PAN
ARCHIVES OF ACOUSTICS
Vol. **46**, No. 1, pp. 67–78 (2021)
DOI: 10.24425/aoa.2021.136561

Research Paper

Acoustic Source Localization Using Kernel-based Extreme Learning Machine in Distributed Microphone Array

Rong WANG, Zhe CHEN, Fuliang YIN*

School of Information and Communication Engineering
Dalian University of Technology
Dalian 116023, China
*Corresponding Author e-mail: flyin@dlut.edu.cn

(*received December 11, 2019; accepted September 23, 2020*)

Acoustic source localization using distributed microphone array is a challenging task due to the influences of noise and reverberation. In this paper, acoustic source localization using kernel-based extreme learning machine in distributed microphone array is proposed. Specifically, the space of interest is divided into some labeled positions, and the candidate generalized cross correlation function in each node is treated as the feature mapped into the hidden nodes of extreme learning machine. During the training phase, by the implementation of kernel function, the output weights of the classifier are calculated and do not need to be tuned. After the kernel-based extreme learning machine (K-ELM) is well trained, the measured generalized cross correlation data are fed into the K-ELM classifier, and the output is the estimated acoustic source position. The proposed method needs less human intervention for both training and testing and it does not need to calibrate the node in advance. Simulation and real-world experimental results reveal that the proposed method has extremely fast training and testing speeds, and can obtain better localization performance than steered response power, K-nearest neighbor, and support vector machine methods.

Keywords: extreme learning machine; acoustic source localization; distributed microphone array; generalized cross correlation function.

1. Introduction

The position of acoustic source is widely used in many audio and multimedia applications. It is a challenging task to localize the acoustic source in reverberant and noisy conditions. Distributed microphone array (DMA) is a promising approach for acoustic capture and processing systems. Due to the unconstrained network structure and flexible deployment, DMA has been widely used in video conference systems, speech enhancement, sound source localization, speaker tracking, security monitor, sniper detection, etc. (Hengy *et al.*, 2016; Tian *et al.*, 2015; Wan, Wu, 2013; Zhang *et al.*, 2016).

The existing acoustic source localization methods can be divided into two categories: the indirect and direct methods. The indirect methods usually utilize the range information between the source and the nodes, e.g. time difference of arrival (TDOA), direction of arrival (DOA), and time of arrival (TOA, Canclini *et al.*, 2015; Kan *et al.*, 2015), then the acoustic source position is found by geometrical derivation, that is, by solving a set of equations to compute the intersection point in the space. Traditional signal processing methods such as the least square (LS) and maximum likelihood (ML) can be used to locate the acoustic source position, but they strongly rely on the assumptions of signal models and accurate estimations of TDOA, DOA, and TOA. The key for obtaining an accurate source position while using these methods is the exact estimation of the range parameters. They only work well in relatively moderate environments, i.e. low reverberation and weak background noise conditions, otherwise their performance will degrade rapidly. These indirect methods can localize the acoustic source position by using complex algorithms at the expense of huge computational cost. In addition, the indirect methods for acoustic source localization usually need

to know the node microphone positions in advance, leading to the cumbersome process of node calibration (Crocco *et al.*, 2012; Khanal *et al.*, 2013).

The direct methods include received signal strength (RSS), steered response power (SRP), and neural network methods. The direct acoustic source localization methods usually divide the space of interest into some candidate positions. The acoustic source position is estimated by searching all the candidate positions and finding out the one that best explains the measured data (Ho, 2012; Lim *et al.*, 2015; Nunes *et al.*, 2014). Compared with the indirect localization methods, the direct localization method does not need to locate the nodes in advance, which helps avoid the cumbersome calculation of node calibration. Nakano *et al.* (2009) proposed an artificial neural network (ANN) method to obtain the position and orientation of an acoustic source, where the best combination of parameters of time difference and microphone positions are regarded as the input of the ANN, and the position and orientation of a directional acoustic source are treated as the output. Zhang *et al.* (2013) proposed an acoustic source localization method based on microphone clustering and back-propagation (BP) network, where the clustering technique is first used to divide the microphones into several clusters, then the microphone clusters close to the acoustic source are selected by energy, and finally, the TDOAs in the selected microphone clusters are fed into the BP network for source position estimation. These two direct localization methods can estimate the acoustic source position, but need more complex preprocessing for the input of ANNs. Xiao *et al.* (2015) proposed a learning-based approach to estimate the DOA in noisy and reverberant environments by an 8-channel circular array. The features are first extracted from the generalized cross-correlation (GCC) vectors, and a multilayer perception neural network is used to learn the nonlinear mapping from such features to the DOA. In recent years, researchers investigated a variety of acoustic source location methods based on deep learning. The convolutional neural network (CNN) framework is often used to obtain the source position (Ferguson *et al.*, 2018; Salvati *et al.*, 2018). In (Vera-Diaz *et al.*, 2018), the acoustic source is localized based on CNN approach by directly using the audio signal as the input information. These learning-based methods need many parameters to be tuned, which is cumbersome and takes a long time to train; moreover, the localization accuracy is not good enough. To overcome the weaknesses of huge time consuming and large human intervention in a node, acoustic source localization using kernel-based extreme learning machine (K-ELM) in distributed microphone array is proposed in this paper.

Extreme learning machine (ELM) is a single-hidden layer feedforward neural network (SLFN), where learning is made without iterative tuning (Huang *et al.*, 2012). It can be used for classification and regression. The direct localization methods usually divide the space of interest into some candidate positions, and find the one that best explains the measured data. Acoustic source localization can be viewed as a multi-classification problem. Considering the fact that ELM can be used for classification at extremely fast speed, it is exploited to locate the acoustic source. Different from the traditional feedforward neural network theories that all the parameters of the network need to be turned to minimize the cost function, ELM shows that the hidden node parameters can be initialized randomly and the output weights can be analytically determined by using the least square method (Principi *et al.*, 2015). The learning speed of the ELM is significantly faster than those of traditional learning methods (Kongsorot *et al.*, 2019). As the hidden layer feature mapping is unknown in advance, the kernel-based ELM in (Huang *et al.*, 2012) is used to estimate the acoustic source position in this paper.

The localization performance can be improved both in the feature extraction procedure and in the classifier (Gu *et al.*, 2015). A good feature is important to obtain better performance. TDOA is a characteristic parameter reflecting the relationship between the acoustic source and the node. The GCC is a popular method for calculating the time delay between the microphones in a node. Generally, the real time delay corresponds to the largest peak of GCC under ideal conditions, but may be no longer the largest one due to the influences of noise and reverberation. As the GCC contains the information about the position relationship of source and nodes, it is a good feature to map to the acoustic source position.

