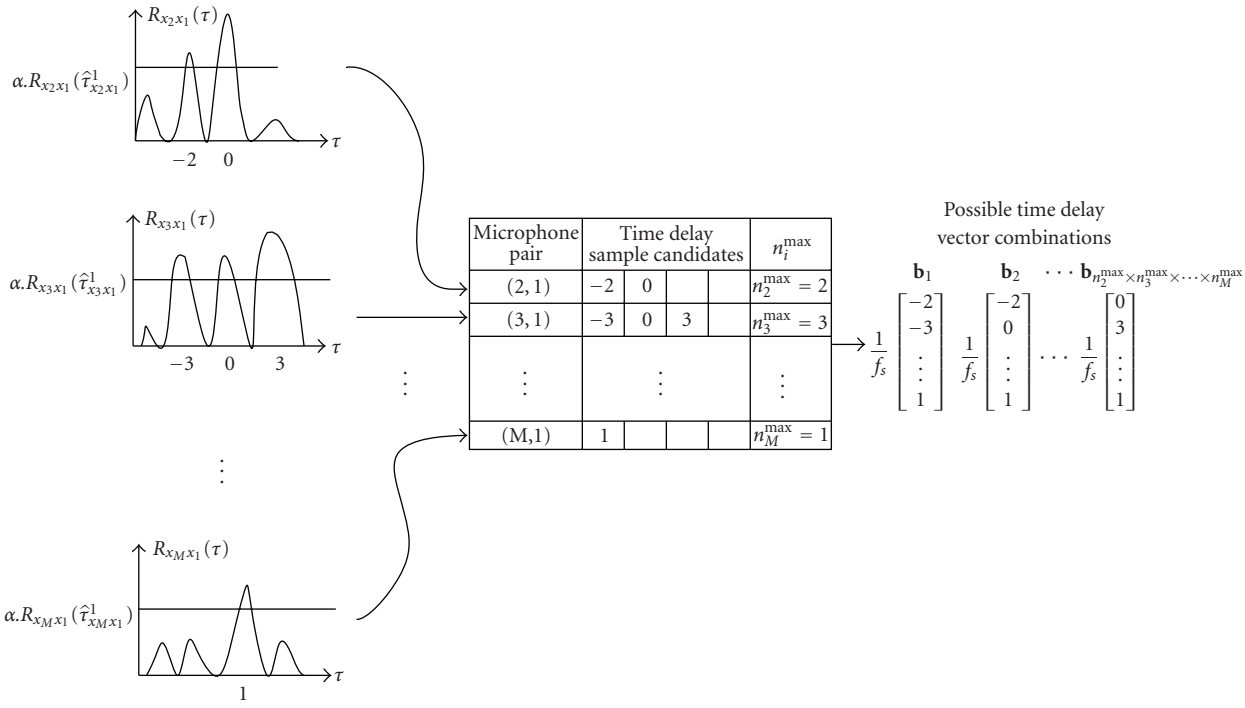


Figure 4: Illustration of the procedure of forming possible time delay vector combinations.

advance for K-means which cannot be met in our scenario since we have no information of the sound source number. To improve the problems of robustness and cluster number, this paper proposes the adaptive K-means++ method based on the K-means [26] and K-means++ [27] methods for clustering. The K-means++ method is a way of initializing K-means by choosing random starting centers with very specific probabilities. It then runs the normal K-means algorithm afterwards. Because the seeding technique of K-means++ method can improve both the speed and accuracy of the K-means method [27], this paper employs the seeding technique of K-means++ method to seed the initial centers for the proposed adaptive K-means++ method.

*4.1. Rejecting Incorrect Time Delay Combinations Using Acceptable Velocity Range.* For multiple sound sources environment, the GCC function should have multiple peaks [28]. Without a 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:

$$R_{x_ix_1}\left(\hat{\tau}^{n_i}_{x_ix_1}\right) > \alpha \cdot R_{x_ix_1}\left(\hat{\tau}^{1}_{x_ix_1}\right), \quad n_i = 2,3,\ldots,n_i^{\max}, \quad i = 2,3,\ldots,M, \tag{34}$$

where $\alpha$ is a gain factor and $\hat{\tau}^1_{x_ix_1}$ and $\hat{\tau}^{n_i}_{x_ix_1}$ are the time delay samples corresponding to the largest and the $n_i$th largest peak in ES-GCC function $R_{x_ix_1}$. If $R_{x_ix_1}$possesses no time delay sample that can meet the constraint above, 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 possible time delay vector $\mathbf{b}_u$ and there should be $D$ correct combinations in those possible combinations. Figure 4 illustrates the procedure of forming the possible time delay vector combinations and $f_s$ is the sampling rate. The relation between estimated time delay and estimated time delay sample is:

$$\hat{t}_i - \hat{t}_1 = \frac{1}{f_s} \times \hat{\tau}_{x_ix_1}, \tag{35}$$

where $\hat{t}_i$ is the estimated time delay from the sound source to the $i$th microphone and $\hat{\tau}_{x_ix_1}$ is the estimated time delay sample between the $i$th microphone and the first microphone. The next issue is how to choose correct combinations and determine the sound source number.

To access whether the delay combination is likely to be a correct one, this work proposes a novel concept of evaluating if the corresponding sound velocity estimation of (31) is within an acceptable range. In other words, each possible combination $\mathbf{b}_u$ is plugged into (31) to compute the sound velocity. It is considered as a correct combination if the following criterion is satisfied.

$$\left| \frac{1}{\left| \left(\mathbf{A}_f^{\mathrm{T}}\mathbf{A}_f\right)^{-1}\mathbf{A}_f^{\mathrm{T}}\mathbf{b_u} \right|} - \overline{v} \right| < \varepsilon, \tag{36}$$

$$u = 1,2,3,\ldots,n_2^{\max} \times n_3^{\max} \times \cdots \times n_M^{\max},$$

where $\overline{v} = 34300$ is the sound velocity in cm/sec and $\varepsilon$ is a threshold representing the acceptable range. Assume that

there are $\widetilde{D}$ combinations $(\widetilde{\mathbf{b}}_1, \widetilde{\mathbf{b}}_2, \ldots, \widetilde{\mathbf{b}}_{\widetilde{D}})$ satisfying (36) and the corresponding sound sources direction can be obtained by

$$\widetilde{\mathbf{r}}_u = \begin{bmatrix} \widetilde{x}_u & \widetilde{y}_u & \widetilde{z}_u \end{bmatrix} = \frac{\left(\mathbf{A}_f^{\mathrm{T}}\mathbf{A}_f\right)^{-1}\mathbf{A}_f^{\mathrm{T}}\widetilde{\mathbf{b}}_u}{\left|\left(\mathbf{A}_f^{\mathrm{T}}\mathbf{A}_f\right)^{-1}\mathbf{A}_f^{\mathrm{T}}\widetilde{\mathbf{b}}_u\right|},$$
$$\theta_u = \tan^{-1}\left(\frac{\widetilde{y}_u}{\widetilde{x}_u}\right), \qquad \phi_u = \tan^{-1}\left(\frac{\widetilde{z}_u}{\sqrt{\widetilde{x}_u^2 + \widetilde{y}_u^2}}\right), \tag{37}$$
$$u = 1, 2, 3, \ldots, \widetilde{D},$$

