Sound Source Localization Based on SRP-PHAT Spatial Spectrum and Deep Neural Network

Xiaoyan Zhao[1, *], Shuwen Chen[2], Lin Zhou[3] and Ying Chen[3, 4]

Abstract: Microphone array-based sound source localization (SSL) is a challenging task in adverse acoustic scenarios. To address this, a novel SSL algorithm based on deep neural network (DNN) using steered response power-phase transform (SRP-PHAT) spatial spectrum as input feature is presented in this paper. Since the SRP-PHAT spatial power spectrum contains spatial location information, it is adopted as the input feature for sound source localization. DNN is exploited to extract the efficient location information from SRP-PHAT spatial power spectrum due to its advantage on extracting high-level features. SRP-PHAT at each steering position within a frame is arranged into a vector, which is treated as DNN input. A DNN model which can map the SRP-PHAT spatial spectrum to the azimuth of sound source is learned from the training signals. The azimuth of sound source is estimated through trained DNN model from the testing signals. Experiment results demonstrate that the proposed algorithm significantly improves localization performance whether the training and testing condition setup are the same or not, and is more robust to noise and reverberation.

Keywords: Sound source localization, microphone array, steered response power-phase transform (SRP-PHAT) spatial spectrum, deep neural network.

1 Introduction

In many applications, such as speech enhancement [Kim (2014)], sound source separation [Nikunen and Virtanen (2014)], autonomous robots [Salvati, Drioli and Foresti (2016)] and video conferencing [Zhao, Ahmed, Liang et al. (2012)], location information of the sound source is very important. Microphone array-based sound source localization (SSL) aims at estimating the location information of the sound source by using the received multichannel signals.

[1] School of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing, 211167, China.

[2] School of Mathematics and Information Technology, Jiangsu Second Normal University, Nanjing, 210013, China.

[3] School of Information Science and Engineering, Southeast University, Nanjing, 210096, China.

[4] Department of Psychiatry, Columbia University and NYSPI, New York, 10032, USA.

[*] Corresponding Author: Xiaoyan Zhao. Email: xiaoyanzhao205@163.com.

Received: 22 January 2020; Accepted: 02 March 2020.

CMC. doi:10.32604/cmc.2020.09848 www.techscience.com/journal/cmc

Over the past few decades, a number of approaches have been developed for the task of SSL which can perform fairly well under certain acoustic conditions. However, in adverse acoustic scenarios, robust SSL is still a challenging task because the noise and reverberation would degrade the localization performance. The existing approaches for SSL can be generally categorized into two classes: indirect approaches and direct approaches [Salvati, Drioli and Foresti (2018)]. The indirect approaches first estimate time difference of arrivals (TDOAs) between microphone pairs, which are usually estimated by generalized cross correlation (GCC) methods [Knapp and Carter (1976)], and then achieve the sound source location by exploiting TDOAs and the knowledge of microphones' positions. The direct approaches include the methods based on the multiple signal classification (MUSIC) [Zhao, Saluev and Jones (2014)], the estimation of signal parameters via rotational invariance techniques (ESPRIT) [Roy and Kailath (1989)], the steered response power (SRP) [Dibiase (2001)] and so on.

Along with the development of artificial neural network, deep learning methods have been successfully applied in SSL task. There are two main schemes for applying the deep learning method to SSL: the first one is to combine deep learning with the traditional SSL methods, for example, time-frequency (T-F) masking is predicted by deep learning, then the sound source position is estimated by SRP with T-F masking; the second one is to establish the relationship between input features and sound source position by deep learning, which we call model-based methods. The related research for the first scheme is as follows. In Wang et al. [Wang, Zhang and Wang (2018, 2019)], speech dominant T-F units were identified by deep neural networks (DNNs), and the conventional crosscorrelation, beamforming and subspace-based approaches were weighted by T-F masking. In order to emphasize the direct path speech signal in time-varying interference, Pertila et al. [Pertila and Cakir (2017)] used convolutional neural network (CNN) to learn the mapping between a noisy input log-magnitude spectrogram and a corresponding desired T-F masking, and then estimated the azimuth by steered response power-phase transform (SRP-PHAT) with T-F masking. Salvati et al. [Salvati, Drioli and Foresti. (2018)] presented a deep learning-based SRP beamformer that fuses the narrowband response power of incoherent frequency by the weighting factors provided by CNNs.

