Microphone Clustering and BP Network based Acoustic Source Localization in Distributed Microphone Arrays

Qiaoling ZHANG [1], Zhe CHEN[2], Fuliang YIN[3]
[1,2,3]*School of Information and Communication Engineering*
Dalian University of Technology, 116024 Dalian, China
[1]*qiaoling880907@mail.dlut.edu.cn,* [2]*zhechen@dlut.edu.cn,* [3]*flyin@dlut.edu.cn*

***Abstract*—A microphone clustering and back propagation (BP) neural network based acoustic source localization method using distributed microphone arrays in an intelligent meeting room is proposed. In the proposed method, a novel clustering method is first used to divide all microphones into several clusters where each one corresponds to a specified BP network. Afterwards, the energy-based cluster selecting scheme is applied to the select the clusters which are small and close to the acoustic source. In each chosen cluster, the time difference of arrival of each microphone pair is estimated, and then all estimated time delays act as input of the corresponding BP network for position estimation. Finally, all estimated positions from the chosen clusters are fused for global position estimation. Only subsets rather than all the microphones are responsible for acoustic source localization, which leads to less computational cost; moreover, the local estimation in each chosen cluster can be processed in parallel, which expects to improve the localization speed potentially. Simulation results from comparison with other related localization methods confirm the validity of the proposed method.**

***Index Terms*—acoustic source localization, BP neural network, microphone clustering, GCC-PHAT, TDOA.**

I. INTRODUCTION

Acoustic source localization using microphone arrays has been widely used in various applications, ranging from teleconference, audio/video supervision, human computer interaction to hearing aids [1-3], etc. Correspondingly, various localization methods [4-7] have been proposed for different applications. Altogether, localization methods may be broadly grouped into indirect and direct approaches. The generalized cross-correlation and phase transform (GCC-PHAT) based localization method is a popular indirect approach, which first estimates the time difference of arrival (TDOA) of received signals at different microphones, and then estimates the source position. The steered response power and phase transform (SRP-PHAT) method [8] is a typical direct approach, which can perform the TDOA estimation and acoustic source localization in one step. However, most localization literatures focus on a single regular microphone array [9]. In fact, in some special scenarios with distributed microphone arrays which have large microphone spacing, numerous microphones and irregular topology, these traditional methods can not be applied directly. With large microphone spacing, some localization methods may fail due to the spatial aliasing effect. Moreover, signal qualities received at different microphones may rather different. Generally speaking, the microphones closer to the speaker may yield more accurate position estimation than those far away [10]. Therefore, it is necessary to develop localization methods suitable for distributed microphone arrays.

There have been some valuable works on acoustic source localization using distributed microphone arrays. In the Aarabi approach [10], each microphone subarray produces an individual spatial likelihood function (SLF), and then fuses all the SLFs by a weighted summation for the final source position estimation, but its computational load is large. Elahi [11], proposed a similar method which fuses all the SLFs by simple summation to estimate the speaker position, and modeled prior information about the speaker position to reduce the computational complexity. Takagi *et al*. [12] divided the whole microphone network into several microphone subarrays (considered each subarray as a node), and presented a hierarchical localization method. In the Takagi method, in the node layer, the multiple signal classification (MUSIC) algorithm was employed for direction estimation; in the network layer, the intersections were calculated from estimated data and microphone coordinates, and finally the intersections were fused for robust position estimation. As described in [10-12], to circumvent the spatial aliasing due to large microphone spacing, it is sensible to perform traditional localization methods on the small subarrays, and then fuse the estimated results from each subarray for robust position estimation. However, in these literatures, the microphone subarrays are naturally implied and each one has regular topology, they did not mention how to partition the microphones into subarrays. Valenzise *et al*. [13], from the perspective of resource consumption, proposed a resource constrained efficient acoustic source localization approach where according to the performance constrains (i.e., Cramer-Rao low bound CRLB) and resource constrains, the optimal subset of microphones rather than the whole array is selected for acoustic source localization.

In this paper, a microphone clustering and BP neural network based localization method is proposed for acoustic source localization in an intelligent meeting room with distributed microphone arrays. The main idea of this method

This work was supported by National Natural Science Foundation of China (No.61172107, No.61172110, No.60772161), Dalian Municipal Science and Technology Fund Scheme of China (No. 2008J23JH025), Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 200801410015) and Fundamental Research Funds for the Central Universities of China (DUT13LAB06).

Digital Object Identifier 10.4316/AECE.2013.04006

is to divide the distributed microphone arrays into several small microphone clusters (i.e., subarrays); each cluster may carry on local estimation of the acoustic source position in parallel, and all local estimated results from these clusters are fused for the global acoustic source position. In order to reduce the computational cost and gain robust estimation, the energy based cluster selecting method is used to select the clusters which are tight and close to the speaker for acoustic source localization. Each microphone cluster corresponds to a specified BP neural network whose inputs are the estimated TDOAs from the cluster and output is the estimated source position. During the acoustic source localization, in each selected cluster, the TDOA estimation is first carried on with the `GCC-PHAT` method; then, all the estimated TDOAs act as the inputs to the corresponding BP network for position estimation.

This paper is organized as follows. In Section II the basic theory for acoustic source localization is presented. In Section III, the proposed acoustic source localization method using distributed microphone arrays is described in detail, including the microphone clustering, cluster selecting, BP networks for position estimation, and estimated results fusion. Simulated results are given in Section IV. Finally, some conclusions are drawn in Section V.