In this paper, the extreme learning machine is exploited to locate acoustic source, and the acoustic source localization using kernel-based extreme learning machine in DMA is proposed. Each node in the DMA is only equipped with one pair of microphones and does not need to be calibrated in advance. First, the space of interest is split into some candidate positions. The candidate GCCs calculated in each node are concatenated and chosen as a feature to map the acoustic source position. Then the K-ELM classifier is trained and the optimal parameters are chosen by cross-validation. Finally, the measured data is fed into the trained K-ELM classifier and the output is the estimated acoustic source position. The proposed method has a lower time consumption and needs less human intervention for training and testing. Besides, it can obtain enough localization accuracy and is robust against noise and reverberation.

The rest of the paper is structured as follows. Section 2 presents the related works of GCC function and K-ELM. Section 3 describes the proposed acoustic source localization method using K-ELM. In Sec. 4, the results of the simulation and real-world experiments

are presented and discussed. Finally, some conclusions are drawn in Sec. 5.

2. Generalized cross-correlation function and kernel-based extreme learning machine

2.1. Generalized cross-correlation function

The generalized cross-correlation function has been widely used for estimating time difference of arrival, and it is the basis for many localization algorithms.

Considering the direct path, the signal emitted from an acoustic source in the presence of noise at two spatially separated microphones in the i-th node can be mathematically modeled as (Knapp, Carter, 1976)

$$\begin{cases} x_{i,1}(t) = s(t) + n_{i,1}(t), \\ x_{i,2}(t) = \alpha s(t+\tau) + n_{i,2}(t), \end{cases} \tag{1}$$

where $s(t)$ is the source signal received in the microphone one of the i-th node, α is the attenuation coefficient relative to the first microphone, τ is the time difference between the two microphones in the i-th node, and $n_{i,1}(t)$ and $n_{i,2}(t)$ are real, jointly stationary noises. Signal $s(t)$ is assumed to be uncorrelated with noises $n_{i,1}(t)$ and $n_{i,2}(t)$.

The GCC function of the signals recorded by two microphones in the i-th node is expressed as

$$R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Psi_{i,1,2}(\omega) X_{i,1}(\omega) X_{i,2}^*(\omega) e^{j\omega\tau} \, \mathrm{d}\omega, \tag{2}$$

where $X_{i,1}(\omega)$ and $X_{i,2}(\omega)$ denote the Fourier transform of signals $x_{i,1}(t)$ and $x_{i,2}(t)$ recorded by microphones 1 and 2 in the i-th node, respectively; ω is the angular frequency, and $[\cdot]^*$ stands for the complex conjugate operation. The weighting function $\Psi_{i,1,2}(\omega)$ is designed to optimize a given performance criteria.

Different functions were proposed in the literature, and among all of them, the phase transform (PHAT), defined as

$$\Psi_{i,j}(\omega) = \frac{1}{|X_i(\omega) X_j^*(\omega)|}, \tag{3}$$

has been found to perform well for acoustic localization in reverberant environments, leading to the GCC-PHAT method (Knapp, Carter, 1976).

2.2. Brief description of ELM and K-ELM

ELM was originally developed for the feedforward neural networks (SLFNs) and then extended to the "generalized" SLFNs. It has been proven to provide good generalization performance at extremely fast learning speed (Huang *et al.*, 2006). The following is a brief description of ELM given by (Huang *et al.*, 2006).

The ELM architecture is composed of three parts: input, hidden, and output layers. Different from the common understanding of learning, the hidden layer of ELM does not need to be tuned. One of the characteristics of ELM is that all the parameters of the hidden layer can be randomly generated and can be independent of the training samples (Huang *et al.*, 2011). The minimal norm least square method instead of the standard optimization method was used in the implementation of ELM (Huang *et al.*, 2006). Assume there are K arbitrary samples $(\mathbf{x}_j, \mathbf{t}_j)$, where $\mathbf{x}_j = [x_{j1}, x_{j2}, ..., x_{jm}]^{\mathrm{T}} \in \mathbb{R}^m$ and $\mathbf{t}_j = [t_{j1}, t_{j2}, ..., t_{jn}]^{\mathrm{T}} \in \mathbb{R}^n$, here $j = 1, 2, ..., K$; m and n are the number of input nodes and output nodes, respectively. The standard ELM with L hidden neurons can approximate these K samples with zero error such that

$$\sum_{i=1}^{L} \beta_i \mathbf{G}(\mathbf{w}_i, b_i, \mathbf{x}_j) = \mathbf{t}_j, \quad j = 1, 2, ..., K, \tag{4}$$

where $\mathbf{w}_i = [w_{i1}, w_{i2}, ..., w_{im}]^{\mathrm{T}}$ is the weight vector connecting the input nodes and the i-th hidden node, b_i is the threshold of the i-th hidden node, $\mathbf{G}(\mathbf{w}_i, b_i, \mathbf{x}_j)$ is the output of the i-th hidden node in terms of $\mathbf{x}_j$, $\beta_i = [\beta_{i1}, \beta_{i2}, ..., \beta_{in}]^{\mathrm{T}}$ is the weight vector connecting the i-th hidden node and the output nodes. Equation (4) can be rewritten as

$$\mathbf{H}\beta = \mathbf{T}, \tag{5}$$

where

$$\begin{aligned} \mathbf{H} &= [\mathbf{h}(\mathbf{x}_1), ..., \mathbf{h}(\mathbf{x}_K)]^{\mathrm{T}} \\ &= \begin{bmatrix} \mathbf{G}(\mathbf{w}_1, b_1, \mathbf{x}_1) & \cdots & \mathbf{G}(\mathbf{w}_L, b_L, \mathbf{x}_1) \\ \vdots & \cdots & \vdots \\ \mathbf{G}(\mathbf{w}_1, b_1, \mathbf{x}_K) & \cdots & \mathbf{G}(\mathbf{w}_L, b_L, \mathbf{x}_K) \end{bmatrix}, \end{aligned} \tag{6}$$

$$\beta = [\beta(\mathbf{x}_1), ..., \beta(\mathbf{x}_K)]^{\mathrm{T}} \tag{7}$$

and

$$\mathbf{T} = [\mathbf{t}(\mathbf{x}_1), ..., \mathbf{t}(\mathbf{x}_K)]^{\mathrm{T}}. \tag{8}$$

As the hidden nodes can be randomly generated, the only unknown parameter is the output weight vector between the hidden layer and the output layer, which can simply be resolved by ordinary least square directly. The least square solution of the output weight vectors can be solved as

$$\widehat{\beta} = \mathbf{H}^{\dagger}\mathbf{T}, \tag{9}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse of matrix $\mathbf{H}$. In order to improve the stability of ELM, when $\mathbf{H}\mathbf{H}^{\mathrm{T}}$ is non-singular, we can have

$$\beta = \mathbf{H}^{\mathrm{T}} \left(\frac{\mathbf{I}}{\lambda} + \mathbf{H}\mathbf{H}^{\mathrm{T}} \right)^{-1} \mathbf{T} \tag{10}$$

where λ is a positive value. The corresponding output function of ELM is

$$f(\mathbf{x}) = \mathbf{h}(\mathbf{x})\beta = \mathbf{h}(\mathbf{x})\mathbf{H}^{\mathrm{T}}\left(\frac{\mathbf{I}}{\lambda} + \mathbf{H}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{T}. \tag{11}$$