The related research for the second scheme of the application of deep learning in SSL is as follows. Takeda et al. [Takeda and Komatani (2016)] extracted the directional image by using the orthogonality of eigenvectors in the frequency domain, and then propagated and integrated the directional image at each sub-band through a hierarchical structure. Chakrabarty et al. [Chakrabarty and Habets (2017)] utilized CNN to learn the information required for direction of arrival (DOA) estimation from the phase components of the spectrogram. Methods in Xiao et al. [Xiao, Zhao, Zhong et al. (2015); Sun, Chen, Yuen et al. (2018); Vesperini, Vecchiotti, Principi et al. (2018)] arranged GCC of each microphone pair together to form the input feature vector or matrix. Xiao et al. [Xiao, Zhao, Zhong et al. (2015)] trained a multi-layer perceptron (MLP) model to learn the mapping regularity between the GCC and DOAs. Sun et al. [Sun, Chen, Yuen et al. (2018)] presented a probabilistic neural network-based SSL algorithm. Vesperini et al. [Vesperini, Vecchiotti, Principi et al. (2018)] studied the application of multi-layer perceptron (MLP) and convolutional neural network (CNN) in speaker localization under multi-room domestic environment. Adavanne et al. [Adavanne, Politis and Virtanen

(2018)] extracted the magnitudes and phases of the spectrograms as input feature to train a convolutional recurrent neural network (CRNN). The issue of binaural SSL is studied in Roden et al. [Roden, Moritz, Gerlach et al. (2015); Ma, May and Brown (2017); Yiwere and Rhee (2017); Ma, Gonzalez and Brown (2018); Zhou, Ma, Wang et al. (2019)]. Roden et al. [Roden, Moritz, Gerlach et al. (2015)] extracted inter-aural time difference (ITD) and inter-aural level difference (ILD) as input features for DNNs. Ma et al. [Ma, May and Brown (2017)] combined cross-correlation function (CCF) and ILD as input features for DNNs. Yiwere et al. [Yiwere and Rhee (2017)] trained the DNN model by the input features consisting of CCF and ILD, which can predict the direction and distance of sound source. Ma et al. [Ma, Gonzalez and Brown (2018)] trained the DNN model of each sub-band with CCF as input feature, and modeled each sound source using GMM with ratemap as feature in the training stage; and then integrated the posterior probabilities of azimuth at each sub-band estimated from DNN by the weights provided by GMM in the localization stage. Zhou et al. [Zhou, Ma, Wang et al. (2019)] extracted CCF of each Gammatone filter and combined the CCFs of all sub-bands to assemble a two dimensional feature matrix, then utilized CNN to establish the relationship between feature matrix and sound azimuth.

In this paper, we propose a DNN-based localization algorithm using SRP-PHAT spatial spectrum as input feature. Unlike the existing algorithms, SRP-PHAT spatial power spectrum is adopted as the input feature due to its robust representation of spatial location information. DNN can effectively extract high-level features and learn the relationship between input and output. Therefore, we consider the SSL problem as multi-classification task and introduce DNN to learn the mapping regularity between the SRP-PHAT spatial power spectrum and the azimuth of sound source. The trained DNN model is used to predict probabilities of each azimuth with the testing signals, and the azimuth with maximum probability is considered as the estimated azimuth. Through experimental evaluation, the proposed algorithm has been shown to significantly improve localization performance under various acoustic conditions.

The rest of this paper is organized as follows. Section 2 describes the proposed approach based on SRP-PHAT spatial spectrum and DNN. The experimental results and analysis are presented in Section 3. Finally, the conclusion is drawn in Section 4.

2 DNN-based localization algorithm using SRP-PHAT spatial spectrum

2.1 System architecture

The core idea of our algorithm is to use DNN to extract effective location information from the spatial power spectrum. Fig. 1 illustrates the overall system architecture. The system includes two stages, the training stage and the localization stage. The system input is the signals received by microphone array. Since the reverberation and environmental noise alter, signals in diverse reverberation and noise environments are taken together as training data during the training stage for robustness. Then, SRP-PHAT at each steering position within a frame is combined into a vector, which is treated as DNN input. In the training stage, a DNN model is trained to establish the mapping regularity between the SRP-PHAT spatial power spectrum and the azimuth of sound source. In the localization stage, the probability of each azimuth is predicted through the trained DNN model on the testing signals. Afterwards, the azimuth with maximum probability is considered as the estimated azimuth.