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

*4.2. Proposed Adaptive K-means++ for Sound Source Number and Directions Estimation.* For the robustness consideration, the final sound source number and directions will be determined over $Q$-times results from (37). Define all the accumulated estimation angle results over $Q$-times of (37) estimation as

$$\widetilde{\boldsymbol{\theta}} = \begin{bmatrix} \widetilde{\theta}_1 & \widetilde{\theta}_2 & \cdots & \widetilde{\theta}_G \end{bmatrix},$$
$$\widetilde{\boldsymbol{\varphi}} = \begin{bmatrix} \widetilde{\phi}_1 & \widetilde{\phi}_2 & \cdots & \widetilde{\phi}_G \end{bmatrix}, \tag{38}$$
$$G = Q \times \left(\widetilde{D}_1 + \widetilde{D}_2 + \cdots + \widetilde{D}_Q\right),$$

where $\widetilde{D}_q$ represents the combination number which meets (36) constraint at the $q$th testing. So far, we have $G$ data and each data has two features ($\widetilde{\theta}_g$ and $\widetilde{\phi}_g$). Our goal is to divide these data into $\widehat{D}$ clusters based on the two features. A cluster is defined as a set of sound source direction data points. For a cluster, the data within this cluster should be similar to one another and it means that the data within this cluster should come from the same sound source direction. The number $\widehat{D}$ is defined as the sound source number. Therefore, among the set of $G$ sound source direction data points, we wish to choose $\widehat{D}$ cluster centers so as to minimize the potential function:

$$\min \sum_{\hat{d}=1}^{\widehat{D}} \sum_{\boldsymbol{\sigma}_g \in C_{\hat{d}}} \left\|\boldsymbol{\sigma}_g - \boldsymbol{\mu}_{\hat{d}}\right\|^2, \quad \boldsymbol{\sigma}_g = \begin{bmatrix} \widetilde{\theta}_g & \widetilde{\phi}_g \end{bmatrix}, \tag{39}$$
$$g = 1, 2, 3, \ldots, G,$$

where there are $\widehat{D}$ clusters $\{C_1, C_2, \ldots, C_{\widehat{D}}\}$ and $\boldsymbol{\mu}_{\hat{d}}$ is the center of all the points $\boldsymbol{\sigma}_g \in C_{\hat{d}}$. The sound source direction data $\boldsymbol{\sigma}_g$ is assigned to $C_{\hat{d}}$, if $\boldsymbol{\mu}_{\hat{d}}$ is the closet cluster center to $\boldsymbol{\sigma}_g$. Because the sound source number is unknown, we set the cluster number $\widehat{D}$ to be one and initial center $\boldsymbol{\mu}_1$ to be the median of $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\varphi}}$ as the initial condition to execute K-means. When the K-means algorithm converges, the constraint below is checked:

$$E\left[\left|\boldsymbol{\sigma}_g - \boldsymbol{\mu}_{\hat{d}}\right|^2\right] < \delta, \quad \boldsymbol{\sigma}_g \in C_{\hat{d}},\ \hat{d} = 1, 2, \ldots, \widehat{D}, \tag{40}$$

where $E(\cdot)$ is the expectation operation and $\delta$ is a specified threshold. Equation (40) 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 $\delta$, the value of $\widehat{D}$ is increased by one. Then the other initial center $\boldsymbol{\mu}_{\widehat{D}}$ is found by using the seeding technique of K-means++ [27] defined in (41) and the K-means algorithm is computed again.

Find the integer $\widehat{G}$ that

$$\sum_{g=1}^{\widehat{G}} \mathrm{DIS}\left(\boldsymbol{\sigma}_g\right) \geq \overline{\mathrm{DIS}} > \sum_{g=1}^{\widehat{G}-1} \mathrm{DIS}\left(\boldsymbol{\sigma}_g\right), \tag{41}$$
$$\boldsymbol{\mu}_{\widehat{D}} = \boldsymbol{\sigma}_{\widehat{G}},$$

where $\mathrm{DIS}(\boldsymbol{\sigma}_g)$ represents the distance between $\boldsymbol{\sigma}_g$ and the nearest center we have already chosen; $\overline{\mathrm{DIS}}$ is the real number chosen uniformly at random between 0 and $\sum_{g=1}^{G} \mathrm{DIS}(\boldsymbol{\sigma}_g)$.

Otherwise, the final sound source number is $\widehat{D}$ and the sound source directions are

$$\begin{bmatrix} \widehat{\theta}_{\hat{d}} & \widehat{\phi}_{\hat{d}} \end{bmatrix} = \boldsymbol{\mu}_{\hat{d}} \quad \hat{d} = 1, 2, \ldots, \widehat{D}. \tag{42}$$

For the adaptive K-means++ algorithm, the inputs are $\boldsymbol{\sigma}_g$ and the outputs are $\boldsymbol{\mu}_{\hat{d}}$ and $\widehat{D}$. The flowchart of the adaptive K-means++ algorithm for estimating the sound sources number and directions is shown in Figure 5 and is summarized as follows.

*Step 1.* Calculate ES-GCC function $R_{x_i x_1}(\tau)$. Pick the peaks satisfying (34) from $R_{x_i x_1}(\tau)$ for each microphone pair and list all the possible time delay vector combinations $\mathbf{b}_u$.

*Step 2.* Select $\widetilde{D}$ time delay vector from $\mathbf{b}_u$ using (36) and estimate the corresponding sound source direction using (37).

*Step 3.* Repeat Steps 1 to 2 $Q$ times and accumulate the results. Before each repeat, shift the start frame of Step 1 with $\overline{K}$ frames.

*Step 4.* Cluster the accumulated results using adaptive K-means++ algorithm and the final cluster number and centers are sound source number and directions, respectively.

