Sound Source Localization for HRI Using FOC-Based Time Difference Feature and Spatial Grid Matching

Xiaofei Li and Hong Liu

***Abstract*—In human–robot interaction (HRI), speech sound source localization (SSL) is a convenient and efficient way to obtain the relative position between a speaker and a robot. However, implementing a SSL system based on TDOA method encounters many problems, such as noise of real environments, the solution of nonlinear equations, switch between far field and near field. In this paper, fourth-order cumulant spectrum is derived, based on which a time delay estimation (TDE) algorithm that is available for speech signal and immune to spatially correlated Gaussian noise is proposed. Furthermore, time difference feature of sound source and its spatial distribution are analyzed, and a spatial grid matching (SGM) algorithm is proposed for localization step, which handles some problems that geometric positioning method faces effectively. Valid feature detection algorithm and a decision tree method are also suggested to improve localization performance and reduce computational complexity. Experiments are carried out in real environments on a mobile robot platform, in which thousands of sets of speech data with noise collected by four microphones are tested in $3D$ space. The effectiveness of our TDE method and SGM algorithm is verified.**

***Index Terms*—Fourth-order cumulant spectrum, human–robot interaction (HRI), spatially correlated Gaussian noise, spatial grid matching (SGM), speech sound source localization.**

I. Introduction

AS A NATURAL and effective way, auditory function, including sound source localization (SSL) and separation, automatic speech recognition, speaker recognition, and so on, is widely used for human–robot interaction (HRI), many attentions to which had been paid over the last decades. SSL for HRI means that a robot can compute the relative position of sound source through sound signals collected by a microphone array fixed on the robot, where sound signal is speech in most HRI cases. Robert from MIT installs a simple auditory system for robots in 1995 [1]. Wang from Toronto University presents a SSL system for robot localization and navigation based on steered response power-phase transformation algorithm [2]. Hornstein implements a SSL system for humanoid robots based on two microphones [3]. Honda Co. states an open source software system HARK for robot audition, which consists of SSL, separation and speech recognition [4]. Ishi evaluates a MUSIC-based real-time sound localization system in real noisy environments, which is available for multiple sound sources [5]. The summed generalized cross correlation (GCC) method is applied to sound localization by Kwon [6]. Hu localizes the position a the mobile robot and multiple sound sources simultaneously [7].

There are three kinds of well-known methods for SSL based on microphone array, including: 1) Directional technology based on high-resolution spectral estimation [8]; 2) Controllable beamforming technology based on the biggest output power [9], [10]; 3) Technology based on time difference of arrival (TDOA) [11], [12], which needs low time consumption, and is effective for single SSL. Therefore, TDOA-based SSL method is more suitable for HRI, in which the azimuth and horizontal distance of a single speaker should be localized in real time.

TDOA is a two-step algorithm, and the localization accuracy depends on the performance of time delay estimation (TDE). Second-order statistics has been extensively studied for TDE, such as GCC algorithm with many weighting functions proposed by Knapp [11], eigenvalue decomposition [13], generalized eigenvalue decomposition [14], acoustic transfer functions ratio [15], and so on. Higher order statistics has been widely used for many applications [16]–[20], which is proposed for TDE to suppress spatially correlated Gaussian noise, since the higher order cumulant of Gaussian signal equals zero. Nikias estimates the time delay using bispectrum between two sensor signals [21]. Employing triple sensor signals, Zhang simultaneously extracts two time delays based on bispectrum [22]. However, the third-order cumulant and bispectrum of speech signal equal zero because of the zero skewness of speech signal. Consequently, third-order cumulant and bispectrum are invalid for TDE of speech data. Wang proposed a TDE method based on fourth-second-order normalized cumulant [25]. Time delay is estimated through maximizing the estimator based on

Manuscript received December 14, 2011; revised July 16, 2012; accepted October 9, 2012. This work is supported by National Natural Science Foundation of China (NSFC, 60875050, 60675025), National High Technology Research and Development Program of China (863 Program, 2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (JC201005280682A, CXC201104210010A). This paper was recommended by Associate Editor M. Carvalho.

X. Li is with the Key Laboratory of Integrated Micro-system, Shenzhen Graduate School, Peking University, Shenzhen 518055, China (e-mail: lixiaofei0111@sohu.com).

H. Liu is with the Key Laboratory of Machine Perception and Intelligence, Peking University, Shenzhen Graduate School, Shenzhen 518055 China (e-mail: hongliu@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2012.2226443

fourth-order cumulant (FOC) in time domain by Tugnait [23] and Liang [24]. However, just like the cross correlation method, for TDE, the periodicity of speech signal and non-Gaussian noise often bring about confused peak of these time domain estimators, which worsen their performance. To avoid the shortage of time domain algorithm, we derive FOC spectrum and cross spectrum in this paper. The multiplicative relationship of each frequency spectrum in FOC spectrum indicates the independence of multiple time delays from one signal to the others. FOC spectrum and cross spectrum are subsequently applied to estimate the time delay of speech sensor signals. This TDE method is immune to spatially correlated Gaussian noise and can estimate the time delay between two or multiple sensor signals. Moreover, just like SCOT weighting function in [11], a whitening function is proposed to suppress noise of each channel and weaken confused peaks that time domain estimators suffer. The method mentioned in this paragraph is a refined and expanded version of the conference proceedings paper [34].

Geometric positioning method is always used for the step of localization with given time delays, which needs the solution of hyperbolic equations that is a nonlinear optimization problem. Foy proposes Taylor Series method to locate sound source iteratively [26]. Maximum likelihood estimator [27] and least square estimator [28]–[32] are two primary localization methods. The former requires lots of experimental samples to obtain statistical properties of measurement noise. The latter solves overdetermined nonlinear equations or approximative linear equations to get the coordinates of sound source. Brandstein proposed linear intersection method to localize distant sources [33]. In this paper, time difference feature extracted by FOC spectrum and its spatial distribution are analyzed, and it can be concluded that the relationship between time difference features and sound sources is one-to-one correspondence. Moreover, the farther the distance between two sound sources is, the greater the difference of two features becomes. Based on properties of spatial distribution, a novel localization algorithm called spatial grid matching (SGM) is proposed. The best matched grid with the time difference feature of a unknown sound source is chosen as the position of sound source. By adjusting the size of grid, SGM method can handle the problem of switching between far field and near field easily and weaken some dimensions selectively. In addition, making use of one or multiple recordings from the same sound source position, valid feature detection algorithm eliminates those wrong time differences, and generates a new valid feature vector. Decision tree is suggested for reducing the number of times of template matching. These methods mentioned in this paragraph are a refined and expanded version of the conference proceedings paper [35].

The rest of this paper is organized as follows: In Section II, FOC spectrum is derived, and the TDE algorithm is also given. In Section III, a microphone array is first constructed, based on which time difference feature is analyzed, and then SGM method is presented. A SSL system for HRI is constructed in Section IV. Then, experiments and analysis are provided in Section V. Finally, the conclusion is given in Section VI.

II. FOC-Based TDE

A. FOC Spectrum

The Fourier transform of FOC will be derived in this section, in which a non-Gaussian signal $x(t)$ is considered. Unlike third-order cumulant, the skewness of $x(t)$ is unrestricted, which can be zero or nonzero. The FOC of signal $x(t)$ is defined as

$$\begin{aligned} c_{4x}(\tau_1,\tau_2,\tau_3) = & E\left\{x(t)x(t+\tau_1)x(t+\tau_2)x(t+\tau_3)\right\} \\ & - R_x(\tau_1)R_x(\tau_3-\tau_2) \\ & - R_x(\tau_2)R_x(\tau_3-\tau_1) \\ & - R_x(\tau_3)R_x(\tau_2-\tau_1). \end{aligned} \tag{1}$$

In practical applications, $x(1),\cdots,x(N)$ denote discrete samples of signal $x(t)$, and N is sample length. The biased estimation of FOC is

$$\begin{aligned} \hat{c}_{4x}(\tau_1,\tau_2,\tau_3) = & \hat{m}_{4x}(\tau_1,\tau_2,\tau_3) - \hat{R}_x(\tau_1)\hat{R}_x(\tau_3-\tau_2) \\ & -\hat{R}_x(\tau_2)\hat{R}_x(\tau_3-\tau_1) - \hat{R}_x(\tau_3)\hat{R}_x(\tau_2-\tau_1) \end{aligned} \tag{2}$$

where $\hat{m}$ and $\hat{R}$ are the biased estimation of fourth-order moment and correlation function

$$\hat{m}_{4x}(\tau_1,\tau_2,\tau_3) = \frac{1}{N}\sum_{n=1}^{N} x(n)x(n+\tau_1)x(n+\tau_2)x(n+\tau_3) \tag{3}$$

$$\hat{R}_x(\tau) = \frac{1}{N}\sum_{n=1}^{N} x(n)x(n+\tau) \tag{4}$$

FOC spectrum is defined as the $3D$ Fourier transform of $\hat{c}_{4x}$

$$P_{4x}(\omega_1,\omega_2,\omega_3) = \sum_{\tau_1}\sum_{\tau_2}\sum_{\tau_3} \hat{c}_{4x} e^{-j(\omega_1\tau_1+\omega_2\tau_2+\omega_3\tau_3)}. \tag{5}$$