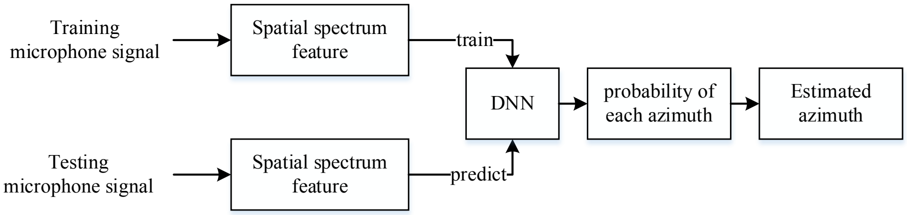


Figure 1: Overall architecture of the proposed sound source localization system

2.2 Feature extraction

In indoor scenarios, the signal received by *m*th microphone at time *t* can be modeled as follows:

$$x_m(t) = h_m(\mathbf{r}_s, t) * s(t) + v_m(t), \quad m = 1, 2, \ldots, M \tag{1}$$

where $s(t)$ denotes the clean sound source signal, $h_m(\mathbf{r}_s\ ,t)$ denotes the room impulse response between the source position $\mathbf{r}_s$ and the *m*th microphone, "*" denotes the linear convolution operator, $v_m(t)$ denotes the uncorrelated additive noise for the *m*th microphone, and M is the number of microphones. In the frequency domain, Eq. (1) can be formulated as:

$$X_m(\omega) = H_m(\mathbf{r}_s, \omega) S(\omega) + V_m(\omega), \quad m = 1, \ldots M \tag{2}$$

where $X_m(\omega)$, $S(\omega)$, $H_m(\mathbf{r}_\mathrm{s}, \omega)$ and $V_m(\omega)$ are the Fourier transforms of $x_m(t)$, $s(t)$, $h_m(\mathbf{r}_s, t)$ and $v_m(t)$.

The SRP can be calculated by summing the GCCs of the chosen pairs of microphones [Zhao, Tang, Zhou et al. (2013)], and the response power when the array is steered to the position r is expressed as:

$$\begin{aligned} P(\mathbf{r}) &= \sum_{m=1}^{M} \sum_{n=m+1}^{M} R_{m,n}\left(\Delta\tau_{mn}(\mathbf{r})\right) \\ &= \sum_{m=1}^{M} \sum_{n=m+1}^{M} \int_{-\infty}^{\infty} \psi_{m,n}(\omega) X_m(\omega) X_n^*(\omega) e^{\mathrm{j}\omega\Delta\tau_{mn}(\mathbf{r})} d\omega \end{aligned} \tag{3}$$

where $P(\mathbf{r})$ represents the response power when the array is steered to the position $\mathbf{r}$, $\Delta\tau_{mn}(\mathbf{r})=\tau_m(\mathbf{r})-\tau_n(\mathbf{r})$, $\tau_m(\mathbf{r})$ is the propagation delay between the steering position $\mathbf{r}$ and the *m*th microphone, in the far-filed case, $\Delta_{mn}(\mathbf{r})$ is independent of the distance from the steering position $\mathbf{r}$ to microphone array and only related to the azimuth of the steering position $\mathbf{r}$, $R_{m,n}(\Delta_{mn}(\mathbf{r}))$ is the GCC of the *m*th and *n*th microphone signals, $\psi_{m,n}(\omega)$ is the weighting function. The phase transform (PHAT) weighting function is defined as:

$$\psi_{m,n}(\omega) = \frac{1}{\left|X_m(\omega) X_n^*(\omega)\right|} \tag{4}$$

Applying the PHAT weighting function to the response power in Eq. (3), we get SRP-PHAT as follows:

$$P(\mathbf{r})=\sum_{m=1}^{M}\sum_{n=m+1}^{M}\int_{-\infty}^{\infty}\frac{X_m(\omega)X_n^*(\omega)}{\left|X_m(\omega)X_n^*(\omega)\right|}e^{\mathrm{j}\omega\Delta\tau_{mn}(\mathbf{r})}d\omega \tag{5}$$