## 5. Experimental Results

The experiments were performed in a real room approximately of the size 10.5 m × 7.2 m and height of 3.6 m and its reverberation time at 1000 Hz is 0.52 second. The reverberation time was measured by playing a 1000 Hz tone and then estimating the time of the direct sound to decay by 60 dB below the level of the direct sound. An 8-channel digital microphone array platform is installed on the robot for the experiment shown in Figure 6 and the microphone positions are marked with the circle symbol. The room temperature is approximately 22°C and the sampling rate is 16 kHz. The experimental condition is shown in Figure 7 and

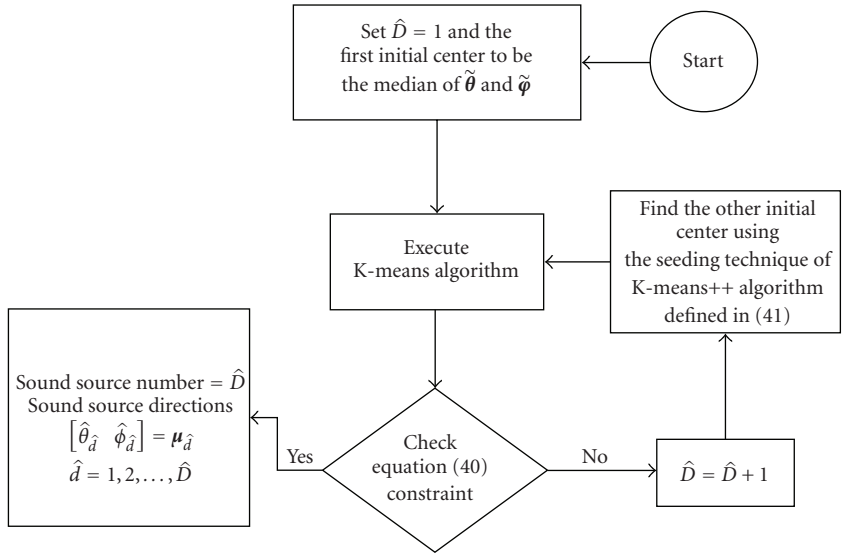


Figure 5: The flowchart of adaptive K-means++ algorithm.

the distance from each sound source to the origin is 270 cm. The sound sources are Chinese and English conversational speech in female and male. Each conversational speech source is different and is spoken by different people. In Figure 7, the microphone and sound source locations are set to (cm)

$$\begin{aligned}
&\text{Mic.1} = \begin{bmatrix} 20 & 20 & 0 \end{bmatrix}, \quad \text{Mic.2} = \begin{bmatrix} 20 & -20 & 0 \end{bmatrix},\\
&\text{Mic.3} = [-20 \quad -20 \quad 0], \quad \text{Mic.4} = \begin{bmatrix} -20 & 20 & 0 \end{bmatrix},\\
&\text{Mic.5} = \begin{bmatrix} 0 & 20 & 30 \end{bmatrix}, \quad \text{Mic.6} = \begin{bmatrix} 0 & 20 & -30 \end{bmatrix},\\
&\text{Mic.7} = \begin{bmatrix} 0 & -20 & 30 \end{bmatrix}, \quad \text{Mic.8} = \begin{bmatrix} 0 & -20 & -30 \end{bmatrix},\\
&S1 = \begin{bmatrix} 190 & -190 & 0 \end{bmatrix}, \quad S2 = \begin{bmatrix} 190 & 190 & 24 \end{bmatrix},\\
&S3 = -\begin{bmatrix} 188 & 188 & 47 \end{bmatrix}, \quad S4 = \begin{bmatrix} -190 & -190 & 0 \end{bmatrix},\\
&S5 = \begin{bmatrix} 0 & 269 & -24 \end{bmatrix}, \quad S6 = \begin{bmatrix} 0 & -266 & -47 \end{bmatrix}.
\end{aligned} \tag{43}$$

The dehumidifier which is 430 cm from the first microphone is turned on during this experiment (Noise 1 in Figure 7). The parameters of $\alpha$, $\varepsilon$, and $\delta$ are determined by our experience and are empirically set to be 0.7, 5000, and 23. The accumulation parameters $Q$ and $\overline{K}$ are set to be 20 and 25.

*5.1. ES-GCC Time Delay Estimation Performance Evaluation.* Two GCC-based TDE algorithms, GCC-PHAT and GCC-ML [2], are computed to compare with the proposed ES-GCC algorithm. Seven microphone pairs ((1,2), (1,3), (1,4), (1,5), (1,6), (1,7), and (1,8) ) and six sound source positions in Figure 7 are selected for this TDE experiment. For each test, only one speech source is active and seven microphone

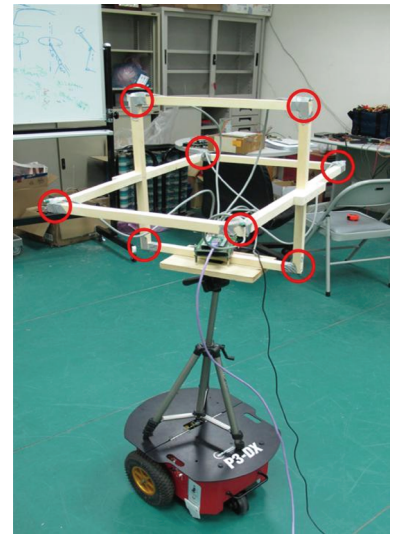

Figure 6: Digital microphone array mounted on the robot.

pairs are all chosen to test. The STFT size is set to be 512 with 50% overlap and mutually independent white Gaussian noise is properly scaled and added to each microphone signal to control the signal-to-noise ratio (SNR). The performance index, Root Mean Square Error (RMSE), is defined below to evaluate the performance of the suggested method:

$$\text{RMSE} = \sqrt{\frac{1}{N_T}\sum_{i=1}^{N_T}\left(\hat{D}_i - D_i\right)^2}, \tag{44}$$

where $N_T$ is the total number of estimation, $\hat{D}_i$ is the $i$th time delay estimation, and $D_i$ is the $i$th correct delay sample with a integer. Figure 8 shows the RMSE results as a function of SNR for three different TDE algorithms. The total number of

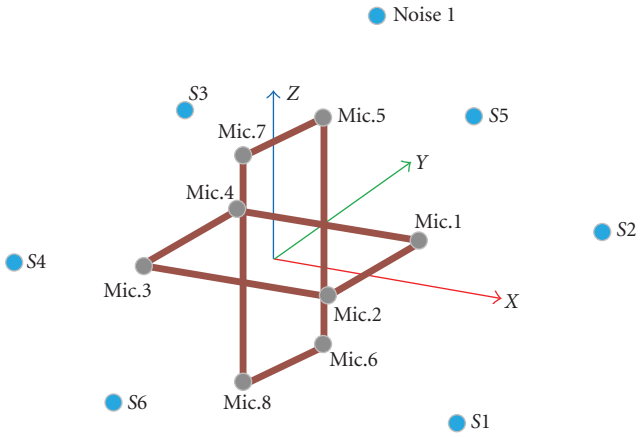


Figure 7: Arrangement of microphone array and sound sources.