Huang *et al.* (2012) also studied the kernel-based ELM. If the hidden layer feature mapping $\mathbf{h}(\mathbf{x})$ is unknown, instead its corresponding kernel $K(u,v)$, the kernel matrix for ELM is denoted as

$$\Omega = \mathbf{H}\mathbf{H}^{\mathrm{T}} : \Omega_{i,j} = \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j). \tag{12}$$

In this special kernel implementation of ELM, the feature mapping $\mathbf{h}(\mathbf{x})$ does not need to be known to users. It is assumed as a radial basis function (RBF) in this work, i.e. $K(u,v) = \exp(-\gamma\| u - v\|^2)$, where γ is the kernel parameter. The output function of the kernel-based ELM can be written compactly as

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{h}(\mathbf{x})\mathbf{H}^{\mathrm{T}}\left(\frac{\mathbf{I}}{\lambda} + \mathbf{H}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{T} \\ &= \begin{bmatrix} K(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ K(\mathbf{x}, \mathbf{x}_K) \end{bmatrix}^{\mathrm{T}} \left(\frac{\mathbf{I}}{\lambda} + \Omega\right)^{-1}\mathbf{T}. \end{aligned} \tag{13}$$

The kernel function can be applied to train the kernel-based ELM model when the hidden layer feature mapping $\mathbf{h}(\mathbf{x})$ is unknown. The K-ELM algorithm does not need to consider the number of hidden nodes and only concerns the selection of the kernel function and the input data (Cheng *et al.*, 2019). Let $\mathbf{W}^k$ denote the output weight, which is expressed as

$$\mathbf{W}^k = \left(\frac{\mathbf{I}}{\lambda} + \Omega\right)^{-1}\mathbf{T}. \tag{14}$$

It can be obtained during the training phase. After the model training is completed, the trained K-ELM classifier can be used for online classification. The measured input data are fed into the classifier and the output can be expressed as

$$f(\mathbf{x}) = \begin{bmatrix} K(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ K(\mathbf{x}, \mathbf{x}_K) \end{bmatrix}^{\mathrm{T}} \mathbf{W}^k. \tag{15}$$

As the feature mapping is usually unknown to users, the kernel-based ELM is suitable to real applications and is used in this paper.

3. Acoustic source localization using K-ELM

3.1. Problem formulation

In order to estimate the acoustic source position, the area of interest is divided into some grids. The goal is to find the position that best explains the GCC feature concatenated from the nodes.

There are N nodes in a DMA placed around the region of interest, where each node contains a pair of synchronized microphones, the intra-node distance L_j is known, and the node positions may be unknown. Assume that the coordinate of the unknown source position is $S = [x, y, z]^{\mathrm{T}}$, the coordinate of the j-th node is $P_j = [x_{j,0}, y_{j,0}, z_{j,0}]^{\mathrm{T}}$, the h-th microphone in the j-th node is is $P_{j,h} = [x_{j,h}, y_{j,h}, z_{j,h}]^{\mathrm{T}}$, the node index $j = 1, 2, ..., N$; the microphone index in each node $h = 1, 2$. The distances from S to $P_{j,1}$ and $P_{j,2}$ are denoted as $r_{j,1}$ and $r_{j,2}$, respectively. The acoustic source and node positions are shown in Fig. 1. The sampling frequency is f_s, the sound speed is c, the received signal in microphone $P_{j,h}$ is $f_{j,h}(t)$, and the GCC in the j-th node is $R_j(\omega)$.

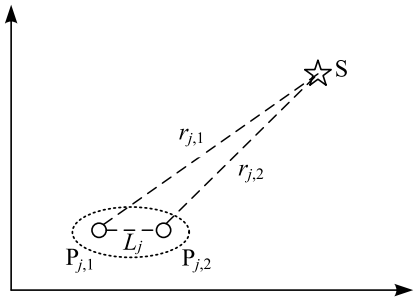

Fig. 1. Acoustic source and node microphone positions.

3.2. Localization by K-ELM

The TDOA reflects the relationship between the acoustic source and node microphone positions. The GCC used for estimating the TDOA is treated as a feature to map to the source position. According to the triangular relationship, the absolute value of the range difference between $r_{j,1}$ and $r_{j,2}$ should be less than L_j. Hence, the candidate TDOAs should satisfy the following constraint, i.e.

$$\tau \in \{-\tau_{\max}, \tau_{\max}\}, \tag{16}$$

where τ is the candidate TDOA, and $\tau_{\max} = L_j f_s / c$ is the maximum possible TDOA of the j-th node. Accordingly, the GCCs correspond to the candidate TDOAs at the j-th node satisfying the following constraint, i.e.

$$R_j^c \in \{R_{j,\tau}\}, \tag{17}$$

where R_j^c is the set of candidate GCCs at the j-th node, and $R_{j,\tau}$ is the correlation coefficient corresponding to the TDOA in set τ. Hence, the $2\tau_{\max} + 1$ correlation coefficients, i.e., candidate GCCs, contain useful information of TDOA. Let us define

$$\mathbf{R}^c = \mathrm{vec}\left(\left\{R_j^c\right\}\right), \tag{18}$$

where the vec operation denotes the concatenation of candidate GCCs of all nodes. The concatenated GCCs $\mathbf{R}^c$ is treated as an input and fed into the network.

The K-ELM based acoustic source localization consists of two phases: offline training and online localization. The acoustic source region of interest is first divided into some grids, where each grid is labeled by its coordinate. The set of grid points is denoted as S_g, and the coordinate of grid points is denoted as S_i^g, $i = 1, ..., |S_g|$, where $|\cdot|$ denotes the cardinality of a set and the number of the grid points is $|S_g|$. The source localization boils down to finding the most probable point from all the labeled points. The nodes in DMA are placed and fixed at some positions. For convenience, the nodes are usually placed around the concerned region.

During the offline training phase, the acoustic source signals are played at each labeled point in different noise and reverberation conditions. The received signals from all the labeled positions at the j-th node are expressed as $f_{j,h}^g(t)$, and are treated as training audio data. The input of the K-ELM is the concatenated candidate GCCs calculated from the training audio data, and the expected output of the K-ELM classifier is the corresponding coordinate of the acoustic source. Multiple candidate GCC vectors calculated in each node are generated to form a training sample vector. After that, all the training samples calculated from the training audio data are treated as input to feed into the K-ELM model. The training data set is $\Phi_g = \{p_m, t_m | m = 1, ..., M\}$, where $p_m = \mathbf{R}^c$ is the m-th column of the input matrix $\mathbf{P}$, $t_m = S_m^g$ is the m-th column of the output matrix $\mathbf{T}$, and M is the number of samples in the training set. As the cross-validation can be used to assess the performance of the classifier when the trained system is generalized to an independent data set, the parameters λ and γ are selected by the cross-validation technique (Stone, 1974). Then, the output weight of K-ELM classifier can be calculated by Eq. (14). After the K-ELM classifier has been well trained, it can be used for online source localization.