Without considering the environment noise, i.e., $V_m(\omega)$=0 in Eq. (2), Eq. (2) can be simplified as

$$X_m(\omega)=H_m(\mathbf{r}_s,\omega)S(\omega),\ \ m=1,\ldots M \tag{6}$$

Substituting Eq. (6) into Eq. (5), Eq. (5) can be rewritten as

$$\begin{aligned}P(\mathbf{r})&=\sum_{m=1}^{M}\sum_{n=m+1}^{M}\int_{-\infty}^{\infty}\frac{H_m(\mathbf{r}_s,\omega)S(\omega)H_n^*(\mathbf{r}_s,\omega)S^*(\omega)}{\left|H_m(\mathbf{r}_s,\omega)S(\omega)H_n^*(\mathbf{r}_s,\omega)S^*(\omega)\right|}e^{\mathrm{j}\omega\Delta\tau_{mn}(\mathbf{r})}d\omega\\&=\sum_{m=1}^{M}\sum_{n=m+1}^{M}\int_{-\infty}^{\infty}\frac{H_m(\mathbf{r}_s,\omega)H_n^*(\mathbf{r}_s,\omega)}{\left|H_m(\mathbf{r}_s,\omega)H_n^*(\mathbf{r}_s,\omega)\right|}e^{\mathrm{j}\omega\Delta\tau_{mn}(\mathbf{r})}d\omega\end{aligned} \tag{7}$$

From Eq. (7), we note that the SRP-PHAT function is independent of the content of the sound source signal, and it is dependent of the room impulse response. The room impulse response is related to the source position, microphone position, room size, reverberation time and so on. Therefore, the SRP-PHAT spatial power spectrum contains spatial location information, which is closely related to the source position and acoustic environment. Thus, the SRP-PHAT spatial power spectrum is exploited as the feature for sound source localization in this paper.

In practical situation, there is not only environment noise but also spectrum leakage due to signal framing. The SRP-PHAT suffers from a deviation between the actual calculated value and the theoretical value resulting from the presence of environment noise and spectrum leakage. An example is given for normalized SRP-PHAT spatial spectrum of different frames in Fig. 2, where the sound source is located at 60° azimuth with 1.5 m from the array. Six omnidirectional microphones form a uniform circular array with radius of 0.1 m. The reverberation time is 0.5 s, and the SNR is 10 dB. The sample rate is 16 kHz, and the microphone signals are segmented into 32-ms frame length. Fig. 2 shows that the SRP-PHAT spatial spectrum has a main peak at 60° azimuth due to direct path, and many pseudo peaks at other azimuths resulting from reflection and noise. From Fig. 2, we also note that SRP-PHAT spatial spectrum of each frames are not identical, which is due to the presence of environment noise and spectrum leakage. In this paper, the DNN is used to extract the efficient location information from SRP-PHAT spatial power spectrum.

The SRP-PHAT spatial spectrum is treated as the input feature for DNN. SRP-PHAT by Eq. (5) is the response power at one steering position, and then SRP-PHAT at each steering position within a frame is arranged into a vector, which can be expressed as follows:

$$\boldsymbol{P}(k)=\left[P(\mathbf{r}_1,k),P(\mathbf{r}_2,k),\ldots P(\mathbf{r}_L,k)\right]^T \tag{8}$$

where $\boldsymbol{P}$ (k) is the feature vector of kth frame, and $\boldsymbol{P}$ ($\mathbf{r}_l$, k) with l=1…L is the SRP-PHAT at $\mathbf{r}_l$ in kth frame which is calculated by Eq. (5). In this paper, we consider the far-filed case, and then the argument $\mathbf{r}_l$ is simplified to the azimuth with a distance of 1.5 meters from the steering position to the microphone array. The azimuth ranges from 0° to 360°

with a step of 5°, corresponding to 72 steering positions, which means that the dimension of SRP-PHAT feature vector is 72.

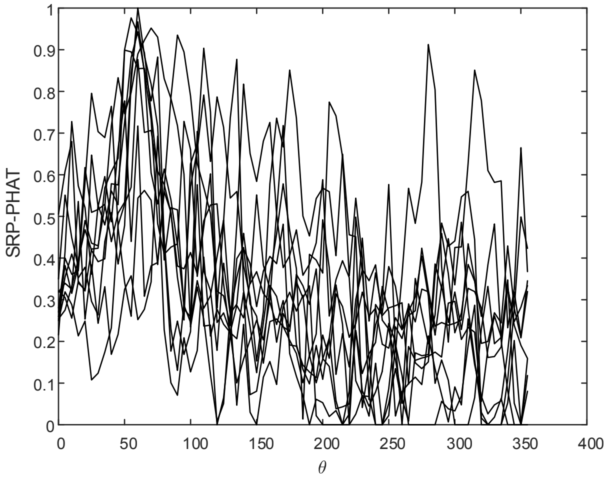


Figure 2: SRP-PHAT spatial spectrum for 10 frames

2.3 The architecture of DNN

DNN model is trained on a set of feature vectors ***P*** (*k*)s with given azimuths. The training azimuth ranges from 0° to 360° with a step of 10°, corresponding to 36 training positions.