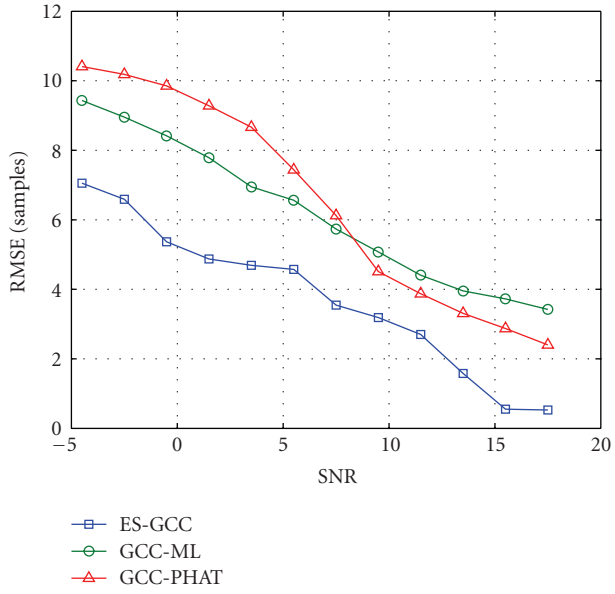


Figure 8: TDE RMSE results versus SNR.

estimation $N_T$ is 294. As seen from Figure 8, the GCC-PHAT yields better TDE performance than GCC-ML at higher SNR. This is because the experimental environment is reverberant and the GCC-ML suffers significant performance degradation under reverberation.

Comparing to GCC-ML, the GCC-PHAT has robustness with respect to reverberation. However, the GCC-PHAT method neglects the noise effect, and hence, it begins to exhibit dramatic performance degradation as the SNR is decreased. Unlike GCC-PHAT, GCC-ML does not exhibit this phenomenon since it has a priori knowledge about the noise power spectra which can help estimator to cope with distortion. The ES-GCC achieves the best performance, because the ES-GCC method does not focus on the weighting function process of GCC-based method and it directly takes the principal component vector as the microphone received signal for further signal processing. The appendix

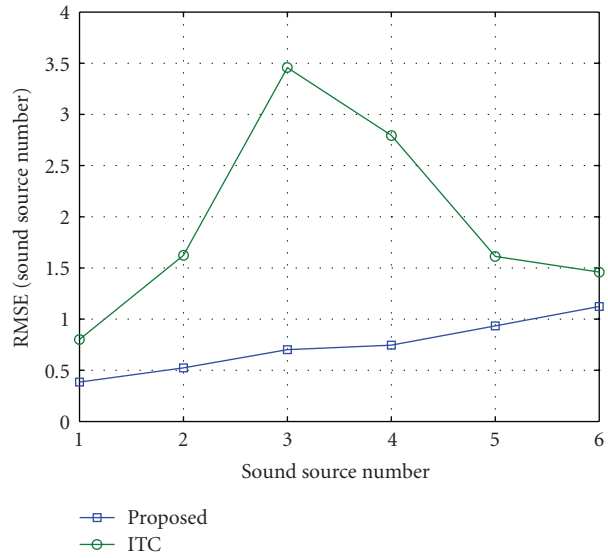


Figure 9: Sound source number estimation result.

provides the proof that the principal component vector can be considered as the approximation of speech-only signal and this is the reason why the ES-GCC method is robust to the SNR.

*5.2. Evaluation of Sound Source Number and Directions Estimation.* The wideband incoherent MUSIC algorithm [9] with arithmetic mean is adopted to compare with the proposed algorithm. Ten major frequencies, ranging from 0.1 KHz to 3.4 KHz, were adopted for the MUSIC algorithm. Outliers were removed from the estimated angles by utilizing the method provided in [29]. In addition, the sound source number should be known first for MUSIC algorithm to construct the noise projection matrix. Therefore, the eigenvalues-based information theoretic criteria (ITC) method [21] is employed to estimate the sound source number. The sound source number estimation RMSE result is shown in Figure 9 and the averaged SNR is 17.23 dB. The RMSE is defined similar to (44) with a different measurement unit. The sound source positions are chosen randomly from six positions shown in Figure 7 and the number of estimation $N_T$ for each condition is 100. The noise 1 in Figure 7 is active in this experiment. As can be seen, the proposed sound source number estimation method yields better performance than the ITC method. One of the reasons is that the eigenvalue distribution is sensitive to reverberation and background noise. When the sound source number is larger than or equal to three, the ITC method often estimates a higher sound source number (5, 6, or 7).

The sound source direction estimation RMSE result is shown in Figure 10. For fair comparison, the RMSE is calculated when the sound source number estimation is correct. Figure 10 shows that the MUSIC algorithm becomes worse as the sound source number is increased since the MUSIC algorithm is sensitive to coherent signal especially when the environment is multiple sound sources and reverberant. The

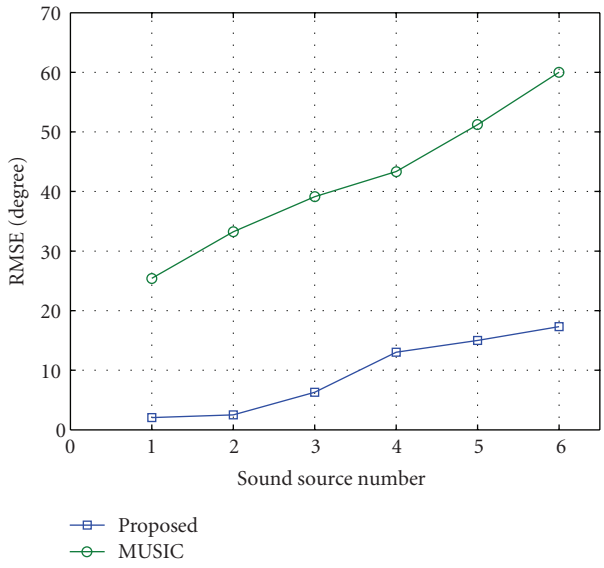


Figure 10: Sound source directions estimation result.

proposed method uses sound velocity as the criterion for time delay candidate selection and the adaptive K-means++ is employed at final stage to cluster the sound source number and directions. The other advantage of the proposed method is that there is no a priori knowledge for sound source number and we use the adaptive K-means++ to estimate the sound source number and directions simultaneously. An incorrect sound source number for MUSIC algorithm would cause an even worse performance than Figure 10. In addition, in multiple sound sources case, if we take all time delay combinations to estimate the sound source direction without sound velocity selection mechanism, the result becomes very poor. We find that the wrong combination of time delay vector $\mathbf{b}_u$ will cause the estimated sound speed to range between 9000 and 15000 or more than 50000.