II. Basic Theory for Acoustic Source Localization

A. Signal Model in Distributed Microphone Arrays

Consider a single acoustic source in a typical noisy and reverberant environment with distributed microphone arrays of M microphones elements. The signal received at the microphone i is modeled as

$$x_i(t) = h_i(t) \otimes s_r(t) + n_i(t), \quad i = 1,2,\cdots,M \tag{1}$$

where $h_i(t)$ denotes the impulse response from the acoustic source $s_r(t)$ to the microphone i, '$\otimes$' denotes the linear convolution operator; and $n_i(t)$ is an additive noise at microphone i and assumed to be uncorrelated with each other as well as $s_r(t)$.

B. `GCC-PHAT` Method for TDOA Estimation

Assume that $x_i(t)$ and $x_j(t)$ are signals acquired by microphones i and j ($i, j = 1,2,\cdots,M$) respectively. The generalized cross-correlation (GCC) function [14] $r(\tau_{ij})$ between microphone pair (i, j) is

$$r(\tau_{ij}) = \int_{-\infty}^{+\infty} \Phi(f) X_i(f) X_j(f) e^{j2\pi f \tau_{ij}} df \tag{2}$$

where $\Phi(f)$ is the frequency-domain weighting function, and different patterns of $\Phi(f)$ may lead to various GCC methods, among which the phase transform (PHAT) method is widely used. In the `GCC-PHAT` method [15], the cross-correlation function is

$$r_{ij}^{GCC-PHAT}(\tau) = \int \frac{X_i(f) X_j^*(f)}{\left| X_i(f) X_j^*(f) \right|} \mathrm{e}^{j2\pi f \tau} \mathrm{d}f \tag{3}$$

where '$*$' denotes the conjugation operator, and the weighting function is $\Phi(f) = \dfrac{1}{\left| X_i(f) X_j^*(f) \right|}$.

The relative time difference of arrival (TDOA) τ_{ij} between microphones i and j corresponds to the time tag when the cross-correlation $r_{ij}^{GCC-PHAT}(\tau)$ reaches the peak

$$\tau_{ij}^{GCC-PHAT} = \arg\max_{\tau}(r_{ij}^{GCC-PHAT}(\tau)) \tag{4}$$

C. TDOA based Acoustic Source Localization

Assume that the position coordinates of microphones i and j in 3-D space are $r_i = (x_i, y_i, z_i)^{\mathrm{T}}$ and $r_j = (x_j, y_j, z_j)^{\mathrm{T}}$, respectively, and the Euclidian distance between them is $d_{ij} = \left\| r_i - r_j \right\|$. For an acoustic source at $r_s = (x_s, y_s, z_s)^T$, the relative time delay τ_{ij}, i.e., TDOA between microphones i and j satisfies

$$\tau_{ij} = \frac{1}{c}(\left\| r_i - r_s \right\| - \left\| r_j - r_s \right\|) \tag{5}$$

where c is the sound propagation speed, and (5) is an equation in vector form which can be also expressed as

$$\tau_{ij} = \frac{1}{c}(\sqrt{(x_i - x_s)^2 + (y_i - y_s)^2 + (z_i - z_s)^2} - \sqrt{(x_j - x_s)^2 - (y_j - y_s)^2 + (z_j - z_s)^2}) \tag{6}$$

Obviously, given the positions and TDOA τ_{ij} of microphone pair (i, j), (6) is an equation with three unknown variables x_s, y_s, z_s. Thus it needs at least three different equations for an acoustic source position. In other words, at least three microphone pairs are required for TDOA based acoustic localization.

D. `Adaptive K-Means++` Clustering method

Assume that the k-th microphone cluster contains n_k microphones, whose coordinates in vector form are $\{r_i^k, i = 1,2\cdots, n_k\}$. The clustering procedure of the distributed microphone arrays using the `Adaptive K-Means++` method is as follows [15]:

1) Set k = 1 and the first initial center to be the median of the microphone coordinates.

2) Carry on the `K-Means` clustering method.

3) Calculate the variance of each microphone cluster k.

$$v(k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \left\| r_i^k - r_0^k \right\|^2 \tag{7}$$

where r_0^k is the center coordinates of cluster k.

4) Check the stop condition for clustering

$$v(k) < \delta_{cluster} \tag{8}$$

where the threshold $\delta_{cluster}$ is estimated in the practice.

5) If the variances of all the clusters satisfy the condition (8), the clustering is stopped; otherwise, k = k +1, and calculate the initial center of the new cluster according to the `K-Means++` algorithm [15], i.e.

$$r_0^k = \arg\max_i \frac{D^2(i)}{\sum_{j=1}^{M} D^2(j)} \tag{9}$$

where $D(i) = \min_k(\left\| r_i - r_0^k \right\|)$ denotes the shortest distance between microphone i and all the current cluster centers.

The schedule diagram of the `Adaptive K-Means+ +` clustering method is shown in Fig. 1.

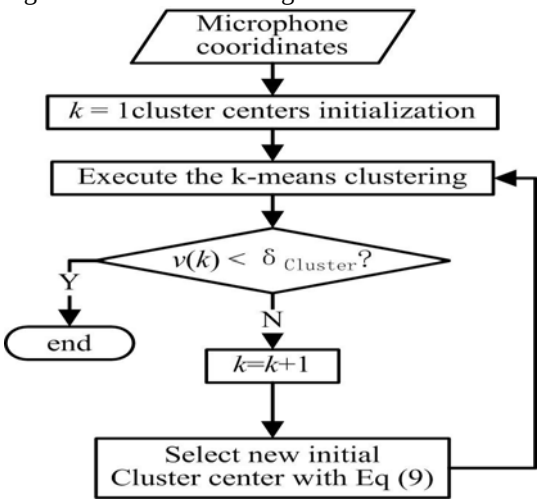

Figure 1. Adaptive K-Means++ clustering method

E. BP Neural Network

The Artificial Neural Network (ANN) has been introduced into microphone array processing for acoustic source localization [17-20]. BP neural network is a classic artificial neural network, and may be used for mathematical modeling and prediction. In the learning process of BP network, the training pattern's input propagates forward across the network, whereas the errors propagate backward from the output nodes to the inner nodes. BP network typically has a three-layer structure: input layer, hidden layer and output layer, as shown in Fig. 2.