The DNN consists of an input layer, multiple hidden layers, and an output layer. For input layer, the input is the feature vector ***P*** (*k*) as described in Section 2.2. The number of input layer nodes is equal to the feature dimension, which is the dimension of the SRP-PHAT feature vector in of our algorithm. The system of our algorithm uses two hidden layers, each layer containing 50 nodes. The hidden layers use the sigmoid activation function.

The system of our algorithm treats the sound source localization problem as a multi-classification problem, and the output layer adopts the softmax regression model, which can be regarded as the generalization of the logistic model on the multi-classification problem. The number of class labels is the number of training positions, which is 36 in this paper. The SRP-PHAT feature vector of the test signal is extracted according to Eq. (8), and the probability that $\boldsymbol{P}^{\text{test}}\left(k\right)$ belongs to each azimuth is estimated by softmax regression model as follows:

$$h_\theta\left(\boldsymbol{P}^{\text{test}}\left(k\right)\right)=\frac{1}{\sum_{j=1}^{N} e^{\theta_j^T \boldsymbol{P}^{\text{test}}(k)}}\begin{bmatrix} e^{\theta_1^T \boldsymbol{P}^{\text{test}}(k)} \\ e^{\theta_2^T \boldsymbol{P}^{\text{test}}(k)} \\ \vdots \\ e^{\theta_N^T \boldsymbol{P}^{\text{test}}(k)} \end{bmatrix} \tag{9}$$

where *N* is the number of class labels, *N*=36 in this paper, *θ* is the model parameter derived from training stage, and ***P***$^{\text{test}}$ (*k*) is the feature vector of *k*th frame test signals.

2.4 The training of DNN

The training of DNN includes two phases: pre-training and fine-tuning. In the pre-training phase, the model parameters are initialized using deep belief network (DBN) formed by layered stacked restricted Boltzmann machines (RBMs). Since the acoustic signal is a continuous signal, a Gaussian restricted Boltzmann machine (GRBM) [Wang and Wang (2013)] model is suitable between the input layer and the first hidden layer, and a RBM is adopted between the first hidden layer and the second hidden layer. The GRBM and RBM are stacked to form the DBN. The contrastive divergence (CD) algorithm [Hinton (2002)] is used in the pre-training phase. In the fine-tuning phase, a Softmax output layer for classification is added on the top of the DBN, and the network parameters are finely adjusted by the Back Propagation (BP) algorithm.

In the pre-training phase, the mini-batch [Hinton (2010)] is set to 100, the number of epochs is set to 10, the momentum is set to 0.5, and the learning rate is set to 0.01. The fine-tuning phase is composed of two processes, and the mini-batch is also set to 100. In the first process, the number of epochs is set to 50, the momentum is set to 0.9, and the learning rate is set to 0.1; in the second process, the number of epochs is set to 50, the momentum is set to 0.9, and the learning rate is set to 0.01. DropOut and Cross Validation are used to prevent over-fitting in the proposed system. The Dropout ratio is set to 0.5, and a 1:10 Cross Validation is used to monitor the model errors in the fine-tuning phase.

3 *Simulation and result analysis*

3.1 Simulation setup

The dimensions of the simulated room are 7 m×7 m×3 m. Six omnidirectional microphones form a uniform circular array with radius of 0.1m, which is located at (3.5 m, 3 m, 1.6 m). The room impulse response is generated by the image method [Allen and Berkley (1979)]. The sound source signals with 16 kHz sampling rate are taken randomly from the TIMIT database. The microphone signal is generated by convolving the sound source signal with the room impulse response and then adding uncorrelated Gaussian white noise. The microphone signals are segmented into 32-m frame length and windowed by Hanning window.