It is well-known that the Fourier transform of $\hat{m}_{4x}$ is

$$M_{4x}(\omega_1,\omega_2,\omega_3) = \frac{1}{N} X(-\omega_1-\omega_2-\omega_3)X(\omega_1)X(\omega_2)X(\omega_3) \tag{6}$$

where $X(\omega)$ is the Fourier transform of $x(n)$. In addition, the Fourier transform of the second term of $\hat{c}_{4x}$ can be derived as

$$\begin{aligned} P_1 = & \sum_{\tau_1}\sum_{\tau_2}\sum_{\tau_3} \hat{\mathbf{R}}_x(\tau_1)\hat{\mathbf{R}}_x(\tau_3-\tau_2) e^{-j(\omega_1\tau_1+\omega_2\tau_2+\omega_3\tau_3)} \\ = & \sum_{\tau_1}\sum_{\tau_2}\sum_{\tau_3} \left\{\frac{1}{N}\sum_{n=1}^{N} x(n)x(n+\tau_1)\right\} \\ & \times \left\{\frac{1}{N}\sum_{n=1}^{N} x(n+\tau_2)x(n+\tau_3)\right\} e^{-j(\omega_1\tau_1+\omega_2\tau_2+\omega_3\tau_3)} \\ = & \frac{1}{N^2}\left\{\sum_{n=1}^{N} x(n)\sum_{\tau_1} x(n+\tau_1)e^{-j\omega_1\tau_1}\right\} \\ & \times \left\{\sum_{n=1}^{N}\left(\sum_{\tau_2} x(n+\tau_2)e^{-j\omega_2\tau_2}\right)\right. \end{aligned}$$

$$\times \left(\sum_{\tau_3} x(n+\tau_3) e^{-j\omega_3\tau_3} \right) \Bigg\}$$

$$= \frac{1}{N^2} \left\{ \sum_{n=1}^{N} x(n) e^{j\omega_1 n} X(\omega_1) \right\} \times \left\{ \sum_{n=1}^{N} e^{j(\omega_2+\omega_3)n} X(\omega_2) X(\omega_3) \right\}$$

$$= \frac{1}{N^2} X(-\omega_1) \Bigg(e^{-j(-\omega_2-\omega_3)(N+1)/2} \times \frac{sin\left((-\omega_2-\omega_3)N/2\right)}{sin\left((-\omega_2-\omega_3)/2\right)} \Bigg) \times X(\omega_1)X(\omega_2)X(\omega_3)$$

$$\approx \frac{1}{N} X(-\omega_1)\delta(-\omega_2-\omega_3)X(\omega_1)X(\omega_2)X(\omega_3) \tag{7}$$

where $\delta(\omega)$ is the unit sample signal, and the relationship $\approx$ is valid when N is big enough. Similarly, the Fourier transform of the third and fourth terms of $\hat{c}_{4x}$ can be obtained as P_2 and P_3

$$P_2 = \frac{1}{N} X(-\omega_2)\delta(-\omega_1-\omega_3)X(\omega_1)X(\omega_2)X(\omega_3) \tag{8}$$

$$P_3 = \frac{1}{N} X(-\omega_3)\delta(-\omega_1-\omega_2)X(\omega_1)X(\omega_2)X(\omega_3). \tag{9}$$

Finally, FOC spectrum can be represented as:

$$P_{4x}(\omega_1,\omega_2,\omega_3) = M_{4x}(\omega_1,\omega_2,\omega_3) + P_1 + P_2 + P_3 = X'(\omega_1,\omega_2,\omega_3)X(\omega_1)X(\omega_2)X(\omega_3). \tag{10}$$

where

$$X'(\omega_1,\omega_2,\omega_3) = \frac{1}{N}\{X(-\omega_1-\omega_2-\omega_3) + X(-\omega_1)\delta(-\omega_2-\omega_3) + X(-\omega_2)\delta(-\omega_1-\omega_3) + X(-\omega_3)\delta(-\omega_1-\omega_2)\} \tag{11}$$

Similarly, the cross FOC spectrum of two different signals $x(n)$ and $y(n)$ can be computed as

$$P_{xyxx}(\omega_1,\omega_2,\omega_3) = X'(\omega_1,\omega_2,\omega_3)Y(\omega_1)X(\omega_2)X(\omega_3). \tag{12}$$

Formula (10) and (12) show that the relationship of each frequency spectrum is multiplicative in FOC spectrum, which causes the additive relationship of each phase spectrum. This property indicates the independence of multiple time delays from one signal to the others.

B. TDE

Second-order statistics is widely used for estimating the time delay of two sensor signals. However, in the case of non-Gaussian sound source, higher order cumulant can suppress spatially correlated Gaussian noise. For example, as Gaussian noise, air-conditioning noise is common in practical applications. FOC can deal with those signals with zero skewness, such as speech signal, for which the third-order cumulant is invalid.

Two discrete sensor signals can be written as

$$\begin{aligned} x(n) &= s(n) + v_0(n) \\ y(n) &= s(n-D) + v_1(n) \end{aligned} \tag{13}$$

where $s(n)$ is the sound source signal, and $v(n)$ denotes spatially correlated Gaussian noise whose FOC equals zero. Signal $x(n)$ denotes reference signal that is considered to have zero time delay with source signal. Signal $s(n)$ is independent with $v(n)$. D denotes the time difference between $x(n)$ and $y(n)$. Because of the semi-invariance of cumulant, the FOC and spectrum of $x(n)$ are

$$\begin{aligned} c_{4x}(\tau_1,\tau_2,\tau_3) &= c_{4s}(\tau_1,\tau_2,\tau_3) + c_{4v_0}(\tau_1,\tau_2,\tau_3) \\ &= c_{4s}(\tau_1,\tau_2,\tau_3) \end{aligned} \tag{14a}$$

$$P_{4x}(\omega_1,\omega_2,\omega_3) = P_{4s}(\omega_1,\omega_2,\omega_3) \tag{14b}$$

where the semi-invariance causes that the cumulant of the sum of two independent signals equals the sum of two cumulants of these two signals. And FOC of Gaussian signal c_{4v_0} equals zero. Similarly, cross FOC and spectrum of these two signals can be calculated by (2) and (12) as

$$c_{xyxx}(\tau_1,\tau_2,\tau_3) = c_{4s}(\tau_1-D,\tau_2,\tau_3) \tag{15a}$$

$$\begin{aligned} P_{xyxx}(\omega_1,\omega_2,\omega_3) &= S'(\omega_1,\omega_2,\omega_3)\left\{S(\omega_1)e^{j\omega_1 D}\right\}S(\omega_2)S(\omega_3) \\ &= P_{4s}(\omega_1,\omega_2,\omega_3)e^{j\omega_1 D}. \end{aligned} \tag{15b}$$

Define function

$$I(\omega_1,\omega_2,\omega_3) = \psi\frac{P_{xyxx}(\omega_1,\omega_2,\omega_3)}{P_{4x}(\omega_1,\omega_2,\omega_3)} = e^{j\omega_1 D} \tag{16}$$

where ψ is a weighting function to whiten the spectrum and suppress noise of each channel, which is

$$\psi(\omega_1,\omega_2,\omega_3) = \frac{|P_{4x}(\omega_1,\omega_2,\omega_3)|}{|P_{xyxx}(\omega_1,\omega_2,\omega_3)|} \tag{17}$$

where $|\cdot|$ denotes the amplitude spectrum. Let ω_2 and ω_3 take arbitrary constants, such as $\omega_2 = \omega_3 = 0$. Then, function

$$T(\tau) = \sum_{\omega_1=1}^{N} I(\omega_1,\omega,\omega)e^{-j\omega_1\tau} = \delta(\tau - D) \tag{18}$$

takes the peak at $\tau = D$, which is the time delay sample of two sensor signals. Through whitening the spectrum, ψ improves the time-delay resolution and suppresses the energy of sidelobe. In theory, the values of ω_2 and ω_3 do nothing about the estimation of time delay.

Algorithm 1 shows the procedure of TDE based on FOC spectrum, where $x(n)$ and $y(n)$ are defined by (13). Symbol K in line 4 and 5 denote the number of frames. Function FT

denotes Fourier transform. Function $FOCS$ represents the calculation of FOC spectrum and cross spectrum shown as (10) and (12).

Algorithm 1 TDE

1: **Input** $x(n), y(n)$
2: **Procedure** $Enframe(x, y)$
3: Set the length of each frame as M and overlap as $M/2$
4: $x^{(k)}(n) \leftarrow x((k-1)M/2+n)k = 1, \cdots, K$
5:$y^{(k)}(n) \leftarrow y((k-1)M/2+n)k = 1, \cdots, K$
6: **Procedure** $FourierTransform(x^{(k)}, y^{(k)})$
7: $X^{(k)}(\omega) \leftarrow FT(x^{(k)}(n))$
8: $Y^{(k)}(\omega) \leftarrow FT(y^{(k)}(n))$
9: **Procedure** $FOCSpecturm(X^{(k)}, Y^{(k)})$
10: $\hat{P}_{4x}^{(k)}(\omega_1, \omega_2, \omega_3) \leftarrow FOCS(X^{(k)})$
11: $\hat{P}_{xyxx}^{(k)}(\omega_1, \omega_2, \omega_3) \leftarrow FOCS(X^{(k)}, Y^{(k)})$
12: **Procedure** $Mean(P_{4x}^{(k)}, P_{xyxx}^{(k)})$
13: $\hat{P}_{4x}(\omega_1, \omega_2, \omega_3) \leftarrow \frac{1}{K}\sum_{k=1}^{K} P_{4x}^{(k)}(\omega_1, \omega_2, \omega_3)$
14: $\hat{P}_{xyxx}(\omega_1, \omega_2, \omega_3) \leftarrow 1/K\sum_{k=1}^{K} P_{xyxx}^{(k)}(\omega_1, \omega_2, \omega_3)$
15: **Procedure** $I(\hat{P}_{4x}, \hat{P}_{xyxx})$
16: $I(\omega_1, \omega_2, \omega_3) \leftarrow \frac{|\hat{P}_{4x}(\omega_1,\omega_2,\omega_3)|}{|\hat{P}_{xyxx}(\omega_1,\omega_2,\omega_3)|}\frac{\hat{P}_{xyxx}(\omega_1,\omega_2,\omega_3)}{\hat{P}_{4x}(\omega_1,\omega_2,\omega_3)}$
17: **Procedure** $TDE(I)$
18: $\omega_2 \leftarrow 0$
19: $\omega_3 \leftarrow 0$
20: $T(\tau) \leftarrow FT(I(\omega_1, \omega_2, \omega_3))$
21: $D \leftarrow \text{argmax}_\tau\{T(\tau)\}$
22: **return** D

III. Time Difference Feature and SGM

In this section, a microphone array model is constructed in Part A, which is suitable for HRI that the azimuth and horizontal distance of a single speaker can be localized in real time. Base on this microphone array, time difference feature of a sound source and its spatial distribution are presented in Part B. Then, a novel localization method based on the spatial properties of time difference feature, termed SGM, is proposed in Part C. Finally, valid feature detection algorithm is proposed in Part D.