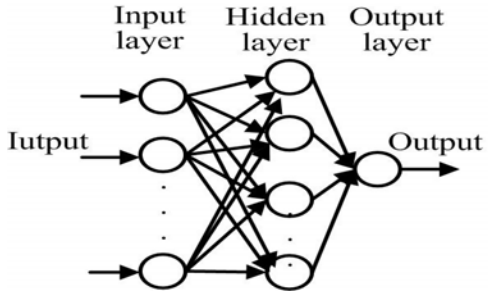

Figure 2. BP neural network

In this paper, several BP networks have been built with each one corresponding to a specific microphone cluster. In the training phase of BP network, the positions of acoustic source are used as the output and the true TDOAs from the microphone cluster act as the input; in the simulation phase, the estimated TDOAs from the microphone cluster act as the input of the built BP network for an estimated acoustic source position.

III. Acoustic Source Localization with Distributed Microphone Arrays

In the noisy and reverberant environment with distributed microphone arrays, the acoustic signal qualities acquired by microphones at different positions are rather different due to the large coverage of the whole array. Microphones which are close to the acoustic source can provide more reliable estimates than those far away [10]. Additionally, when the number of microphones is very large, it will have large computational and communication cost if using all the microphones for the acoustic source localization.

In this paper, the main goal is to select the microphone subsets which are close to the speaker for acoustic source localization. In order to achieve the goal, four essential aspects need to be considered: 1) how to group the microphones into several small clusters (i.e., subsets); 2) how to determine which microphone cluster or clusters are close to the acoustic source and then select it or them for localization; 3) what method to use for the acoustic source localization in the selected cluster or clusters; 4) if more than one microphone cluster is selected, how to fuse the estimated results from each cluster. Correspondingly, the implementation of the proposed approach herein mainly contains four modules as Fig. 3.

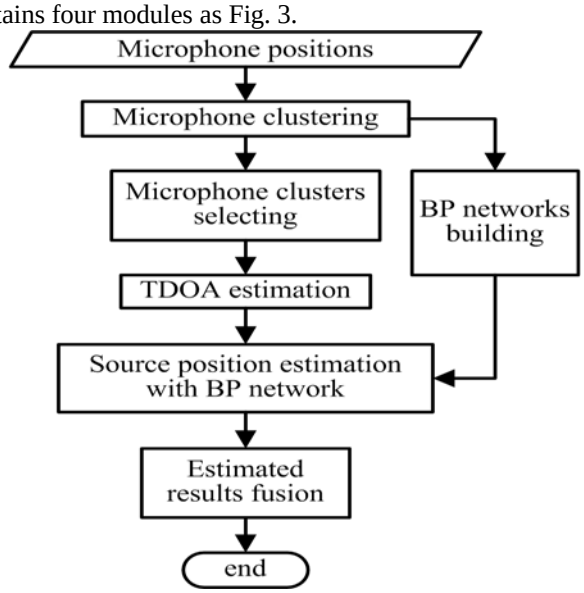

Figure 3. Microphone clustering and BP network-based acoustic source localization method

As shown in Fig. 3, given the microphone positions, a neighborhood measure based microphone clustering is first applied to group the microphones into several clusters; afterwards several BP networks are built with each one corresponding to a specific cluster. During the localization, one or more microphone clusters are first selected on basis of some energy scheme; then the TDOA estimation is carried on in each selected cluster; next the TDOAs from each cluster act as the inputs of the corresponding BP network whose output is the estimated acoustic source position; finally, estimated positions from all the selected clusters are fused for a robust result of the acoustic source localization.

A. Clustering of Distributed Microphone Arrays

Many clustering methods have been applied for data classification [21-25]. We herein introduce the `Adaptive K-Means++` method [17] into microphone clustering and propose an alternative clustering method according to the characteristic of the localization method with the distributed microphone arrays.

The proposed clustering method aims to divide the distributed microphone arrays into several small microphone clusters, and each cluster is available for acoustic source localization. In this method, the Euclidian distance d_{ij} between microphones i and j is used as the proximity measure, and the variance $v(k)$ is adopted as the intensive measure of microphone cluster k.

As analyzed above, at least three microphone pairs are required for the acoustic source localization. Thus during the clustering procedure, the smallest cluster may contain at least three pairs of microphones.

The clustering procedure is given as follows:

1) Any two microphones i and j may constitute a pair (i, j), and keep all the pairs in which the microphones are close enough to each other into a set P

$$P = \{(i, j) \mid d_{ij} < d_{spa}\},\ i, j = 1,2,\cdots,M,\ \ i \neq j \tag{10}$$

where d_{spa} is the distance threshold.

2) In the microphone set P generated from step 1), any three microphone pairs $M_k \subseteq P$ may constitute a potential microphone cluster k.

$$M_k = \{(i^k, j^k), (l^k, m^k), (p^k, q^k)\} \tag{11}$$

3) Calculate the variance v(k) of each potential cluster k and check

$$v(k) < \delta_{clu} \tag{12}$$

where the threshold δ_{clu} is estimated in the practice.

4) If microphone pair set M_k of cluster k satisfies condition (12), its microphones may form an initial microphone cluster k.