## 6. Conclusion

This work explains a sound source number and directions estimation algorithm. The multiple source time delay vector combination problem can be solved by the proposed reasonable 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 and the experimental results show that the proposed algorithm is robust to real environment and can provide reliable information for further robot audition research.

The accuracy of adaptive K-means++ may be influenced by outliers if there is no outlier rejection. Therefore, the outlier rejection method may be incorporated to improve the performance. Moreover, the parameters of $\alpha$, $\varepsilon$, and $\delta$ are determined by our experience. In our experience, the parameter $\varepsilon$ is not as sensitive as $\alpha$ and $\delta$ to influence the results. The sensitivity of these parameters to influence the results is the other issue and this is left as a further research topic.

## Appendix

Equation (2) can also be written as a square matrix form:

$$\mathbf{X}(\omega,k) = \mathbf{A}_s(\omega)\mathbf{S}_s(\omega,k) + \mathbf{N}(\omega,k), \tag{A.1}$$

where

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

$$\mathbf{N}(\omega,k) = \left[N_1(\omega,k),\ \ldots,\ N_M(\omega,k)\right]^{\mathrm{T}} \in C^{M\times 1},$$

$$\mathbf{S}_s(\omega,k) = \left[S_1(\omega,k),\ldots,\ S_D(\omega,k)\ \ 0,\ldots,0\right]^{\mathrm{T}} \in C^{M\times 1},$$

$$\mathbf{A}_s(\omega) = \begin{bmatrix} A_{11}(\omega)\cdots & A_{1D}(\omega) & 0\cdots 0 \\ \vdots & \vdots & \vdots \\ A_{M1}(\omega)\cdots & A_{MD}(\omega) & 0\cdots 0 \end{bmatrix} \in C^{M\times M}. \tag{A.2}$$

Suppose that the noises are spatially white, and the noise correlation matrix is diagonal matrix $\sigma_n^2\mathbf{I}$. Therefore, the received signal correlation matrix with EVD can be described as

$$\mathbf{R}_{xx}(\omega) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{X}(\omega,k)\mathbf{X}^{\mathrm{H}}(\omega,k) = \mathbf{A}_s(\omega)\mathbf{R}_{ss}(\omega)\mathbf{A}_s^{\mathrm{H}}(\omega) + \sigma_n^2\mathbf{I}$$

$$= \sum_{m=1}^{M}\lambda_m(\omega)\mathbf{V}_m(\omega)\mathbf{V}_m^{\mathrm{H}}(\omega), \tag{A.3}$$

where $\mathbf{R}_{ss}(\omega) = (1/\mathrm{K})\sum_{k=1}^{K}\mathbf{S}_s(\omega,k)\mathbf{S}_s^{\mathrm{H}}(\omega,k)$; $\lambda_m(\omega)$ and $\mathbf{V}_m(\omega)$ are eigenvalues and corresponding eigenvectors with $\lambda_1(\omega) \geq \lambda_2(\omega) \geq \cdots \geq \lambda_M(\omega)$. Since the $M$ eigenvectors are orthogonal to one another, they form a basis and can be used to express an arbitrary vector $\mathbf{v}(\omega)$ in the following

$$\mathbf{v}(\omega) = \sum_{m=1}^{M}\lambda_m(\omega)\mathbf{V}_m(\omega) \in C^{M\times 1}. \tag{A.4}$$

Since $\mathbf{V}_m^{\mathrm{H}}(\omega)\mathbf{V}_i(\omega) = 0$ for $m \neq i$ and $\mathbf{V}_m^{\mathrm{H}}(\omega)\mathbf{V}_i(\omega) = 1$ for $m = i$. Therefore, the dot product of $\mathbf{v}(\omega)$ and $\mathbf{V}_i(\omega)$ is

$$\mathbf{v}^{\mathrm{H}}(\omega)\mathbf{V}_i(\omega) = \sum_{m=1}^{M}\lambda_m^{\mathrm{H}}(\omega)\mathbf{V}_m^{\mathrm{H}}(\omega)\mathbf{V}_i(\omega) = \lambda_i^{\mathrm{H}}(\omega). \tag{A.5}$$

Substituting (A.5) into (A.4), we have

$$\mathbf{v}(\omega) = \sum_{m=1}^{M}\mathbf{V}_m^{\mathrm{H}}(\omega)\mathbf{v}(\omega)\mathbf{V}_m(\omega) = \sum_{m=1}^{M}\mathbf{V}_m(\omega)\mathbf{V}_m^{\mathrm{H}}(\omega)\mathbf{v}(\omega). \tag{A.6}$$

Therefore, $\mathbf{I} = \sum_{m=1}^{M} \mathbf{V}_m(\omega)\mathbf{V}_m^{\mathrm{H}}(\omega)$. Because $\sigma_n^2\mathbf{I} = \sum_{m=1}^{M} \sigma_n^2\mathbf{V}_m(\omega)\mathbf{V}_m^{\mathrm{H}}(\omega)$, we have the signal-only correlation matrix:

$$\begin{aligned}\mathbf{C}_{xx}(\omega) &= \mathbf{A}_s(\omega)\mathbf{R}_{ss}(\omega)\mathbf{A}_s^{\mathrm{H}}(\omega)\\ &= \sum_{m=1}^{M} (\lambda_m(\omega) - \sigma_n^2)\mathbf{V}_m(\omega)\mathbf{V}_m^{\mathrm{H}}(\omega)\\ &= \mathbf{V}_s(\omega)\mathbf{\Lambda}_s(\omega)\mathbf{V}_s^{\mathrm{H}}(\omega),\end{aligned} \tag{A.7}$$

where

$$\begin{aligned}\mathbf{V}_s(\omega) &= \begin{bmatrix}\mathbf{V}_1(\omega) & \cdots & \mathbf{V}_M(\omega)\end{bmatrix} \in C^{M\times M},\\ \mathbf{\Lambda}_s(\omega) &= \begin{bmatrix}\lambda_1(\omega) - \sigma_n^2 & & 0\\ & \ddots & \\ 0 & & \lambda_M(\omega) - \sigma_n^2\end{bmatrix} \in C^{M\times M}.\end{aligned} \tag{A.8}$$