During the online localization phase, the concatenated candidate GCCs are calculated from the received audio data and fed into the trained K-ELM classifier. The output of the K-ELM classifier is the estimated source position and can be obtained by Eq. (15). The detailed steps of acoustic source localization with K-ELM are summarized in Algorithm 1.

Algorithm 1: Acoustic source localization by using K-ELM.

N nodes of a DMA are placed around the region of interest. The region of interest is divided into some grids, and labeled by the coordinate S_i^g, $i = 1, ..., |S_g|$. The training audio data of the received signals from the labeled positions at the j-th node are $f_{j,h}^g(t)$.

Step 1: Construct the training set and choose the kernel parameters.

(1) Calculate the candidate TDOAs of $f_{j,h}^g(t)$ under different SNR and reverberation conditions by using GCC function at each node

$$R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Psi_{j,1,2}(\omega) X_{j,1}(\omega) X_{j,2}^*(\omega) e^{j\omega\tau} \, \mathrm{d}\omega,$$

where

$$\Psi_{j,1,2}(\omega) = \frac{1}{|X_{j,1}(\omega) X_{j,2}^*(\omega)|}.$$

Concatenating the candidate GCCs of all nodes, we have

$$\mathbf{R}^c = \mathrm{vec}(\{R_j^c\}),$$

where

$$R_j^c \in \{R_{j,\tau}\},\ \tau \in \{-\tau_{\max}, \tau_{\max}\},\ \tau_{\max} = L_j f_s / c,\ j = 1, 2, ..., N,$$

(2) Construct the training data set $\Phi_g = \{p_m, t_m | m = 1, ..., M\}$, where $p_m = \mathbf{R}^c$ is the m-th column of input matrix $\mathbf{P}$, $t_m = S_m^g$ is the m-th column of output matrix $\mathbf{T}$, and M is the number of training samples.

(3) The input of the classifier is $\mathbf{R}^c$, and the output is the corresponding acoustic source position. Let the RBF kernel $K(u, v) = \exp(-\gamma \|u - v\|^2)$ be the kernel function of K-ELM. Choose the optimal parameters λ and γ by cross-validation technique.

Step 2: Train K-ELM classifier.

Input: Φ_g, λ, γ and $K(u, v) = \exp(-\gamma \|u - v\|^2)$.

Output: $\mathbf{W}^k$.

$$\Omega = \mathbf{H}\mathbf{H}^{\mathrm{T}} : \Omega_{i,j} = \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j),$$

$$\mathbf{W}^k = \left(\frac{\mathbf{I}}{\lambda} + \Omega\right)^{-1} \mathbf{T}.$$

Step 3: Online acoustic source localization.

The measured concatenated candidate GCCs $\mathbf{x}$ is fed into the trained K-ELM classifier, and the output is the estimated acoustic source position.

$$f(\mathbf{x}) = [K(\mathbf{x}, \mathbf{x}_1), ..., K(\mathbf{x}, \mathbf{x}_M)] \mathbf{W}^k.$$

The extreme learning machine based acoustic source localization method in a distributed microphone array can obtain a good localization performance with an extremely fast learning speed. Moreover, it does not need to calibrate the node positions of the microphone array in advance, and is robust against noise and reverberation.

4. Experimental results and discussion

In order to evaluate the performance of the proposed acoustic source localization method based on K-ELM, the simulation and real-world experiments are carried out to compare different localization methods. The steered response power is a rather common source localization method due to its robustness against reverberation and noise. The classical steered response power (C-SRP) method can be implemented in two steps (Lima *et al.*, 2015):

(1) computation of the GCC function between the signals acquired by each microphone pair;

(2) exhaustive search for the source location over a grid of points.

The K-nearest neighbor (KNN) classification is one of the most fundamental and simple classification methods, and support vector machine (SVM) is one of the most popular classifiers. They both have been widely used in machine learning applications. Hence, the K-ELM acoustic source localization is compared with C-SRP, KNN, and SVM methods. Variations of the simulated data are made with a different number of training samples, reverberation times, and SNRs. Besides, the computational times of these compared methods are given to evaluate the efficiency. The simulation and real-world experiments are performed and a discussion of the results is presented in this section.

4.1. Simulation setup

The Image model has been widely used due to its ability to simulate the indoor sound field (Allen, Berkley, 1979). Hence, in this simulation, a reverberant room of $6 \times 5 \times 3$ m is simulated with the Image model. The room setup is shown in Fig. 2, which is composed of 8 nodes, and each node consists of two microphones with a known distance of 0.2 m. The source positions are confined to a 4.5×3.5 m region in the center of the room, and the height of the source is set at 1.6 m. The source positions are divided into two sets: the training source positions for training the K-ELM model and the testing source positions to evaluate the localization performance. The 63 training source locations are labeled by their coordinates and uniformly selected, such that the space between adjacent locations is 0.5 m. The training sample set consists of the concatenated candidate GCCs from the 63 labeled source

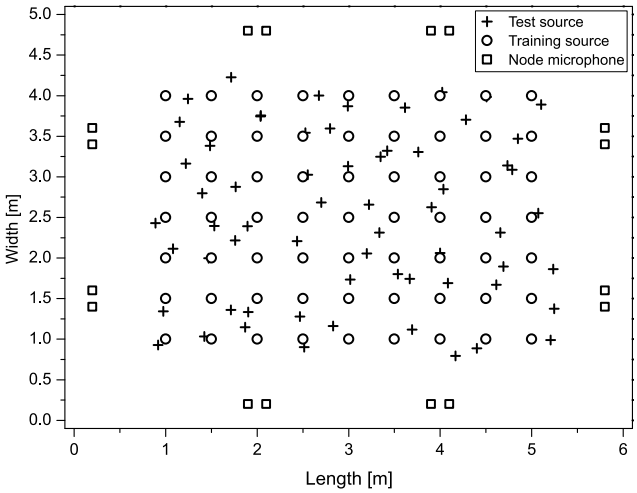

Fig. 2. Acoustic source and node microphone layout.

positions. Besides, there are 63 testing sources with additional deviations in the same region, where the deviation of the testing source positions is randomly generated with a uniform distribution. The positions of testing source are the sum of training source positions and the corresponding biases. The sampling frequency is 48 kHz and the frame size L_f is 1024 samples, about 21 ms. The reverberation time T_{60} is set from 0.1 s to 0.6 s, and the SNR is set from 5 dB to 30 dB. The acoustic source is a female voice. All the evaluations are carried out in MATLAB (2014a) environment running on an Intel Core $i7$, 2.2G CPU with 16G RAM.