5) Merge the initial microphone clusters generated from step 4). Clusters k_1 and k_2 are merged into a single cluster k, if they share the same microphones, and the corresponding microphone pair sets M_{k_1} and M_{k_2} are merged as follows:

$$M_k = M_{k_1} \cup M_{k_2} \tag{13}$$

where M_k is the microphone pair set of the merged cluster k, and the '$\cup$' represents the union operation of set.

B. Microphone Cluster Selecting

As mentioned above, in distributed microphone arrays, large microphone spacing may lead to erroneous TDOA estimation due to the spatial aliasing, which further affects the acoustic source position estimation. It is logical to use the subset rather than all microphones for acoustic source localization. Additionally, the microphones or arrays which are close to the speaker can obtain better signal quality. Generally speaking, given the same background SNR and geometry structures for two different microphone arrays, the array closer to the speaker will ordinarily yield more accurate location estimation than the array far away [10]. Theoretically, the acoustic energy is inversely proportional to the square of the distance between the sound source and the microphone [26], thus it may be rational to assume that the microphone cluster with high received acoustic energy is closer to the speaker than the one with low acoustic energy. Herein two potential energy-based proposals are presented for cluster selection.

Cluster selecting method-I: Individual microphone signal energy based method

Assume that a person is speaking in a specific position of the room. The procedure of cluster selecting is as follows.

1) Assign a specified wall number w_l, l=1,2, 3,4,5, to each microphone i, and $i(w_l)$ indicates microphone i is on wall w_l.

2) Calculate the short-time energy $s(i)$ of the signal received at each microphone i:

$$s(i) = \sum_{i=1}^{L} x_i^2(t),\ \ i = 1,2,\cdots,M \tag{14}$$

where L is the signal length in samples.

3) Search M_0 microphones with the largest short-time energies, and determine the corresponding wall numbers.

4) As to each selected wall from step 3), calculate the average signal energy $S_e(k)$ of all microphones for each cluster k on it.

$$S_e(k) = \frac{1}{n_k} \sum_{i=1}^{n_k} s(i^k) \tag{15}$$

where i^k denotes the microphone i in cluster k, and n_k is the microphone number of cluster k.

5) For each selected wall from step 3), select the microphone cluster k_0 with the largest average energy on it.

$$k_0 = \arg\max_k S_e(k) \tag{16}$$

In this cluster selecting method, the walls which are close to the speaker are first determined, and then the microphone cluster with largest average signal energy corresponding to each selected wall is further selected for acoustic source localization.

Cluster selecting method-II: Average energy to energy variance ratio based cluster selecting method

Alternatively, another cluster selecting method is presented as well, and the cluster selecting process is described as follows.

1) Calculate the short-time signal energy $s(i)$ of each microphone i using equation (14).

2) Calculate the average signal energy $S_e(k)$ of all microphones for each cluster k using equation (15).

3) Initially select the microphone clusters with large energy

$$\{k \mid S_e(k) > S_E\} \tag{17}$$

where S_E is the energy threshold, and $S_E = \gamma S_{emax}$, and S_{emax} is the maximum average energy of all clusters, and γ is a ratio factor.

4) Calculate energy variance $e_v(k)$ of each cluster k generated from step 3)

$$e_v(k) = \frac{1}{n_k} \sum_{i=1}^{n_k} (s(i^k) - S_e(k))^2 \tag{18}$$

5) Calculate the ratio of average energy $S_e(k)$ to energy variance $e_v(k)$ for each selected cluster k from step 3)

$$e_{sv}(k) = \frac{S_e(k)}{e_v(k)} \tag{19}$$

6) Further select microphone cluster k, if it satisfies

$$e_{sv}(k) < E_{sv} \tag{20}$$

where $E_{sv} = \beta e_{vmin}$, and e_{vmin} is the minimum energy variance of all microphone clusters, and β is a ratio factor.

In the cluster selecting method-I, each selected wall can contribute only one cluster for acoustic localization, whereas in the cluster selecting method-II, more than one cluster on one wall may be selected. Moreover, the cluster selecting method-II is more computation consuming as it requires the average signal energies of all available clusters.

C. BP Networks for Source Position Estimation

Each BP network in this paper has a three-layer network

structure, but different BP networks have dissimilar parameter setups. The input layer has as many neurons as the dimension of the input vector. In the hidden layer, the number of hidden units is optimized in the experiment. The output layer has three neurons corresponding to the dimension of the estimated speaker position in 3-D space.

Input definition

The estimated TDOAs of each microphone cluster act as the input vector of the corresponding BP network. Consider a microphone cluster p with n estimated TDOA values from n pairs of microphones, the input vector to the specified BP network is

$$[\tau_{p,1}, \tau_{p,2}, \cdots \tau_{p,n}]^{\mathrm{T}} \tag{21}$$

where $\tau_{p,i}$ is the relative time delay from microphone pair i of the cluster p, and '$[\cdot]^{\mathrm{T}}$' denotes the vector or matrix transposition. Obviously, microphone clusters with different numbers of microphone pairs may have different input vector dimensions for specified BP networks.

Output definition

The output of each BP network is the estimated position coordinates which act as the acoustic position estimation of the specific microphone cluster. In 3-D space the output vector format is:

$$[x_s, y_s, z_s]^{\mathrm{T}} \tag{22}$$

Microphone cluster table

When all the BP networks are trained, a microphone cluster table has been completely built. In the cluster table, each BP network item corresponds to a specified microphone cluster. During the acoustic source localization stage, once the microphone clusters are selected, the corresponding BP networks in the microphone cluster table are then taken for source position estimation.

D. Position Estimation of Acoustic Source

As mentioned above, more than one microphone clusters may be selected for acoustic source localization. Thus, data fusion is adopted for final global estimated result.