A. Microphone Array Model

A microphone array for HRI has been constructed in [35], [36]. The following issues should be considered for designing the microphone array:

1) The SSL task to be solved.
2) The cost of equipment.
3) Computational complexity.
4) The shape of platform where microphone array is fixed.

The SSL system is used for HRI, in which azimuth and horizontal distance of speaker should be localized, and it is unimportant for HRI to localize the vertical height of speaker in most cases. Localizing the azimuth and the horizontal distance of sound source in $3D$ space needs more than two microphones. For omnidirectional azimuth localization, microphone array should be approximately isotropic in horizontal plane. The aperture of microphone array should be big enough for a suitable localization resolution. In order to reduce the cost of equipment and computational complexity, the less microphones are, the better. Microphone array is installed on a horizontal plane of a robot with a specific height, so the shape of the robot should be taken into account. A microphone array with a cruciform plane is shown in Fig. 1, which includes four microphones.

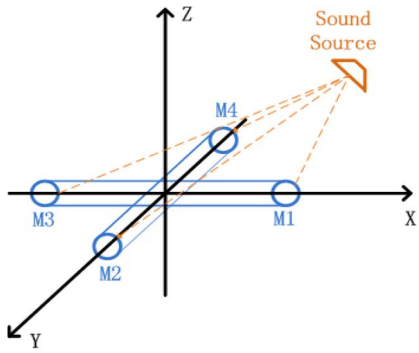


Fig. 1. Microphone array model for HRI.

B. Spatial Distribution of Time Difference Feature

The symbol τ_{mn} denotes time difference of two signals recorded by m-microphone and n-microphone. The number of microphones used in the SSL system is M. Then, $M(M-1)/2$ pairs of time differences can be obtained, and only $M-1$ pairs of them are independent mutually. However, the measurement of time difference always exists deviation, even mistake, and more time differences are, more robust. Thence, all time differences can be combined into a feature vector as

$$\boldsymbol{\tau} = [\tau_{12}, \tau_{13}, \ldots, \tau_{mn}, \ldots, \tau_{(M-1)M}]. \tag{19}$$

The coordinates of sound source S_i and microphone R_m are defined as $\boldsymbol{s_i} = [x_{s_i}, y_{s_i}, z_{s_i}] = [d_i cos(\alpha_i), d_i sin(\alpha_i), h_i]$ and $\boldsymbol{r_m} = [x_{r_m}, y_{r_m}, z_{r_m}]$, where α_i, d_i, and h_i are the azimuth, horizontal distance, and height of sound source S_i, respectively.

Then, the feature of sound source S_i can be computed as

$$\begin{aligned} d_{S_iR_m} &= |\boldsymbol{s_i} - \boldsymbol{r_m}| \quad (\text{and } d_{S_iR_n} = |\boldsymbol{s_i} - \boldsymbol{r_n}|) \\ \tau_{S_iR_{mn}} &= (d_{S_iR_m} - d_{S_iR_n})/c \\ \boldsymbol{\tau_{S_i}} &= [\tau_{S_iR_{12}}, \tau_{S_iR_{13}}, \ldots, \tau_{S_iR_{mn}}, \ldots, \tau_{S_iR_{(M-1)M}}] \end{aligned} \tag{20}$$

where $d_{S_iR_m}$ represents the distance between S_i and R_m, and c denotes the speed of sound. The difference of features of two sound sources S_i and S_j is

$$\tau_d = |\boldsymbol{\tau_{S_i}} - \boldsymbol{\tau_{S_j}}| \tag{21}$$

where Euclidean distance is used.

Corresponding to the microphones array mentioned in the last part of this section, the following properties of τ_d can be obtained as:

1) If $\tau_d = 0$ i.e., $\boldsymbol{\tau_{S_i}} = \boldsymbol{\tau_{S_j}}$, then $\alpha_i = \alpha_j$, $d_i = d_j$ and $h_i = h_j$ or $-h_j$.

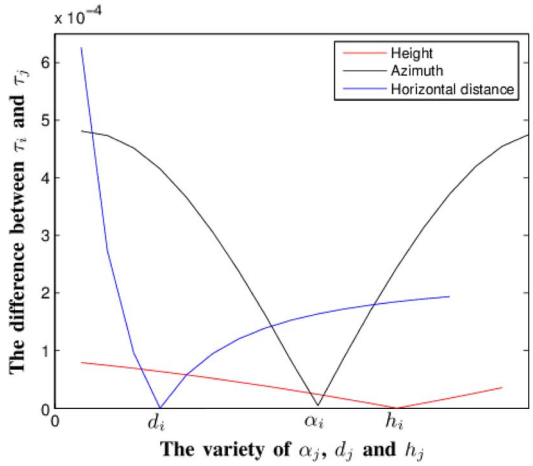


Fig. 2. Relationship between $\boldsymbol{\tau_{S_i}}$ and $\boldsymbol{\tau_{S_j}}$.

2) The relationship between τ_d and $|\boldsymbol{s_i} - \boldsymbol{s_j}|$:

$$\tau_d \propto \begin{cases} |\alpha_i - \alpha_j|, & \text{for } |\alpha_i - \alpha_j| \geq 180^\circ \\ 360 - |\alpha_i - \alpha_j|, & \text{for } |\alpha_i - \alpha_j| \geq 180^\circ \end{cases} \tag{22}$$

$$\tau_d \propto \begin{cases} |d_i - d_j|, & \text{for } d_i \geq d_j \\ d_i - d_j, & \text{for } d_i \geq d_j \end{cases} \tag{23}$$

$$\tau_d \propto \begin{cases} |h_i - h_j|, & \text{for } 0 \geq h_i \geq h_j \\ h_i - h_j, & \text{for } h_i \geq h_j > 0. \end{cases} \tag{24}$$

Property 1) means a specific time difference feature corresponds to two sound sources that are symmetric with respect to the plane of microphone array; however, the negative h_j is discarded in applications of HRI. In other words, therefore, the relationship between time difference feature and sound source is one-to-one correspondence. Property 2) reveals the positive relationship between the difference of features and the difference of azimuth, horizontal distance, and height, respectively. It can be concluded that the farther the distance of two sound sources is, the greater the difference of two features of these two sound sources becomes. This relationship is shown in Fig. 2, where α_i, d_i and h_i are selected randomly, and it can be intuitively confirmed that they are representative.

C. SGM Localization Algorithm

As mentioned in Part B of this section, one sound source corresponds to one time difference feature, vice versa. Moreover, two adjacent sound sources own a pair of similar features. The horizontal plane can be divided into many grids with a certain size. The partition of horizontal space is shown in Fig. 3. Those sound sources in the same grid are adjacent, whose features are similar. On the contrary, features of two different grids that are far from each other will differ largely.

Using Monte Carlo method, a Gaussian mixture model (GMM) will be constructed as the template for each grid based on time difference feature. For an arbitrary grid, its azimuth distributes from α_1 to α_2, its horizontal distance distributes from d_1 to d_2, and its height distributes from h_1 to h_2. The GMM of this grid can be trained offline as Algorithm 2, where N denotes the number of random sound sources in a grid, which

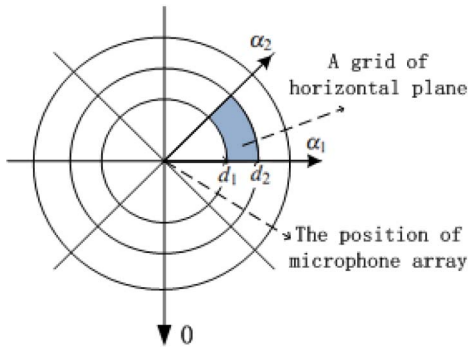


Fig. 3. Partition of horizontal space.

should be big enough to guarantee that random sound sources can cover the whole grid in probability.

Algorithm 2 Training GMM

1: Initializing a GMM
2: **for** $n = 1$ to N **do**
3: Generating an azimuth α randomly, $\alpha \sim U(\alpha_1, \alpha_2)$.
4: Generating a horizontal distance d randomly, $d \sim U(d_1, d_2)$.
5: Generating a height h randomly, $h \sim U(h_1, h_2)$.
6: The coordinate of this random sound source $[x_n, y_n, z_n] = [dcos(\alpha), dsin(\alpha), h]$.
7: Time difference feature τ_n can be computed using (20).
8: **end for**
9: Training GMM using $\tau_1, \tau_2, \cdots, \tau_N$.

Then, the problem of localization can be changed to find which grid does the sound source distribute in, with time difference feature τ given. This problem can be described as

$$\begin{aligned} G_s &= \arg\max_G P(G|\tau) \\ &= \arg\max_G \frac{P(\tau|G)P(G)}{P(\tau)} \\ &\propto \arg\max_G P(\tau|G) \end{aligned} \tag{25}$$

where G denotes a grid, and G_s represents the grid of a sound source. This equation indicates that the solution of localization is the grid which has the greatest likelihood value. All of the likelihood values between the GMM of each grid and the time difference feature of an unknown sound source should be computed, then the greatest corresponds to the sound source grid.