The root-mean-square error (RMSE), classification accuracy, and computational time are used to evaluate the performance of the acoustic source localization methods. Let J denotes the number of testing sources. The RMSE reflects the difference between the real location coordinates $\mathbf{x}$ and the estimated location coordinates $\widehat{\mathbf{x}}$, and is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J} \left\|\mathbf{x}_j - \widehat{\mathbf{x}}_j\right\|^2}. \tag{19}$$

The classification accuracy is also used to measure the localization accuracy of the system. It is treated as a correct classification while the estimated testing source position is classified into the nearest labeled training source position. Assume that the number of correct classifications is Z_r, and the total number of testing samples is Z. The classification accuracy r is expressed as

$$r = \left(\frac{Z_r}{Z}\right) \cdot 100\%. \tag{20}$$

4.2. Synthetic data

The simulated data is synthesized by convolving clean speech signals with the room impulse responses

(RIRs) measured from the array based on the Image method, and the supplementary noises are added to the received signals in each node. The acoustic source produces a phonetically rich speech signal of 0.23 s in each training acoustic source position. The training data is the candidate GCCs of the received signals in each node from the labeled source positions. The testing data is the candidate GCCs of the received signals in each node from the test source positions. In order to evaluate the performance of the proposed method, two kinds of training sets are used to train the system, where one training set includes the candidate GCCs of received signals in some given SNR cases; the other training data set includes the candidate GCCs under different SNRs. As the silence portions are not suitable to train the K-ELM model, only the GCC vectors obtained from the speech segments are used for training.

4.3. Model construction

In K-ELM, the number of input neurons is equal to the number of features in the input data, i.e., 440 in our case. The training set consists of the candidate GCCs of the received signals from 63 different labeled positions, and the labeled source coordinates are treated as output. Hence, the number of output neurons is equal to the number of classes, i.e., 63 in this experiment. Since the RBF kernel provides a lower error for the correct maps (Salvati *et al.*, 2016), it is used as activation function for K-ELM in all experiments.

According to Huang *et al.* (2012), the accuracy can be improved with the regularization factor λ and kernel parameter γ, which can optimize the architecture of the learning model. Hence, the parameters λ and γ should be selected properly. The training data are divided into two subsets with 80% of the data for training and 20% for testing, with no overlap. The cross-validation technique is used to determine user-specified parameters λ and γ. Then the best parameter setting of λ and γ with optimal performance can be selected. To tradeoff between accuracy and computation time, the parameters λ and γ are set to $\lambda = 0.2$ and $\gamma = 1$, and are used in the next simulation experiments.

The classical steered response power (C-SRP) method searches the grid of predefined spatial points to estimate an acoustic source position, as shown in Fig. 2, where each training source position represents the spatial point.

The parameters of K-nearest neighbor (KNN) classification method include the number of nearest neighbors and the type of measure distance. In the following experiments, the number of nearest neighbors and the type of measure distance are set to three and 2-norm, respectively.

The RBF kernel is used for the support vector machine (SVM) classifier in the subsequent tests. The likelihood parameter μ and kernel parameter γ are obtained by using a cross-validation technique. Hence, a good choice is given by setting $\mu = 0.5$ and $\gamma = 1.2$ with the cross-validation rates of 80% and 20%. This setup is used in the following experiments. The SVM classifier is implemented using '*libsvm*' software package (Chang, Lin, 2011).

4.4. Comparison of C-SRP, KNN, SVM, and K-ELM method

To evaluate the performance of the localization methods in noisy conditions, the experiments of four methods with a different number M of training samples when T_{60} is 0.3 s are carried out, and the localization results are shown in Figs 3 and 4.

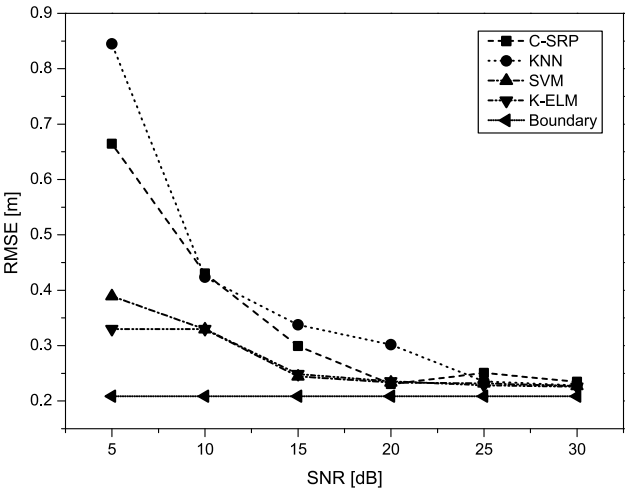

Fig. 3. RMSE of different methods for different SNRs with $T_{60} = 0.3$ s and $M = 315$.

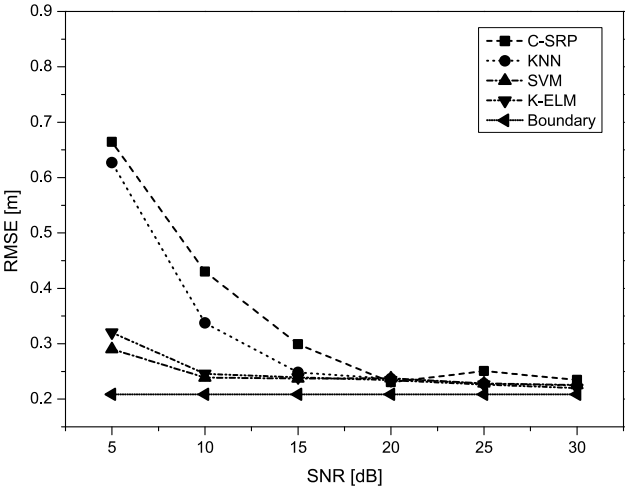

Fig. 4. RMSE of different methods for different SNRs with $T_{60} = 0.3$ s and $M = 630$.

By observing Figs 3 and 4, we can see that the RMSEs of the learning-based localization methods decrease with the increasing number of training samples. When $M = 315$, the C-SRP method performs slightly better than KNN method, while when $M = 630$, the

RMSE of KNN decreases and is lower than that of C-SRP method. The K-ELM method can achieve comparable performance with SVM method, and the RMSE of K-ELM decreases to the boundary gradually.

To evaluate the localization methods in reverberation conditions, when SNR is 15 dB, the compared methods for different reverberation times and number M of training samples are executed and the results are shown in Figs 5 and 6.