For the training signals, the distance between the training position and the array is set to 1.5 m, the training azimuth ranges from 0° to 360° with a step of 10°, the SNR is varied from 0 dB to 20 dB with a step of 5 dB, and the reverberation time T60 is set to 0.2 s, 0.5 s and 0.8 s. Microphone signals in different reverberation and noise environments are taken together as training data during the training stage for robustness. The number of training data at each training position under each reverberation and noise environment is 400 frames.

The performance of the proposed algorithm is compared with two baseline methods, namely the SRP-PHAT [Dibiase (2001)] and SRP-PHAT based on principal eigenvector (SRP-PHAT-PE) [Wan and Wu (2010)]. We evaluate the localization performance by the percentage of correct estimates, which is calculated as:

$$p = n_{\mathrm{c}} / N_{\mathrm{all}} \tag{10}$$

where N_{all} is the total number of testing frames, n_c is the number of correct estimate frames, and the correct estimate is defined that the estimated azimuth is the true azimuth.

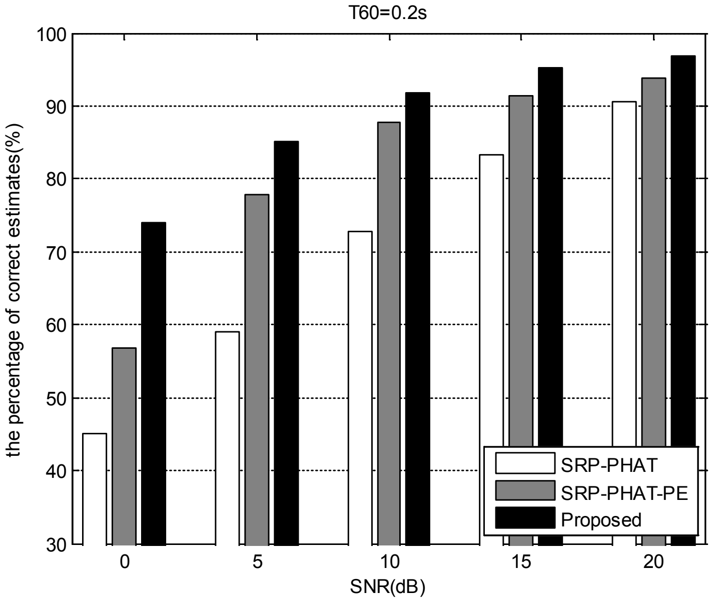
3.2 Evaluation in setup-matched environments

In this section, we investigate and analyze the localization performance when the training and testing condition setup are the same. Fig. 3 depicts the localization performance as a function of SNR for SRP-PHAT, SRP-PHAT-PE and the proposed algorithm under various reverberation environments.

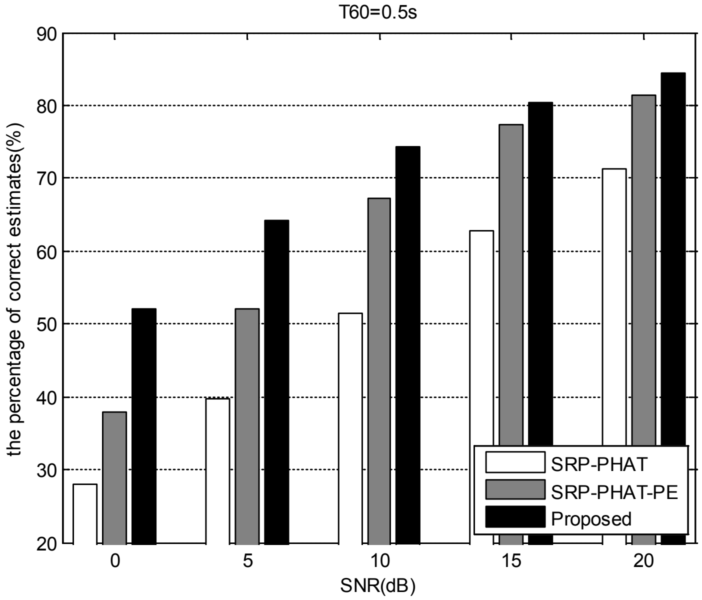
In the high SNR and low reverberation environments, each of the methods performs well, and the percentage of correct estimates of the proposed method is slightly higher than that of the SRP-PHAT and SPR-PHAT-PE. As the SNR decreases and the reverberation time increases, the performance of SRP-PHAT and SRP-PHAT-PE deteriorates, and the proposed method outperforms the baseline methods significantly. The reason is that the proposed algorithm adopts DNN to extract the efficient spatial location information from the SRP-PHAT spatial power spectrum in diverse environments.