How to determine the size of a grid? Unlike traditional clustering algorithm, FOC-based time difference features uniformly distribute in the whole space ignoring the effect of feature resolution. Therefore, each grid must have identical size in the sense of feature resolution. Obviously, there is no upper bound for the size of a grid. Theoretically, any size is correct. However, measurement error of time difference feature is inevitable. The feature difference of two opposite boundary of a grid indicates the sensitivity of this grid to measurement error, and the smaller the difference is, the more sensitive. In

other words, the minimal size of this grid depends on the level of average measurement. If the size is smaller than the valid minimal size, localization results will deviate from real value. Hence, the feature difference of two opposite boundary should be greater than the average measurement error. In particular, the sensitivities are different between each dimension, furthermore, between different areas of one dimension.

As well-known, geometric positioning method solves localization problem from time difference to location inversely. The solution of inverse problem has many problems such as nonlinear and high computational complexity. In comparison, SGM method avoids the solution of inverse problem. In addition, it weakens those dimensions that we are not interested in, such as the height dimension. Furthermore, it does not need the assumption of far field or near field. This method is more powerful and efficient than geometric positioning method in some way.

D. Valid Feature Detection

The wrong time difference will deteriorate the accuracy of localization, which should be removed. Furthermore, in applications of HRI in noisy environment, if a wrong localization takes place, it is reasonable that the speaker calls the robot again. Valid feature can be detected from one recording or multiple recordings generated at the same sound source position.

Here, τ_{imn} denotes the time difference of the $i$$th$ recording between $m-$microphone and $n-$microphone. Theoretically, the following equation can be established:

$$\tau_{imn} = \tau_{jmk} + \tau_{jkn} \quad k \in [1, M], j \in [1, I] \tag{26}$$

$$\tau_{imn} = \tau_{jmn} \quad j \in [1, I] \tag{27}$$

where M denotes the number of microphones and j denotes the number of recordings.

Set Γ_{imn} is defined as

$$\begin{aligned} e_{ijmnk} &= |\tau_{jmk} + \tau_{jkn} - \tau_{imn}| \quad k \in [1, M], j \in [1, I] \\ e_{ijmn} &= |\tau_{imn} - \tau_{jmn}| \quad j \in [1, I] \\ \mathbf{\Gamma_{imn}} &= (e_{ijmnk} | e_{ijmnk} > th) \cup (e_{ijmn} | e_{ijmn} > th) \end{aligned} \tag{28}$$

where th is a threshold empirically determined by the average measurement error of time delay. If e_{ijmnk} is greater than th, (26) will have discrepancy. Similarly, if e_{ijmn} is greater than th, (27) will have discrepancy. The element number of set $\mathbf{\Gamma_{imn}}$ is q_{imn}. Then, the validity of τ_{imn} can be defined as

$$\tau_{imn} \text{ is } \begin{cases} \text{valid}, & \text{for } q_{imn} \geq TH \\ \text{invalid}, & \text{for } q_{imn} \geq TH. \end{cases} \tag{29}$$

where TH is a threshold decided by the number of microphones M and the number of recordings I, and is set to $M \times I/2$. If the number of unmatched pairs defined by (28) is greater than TH, the time difference τ_{imn} is considered to be invalid.

Finally, computing the mean value of the valid time differences of multiple recordings, these time differences correspond to the same microphone pair, which is shown as

$$\tau'_{mn} = \frac{1}{J} \sum_{valid \tau_{imn}} \tau_{imn} \tag{30}$$

where J represents the number of valid τ_{imn}. If J equals 0, τ_{mn} is invalid, and it will be removed. All valid τ'_{mn} are combined into a new feature vector τ'. The procedure of detecting the valid feature is shown in Algorithm 3. The input of this algorithm is time difference features of multiple recordings generated at the same position, and the output is the combined valid feature.

Algorithm 3 Valid Feature Detection

Input $\tau_{imn} i = 1, \cdots, I; m, n = 1, \cdots, M$
$q_{imn} \leftarrow 0$
for $j = 1$ to I **do**
  for $k = 1$ to M **do**
    if $|\tau_{imn} - \tau_{jmn}| > th$ **then**
      $q_{imn} \leftarrow q_{imn} + 1$
    end if
    if $|\tau_{jmk} + \tau_{jkn} - \tau_{imn}| > th$ **then**
      $q_{imn} \leftarrow q_{imn} + 1$
    end if
  end for
end for
$J \leftarrow 0$
for $i = 1$ to I **do**
  if $q_{imn} \geq TH$ **then**
    τ_{imn} is valid
    $J \leftarrow J + 1$
  else
    τ_{imn} is invalid
  end if
end for
if $J \neq 0$ **then**
  $\tau'_{mn} \leftarrow 1/J \sum_{valid \tau_{imn}} \tau_{imn}$
else
  τ_{mn} will be removed
end if
return τ'_{mn}

IV. SSL System for HRI

In this section, a speaker localization system in applications of HRI is constructed. Some issues of SSL for HRI are given in Part A. Then, the size of spatial grid is determined by the average measurement error of time difference features in Part B. In Part C, decision tree is proposed to reduce the number of times of template matching.

A. SSL for HRI

A microphone array with a cruciform plane has been shown in Fig. 1. In HRI, the microphone array is placed on a horizontal plane of a mobile robot with a certain height, and the distance between two adjacent microphones is set to 0.4 m. The position of sound source relative to the mobile robot should be localized, including azimuth and horizontal distance of a speaker, while the vertical height of a speaker is not taken into account. The azimuth should be localized as accurately as possible,

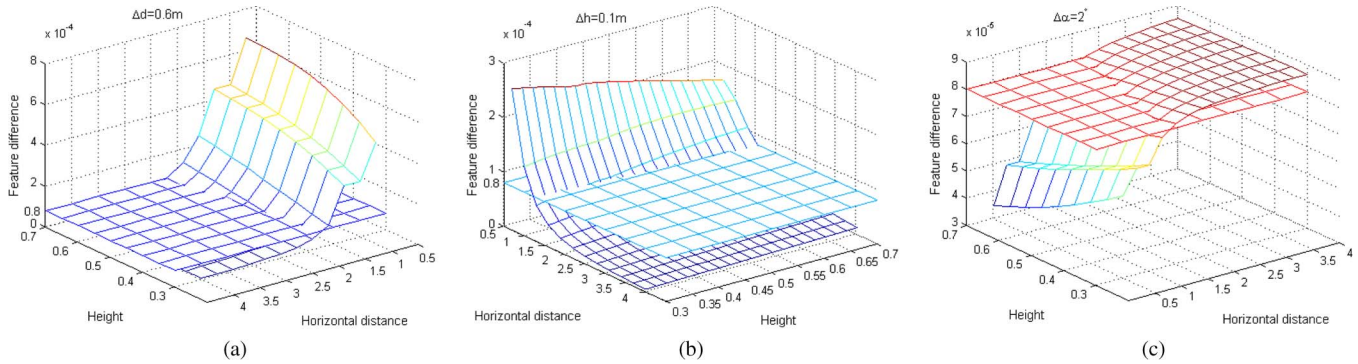


Fig. 4. Spatial distribution of feature difference for horizontal distance, height and azimuth. (a) Horizontal distance. (b) Height. (c) Azimuth

and the dangerous area and safe area of the robot should be distinguished well, where the boundary between dangerous area and safe area is set to 1.5 m roughly in our system.

B. Size of Spatial Grid

In order to determine the minimal size of the spatial grid, the average measurement error of time difference feature should be taken into account. Considering the microphone array model mentioned in Part A of this section, here, the average measurement error is surveyed as $\varepsilon = 0.8 \times 10^{-4} s$ based on lots of sensor data. Fig. 4 shows the feature difference of two opposite boundary of a grid with a specific azimuth, horizontal distance, and height size, respectively. Fig. 4(a) shows the feature difference in the whole horizontal distance and height space with a horizontal distance size $\Delta d = 0.6m$ as an example. Obviously, it is approximately isotropic in azimuth dimension, so an arbitrary azimuth is representative, such as $\alpha = 30^\circ$ here, the same below. In each area, the plane with feature difference ε can judge whether it is reasonable that the size of horizontal distance is set to Δd or not. The size of horizontal distance of those areas below the plane cannot be equal to or less than Δd. Fig. 4(b) shows the feature difference with a height size $\Delta h = 0.1m$. Fig. 4(c) shows the feature difference with an azimuth size $\Delta\alpha = 2^\circ$. Similarly, the size of height and azimuth of those areas below the plane cannot be equal to or less than Δh and $\Delta\alpha$. It can be seen that horizontal distance and height are more sensitive in far horizontal distance area, azimuth is more sensitive in near horizontal distance area, and the variety of sensitivity is small between different heights. A reasonable size of a grid in each area can be determined by many feature difference figure with different size of each dimension. For example, considering horizontal distance, the feature difference with different horizontal distance size is shown in Fig. 5, where azimuth and height are set to 30° and 0.5 m, respectively.

In addition, for HRI applications, the azimuth size of spatial grid should be set as small as possible, and the horizontal distance size of spatial grid should be determined based on the dangerous area of the mobile robot. It is isotropic in azimuth dimension, which means the feature resolution has no difference in azimuth space. Hence, the azimuth of each grid should be identical. Considering the aperture of microphone array, the boundary between far-field and near-field is in the range of 1.5 $\sim$ 2.5 m. Different from the azimuth dimension, the

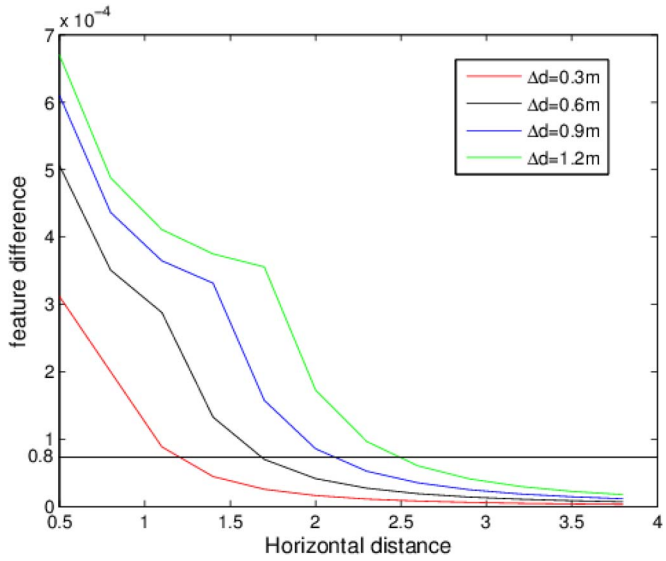


Fig. 5. Feature difference of horizontal distance.