Applying $QR$ factorization to $\mathbf{A}_s(\omega)$, we have

$$\mathbf{A}_s(\omega) = \mathbf{Q}(\omega)\mathrm{R}(\omega), \tag{A.9}$$

where

$$\begin{aligned}\mathbf{Q}(\omega) &= \begin{bmatrix}q_{11}(\omega) & \cdots & q_{1M}(\omega)\\ \vdots & \ddots & \vdots\\ q_{M1}(\omega) & \cdots & q_{MM}(\omega)\end{bmatrix} \in C^{M\times M},\\ \mathbf{R}(\omega) &= \begin{bmatrix}r_{11}(\omega) & \cdots & r_{1M}(\omega)\\ 0 & & \\ \vdots & \ddots & \vdots\\ 0 & 0\cdots 0 & r_{MM}(\omega)\end{bmatrix} \in C^{M\times M}.\end{aligned} \tag{A.10}$$

Hence,

$$\begin{aligned}\mathbf{C}_{xx}(\omega) &= \mathbf{A}_s(\omega)\mathbf{R}_{ss}(\omega)\mathbf{A}_s^{\mathrm{H}}(\omega)\\ &= \mathbf{Q}(\omega)\mathbf{R}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{R}^{\mathrm{H}}(\omega)\mathbf{Q}^{\mathrm{H}}(\omega)\\ &= \mathbf{Q}(\omega)\mathbf{R}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{R}^{\mathrm{H}}(\omega)\mathbf{Q}^{-1}(\omega).\end{aligned} \tag{A.11}$$

$\mathbf{A}_s(\omega)\mathbf{R}_{ss}(\omega)\mathbf{A}_s^{\mathrm{H}}(\omega)$ and $\mathbf{R}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{R}^{\mathrm{H}}(\omega)$ are similar matrix and they have the same eigenvalues. Decompose $\mathbf{R}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{R}^{\mathrm{H}}(\omega)$ using EVD, and we have

$$\mathbf{R}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{R}^{\mathrm{H}}(\omega) = \mathbf{\Delta}(\omega)\mathbf{\Lambda}_s(\omega)\mathbf{\Delta}^{\mathrm{H}}(\omega), \tag{A.12}$$

where $\mathbf{\Delta}(\omega)$ is the eigenvector matrix of $\mathbf{R}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{R}^{\mathrm{H}}(\omega)$ defined as

$$\begin{aligned}\mathbf{\Delta}(\omega) &= \begin{bmatrix}\mathbf{\Delta}_1(\omega) & \cdots & \mathbf{\Delta}_M(\omega)\end{bmatrix} \in C^{M\times M},\\ \mathbf{\Delta}_m(\omega) &= \begin{bmatrix}\Delta_{1m}(\omega) & \cdots & \Delta_{Mm}(\omega)\end{bmatrix}^{\mathrm{T}} \in C^{M\times 1}.\end{aligned} \tag{A.13}$$

Therefore, substituting (A.12) into (A.11), we have the relationship between $\mathbf{V}_m(\omega)$ and $\mathbf{\Delta}_m(\omega)$:

$$\mathbf{V}_m(\omega) = \mathbf{Q}(\omega)\mathbf{\Delta}_m(\omega) \quad m = 1, 2, \ldots, M \tag{A.14}$$

Next, we need to represent $\mathbf{\Delta}_m(\omega)$ using $\mathbf{R}(\omega)$ and $S_d(\omega)$ for further process. The matrix $\mathbf{R}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{R}^{\mathrm{H}}(\omega)$ can also be expressed as

$$\begin{aligned}\mathbf{R}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{R}^{\mathrm{H}}(\omega) &= E\left\{\mathbf{R}(\omega)\mathbf{S}_s(\omega)\mathbf{S}_s^{\mathrm{H}}(\omega)\mathbf{R}^{\mathrm{H}}(\omega)\right\}\\ &= E\left\{\begin{bmatrix}\sum_{d=1}^{D} r_{1d}(\omega)S_d(\omega)\\ \sum_{d=2}^{D} r_{2d}(\omega)S_d(\omega)\\ \vdots\\ \sum_{d=D}^{D} r_{Dd}(\omega)S_d(\omega)\\ 0\\ \vdots\\ 0\end{bmatrix}\begin{bmatrix}\sum_{d=1}^{D} r_{1d}(\omega)S_d(\omega)\\ \sum_{d=2}^{D} r_{2d}(\omega)S_d(\omega)\\ \vdots\\ \sum_{d=D}^{D} r_{Dd}(\omega)S_d(\omega)\\ 0\\ \vdots\\ 0\end{bmatrix}^{\mathrm{H}}\right\}\\ &= \begin{bmatrix}E\left(z_1(\omega)z_1^{\mathrm{H}}(\omega)\right) & E\left(z_1(\omega)z_2^{\mathrm{H}}(\omega)\right)\cdots & E\left(z_1(\omega)z_D^{\mathrm{H}}(\omega)\right) & 0\cdots & 0\\ E\left(z_2(\omega)z_1^{\mathrm{H}}(\omega)\right) & E\left(z_2(\omega)z_2^{\mathrm{H}}(\omega)\right)\cdots & E\left(z_2(\omega)z_D^{\mathrm{H}}(\omega)\right) & 0\cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ E\left(z_D(\omega)z_1^{\mathrm{H}}(\omega)\right) & E\left(z_D(\omega)z_2^{\mathrm{H}}(\omega)\right)\cdots & E\left(z_D(\omega)z_D^{\mathrm{H}}(\omega)\right) & 0\cdots & 0\\ 0 & 0\cdots & 0 & 0\cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0\cdots & 0 & 0\cdots & 0\end{bmatrix} \in C^{M\times M},\end{aligned} \tag{A.15}$$