Furthermore, at the same reverberation time, the performance improvement of the proposed algorithm compared with the baseline methods tends to increase gradually as the SNR decreases. For example, in the T60=0.5 s scenario, the improvement of the percentage of correct estimates of the proposed algorithm compared with the SRP-PHAT algorithm increases from 13.27% to 23.96% as the SNR decreases from 20 dB to 0 dB. This shows that performance improvement is especially in low SNR environments.

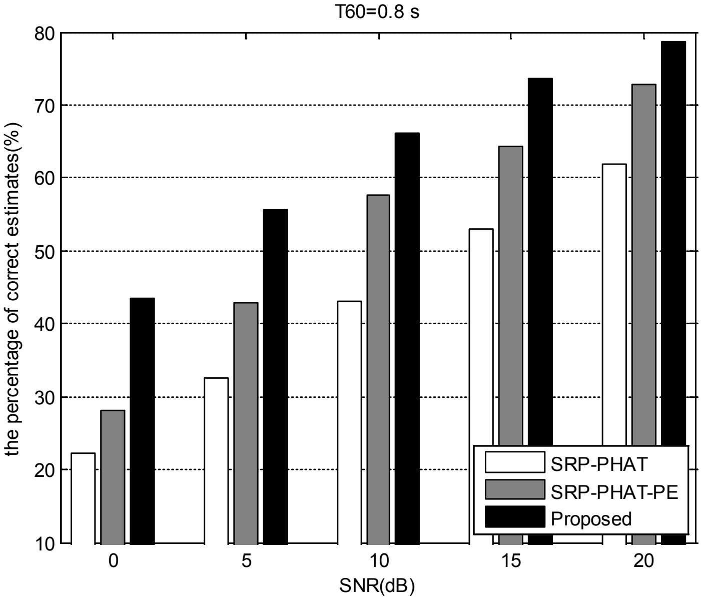
It can also be seen that, for moderate to high SNR (10-20 dB), the performance improvement increases as the reverberation time increases with the same SNR. For example, the performance improvement of the proposed algorithm compared with the SRP-PHAT algorithm increases from 11.88% to 20.69% as the reverberation time increases from 0.2 s to 0.8 s with SNR=15 dB. And for low SNR (below 5 dB), the performance improvement is slightly reduced as the reverberation time increases in the same SNR scenario. For example, the performance improvement of the proposed algorithm compared with the SRP-PHAT algorithm is reduced from 29.04% to 21.18% as the reverberation time increases from 0.2 s to 0.8 s with SNR=0 dB.



(a) Percentage of correct estimates with T_{60}=0.2 s



(b) Percentage of correct estimates with T_{60}=0.5 s



(c) Percentage of correct estimates with T_{60}=0.8 s

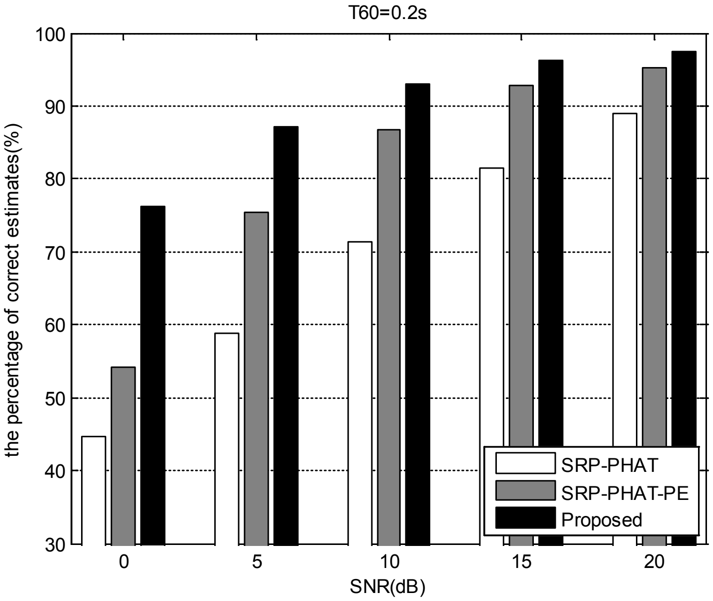
Figure 3: Performance comparison of different algorithm in setup-matched environments