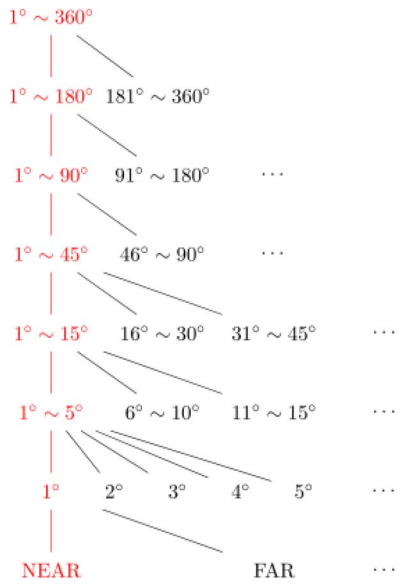


Fig. 6. Decision tree for SGM method.

greater horizontal distance is, the smaller the feature resolution becomes. Thence, the grid that has greater horizontal distance should be bigger, and vice versa. In the far-field area, because

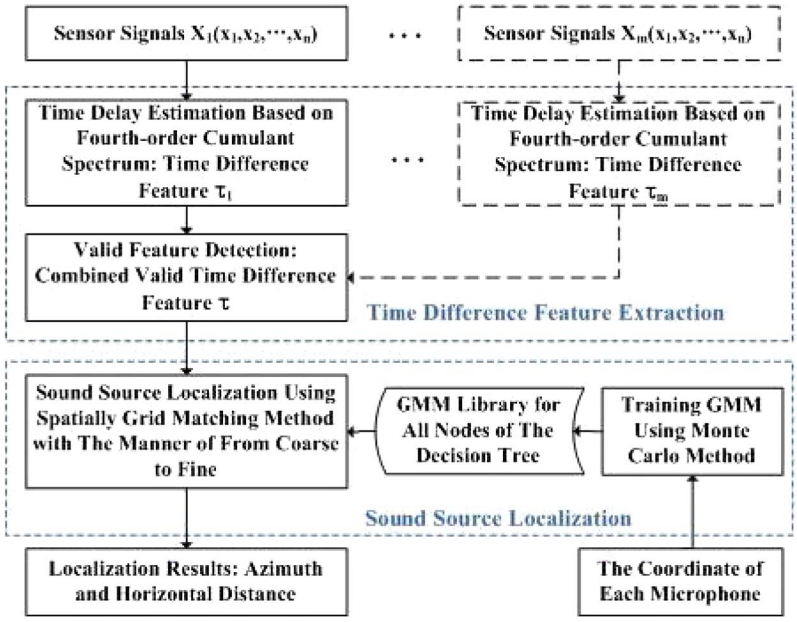


Fig. 7. Flowchart of SSL algorithm.

of the planar wavefronts, time differences of those sources with different horizontal distance will be almost the same. Therefore, the far-field area should be in the same gird. We are not very interested in the height of the sound source, which is unimportant for most applications. In summary, the minimal size of azimuth is set to 1°. Taking into account the body size of a robot, the horizontal distance is divided into two parts: *NEAR* $0.5 \sim 1.5$ m and FAR $> 1.5m$, which correspond to dangerous region and safe region for human, respectively. The whole height space is treated as one part.

C. Decision Tree for SGM

The computational complexity would be very high if too many grids are divided, since all of the grids must be matched with the time difference feature of the unknown sound source. As mentioned above, the azimuth is divided into 360 parts, and the horizontal distance is divided into two parts. A total of $360 \times 2 = 720$ likelihood values should be computed. In contrast, as shown in Fig. 6, decision tree is used. The number of likelihood values computed becomes $2 + 2 + 2 + 3 + 3 + 5 + 2 = 19$. First, a GMM should be trained for each node of the tree offline, which is highly time consuming. For those nodes of the first seven layers, the horizontal distance is treated as one part. In the $8th$ layer, the azimuth size of each grid is 1°. Then, in the stage of localization, the likelihood value is computed layer by layer from the root of the tree to the leaf, just like the trajectory from the root to the leaf of *NEAR* as shown in Fig. 6. In each layer, all children of the current node are matched with the time difference feature of an unknown sound source, then the sound source will be located to the subgrid whose likelihood value is the greatest. The time consumption of localization step would greatly benefit from this localization manner of from coarse to fine.

Overall, the flowchart of SSL algorithm mentioned above can be concluded in Fig. 7. Valid feature detection method uses one time difference feature τ_1 or multiple features $\tau_1, \cdots, \tau_m$. The GMM library is trained offline when the coordinate of each microphone is given. This library includes those templates corresponding to all of the nodes of decision tree.

V. Experiments and Analysis

In order to evaluate the effectiveness of FOC-based TDE algorithm and SGM algorithm, lots of experiments of TDE and SSL for HRI are carried out. First, experiments of TDE are presented in Part A, in which speech signals with spatially correlated Gaussian noise are tested. Then, in Part B, SSL experiments for HRI are given, which are achieved in real noise environments.

A. TDE

In each experiment, two sensor speech signals collected in office environment are tested. Speech signals are collected by two microphones with the sampling rate 44.1 kHz. The time delay between the first signal and other is 50 sampling points. Spatially correlated Gaussian noise are added into speech signals with certain SNR. Time delay of Gaussian noise between the first signal and other is 30 sampling points. In addition, another kinds of environmental noise, namely non-Gaussian noise or spatially uncorrelated noise, also exist in office environment, such as computer fans noise. The average signal to environmental noise ratio is detected as 12 dB. For convenience, SNR denotes the signal to spatially correlated Gaussian noise ratio below. In total, 600 sets of different speech signals are tested, and Gaussian noise are also different in each set. The duration of each speech signal is about $1s$, which are segmented with length $93ms$ and 50% overlap. For comparison, GCC method with PHAT weighting function in [11] and time domain FOC method in [24] are also tested, which are, respectively, named

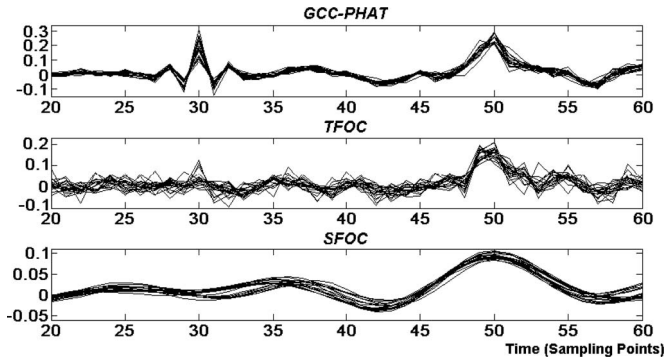


Fig. 8. 20 overlapped time delay plots of *GCC*, *TFOC*, and *SFOC* versus time (sampling point) with SNR 5 dB.

TABLE I
TIME DELAY ESTIMATION CORRECT RATE

TDE METHOD	Correct rate (%) No nosie	$10dB$	$5dB$	$0dB$	$-5dB$
GCC-PHAT	100.0	83.83	44.33	0	0
TFOC	100.0	100.0	100.0	91.32	83.24
SFOC	100.0	100.0	100.0	91.00	83.60

TABLE II
SELECTION OF MIXTURE NUMBER

GMM MixNum	Azimuth Correct Rate (%) 5°	10°	15°
1	87.12	89.33	92.79
2	87.76	90.23	92.68
4	89.94	90.97	93.40
6	88.91	90.91	93.11
8	88.77	90.29	93.24

as *GCC-PHAT* and *TFOC* here. Our spectrum domain FOC method is named as *SFOC* here.

Fig. 8 shows the 20 overlapped plots of *GCC-PHAT*, *TFOC* and *SFOC* versus time (sampling point) with SNR 5dB. These 20 plots is randomly selected from 600 time delays. Obviously, *GCC-PHAT* has a wrong peak at the time delay of Gaussian noise with sampling point 30. *TFOC* and *SFOC* have correct peak at the $50th$ sampling point. However, *TFOC* has higher variance and confused peaks at neighboring sampling point of correct peak, which bring about bigger estimation error.

Wrong estimations brought by Gaussian noise are always around the peak of noise time delay, thus estimation correct rate is tested: If estimation result is around the time delay of speech signal, it is correct. Otherwise, it is wrong. Tables I and II shows the estimation correct rate versus SNR based on 600 sets of signals. It can be seen that the performance of *GCC-PHAT* dramatically declines as the intensity of Gaussian noise increases. *TFOC* and our method have similar correct rate. Because of the limited sample number, the probability distribution of noise deviates from Gaussian distribution more or less [37], which leads FOC of noise unequal to zero. Therefore, the performance of *TFOC*, *SFOC* will also slightly decline when SNR is less than 0 dB. The good performance shows that FOC-based method can suppress spatially correlated Gaussian noise effectively.

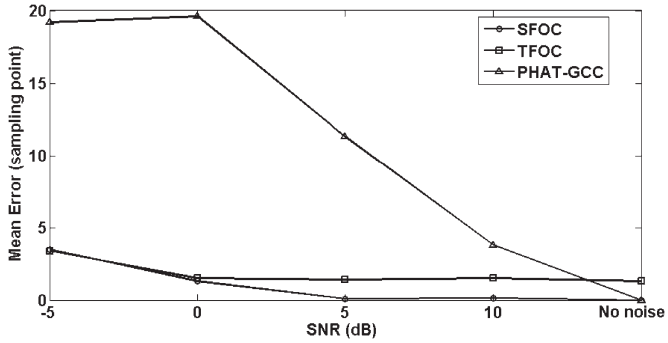


Fig. 9. Estimation mean error of *TFOC* and *SFOC*.