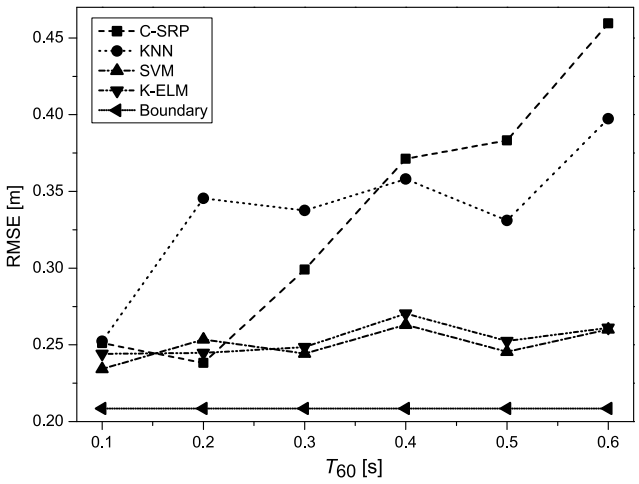

Fig. 5. RMSE of different methods for different T_{60}s with SNR = 15 dB and M = 315.

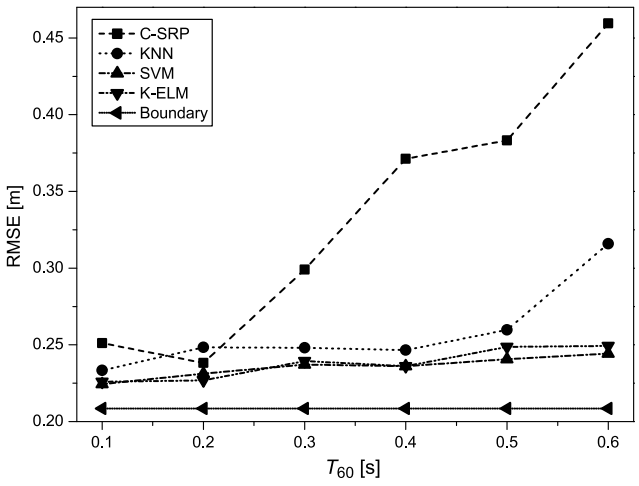

Fig. 6. RMSE of different methods for different T_{60}s with SNR = 15 dB and M = 630.

By comparing Fig. 5 and Fig. 6, we can observe that the RMSE of KNN decreases with the increasing number of training samples. When the reverberation time is less than 0.2 s, the C-SRP method can localize the source position with small error. With the increase of T_{60}, the localization error becomes larger gradually. The performance of these learning-based methods has small fluctuations. Among them, the RMSEs of K-ELM and SVM are comparable and have similar trends for different T_{60}s, and they both have better performance than that of KNN.

The classification accuracy is another metric to evaluate the localization performance. When T_{60} = 0.3 s, SNR = 15 dB, and M = 630, the RMSE and testing accuracy for the compared methods are given in Table 1.

Table 1. RMSE and classification accuracy for the compared methods with T_{60} = 0.3 s, SNR = 15 dB, and M = 630.

Metric	C-SRP	KNN	SVM	K-ELM
RMSE [cm]	29.91	24.81	23.72	23.95
Testing accuracy [%]	66.67	63.49	73.02	69.84

As seen from Table 1, the RMSE of KNN is lower than that of C-SRP method, and K-ELM outperforms KNN. Moreover, K-ELM has comparable performance with SVM. The testing accuracy of K-ELM is slightly lower than that of SVM and higher than that of C-SRP and KNN methods.

One advantage of K-ELM with respect to other methods is its extremely fast training speed. In order to evaluate the execution efficiency of the compared methods, the training time and testing time of these methods are given in Tables 2 and 3.

Table 2. Training and testing time of the compared methods with SNR = 15 dB, T_{60} = 0.3 s and M = 630.

Metric	C-SRP	KNN	SVM	K-ELM
Training time [ms]	–	–	1417.5	389.0
Testing time [ms]	593.7	167.8	54.5	1.1

Table 3. Training and testing time of the compared methods with SNR = 15 dB, T_{60} = 0.5 s and M = 630.

Metric	C-SRP	KNN	SVM	K-ELM
Training time [ms]	–	–	1498.8	382.5
Testing time [ms]	592.3	172.2	54.9	1.6

From Tables 2 and 3, it can be seen that, when SNR = 15 dB and M = 630, among these learning-based methods, the testing time of KNN is the longest for different T_{60}s, next is the testing time of SVM, and K-ELM has a significantly less testing time consumption. In addition, the training time of K-ELM is far shorter than that of SVM. Hence, the proposed K-ELM acoustic source localization method can obtain a high computational speed in real applications.

To evaluate the performance of source localization methods in noisy and reverberant conditions, the training samples generated from different SNRs are used for training the K-ELM system. To obtain good localization performance under different scenarios, a source signal with a length of 0.128 s is used for testing. The average RMSEs of the compared methods for different T_{60}s and number M of training samples are shown in Fig. 7. The simulation results with T_{60} = 0.3 s are listed in Tables 4 and 5.

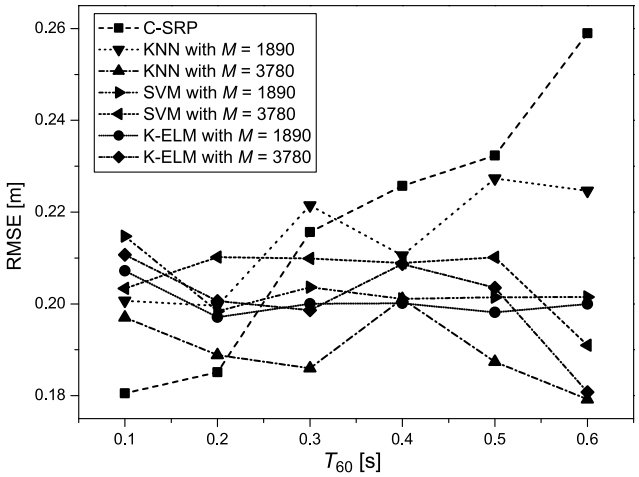

Fig. 7. RMSEs of the compared methods for different T_{60}s when the numbers of training samples M are 1890 and 3780.

Table 4. RMSE, testing accuracy and testing time of the compared methods with T_{60} = 0.3 s and M = 1890.

Metric	C-SRP	KNN	SVM	K-ELM
RMSE [cm]	21.57	22.15	20.36	20.00
Testing accuracy [%]	60.84	67.46	76.98	78.04
Testing time [ms]	3593.5	2353.6	882.6	7.5

Table 5. RMSE, testing accuracy and testing time of the compared methods with T_{60} = 0.3 s and M = 3780.