Weighting factor

Assume that s_k microphone clusters are selected for local acoustic position estimation. For each selected microphone cluster k, a weighting factor $J(k)$ is introduced as

$$J(k) = \alpha S_e(k) + \frac{1-\alpha}{v(k)},\ k = 1,2,\cdots,s_k \tag{23}$$

where α is a ratio factor, and $S_e(k)$ and $v(k)$ are defined as before.

The weighting factor $J(k)$ may be normalized as

$$\bar{J}(k) = \frac{J(k)}{\sum_{i=1}^{s_k} J(i)},\ k = 1,2,\cdots,s_k \tag{24}$$

Next, the normalized weighting factor $\bar{J}(k)$ is adopted for the fusion of local estimated results.

Data fusion

Let r_s denotes the acoustic source position vector, and $\hat{r}_s(k)$ is the estimated source position from cluster k. The final global position estimation is calculated by the weighted fusion

$$\hat{r}_s = \sum_{k=1}^{s_k} \bar{J}(k) \cdot \hat{r}_s(k) \tag{25}$$

E. Computational Cost

In this subsection, we attempt to provide an upper bound estimation of the computational cost in terms of multiplications and accumulations (MACs) for the proposed method and the method using one ANN network for all microphones [20] (we call it single ANN method afterwards in this paper) as well as the SRP-PHAT method [8].

In the proposed method, the microphone clustering and BP networks training can be completed offline before the localization, therefore the associated processing does not involve additional computational cost. Denote M as the number of microphones, N_c as the number of clusters in total, P_c as the number of microphone pairs in one cluster, s_k as the number of selected clusters; and n_i, n_h, n_o as the number of neural units in input-layer, hidden-layer and output-layer, respectively. The upper bound of the parameters above is given as: $M = 36$, $N_c = 39$, $P_c = 6$, $s_k = 4$, $n_i = 6$, $n_h = 14$, $n_o = 3$. In the method of all microphones with one single ANN, the number of microphone pairs is $p^{all} = 34$. The parameters of the ANN network are $n_i^{all} = 34$, $p^{all} = 34$, $n_h^{all} = 68$, $n_o^{all} = 3$, respectively. In the SRP-PHAT method, the number of microphone pairs is $p^{SRP\text{-}PHAT} = 34$, and the number of grid positions is $r = 5556$.

Let m_{ac} denote the numbers of multiplications and accumulation of one TDOA computation. The computational cost of the three methods in terms of MACs is estimated in Table I.

TABLE I. COMPARISSON OF COMPUTATIONAL COST FOR THREE METHODS

Method	Number of MACs
Proposed method	$M + 1.5N_c + (m_{ac} \cdot p_c + n_i \cdot n_h \cdot n_o) \cdot s_k + 1 = 24m_{ac} + 1103.5$
Single ANN	$m_{ac} \cdot p^{all} + n_i^{all} \cdot n_h^{all} \cdot n_o^{all} = 34m_{ac} + 3670$
SRP-PHAT	$m_{ac} \cdot p^{SRP-PHAT} + r = 34m_{ac} + 5556$

As shown in Table I, the proposed method expects to have less computational cost than the other two methods.

IV. SIMULATION AND RESULT DISCUSSION

A. Simulation Setups

Simulation environment setups

The simulation environment is a smart meeting room of size $5\text{m} \times 6\text{m} \times 3\text{m}$ with 36 microphones distributed on four walls (walls $w1$, $w2$, $w3$, $w4$) and the ceiling ($w5$). As shown in Fig. 4, walls $w1$ and $w2$ contain four microphones around the central region, respectively; walls $w3$ and $w4$ contain six microphones around the central region severally; the ceiling $w5$ contains four T-shaped microphone subarrays. The distance between adjacent microphones is 25cm.

BP network training and simulation

Each BP network has a three-layer structure as described in Section III, which has as many neural units as the estimated TDOAs from the corresponding cluster in the input layer, 14 neural units in the hidden layer and 3 neural units in the output layer. The BP networks were simulated in

Matlab-2010a, and the transfer function from the input layer to hidden layer was *tansig*, the transfer function from the hidden layer to the output layer was *purelin*, the training function was *trainbr*, and the learning ratio was 0.0001.

For the BP network training and simulation, 4575 acoustic source positions were uniformly sampled from the rectangular region of $3\text{m}\times 4\text{m}$ with the height from 145cm to 175cm，and the grid size is $10\text{cm}\times 10\text{cm}\times 10\text{cm}$. All the positions were divided into two complementary segments: training data denoted by TD is used for BP network training phase and simulation data denoted by SD is used for BP network simulation phase, with the proportions are 70% and 30%, respectively and with no intersection between them.

Microphones clustering

In order to evaluate the proposed microphones clustering method, a comparison with the `Adaptive K-Means++` clustering method was carried on in a scenario of Fig. 6(a), where six microphones are distributed within an area of $3\text{m}\times 4\text{m}$. The distance between two adjacent microphones is d_{spa} = 25cm, In the acoustic source localization experiment, all the microphone clusters in Fig. 4 were generated using the proposed clustering method and the threshold for clustering is $\delta_{cluster}$ = 657.