where $E(\cdot)$ is the expectation operation and

$$z_i(\omega) = \sum_{d=i}^{D} r_{id}(\omega) S_d(\omega)$$

$$E\left(z_i(\omega) z_j^{\mathrm{H}}(\omega)\right) = \sum_{d=i}^{D} E(r_{id}(\omega) S_d(\omega)) \times \sum_{d=j}^{D} E\left(S_d^{\mathrm{H}}(\omega) r_{jd}^{\mathrm{H}}(\omega)\right) + \sigma_{ij}^2(\omega)$$

$$\sigma_{ij}^2(\omega) = \begin{cases} \sum_{d=i}^{D} r_{id}(\omega) r_{jd}^{\mathrm{H}}(\omega) \mathrm{var}(S_d(\omega)) & \text{if } i = j \\ \sum_{d=\max(i,j)}^{D} r_{id}(\omega) r_{jd}^{\mathrm{H}}(\omega) \mathrm{var}(S_d(\omega)) & \text{if } i \neq j \end{cases} \tag{A.16}$$

where $\mathrm{var}(x)$ is the variance of $x$, $\max(i, j)$ is the maximum value and between $i$ and $j$.

From (A.15) and the eigenvalue equation $(\mathbf{R}(\omega)\mathbf{R}_{ss}(\omega)\mathbf{R}^{\mathrm{H}}(\omega)\boldsymbol{\Delta}_m(\omega) = (\lambda_m(\omega) - \sigma_n^2)\boldsymbol{\Delta}_m(\omega))$, we have the linear equation in $M$ unknowns $(\Delta_{1m}(\omega), \Delta_{2m}(\omega), \ldots, \Delta_{Mm}(\omega))$ shown at (A.17):

$$\begin{aligned} &\sum_{k=1}^{D} E\left(\theta_1(\omega)\theta_k^{\mathrm{H}}(\omega)\right)\Delta_{km}(\omega) = (\lambda_m(\omega) - \sigma_n^2)\Delta_{1m}(\omega) - \mu_{1m}(\omega), \\ &\sum_{k=1}^{D} E\left(\theta_2(\omega)\theta_k^{\mathrm{H}}(\omega)\right)\Delta_{km}(\omega) = (\lambda_m(\omega) - \sigma_n^2)\Delta_{2m}(\omega) - \mu_{2m}(\omega), \\ &\vdots \\ &\sum_{k=1}^{D} E\left(\theta_D(\omega)\theta_k^{\mathrm{H}}(\omega)\right)\Delta_{km}(\omega) = (\lambda_m(\omega) - \sigma_n^2)\Delta_{Dm}(\omega) - \mu_{Dm}(\omega), \\ &0 = (\lambda_m(\omega) - \sigma_n^2)\Delta_{\beta m}(\omega), \end{aligned} \tag{A.17}$$

where $\mu_{pm}(\omega)$ is the variance part which is defined as

$$\begin{aligned} \mu_{pm}(\omega) &= \sum_{k=1}^{D} \Delta_{km}\sigma_{pk}^2(\omega), \quad p = 1, 2, \cdots D \\ \beta &= D+1, D+2, \ldots, M \\ E\left(\theta_i(\omega)\theta_j^{\mathrm{H}}(\omega)\right) &= E\left(z_i(\omega) z_j^{\mathrm{H}}(\omega)\right) - \sigma_{ij}^2(\omega) \end{aligned} \tag{A.18}$$

To solve $\Delta_{dm}(\omega)$, we assume that the variance part $\mu_{pm}(\omega)$ can be neglected. This is possible if $(\lambda_m(\omega) - \sigma_n^2)\Delta_{dm}(\omega) \gg \mu_{dm}(\omega)$. Therefore we chose the maximum eigenvalue $(\lambda_1(\omega) - \sigma_n^2)$ to solve this linear equation. In (A.17), the first row divided by the second row is and we have

$$\begin{aligned} \frac{\sum_{k=1}^{D} E\left(\theta_1(\omega)\theta_k^{\mathrm{H}}(\omega)\right)\Delta_{k1}(\omega)}{\sum_{k=1}^{D} E\left(\theta_2(\omega)\theta_k^{\mathrm{H}}(\omega)\right)\Delta_{k1}(\omega)} &= \frac{\Delta_{11}(\omega)}{\Delta_{21}(\omega)} \\ &= \frac{\sum_{d=1}^{D} E(r_{1d}(\omega)S_d(\omega)) \times \mathcal{A}}{\sum_{d=2}^{D} E(r_{2d}(\omega)S_d(\omega)) \times \mathcal{A}}, \end{aligned} \tag{A.19}$$

where $\mathcal{A}$ denotes $(\sum_{k=1}^{D}\sum_{d=k}^{D} E(S_d^{\mathrm{H}}(\omega) r_{kd}^{\mathrm{H}}(\omega))\Delta_{k1}(\omega))$.

Therefore,

$$\frac{\Delta_{11}(\omega)}{\Delta_{21}(\omega)} = \frac{\sum_{d=1}^{D} E(r_{1d}(\omega)S_d(\omega))}{\sum_{d=2}^{D} E(r_{2d}(\omega)S_d(\omega))}. \tag{A.20}$$

With the similar method, the eigenvector $\boldsymbol{\Delta}_1(\omega)$ associated with the maximum eigenvalue can be obtained:

$$\Delta_{i1}(\omega) = \begin{cases} \beta \cdot \sum_{d=i}^{D} E(r_{id}(\omega)S_d(\omega)) & \text{if } i \leq D \\ 0 & \text{if } i > D, \end{cases} \quad i = 1, 2, \cdots M, \tag{A.21}$$

where $\beta$ is a scalar.