Metric	C-SRP	KNN	SVM	K-ELM
RMSE [cm]	21.57	18.60	20.99	19.86
Testing accuracy [%]	60.84	72.22	78.04	78.84
Testing time [ms]	3593.5	5289.4	1764.9	8.6

According to Fig. 7, by using a longer source signal, the average RMSEs of all the methods decrease. Specifically, the RMSE of C-SRP rises with the increase of T_{60}. When M = 1890, the RMSE of C-SRP is lower than that of KNN with $T_{60} < 0.3$ s, but higher than that of KNN with $T_{60} > 0.3$ s. When M = 3780, the RMSE of KNN under different T_{60}s is lower than that when M = 1890. Compared with SVM, the RMSE of K-ELM is slightly lower than that of SVM in a different number of training samples and T_{60}s except when T_{60} = 0.1 s. The RMSE of SVM with M = 1890 is slightly lower than that with M = 3780 when T_{60} is in the range of 0.2 s to 0.5 s. The RMSE of K-ELM with M = 1890 is similar to that with M = 3780, and it has a small fluctuation for different T_{60}s, which shows that the proposed K-ELM source localization method can obtain enough localization accuracy when the number of training samples M is around 1890.

Localization accuracy and time cost are important measurement indexes for localization methods. From Tables 4 and 5, we can see that the testing time of KNN with M = 3780 is almost twice long with M = 1890. KNN has to calculate similarity sample by sample, which leads to its large time consumption. Moreover, the computation time of KNN is proportional to the number of samples. On the other hand, K-ELM has the least testing time consumption. By the implementation of kernel function, the hidden layer mapping can be unknown and the output weight is easy to calculate and does not need to be tuned. Compared with SVM, K-ELM method can acquire slightly higher testing accuracy with a significantly shorter testing time. Considering the localization accuracy and time consumption, K-ELM has a better performance than other compared methods.

4.5. Real-world experiments

The real-world experiments are carried out in a conference room of a size of $8 \times 6 \times 3$ m, where a rectangular area of the size of 4.5×3.5 m is used to conduct the following experiments as shown in Fig. 8. The layout of the node microphones and the source positions used for training and testing in real-world experiments are given in Fig. 9.

Fig. 8. Real-world experimental environment.

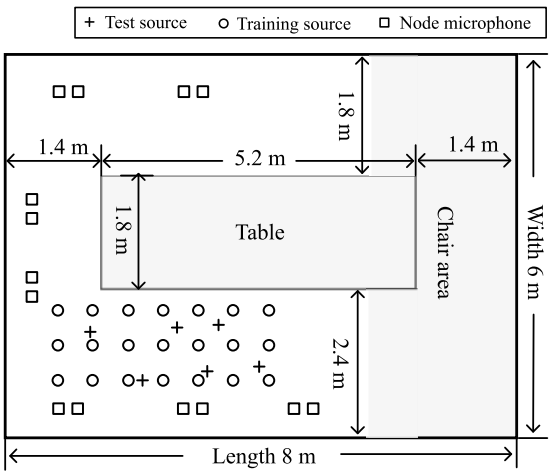

Fig. 9. Layout of node microphones and source positions used for training and testing in real-world experiments.

The distributed microphone network comprises seven pairs of microphones (Model: GK-1000) at the height of 1.2 m, and the intra-node distance is 0.2 m. The acoustic source is a sphere loudspeaker (Model: BSWA-OS002) driven by a power amplifier (Model: FEILE USB-180M). There are 21 training source positions which are labeled by their coordinates, and the space between the adjacent positions is 0.5 m. Besides, 6 testing source positions are randomly chosen from the area of interest and used for testing. The parameters λ and γ used in K-ELM are set as $\lambda = 1$ and $\gamma = 10$. And the parameters of SVM are set to $\mu = 1$ and $\gamma = 0.05$. The number of nearest neighbors and the type of measure distance parameters in the KNN method are set to 3 and 1-norm, respectively. The other configurations are the same as the simulation. The source signal is taken from the TIMIT database and the sampling frequency is 48 kHz. The audio signals for training and testing are collected by a data acquisition card (Model: USB-1608FS-Plus). The signal with a length of 1.07 s is played by the sound source at each training position, and the testing audio signal with a length of 0.21 s is emitted at each testing position. All the received signals are recorded at each node microphone. The reverberation time of the conference room is about 0.45 s, which is measured as the 60 dB decay period for the energy of a high-level white noise signal emitted by a loudspeaker (Model: BSWA-OS002) after it is shut down. The ambient noise mainly comes from the power amplifier and the air conditioner, and the noise level measured by a sound pressure meter (Model: TES1357) is 42 dB (A-weight).

The localization results in the real-world experiments are depicted in Fig. 10. From the results it can be seen that the proposed method can localize the acoustic source successfully in real-world noisy and reverberant environments. Compared with other methods, K-ELM can obtain the smallest RMSE at the fastest speed.

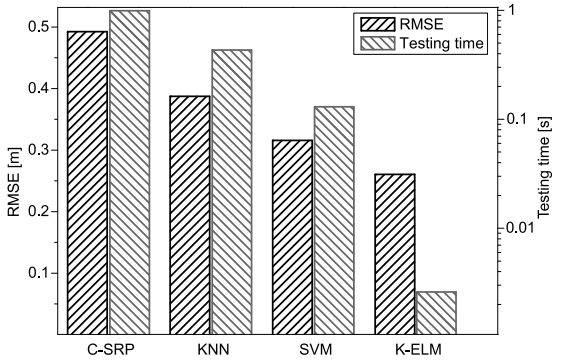

Fig. 10. Localization results in real-world experiments.

5. Conclusion

The acoustic source localization based on K-ELM in distributed microphone array is proposed in this paper. The space of interest is first divided into some labeled positions, and the candidate GCCs calculated in each node are concatenated and treated as a feature to be fed into the K-ELM system, and then mapped into the source position. The output weight is calculated by implementing the kernel function and does not need to be tuned. After the K-ELM model is well trained, the acoustic source position is estimated by feeding the measured candidate GCCs into the K-ELM classifier. This method can obtain a good localization performance with an extremely fast learning speed. Besides, it does not need to know the node positions in advance and is robust against noise and reverberation. Simulation and real-world experimental results indicate that compared with C-SRP, KNN, and SVM methods, the proposed method gives a better performance in terms of both the computational efficiency and the localization accuracy. In the future, the multiple sound source localization methods will be studied. Besides, the deep neural network and semi-supervised learning can be used for acoustic source localization. The multi-modal based sound source location and tracking by fusing audio and video information is also an important issue.

Acknowledgement

This work was supported by National Natural Science Foundation of China (Nos. 61771091, 61871066), National High Technology Research and Development Program (863 Program) of China (No. 2015AA016306), Natural Science Foundation of Liaoning Province of China (No. 20170540159), and Fundamental Research Funds for the Central Universities of China (No. DUT17LAB04).

References

1. Allen J.B., Berkley D.A. (1979), Image method for efficiently simulating small room acoustics, *The Journal of the Acoustical Society of America*, **65**(4): 943–950, doi: 10.1121/1.382599.