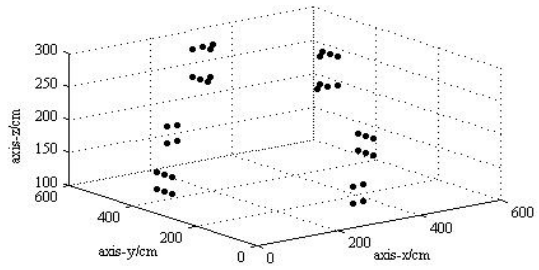

Figure 4. Distributed microphone arrays in a $5\text{m}\times 6\text{m}\times 3\text{m}$ room

Acoustic source localization

The room impulse response (RIR) of the scenario, as shown in Fig. 4, was simulated from the image model method [27]. The acoustic source was a 32 kHz sampled, 600ms female speech utterance with silence removed beforehand. Fig. 5 shows the original speech and the simulated signal received at one microphone in the reverberant environment with SNR = 20dB and RT60 = 0.2s. During the TDOA estimation using `GCC-PHAT` method, a frame length of 125ms with 4000 samples, frame shift of 2000 samples and Hamming window were adopted. In the clusters selecting phase, the two parameters were set as $\gamma = 0.9$ and $\beta = 1.15$, respectively.

During localization experiments, three segments from the positions SD with the female speech as speaker were used for position estimation: SD1 represents that the speaker is in the center region of the room; SD2 represents the speaker is close to one wall and SD3 represents the speaker is positioned in the corner.

In order to evaluate the localization performance of the proposed method, a comparison with the single ANN method [20] was first carried on in two different noise environments: background noise and localized noise source environment. In both scenarios the RT60 = 0.2s and the SNR varies. In the localized noise environment, the noise was positioned at [15,560,170] (cm). As a second comparison, the proposed localization method was also compared with the `SRP-PHAT` method in the reverberant environment where SNR varies and RT60 = 0.2s. Speaker positions were in the *x-y* plane with *z* fixed to150cm. The grid size of `SRP-PHAT` method is $6\text{cm}\times 6\text{cm}$, and 5556 grid positions are considered.

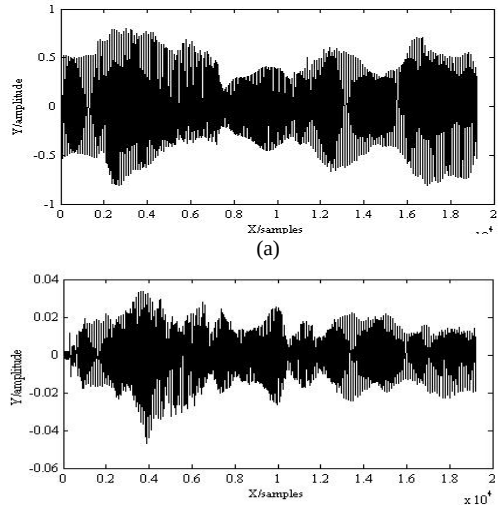

Figure 5. (a) The original female speech (b) The received signal of one microphone using image model method

B. Simulation Results

Microphone clustering

In the scenario of Fig. 6(a), the proposed microphone clustering method was compared with the `Adaptive K-Means++` method. Fig. 6(b) shows that using the `Adaptive K-Means++` method, only two topologies of clusters (i.e., subarrays) are generated and different clusters have no intersections (microphones). Whereas using the proposed clustering method, as shown in Fig. 7, six clusters with three topologies are generated, and different clusters may have intersections. Simulation results show that the proposed method can provide more clusters (i.e., various combinations of microphones) available for localization than the `Adaptive K-Means++` method, which is especially beneficial in the scenario where the speaker is moving. Additionally, the `Adaptive K-Means++` method may be affected by the center initialization, whereas the proposed method does not have this limitation.

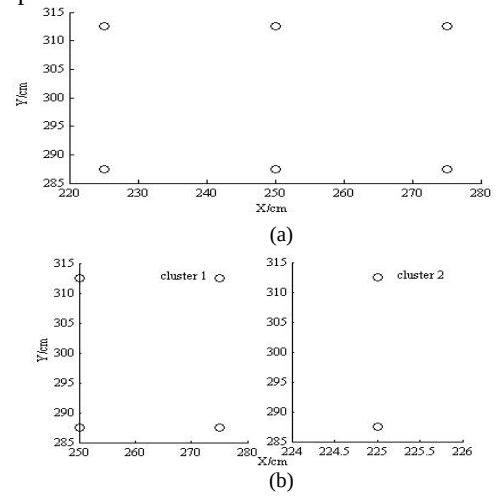

Figure 6. Microphone clusters generated by the Adaptive K-Means++ method with $\delta_{cluster}$ = 480

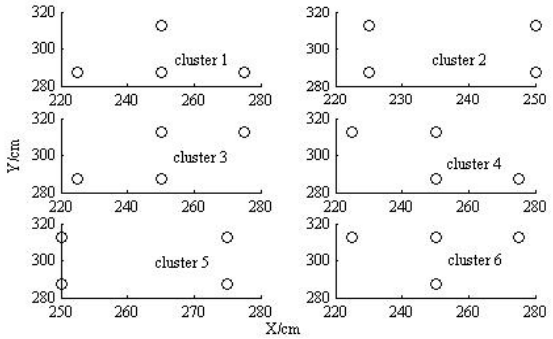

Figure 7. Microphone clusters generated by the proposed clustering method with $\delta_{cluster} = 480$

Acoustic source localization in background noise environment

The estimated results in background noise environment are shown in Table II. From the Table II, we can observe that the proposed localization method with cluster selecting is not as good as the single ANN method in high SNR conditions, whereas it performs better in low SNR conditions. The advantage of the proposed method lies in that the computational cost is less.