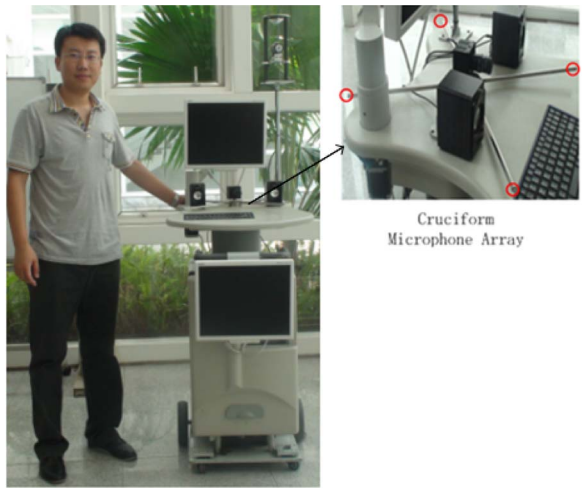


Fig. 10. Mobile robot and microphone array.

The estimation mean error (ME) is tested: the average error between estimation value and real value, which can be computed as: $ME = (\sum^{N} |D^\circ - D|)/N$, here D', D, and N denote the estimation value, real value, and estimation times, respectively. In this experiment, $D = 50$ and $N = 600$. Fig. 9 shows the estimation ME of three methods versus SNR. It is obvious that *SFOC* has smaller estimation error than *TFOC* when they are less affected by Gaussian noise, such as 5 dB, 10 dB and "no noise." The estimation error of *GCC-PHAT* is great because of its poor estimation correct rate. However, if Gaussian noise do not exist, as "No noise," *GCC-PHAT* gets smaller estimation error than *TFOC*.

B. Speech SSL for HRI

1) Configuration of Experimental Environments: A mobile robot and a microphone array designed by our lab are shown in Fig. 10. The microphone array is placed on the shoulder of the robot with a height of 1 m. The mobile robot works in a hall, semi-door environment, with a size of 8 m $\times$ 8 m. There are various kinds of noise in this acoustic environment, such as mechanical noise and electrical noise of the mobile robot, air-conditioning noise, computer fan noise, outdoor noise, etc. Four microphones are used for the sampling of speech sound, and the sampling rate is $44.1kHz$.

2) SSL for HRI: In this experiment, the mobile robot is placed in the center of the hall, and the sound source to be

TABLE III
AZIMUTH LOCALIZATION CORRECT RATE. TIME DIFFERENCE FEATURE CALCULATED BY *GCC-PHAT, TFOC*, AND *SFOC* ARE TESTED WITH DIFFERENT SIGNAL TO SPATIALLY CORRELATED GAUSSIAN NOISE RATES

Experimental Condition		Azimuth Correct Rate (%)								
		Near-Field			*Far-Field*			*MEAN*		
		5°	10°	15°	5°	10°	15°	5°	10°	15°
No noise	*GCC-PHAT*	94.02	95.73	99.01	84.38	86.26	87.71	89.20	91.00	93.36
	TFOC	92.75	95.43	98.20	80.77	82.34	87.41	86.76	88.89	92.81
	SFOC	94.54	95.98	98.72	85.33	85.95	88.07	89.94	90.97	93.40
$25dB$	*GCC-PHAT*	78.43	82.19	85.77	69.48	73.65	76.07	73.96	77.92	80.92
	TFOC	92.31	94.67	98.92	81.19	83.06	86.87	86.75	88.87	92.90
	SFOC	94.33	95.50	99.01	84.62	85.31	87.75	89.48	90.41	93.38
$0dB$	*GCC-PHAT*	37.15	44.79	50.20	24.14	29.51	36.44	30.65	37.15	43.32
	TFOC	85.35	87.21	92.90	73.88	76.21	80.65	83.06	84.56	88.01
	SFOC	87.64	89.31	93.10	80.76	81.90	83.12	84.20	85.61	88.11

localized is placed on another 120 positions in the vicinity of microphones array, with the azimuth of every 15° and the distance of 1, 1.5, 2, 3, 4 m from the center. Where 1 and 1.5 m are in the near-field area, 2, 3, and 4 m are in the far-field area. These sensor data are represented as *Near-Field* and *Far-Field* below, respectively. In real noise environments, owing to the attenuation of source signal, the greater the distance between the sound source and the microphone array is, the lower overall SNR of sensor signals will be. The sensor signal to environments noise ratios of *Near-Field* and *Far-Field* are detected as 13 dB and 10 dB, respectively. In each position, 21 groups of speech data emitted from three speakers are recorded, and the height of these speakers are different. The content of speech is Chinese words "dingwei," "pengpeng," and "guolai" which mean "locate," "Robot's name," and "come here," respectively. In total, $120 \times 21 = 2520$ sets of data are tested for each experiment.

The localization accuracy of azimuth is 1°, and the azimuth localization correct rate has three kinds of situations, the difference between localization result and real value is less than 5°, 10°, and 15°, respectively. The horizontal distance is divided into two parts: *NEAR* $0.5 \sim 1.5$ m and $FAR > 1.5$ m.

The number of mixture components of GMM affects the localization performance. The azimuth localization performance with varying mixture number is shown in Table II. In this experiment, the basic SGM method and time delay feature based on FOC spectrum are used, and those sensor data of *Near-Field* and *Far-Field* are all tested. Obviously, the best performance is obtained with four mixtures. As a result, the number of mixture of GMM is set to 4.

In theory, FOC can completely suppress spatially correlated Gaussian noise. In the scene of HRI, there are two kinds of noise in the hall: spatially correlated Gaussian noise and others. The former includes air-conditioning noise, and so on, and the latter consists of all other kinds of environment noises. A loudspeaker is used as the point Gaussian noise source recorded from an air-conditioning, with azimuth 7.5°, horizontal distance 1.5 m, and height 1.5 m, which is used to generate spatially correlated Gaussian noise, and it is convenient to control the noise intensity. In this experiment, three kinds of noise intensity are tested, namely *No noise*, weak noise with average SNR 25 dB, and strong noise with average SNR 0 dB. Table III shows the azimuth localization results. For comparison, *GCC-PHAT* method and *TFOC* method are also tested. It can be seen that the localization performance of *Far-Field* is worse than *Near-Field* owing to the larger environmental noise. For all the experimental conditions, *TFOC* gets greater localization error than *SFOC* due to its larger TDE error, particularly in *Far-Field* area. Which can be observed from the results of 5° and 10°. Taking into account the results of *No noise*, the correct rates of *GCC-PHAT* and *SFOC* are about the same. When the signal to spatially correlated Gaussian noise rate is 25 dB, the localization performance of *GCC-PHAT* method declines dramatically. On the contrary, the performances of *TFOC* and *SFOC* method are almost equivalent with the situation of *No noise*. It is verified that FOC-based method is more effective for TDE than *GCC-PHAT* method when spatially correlated Gaussian noise exists. As mentioned in Part A, the probability distribution of noise deviates from Gaussian distribution more or less. Consequently, the localization performances of *TFOC* and *SFOC* methods also slightly decline as Gaussian noise increases, which can be seen from the experimental results of 0 dB. SGM localization method is used in this experiment.

By adjusting the size of grid, SGM method can handle the problem of switching between far field and near field easily and can weaken some dimensions selectively. For comparison, the nonlinear least squares error criteria J_{DOA} presented in [33] is tested for DOA estimation, which estimates the direction of arrival of sound source. In this experiment, for J_{DOA}, all the six time differences are used for increasing the robustness of DOA, and the variance of TDOA error of each sensor pair are assumed to be identical. In addition, the closed-form spherical LS source localization method presented in [32] is also tested, which can estimate the location of sound source. In our system, only three time differences of four microphones are used for localizing a source in 3-D space. Therefore, the estimator in [32] is equivalent to spherical intersection estimator in [29]. Table IV shows the azimuth localization correct rate of J_{DOA}, SX and SGM method. The loudspeaker that generates spatially correlated Gaussian noise is turned off in this experiment. It is obvious that J_{DOA} and SX method have greater localization error than SGM, which can be seen from the correct rate of 5°. The speakers have different height with the horizontal plane of microphone array. Therefore, for the 3-D sound source, the localization error will emerge when the 2-D planar array is used. However, it can be observed from the results of 10° that

TABLE IV
AZIMUTH LOCALIZATION CORRECT RATE OF J_{DOA}, SX AND SGM METHOD

Experimental Condition	Azimuth Correct Rate (%)								
	Near-Field			*Far-Field*			*MEAN*		
	5°	10°	15°	5°	10°	15°	5°	10°	15°
J_{DOA}	50.12	89.48	95.24	68.06	83.85	89.94	59.09	86.67	92.59
SX	77.18	97.02	97.22	60.32	82.34	86.51	68.75	89.68	91.87
SGM	94.54	95.98	98.72	85.33	85.95	88.07	89.94	90.97	93.40

TABLE V
HORIZONTAL DISTANCE LOCALIZATION RATE

Experimental Condition		Azimuth Correct Rate (%)								
		Near-Field			*Far-Field*			*MEAN*		
		5°	10°	15°	5°	10°	15°	5°	10°	15°
No noise	*NO VFD*	94.54	95.98	98.72	85.33	85.95	88.07	89.94	90.97	93.40
	1 recording	95.46	96.27	99.53	93.38	94.72	97.60	94.42	95.50	98.57
	2 recordings	95.69	96.25	99.68	94.05	94.91	97.93	94.87	95.58	98.81
	3 recordings	96.00	97.01	99.79	94.35	95.18	97.89	95.18	96.10	98.84
0 *dB*	*NO VFD*	87.64	89.31	93.10	80.76	81.90	83.12	84.20	85.61	88.11
	1 recording	94.25	95.36	99.10	93.52	95.17	97.24	93.89	95.27	98.17
	2 recordings	95.13	95.88	99.33	94.68	95.73	97.99	94.91	95.81	98.66
	3 recordings	95.02	96.32	99.27	94.71	95.92	97.68	94.87	96.12	98.48

these errors is less than 10° in most cases. On the contrary, SGM method can weaken the height dimension selectively. The performance of J_{DOA} for *Near-Field* is worse than SGM. The wavefront of near-field source deviates from the planer wavefront. The performance for *Far-Field* of these two methods are similar. SX method has good performance for *Near-Field* with the localization error 10° and 15°. However, its performance is worse than SGM, which is probably due to the fact that the microphone is too limited and SX method is more sensitive to time delay error in the far-field area. SGM method can switch between far-field and near-field naturally.