2. Canclini A., Bestagini P., Antonacci F., Compagnoni M., Sarti A, Tubaro S. (2015), A robust and low-complexity source localization algorithm for asynchronous distributed microphone networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(10): 1563–1575, doi: 10.1109/taslp.2015.2439040.

3. Chang C.C., Lin C.J. (2011), LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, **2**(3): 27, doi: 10.1145/1961189.1961199.

4. Cheng S., Xu Y., Zong R., Wang C. (2019), A fast decision making method for mandatory lane change

using kernel extreme learning machine, *International Journal of Machine Learning and Cybernetics*, **10**(12): 3363–3369, doi:10.1007/s13042-019-00923-8.

5. Crocco M., Bue A.D., Murino V. (2012), A bilinear approach to the position self-calibration of multiple sensors, *IEEE Transactions on Signal Processing*, **60**(2): 660–673, doi: 10.1109/tsp.2011.2175387.

6. Ferguson E.L., Williams S.B., Jin C.T. (2018), Sound source localization in a multipath environment using convolutional neural networks, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2386–2390, Seoul, South Korea, doi: 10.1109/ICASSP.2018.8462024.

7. Gu Y., Chen Y., Liu J., Jiang X. (2015), Semi-supervised deep extreme learning machine for Wi-Fi based localization, *Neurocomputing*, **166**: 282–293, doi: 10.1016/j.neucom.2015.04.011.

8. Hengy S., Duffne, P., DeMezzo S., Heck S., Gross, L., Naz P. (2016), Acoustic shooter localisation using a network of asynchronous acoustic nodes, *IET Radar, Sonar & Navigation*, **10**(9): 1528–1535, doi: 10.1109/ICASSP.2018.8462024.

9. Ho K.C. (2012), Bias reduction for an explicit solution of source localization using TDOA, *IEEE Transactions on Signal Processing*, **60**(5): 2101–2114, doi: 10.1109/tsp.2012.2187283.

10. Huang G.B., Wang D.H., Lan Y. (2011), Extreme learning machines: a survey, *International Journal of Machine Learning and Cybernetics*, **2**(2): 107–122, doi: 10.1007/s13042-011-0019-y.

11. Huang G.B., Zhou H., Ding X., Zhang R. (2012), Extreme learning machine for regression and multiclass classification, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **42**(2): 513–529, doi: 10.1109/TSMCB.2011.2168604.

12. Huang G.B., Zhu Q.Y., Siew C.K. (2006), Extreme learning machine: Theory and applications. *Neurocomputing*, **70**(1–3): 489–501, doi: 10.1016/j.neucom.2005.12.126.

13. Kan Y., Wang P., Zha F., Li M., Gao W., Song B. (2015), Passive acoustic source localization at a low sampling rate based on a five-element cross microphone array, *Sensors (Basel)*, **15**(6): 13326–13347, doi: 10.3390/s150613326.

14. Khanal S., Silverman H.F., Shakya R.R. (2013), A free-source method (FrSM) for calibrating a large-aperture microphone array, *IEEE Transactions on Audio, Speech, and Language Processing*, **21**(8): 1632–1639, doi: 10.1109/tasl.2013.2256896.

15. Knapp C.H., Carter G.C. (1976), The generalized correlation method for estimation of time delay, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **24**(4): 320–327, doi: 10.1109/TASSP.1976.1162830.

16. Kongsorot Y., Horata P., Musikawan P., Sunat K. (2019), Kernel extreme learning machine based on fuzzy set theory for multi-label classification, *International Journal of Machine Learning and Cybernetics*, **10**(5): 979–989, doi: 10.1007/s13042-017-0776-3.

17. Lim H., Yoo I.-C., Cho Y., Yook D. (2015), Speaker localization in noisy environments using steered response voice power, *IEEE Transactions on Consumer Electronics*, **61**(1): 112–118, doi: 10.1109/TCE.2015.7064118.

18. Lima M.V.S. *et al.* (2015), A volumetric SRP with refinement step for sound source localization, *IEEE Signal Processing Letters*, **22**(8): 1098–1102, doi: 10.1109/lsp.2014.2385864.

19. Nakano A.Y., Nakagawa S., Yamamoto K. (2009), Automatic estimation of position and orientation of an acoustic source by a microphone array network, *The Journal of the Acoustical Society of America*, **126**(6): 3084–3094, doi: 10.1121/1.3257548.

20. Nunes L.O. *et al.* (2014), A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays, *IEEE Transactions on Signal Processing*, **62**(19): 5171–5183, doi: 10.1109/tsp.2014.2336636.

21. Principi E., Squartini S., Cambria E., Piazza F. (2015), Acoustic template-matching for automatic emergency state detection: An ELM based algorithm, *Neurocomputing*, **149**: 426–434, doi: 10.1016/j.neucom.2014.01.067.

22. Salvati D., Drioli C., Foresti G.L. (2016), A weighted MVDR beamformer based on SVM learning for sound source localization, *Pattern Recognition Letters*, **84**: 15–21, doi: 10.1016/j.patrec.2016.07.003.

23. Salvati D., Drioli C., Foresti G.L. (2018), Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions, *IEEE Transactions on Emerging Topics in Computational Intelligence*, **2**(2): 103–116, doi: 10.1109/tetci.2017.2775237.

24. Stone M. (1974), Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**(2): 111–147, doi: 10.1111/j.2517-6161.1974.tb00994.x.

25. Tian Y., Chen Z., Yin F. (2015), Distributed IMM-Unscented Kalman filter for speaker tracking in microphone array networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(10): 1637–1647, doi: 10.1109/taslp.2015.2442418.

26. Vera-Diaz J.M., Pizarro D., Macias-Guarasa J. (2018), Towards end-to-end acoustic localization using deep learning: from audio signals to source posi-

tion coordinates, *Sensors (Basel)*, **18**(10): 3418, doi: 10.3390/s18103418.

27. Wan X., Wu Z. (2013), Sound source localization based on discrimination of cross-correlation functions, *Applied Acoustics*, **74**(1): 28–37, doi: 10.1016/ j.apacoust.2012.06.006.

28. Xiao X., Zhao S., Zhong X., Jones D.L., Chng E.S., Li H. (2015), A learning-based approach to direction of arrival estimation in noisy and reverberant environments, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2814–2818, Brisbane, Australia, doi: 10.1109/ICASSP.2015. 7178484.

29. Zhang Q., Chen Z., Yin F. (2016), Distributed marginalized auxiliary particle filter for speaker tracking in distributed microphone networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(11): 1921–1934, doi: 10.1109/ taslp.2016.2590146.

30. Zhang Q., Chen Z., Yin F. (2013), Microphone clustering and BP network based acoustic source localization in distributed microphone arrays, *Advances in Electrical and Computer Engineering*, **13**(4): 33–40, doi: 10.4316/aece.2013.04006.