TABLE II. THE AVERAGE POSITION ESTIMATION ERRORS OF THE PROPOSED METHOD AND THE SINGLE ANN MEHTOD FOR SPEAKER IN DIFFERENT REGIONS OF THE BACKGROUND NOISE ENVIRONMENT

Data sets	Localization methods	Noise condition (SNR/dB)					
		Average localization errors (cm)					
		30	20	15	10	5	0
SD1	Cluster selecting-I	13.7	13.8	20.3	20.9	49.5	44.7
	Cluster selecting-II	18.6	20.9	17.0	17.1	22.3	49.6
	Single ANN method	4.30	4.94	6.20	7.97	159	313
SD2	Cluster selecting-I	18.5	25.9	20.3	21.7	44.6	46.3
	Cluster selecting-II	16.4	23.1	17.0	17.1	20.7	50.0
	Single ANN method	5.66	5.74	6.20	7.97	159	311
SD3	Cluster selecting-I	13.9	21.9	32.6	43.6	52.5	66.3
	Cluster selecting-II	21.2	26.0	36.4	46.8	56.1	69.5
	Single ANN method	4.50	14.9	7.58	8.75	132	248

TABLE III. THE AVERAGE POSITION ESTIMATION ERRORS OF THE PROPOSED METHOD AND THE SINGLE ANN MEHTOD FOR THE SPEAKER IN DIFFERENT REGIONS OF THE LOCALIZED NOISE SOURCE ENVIRONMENT

Data sets	Localization methods	Noise condition (SNR/dB)		
		Average localization errors (cm)		
		30dB	20dB	15dB
SD1	Cluster selecting-II	15.7	22.6	30.5
	Single ANN method	25.1	36.5	52.0
SD2	Cluster selecting-II	20.2	28.7	43.5
	Single ANN method	31.6	42.1	48.3
SD3	Cluster selecting-II	17.6	20.1	25.2
	Single ANN method	36.1	46.5	50.4

Acoustic source localization in localized noise environment

In this experiment, the noise source is localized in one corner of the room; the acoustic source is positioned in different regions. Simulation results from Table III show that the proposed localization method with cluster selecting outperforms the single ANN method under various SNR conditions.

Additionally, in both noise source environments, when the speaker position lies in different regions in the room, the performances of the proposed localization method are different: when the speaker lies in the central region, the estimated results are best; when the speaker lies in the corner, estimated results are worst.

Acoustic source localization with the proposed method versus the SRP-PHAT method

In this experiment, the proposed localization method was compared with the SRP-PHAT method, Average errors results in Fig. 8 shows that the proposed method is comparable to the SRP-PHAT method in low SNR conditions, whereas it performs better than the SRP-PHAT method as the SNR values becomes larger.

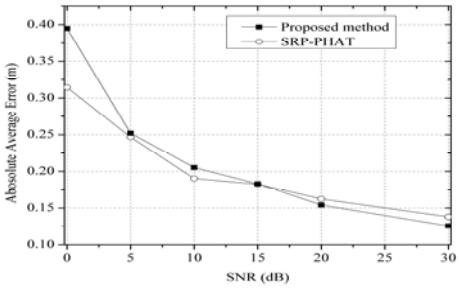

Figure 8. Localization results of the proposed method and the SRP-PHAT method in the background noise environment.

V. CONCLUSION

In this paper, a novel method for acoustic localization in distributed microphone arrays was proposed. In the proposed method, a microphone clustering method is first applied for dividing all the microphones into small clusters and each cluster corresponds to a specified BP neural network. After clustering selecting, for each chosen cluster local position estimation is carries on. Finally, the local estimated positions from the chosen clusters are further fused for robust global acoustic source position.

In the microphone clustering experiment, compared with the Adaptive K-Means++ method, the proposed clustering method may generate more clusters, which means more choices for cluster selecting and localization; meanwhile, the proposed clustering method is not sensitive to the initialization. During the localization experiments, the proposed method with cluster selecting is first compared with the method with one single ANN for all microphones. In the background noise scenario, though the proposed method is not as good as the single ANN method in high SNR conditions, whereas it performs better in low SNR conditions. In the localized noise scenario, the proposed method outperforms the single ANN method under different SNR conditions. Additionally, when the speaker lies in the central region, the estimated results are better than that when the speaker is in other regions of the room. Considering the microphones' placement, we may infer that on the condition of proper microphones' arrangement, the proposed method may yield expected results. In another comparison, the

average localization errors of the proposed localization method are comparable to the SRP-PHAT method in low SNR conditions, and it performs better in high SNR conditions. In addition, the proposed method expects to have less computational cost.

REFERENCES

[1] M. R. Bai, C. Lin, "Microphone array signal processing with application in three-dimensional spatial hearing", The Journal of the Acoustical Society of America, vol. 117, no. 4, pp. 2112-2121, Apr. 2005. [Online]. Available: http://dx.doi.org/10.1121/1.1853242

[2] Z. Zhang, A. G. Andreou, "Slow moving vehicles using the microphone arrays in the Hopkins acoustic surveillance unit", Micro-Nanoelectronics, Technology and Applications, pp. 140-143, Buenos Aires, Argentina, Sept. 2008.

[3] Z. W. Yu, Z. Y. Yu, H. Aoyama, M. Ozeki, Y. Nakamura, "Capture, recognition, and visualization of human semantic interactions in meetings", IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 107-115, Mannheim, German, March-April 2010. [Online]. Available: http://dx.doi.org/10.1109/PERCOM.2010.5466987

[4] C. Zhang, D. Florencio, D. E. Ba, Z. Zhang, "Maximum likelihood sound source localization and beam-forming for directional microphone arrays in distributed meetings", IEEE Transaction on Pervasive Computing and Communications, vol. 10, no. 3, pp. 538-548, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1109/TMM.2008.917406

[5] J. Benesty, Y. Huang, and J. Chen, "Time delay estimation via minimum entropy", IEEE Signal Processing Letters, vol. 14, no. 3, pp. 157-160, Mar. 2007. [Online]. Available: http://dx.doi.org/10.1109/LSP.2006.884038