For two different sound sources, azimuth is the dominant factor, which means the difference of their time difference features is mainly determined by the difference of their azimuth. On the contrary, horizontal distance is the secondary factor, which is more sensitive to the measurement error of time difference feature. Thence, the horizontal distance localization result depends on the azimuth localization result to a certain extent. The horizontal distance localization results of SGM method are shown in Table V. It can be found that the localization performance of horizontal distance is approximately proportional to the performance of azimuth; however, it is more lower because of the high sensitiveness to the measurement error. The grid of *Near* with the horizontal distance $0.5 \sim 1.5$ m, which is equivalent to the statement that the average localization error of this grid is 0.75 m. For a comparison, the horizontal distance error for *Near-Field* of SX method is 0.54 m. In addition, the average horizontal distance error for *Far-Field* of SX method is greater than 2.0 m, which has little practical significance.

Decision tree reduces the matching times between the GMM of each grid and the time difference feature dramatically. As mentioned in Section IV, the matching times are reduced from 720 to 19 for each localization, and the matching stage takes $40ms$ reduced from $420ms$ in our experiments.

Removing those wrong time differences can improve the performance of SGM method. In particular, when a wrong localization takes place, it is reasonable that the speaker calls the robot again. Therefore, multiple features generated by the same sound source position can be used for valid feature detection. Table III shows that the performance of *SFOC* is about the same between *No noise* and 25 dB. Therefore, only the sensor signals of *No noise* and 0 dB are tested in this experiment. Azimuth localization results *SFOC* using valid feature detection method are shown in Table VI and VII, where valid feature detection method is named as *VFD*, and 1, 2, and 3 recordings are tested, respectively. It can be seen that *VFD* method improves the azimuth localization performance markedly, and the more recordings are, the better the performance will be. However, too many recordings are not available in applications of HRI, generally, no more than three recordings are used. It is effective particularly for lower SNR environment. If valid features are used, the performance of azimuth localization with SNR 0 dB is about the same as the case of *No noise*. Table VII shows the horizontal distance localization results using valid feature. Once again, *VFD* method improves the performance effectively. However, unlike azimuth localization results, the improvement

TABLE VI
AZIMUTH LOCALIZATION RESULTS USING VALID FEATURE DETECTION METHOD. 1, 2, AND 3 RECORDINGS GENERATED BY THE SAME SOUND SOURCE AT THE SAME POSITION ARE USED FOR VALID FEATURE DETECTION

Experimental Condition		Horizontal distance Correct Rate (%)		
		Near-Field	*Far-Field*	*MEAN*
No noise	*GCC-PHAT*	90.23	81.81	86.02
	TFOC	90.54	76.48	83.49
	SFOC	91.09	81.11	86.10
$25dB$	*GCC-PHAT*	70.44	59.33	64.89
	TFOC	88.76	76.69	82.73
	SFOC	90.67	83.32	87.00
$0dB$	*GCC-PHAT*	30.61	19.25	24.93
	TFOC	81.37	62.49	71.93
	SFOC	83.10	76.11	79.60

TABLE VII
HORIZONTAL DISTANCE LOCALIZATION RATE USING VFD METHOD

Experimental Condition		Horizontal distance Correct Rate (%) *Near-Field*	*Far-Field*	*MEAN*
No noise	*NO VFD*	91.09	81.11	86.10
	1 recording	93.11	84.19	88.65
	2 recordings	93.23	84.08	88.66
	3 recordings	92.87	84.10	88.49
0 dB	*NO VFD*	83.10	76.11	79.60
	1 recording	89.25	81.62	85.44
	2 recordings	88.78	81.55	85.17
	3 recordings	90.15	82.17	86.16

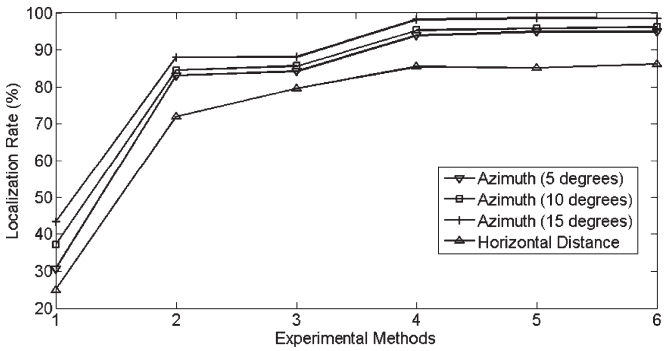


Fig. 11. Localization results with signal to spatially correlated Gaussian noise ratio 0 dB. From 1 to 6 of the abscissa represent *GCC-PHAT, TFOC, SFOC* without *VFD, SFOC* with one recording, two recordings, and three recordings VFD. Mean localization rate is used.

caused by the increase of recordings is not apparent because of the considerable localization error. The threshold th mentioned in Section III is set to $0.23ms$ empirically, and the threshold TH is set to $M \times I/2 = 2I$, here the number of microphones is $M = 4$. Generally speaking, more than 94.87% average azimuth correct rate with error less than 5° and more than 85.17% average horizontal distance correct rate can be obtained, respectively. It can effectively estimate the bearing of sound source and judge whether the sound source is standing in dangerous area.

It is easy to observe localization results with multiple localization methods from Fig. 11. The localization performance depends on the accuracy of time difference feature extracted by these methods. The abscissa 1, 2, and 3 indicate the effectiveness of three TDE method in the case that spatially correlated Gaussian noise exist. The abscissa 3 and 4 indicate the effectiveness of valid feature detection method. The abscissa 4 to 6 show that the increase of the recordings obviously improves the localization performance of azimuth, but not horizontal distance.

Comparing with geometric positioning method, it is proved above that our SGM method can localize far-field sound source more effectively. However, the localization performance will deteriorate along the increasing of environmental noise. To test the localization ability of our system for instant sound source that has lower signal to environmental noise ratio, another 24 positions with the azimuth of every 30° and the distance of 6, 8 m from the robot are tested. Because of the attenuation of source signal, the average signal to environmental noise ratios of these two distance are detected as 6.8 dB and 4.0 dB,

TABLE VIII
AZIMUTH LOCALIZATION RATEF OR DISTANT SOURCE

Methods	Azimuth Correct Rate (%) $4m$	$6m$	$8m$
SFOC+SGM	81.46	63.27	44.26
SW+SFOC+SGM	87.16	66.36	51.48
SW+SFOC+SGM+VFD(1 recording)	95.78	70.25	53.91
SW+SFOC+SGM+VFD(2 recordings)	96.21	78.77	54.25
SW+SFOC+SGM+VFD(3 recordings)	96.32	86.10	54.01

respectively. In each position, ten groups of speech data emitted from two speakers are recorded. In this experiment, *SFOC*, SGM, and *VFD* methods are used. In addition, the spectral weighting (SW) function presented in [35] is also used for suppressing stationary environmental noise. Table VIII shows the azimuth localization performance with localization error less than 15°. Owing to the non-Gaussian environmental noise, the performance of basic SFOC+SGM method is bad for 6 and 8 m. SW can significantly improve the localization rate for all the three distances by reducing the weight of those frequencies affected by stationary noise. VFD method is effective for 4 and 6 m. In particular, for 6 m, the correct rate increase quickly if two or three recordings are used. However, VFD method improves the performance of 8 m sparingly, which is probably due to the fact that the time differences for the new valid feature are too limited. The correct rate of 6 m can meet the demand of HRI roughly; however, it needs multiple recordings. In general, for speaking loudly in a normal way, our localization system works well within 5 m.

3) Application of SSL System: In the scene of HRI, human call the mobile robot to attract its attention, then the auditory system collects speech signals and gives feedback to the speaker. Auditory system consists of two subsystems: Speech recognition subsystem recognizes speaker-independent speech commands in real noise environments, and SSL subsystem localizes the relative direction and range of a speaker with respect to the robot.

In this application, the mobile robot works in the hall mentioned above with $3 \sim 5$ persons around it. Speech commands include Chinese words "zhuyi" and "guolai" which mean "pay attention" and "come here." First, when a speaker call the robot, the meaning of command and the position of sound source can be obtained. Then the robot turns to the speaker, and the vision-based system is also used to detect the direction of human accurately, such as "Body Detection" and "Hands-Raising Detection." In addition, if the horizontal distance is localized as *NEAR* which means that the speaker stands in the dangerous area, robot calls attention to him/her that "Pay attention, you are standing in the dangerous area." Second, the robot will stay put if the command is "zhuyi." Whereas, if "guolai" is called and the horizontal distance is *FAR*, robot will move 1 m toward the speaker. In this step, "guolai" can be called several times to get an appropriate distance between the speaker and the robot. Finally, for another interaction tasks, the robot faces the speaker directly with a suitable distance. For the friendliness of HRI, only one speaker can call the robot within a certain period of time, which indicates that the source speech signals are sparse in time domain. If multiple sound

TABLE IX
LOCALIZATION PERFORMANCE

The average call times	1.26
The average azimuth error	4.72°
The maximum time consumption	0.73s
The failure rate	5.39%

Fig. 12. Application of sound source localization system.

events occur simultaneously, those speech segments that just have one peak in the envelope of formula (18) are used for localization.