Hence, the eigenvector can be represented as

$$\mathbf{V}_1(\omega) = \mathbf{Q}(\omega)\boldsymbol{\Delta}_1(\omega) = \begin{bmatrix} \beta \cdot \sum_{i=1}^{D} \left( q_{1i}(\omega) \times \sum_{d=i}^{D} E(r_{id}(\omega)S_d(\omega)) \right) \\ \beta \cdot \sum_{i=1}^{D} \left( q_{2i}(\omega) \times \sum_{d=i}^{D} E(r_{id}(\omega)S_d(\omega)) \right) \\ \vdots \\ \beta \cdot \sum_{i=1}^{D} \left( q_{Mi}(\omega) \times \sum_{d=i}^{D} E(r_{id}(\omega)S_d(\omega)) \right) \end{bmatrix}. \tag{A.22}$$

If the observation time is sufficiently long, then $S_d(\omega, k) \approx E(S_d(\omega))$. Therefore, the microphone received signal can be

modeled as

$$\begin{aligned}
&\mathbf{X}(\omega,k)\\
&= \mathbf{A}_s(\omega)\mathbf{S}_s(\omega,k) + \mathbf{N}(\omega,k) = \mathbf{Q}(\omega)\mathbf{R}(\omega)\mathbf{S}_s(\omega,k) + \mathbf{N}(\omega,k)\\
&= \begin{bmatrix} \sum_{i=1}^{D}\left(q_{1i}(\omega)\times\sum_{d=i}^{D} r_{id}(\omega)E(S_d(\omega))\right) + N_1(\omega,k)\\ \sum_{i=1}^{D}\left(q_{2i}(\omega)\times\sum_{d=i}^{D} r_{id}(\omega)E(S_d(\omega))\right) + N_2(\omega,k)\\ \vdots \\ \sum_{i=1}^{D}\left(q_{Mi}(\omega)\times\sum_{d=i}^{D} r_{id}(\omega)E(S_d(\omega))\right) + N_M(\omega,k)\end{bmatrix}\\
&= \frac{1}{\beta}\mathbf{V}_1(\omega) + \mathbf{N}(\omega,k).
\end{aligned} \tag{A.23}$$

As can be seen from (A.23), the received speech signal is only the scalar version of the corresponding eigenvector for the maximum eigenvalue. Therefore, we take this eigenvector as the microphone received signal for time delay estimation. Equation (A.23) is obtained by using the maximum eigenvalue to solve (A.17). If other eigenvalues can also neglect the variance as $(\lambda_m(\omega) - \sigma_n^2)\Delta_{dm}(\omega) \gg \mu_{dm}(\omega)$, they can also have the speech signal approximation property. It represents that if the sound source number is one, $\mathbf{V}_1(\omega)$ is the only eigenvector which can represent the received speech signal since $\lambda_1(\omega)$ is the only dominant eigenvalue and the other eigenvectors ($\mathbf{V}_i(\omega)$, $i = 2,3,\ldots,M$) contain the noise information. If the sound source number is larger than one, the other eigenvectors ($\mathbf{V}_i(\omega)$, $i = 2,3,\ldots,D$) may contain some speech signal information. However, the conversational speech sources are asynchronous and contain many short pauses. Some speech sources information may not be represented by $\mathbf{V}_1(\omega)$ in this frame but may be represented in the next frame. Based on this concept, this paper uses eigenvector $\mathbf{V}_1(\omega)$ for time delay estimation since it can represent received speech signal most, accumulates the estimated DOA results, and uses adaptive K-means++ for clustering the accumulated results. The algorithms that use the vectors that lie in the signal subspace are based on a principal components analysis (PCA) of the autocorrelation matrix and are referred to as signal subspace method [24]. This paper further justifies the use of $\mathbf{V}_1(\omega)$ since it can represent the speech signal better than the other eigenvectors from (A.17) and (A.23).

## References


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

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

[3] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, pp. 375–378, Munich, Germany, April 1997.

[4] Q. H. Wang, T. Ivanov, and P. Aarabi, "Acoustic robot navigation using distributed microphone arrays," *Information Fusion*, vol. 5, no. 2, pp. 131–140, 2004.

[5] J. Scheuing and B. Yang, "Correlation-based TDOA-estimation for multiple sources in reverberant environments," in *Speech and Audio Processing in Adverse Environments*, Chapter 11, pp. 381–416, Springer, Berlin, Germany, 2008.

[6] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1110–1124, 2003.

[7] 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.

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

[9] M. Wax, T. Shan, and T. Kailath, "Spatio-Temporal spectral analysis by eigenstructure methods," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 4, pp. 817–827, 1984.

[10] H. Wang and M. Kaveh, "Coherent signal-subspace processing for detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, 1985.

[11] I. Hara, F. Asano, H. Asoh et al., "Robust speech interface based on audio and video information fusion for humanoid HRP-2," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '04)*, pp. 2404–2410, Sendai, Japan, October 2004.

[12] M. Walworth and A. Mahajan, "3D Position sensing using the difference in the time-of-flights from a wave source to various receivers," in *Proceedings of the International Conference on Advanced Robotics (ICAR '97)*, pp. 611–616, Monterey, Calif, USA, July 1997.

[13] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1228–1233, Maui, Hawaii, USA, October 2003.

[14] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[15] A. P. Badali, J. M. Valin, and P. Aarabi, "Evaluating real-time audio localization agorithms for artificial audition on mobile robots," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2033–2038, St. Louis, Mo, USA, 2009.

[16] K. Yao, R. E. Hudson, C. W. Reed, D. Chen, and F. Lorenzelli, "Blind beamforming on a randomly distributed sensor array system," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1555–1566, 1998.

[17] N. Strobel and R. Rabenstein, "Classification of time delay estimates for robust speaker localization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, vol. 6, pp. 3081–3084, March 1999.

[18] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 520–529, 2004.

[19] J.-S. Hu, C.-C. Cheng, and W.-H. Liu, "Robust speaker's location detection in a vehicle environment using GMM models," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 36, no. 2, pp. 403–412, 2006.

[20] A. Cantoni and P. Butler, "Properties of the eigenvectors of persymmetric matrices with applications to communication theory," *IEEE Transactions on Communications*, vol. 24, no. 8, pp. 804–809, 1976.

[21] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[22] K. Yamamoto, F. Asano, W. F. G. Van Rooijen, E. Y. L. Ling, T. Yamada, and N. Kitawaki, "Estimation of the number of sound sources using support vector machine," in *Proceedings of the IEEE International Conference on Accoustics, Speech, and Signal Processing*, pp. 485–488, Hong Kong, April 2003.

[23] J.-S. Hu, C.-H. Yang, and C.-K. Wang, "Estimation of sound source number and directions under a multi-source environment," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '09)*, pp. 181–186, St. Louis, Mo, USA, December 2009.

[24] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons, New York, NY, USA, 1996.

[25] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 26503, 19 pages, 2006.

[26] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100–108, 1979.

[27] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*, New Orleans, La, USA, 2007.

[28] D. Bechler and K. Kroschel, "Considering the second peak in the GCC function for multi-Source TDOA estimation with a microphone array," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC '03)*, pp. 315–318, Kyoto, Japan, September, 2003.

[29] T. Pham and B. M. Sadler, "Adaptive wideband aeroacoustic array processing," in *Proceedings of the IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, pp. 295–298, Corfu, Greece, June 1996.