[6] R. Roy, A. Paulraj, T. Kailath, "Direction-of-arrival estimation by subspace rotation methods - ESPRIT", IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 11, pp. 2495-2498, Tokyo, Japan, Apr. 1986. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.1986.1168673

[7] D. R. Farrier , "Direction of arrival estimation by subspace methods", IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 2651-2654, Albuquerque, New Mexico, USA, Apr. 1990. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.1990.116163

[8] M. Cobos, A. Marti, J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling", IEEE Signal Processing Letters, vol. 18, no. 1, Jan. 2011. [Online]. Available: http://dx.doi.org/10.1109/LSP.2010.2091502

[9] Y. Rui , D. Florencio , W. Lam, J. Su," Sound source localization for circular arrays of directional microphones", IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. iii/93-iii/96, Pennsylvania, USA, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2005.1415654

[10] P. Aarabi, "The fusion of distributed microphone arrays for sound localization", EURASIP Journal on Advances in Signal Processing, pp. 338-347, Jan. 2003. [Online]. Available: http://dx.doi.org/10.1155/S1110865703212014

[11] E. Elahi, "Sound localization and tracking using distributed microphones fusion: maximum likelihood or maximum a-posteriori approach?", The 2nd International Conference on Computer, Control and Communication, pp. 1-6, Karachi, Pakistan, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1109/IC4.2009.4909221

[12] T. Takagi, H. Noguchi, K. Kugata, "Microphone array network for ubiquitous sound acquisition", IEEE International Conference on Acoustics Speech and Signal Processing, pp. 1474-1477, Dallas, Texas, USA, Mar. 2010. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2010.5495490

[13] G. Valenzise, G. Prandi, M. Tagliasacchi, A. Sarti, "Resource constrained efficient acoustic source localization and tracking using a distributed network of microphones", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2581-2584, Las Vegas, USA, March - April 2008. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2008.4518176

[14] C. Knapp, G. Carter, "The generalized correlation method for estimation of time delay", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 24, no. 4, pp. 320-327, Aug. 1976. [Online]. Available: http://dx.doi.org/10.1109/TASSP.1976.1162830

[15] J. S. Hu, C. H. Yang, C. K. Wang, "Estimation of sound source number and directions under a multi-source environment", IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 181-186, St. Louis, USA, Oct. 2009. [Online]. Available: http://dx.doi.org/10.1109/IROS.2009.5354706

[16] D. Arthur, S. Vassilvitskii, "k-means++: the advantages of careful seeding", Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027-1035, PA, USA, 2007.

[17] G. Arslan, F. A. Sakarya, "A unified neural-network-based speaker localization technique", IEEE Transactions on Neural Networks, vol. 11, no. 4, pp. 997-1002, Jul. 2000. [Online]. Available: http://dx.doi.org/10.1109/72.857779

[18] G. Yang, J. Jongdae, S. Donggug, "Sound-source localization system based on neural network for mobile robots", IEEE International Joint Conference on Neural Networks, IEEE World Congress on Computational Intelligence, pp. 3126-3130, Hong Kong, China, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1109/IJCNN.2008.4634240

[19] V. H. Dang, T. P. Phan, B. V. Le, Y. K. Lee, "Clustering based multi-object positioning system", International Conference on Advanced Technologies for Communications (ATC), pp. 40-44, Da Nang, Vietnam, Aug. 2011 [Online]. Available: http://dx.doi.org/10.1109/ATC.2011.6027431

[20] A. Y. Nakano, K. Yamamoto, S. Nakagawa, "Directional acoustic source's position and orientation estimation approach by a microphone array network", IEEE Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, pp. 606-611, Marco Island, Florida, USA, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1121/1.3257548

[21] I. Himawan, I. McCpwam, S. Sridharam, "Clustered blind beamforming from Ad-Hoc microphone arrays", IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 4, pp. 661-676, May. 2011. [Online]. Available: http://dx.doi.org/10.1109/TASL.2010.2055560

[22] T. L. Nwe, H. Sun, B. Ma, H. Li, "Speaker clustering and cluster purification methods for RT07 and RT09 Evaluation meeting data", IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 2, pp. 461-473, Feb. 2012. [Online]. Available: http://dx.doi.org/10.1109/TASL.2011.2159203

[23] M. Souden, K. Kinoshita, M. Delcroix, T. Nakatani, "Distributed microphone array processing for speech source separation with classifier fusion", IEEE International Workshop on Machine Leaning for Signal processing (MLSP), pp. 1-6, Santander, Spain, Sept. 2012. [Online]. Available: http://dx.doi.org/10.1109/MLSP.2012.6349782

[24] H. Chen, C. K. Tse, J. B. Feng, "Minimizing effective energy consumption in multi-cluster sensor network for source extraction", IEEE Transactions on Wireless Communications, vol. 8, no. 3, pp. 1480-1489, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1109/TWC.2008.080319

[25] B. R. Dai, J. W. Huang, M. Y. Yeh, M. S. Chen, "Adaptive clustering for multiple evolving streams", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 9, pp. 1166-1180, Sept. 2006. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2006.137

[26] M. Chen, Z. Liu, L. W. He, P. Chou, Z. Y. Zhang "Energy-based position estimation of microphones and speakers for Ad Hoc microphone arrays", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 22-25, Honolulu, HI, USA, Oct. 2007. [Online]. Available: http://dx.doi.org/10.1109/ASPAA.2007.4393035

[27] E. A. Lehmann, A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model", The Journal of Acoustic Society of America, vol. 124, no. 1, Jun, 2008. [Online]. Available: http://dx.doi.org/10.1121/1.2936367