100 interaction tasks are tested, including 50 near-field tasks and 50 far-field tasks within 5 m. The target of each task is that, through the rotation and motion of the mobile robot, speakers stand in the region with azimuth $-23^\circ \sim 23^\circ$ (camera's detection range of our system) and horizontal distance 1.5 m $\sim$ 2.5 m (a suitable distance for interaction). Near-field interaction task needs once localization to cumulate the rotation angle. Far-field interaction task needs several localization tasks to adjust the distance between the robot and speakers. In total, 167 localization tasks are implemented for these 100 interaction tasks. For each localization, the speaker can call $1 \sim 3$ times if the wrong localization (azimuth localization error greater than 23°) occurs. Table IX shows the performance of localization tasks. The average azimuth error is not very accurate due to the measurement error of real value. The average time consumption just refer to the SSL function. The failure denotes the localization task that the localization results are all wrong for three times. In this experiment, SW, *SFOC*, SGM, and *VFD* methods are used. The scene of this application for HRI is shown in Fig. 12.

VI. CONCLUSION

In this paper, a novel SSL method for HRI based on the FOC-based time difference feature and SGM method is proposed. FOC spectrum and cross spectrum are derived. The multiplicative relationship of each frequency spectrum in FOC spectrum guarantees the additive relationship of each phase spectrum, which indicates the independence of multiple time differences from one signal to the others. Then, a TDE method for speech signal is proposed based on FOC spectrum, which can remove the influence of spatially correlated Gaussian noise. In addition, it is more robust than time domain FOC-based method by weakening confused peaks. However, the performance will decline as the Gaussian noise increases. This method is invalid to spatially uncorrelated noise and non-Gaussian noise.

SGM method is proposed for localization step. First, it avoids the difficulty of the solution of inverse problem, which makes geometric positioning method difficult in some situation. Then, it can handle the problem of the switch between far field and near field easily. In addition, it can weaken some dimensions selectively. Therefore, it can solve some problems that geometric positioning method cannot. However, time difference feature is too sensitive to the measurement error of horizontal distance, which brings about the low resolution and bad localization performance of horizontal distance. Hence, a more effective feature should be investigated, such as amplitude difference. Valid feature detection method removes those wrong time differences and improves localization performance. Decision tree reduces the number of times of template matching greatly. Experiments of TDE and SSL for real-time HRI are presented in real environments, which proves the effectiveness of these algorithms for speech sound source with spatially correlated Gaussian noise.

REFERENCES

[1] I. E. Robert, "Robust sound localization: An application of an auditory perception system for a humanoid robot," M.S. thesis, Massachusetts Inst. of Technol., Department of Elect. Eng. and Comput. Sci., Cambridge, MA, 1995.

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

[3] J. Hornstein, F. Lacerda, M. Lopes, and J. Santos-Victor, "Sound localization for humanoid robots-building audio-motor maps based on the HRTF," in *Proc. IEEE/RSJ Int. Conf. IROS*, Beijing, China, 2006, pp. 1170–1176.

[4] K. Nakadai, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition HARK and its evaluation," in *Proc. IEEE Int. Conf. Humanoid Robots*, Daejeon, Korea, 2008, pp. 561–566.

[5] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments," in *Proc. IEEE/RSJ Int. Conf. IROS*, St. Louis, MO, 2009, pp. 2027–2032.

[6] B. Kwon, Y. Park, and Y.-S. Park, "Sound source localization for robot auditory system using the summed GCC method," in *Proc. Int. Conf. CAS*, Seoul, Korea, 2008, pp. 241–245.

[7] J.-S. Hu, C.-Y. Chan, C.-K. Wang, and C.-C. Wang, "Simultaneous localization of mobile robot and multiple sound sources using microphone array," in *Proc. IEEE ICRA*, Kobe, Japan, 2009, pp. 29–34.

[8] H. Wang and M. Kaveh, "Coherent signal subspace processing for the detection and estimation of angles of arrival of mutiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 4, pp. 823–831, Aug. 1985.

[9] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 5, pp. 1210–1217, Oct. 1983.

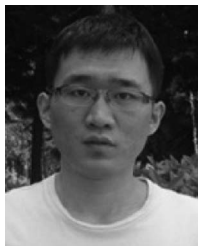
[10] G. Carter, "Variance bounds for passively locating an acoustic source with a sysmmetric line array," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 922–926, Oct. 1977.

[11] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[12] G. C. Cater, "Special issue on time delay estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-49, no. 1, pp. 1–12, Mar. 1981.

[13] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.

[14] S. Doclo and M. Moonen, "Robust time-delay estimation in highly adverse acoustic environments," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Platz, NY, 2001, pp. 59–62.
[15] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Process.*, vol. 85, no. 1, pp. 177–204, Jan. 2005.
[16] P. Chevalier, A. Ferreol, and L. Albera, "High-resolution direction finding from higher order statistics: The 2q-MUSIC algorithm," *IEEE Trans. Signal Process.*, vol. 54, no. 8, pp. 2986–2997, Aug. 2006.
[17] L. D. Lathauwer, J. Castaing, and J. Cardoso, "Fourth-order cumulant-based blind identification of underdetermined mixtures," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2965–2973, Jun. 2007.
[18] M. Swartling, B. Sällberg, and N. Grbic, "Direction of arrival estimation for speech sources using fourth order cross cumulants," in *Proc. IEEE Int. Symp. CAS*, Seattle, WA, 2008, pp. 1696–1699.
[19] H. D. Han and Z. Ding, "A Blind Channel Shortening Criterion Based on High-order Cumulants," in *Proc. IEEE Int. Conf. ASSP*, Dallas, TX, 2010, pp. 3210–3213.
[20] L. N. Sharma, S. Dandapat, and A. Mahanta, "ECG signal denoising using higher order statistics in Wavelet subbands," *J. Biomed. Signal Process. Control*, vol. 5, no. 3, pp. 214–222, Jul. 2010.
[21] C. L. Nikias and R. Pan, "Time delay estimation in unknown Gaussian spatially correlated noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 11, pp. 1706–1714, Nov. 1988.
[22] W. Zhang and M. Raghuveer, "Nonparametric bispectrum-based time-delay estimators for multiple sensor data," *IEEE Trans. Signal Process.*, vol. 39, no. 3, pp. 770–774, Mar. 1991.
[23] J. K. Tugnait, "Time delay estimation with unknown spatially correlated Gaussian noise," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 549–558, Feb. 1993.
[24] Y. C. Liang, A. R. Leyman, and B. H. Soong, "Criteria and algorithms for time delay estimation based on cumulants," in *Proc. IEEE Int. Symp. CAS*, ., HongKong, 1997, pp. 2493–2496.
[25] H. Wang, J. Zhao, and L. Qian, "Research of time-delay estimation based on fourth-second order normalized cumulant," in *Proc. Int. Conf. CMCE*, ., HongKong, 2010, pp. 29–32.
[26] W. H. Foy, "Position-localization solution by Taylor-series estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-12, no. 2, pp. 187–194, Mar. 1976.
[27] W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 5, pp. 608–614, Sep. 1973.
[28] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, 1997, pp. 1–4.
[29] H. Schau and A. Robinson, "Passive source localization employing intersection spherical surfaces from time-of-arrival difference," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 8, pp. 1223–1225, Aug. 1987.
[30] J. Smith and J. Abel, "Closed-form least-square source location estimation from range-different measurement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661–1669, Dec. 1987.
[31] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.
[32] Y. T. Chan and K. C. Ho, "A simple and effcient estimator for hyperbolic location," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 1905–1915, Aug. 1994.
[33] M. S. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *J. Comput. Speech Lang.*, vol. 11, no. 2, pp. 91–126, Apr. 1997.
[34] H. Liu and X. F. Li, "Time delay estimation for speech signal based on FOC-spectrum," in *Proc. Int. Conf. INTERSPEECH*, Portland, OR, 2012, pp. 1–4.
[35] X. F. Li, H. Liu, and X. S. Yang, "Sound source localization for mobile robot based on time difference feature and space grid matching," in *Proc. IEEE/RSJ Int. Conf. IROS*, San Francisco, CA, 2011, pp. 2879–2886.
[36] H. Liu and M. Shen, "Continuous sound source localization based on microphone array for mobile robots," in *Proc. IEEE/RSJ Int. Conf. IROS*, Taipei, Taiwan, 2010, pp. 4332–4339.
[37] M. Feng and K. D. Kammeyer, "Suppression of Gaussian noise using cumulants: A quantitative analysis," in *Proc. IEEE Int. Conf. ASSP*, 1997, pp. 3813–3816.



Xiaofei Li was born in Handan, China, in 1987. He received the B.E. degree in electronic information science and technology from Beijing Institute of Machinery, Beijing, China, in 2007. Currently, he is working toward the Ph.D. degree at the School of Electronics Engineering and Computer Science, Peking University, Shenzhen, China.

His current research interests are speech signal processing, speech recognition, and sound source localization.

Hong Liu received the Ph.D. degree in mechanical electronics and automation, in 1996, and serves as a Full Professor in the School of EE&CS, Peking University, Shenzhen, China.

He is also the Director of Engineering Lab on Intelligent Perception for Internet of Things. His research fields include computer vision and robotics, image processing, and pattern recognition.

Prof. Liu has published more than 100 papers and gained Chinese National Aero-space Award, Excellence Teaching Award, and Candidates of Top Ten Outstanding Professors, Peking University. He is an IEEE Member, Vice Chair of Intelligent Robotics Society of Chinese Association for Artificial Intelligent (CAAI), and also the President of National Youth Committee of CAAI. He has served as cochairs, session chairs, or PC members of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, and IEEE SMC, and also serves as reviewers for many international journals such as Pattern Recognition, IEEE Trans. on Signal Processing, and IEEE Trans. on PAMI.