3.3 Evaluation in untrained distance environment

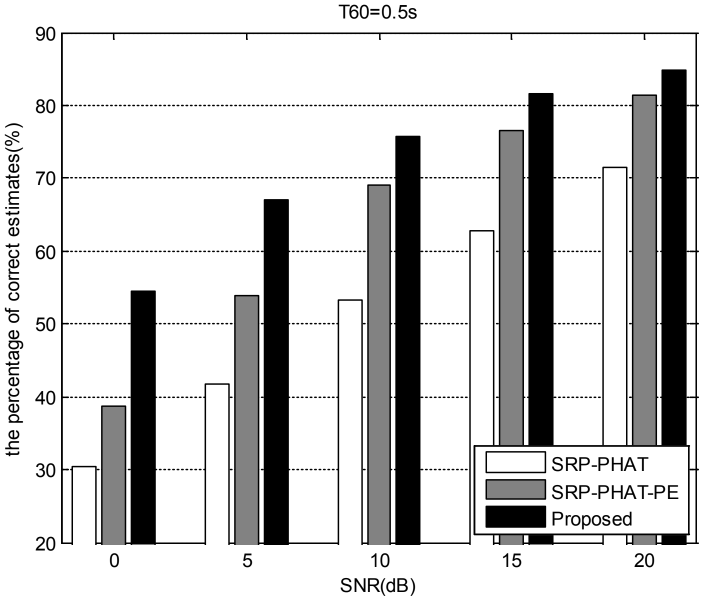
In this section, we investigate the robustness of the proposed method when the distance between the testing position and the array is not the same as the distance between the training position and the array. Two distances are tested: 1.3 m and 1.8 m. The setting of azimuth, reverberation time and SNR of the testing signals are consistent with those of the training signals. Figs. 4 and 5 depict the localization performance of the three methods for a distance of 1.3 m and a distance of 1.8 m, respectively.

First, we have found that the regularity of data variation in Figs. 4 and 5 are consistent with those in Fig. 3, which has been described in Section 3.2. This reflects the robustness of the proposed algorithm to the distance between the testing position and the array.

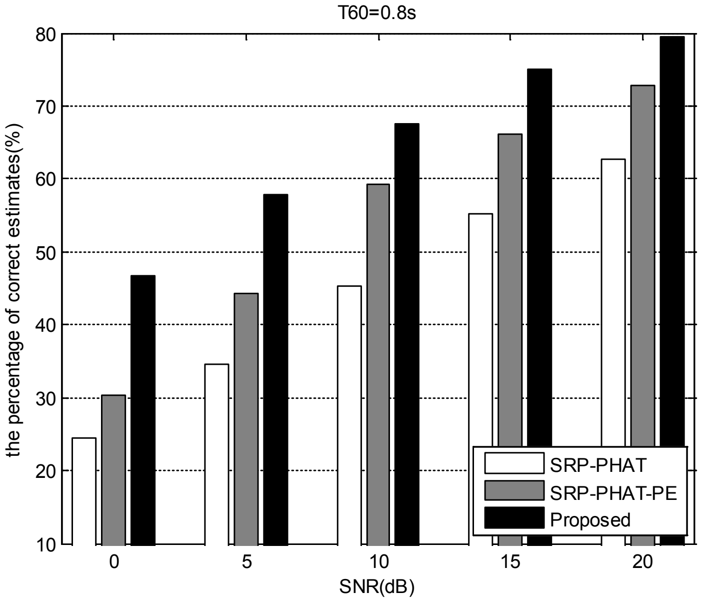
In addition, comparing Figs. 4 and 5 with Fig. 3, in the same reverberation and SNR scenario, the percentage of correct estimates with a testing distance of 1.3 m is slightly higher than that with a testing distance of 1.5 m, and the percentage of correct estimates with a testing distance of 1.8 m is lower than that with a testing distance of 1.5 m. For example, at T60=0.2 s, compared with the testing distance of 1.5 m, the percentage of correct estimates of the proposed algorithm is reduced by about 2~3% and that of the SRP-PHAT algorithm is reduced by about 3~4% when the testing distance is 1.8 m. The main reason is that the adverse effect of reverberation on localization performance is related to the distance between the sound source and the microphone array. The adverse effect of reverberation on localization performance increase as the distance increases.



(a) Percentage of correct estimates with T_{60}=0.2 s

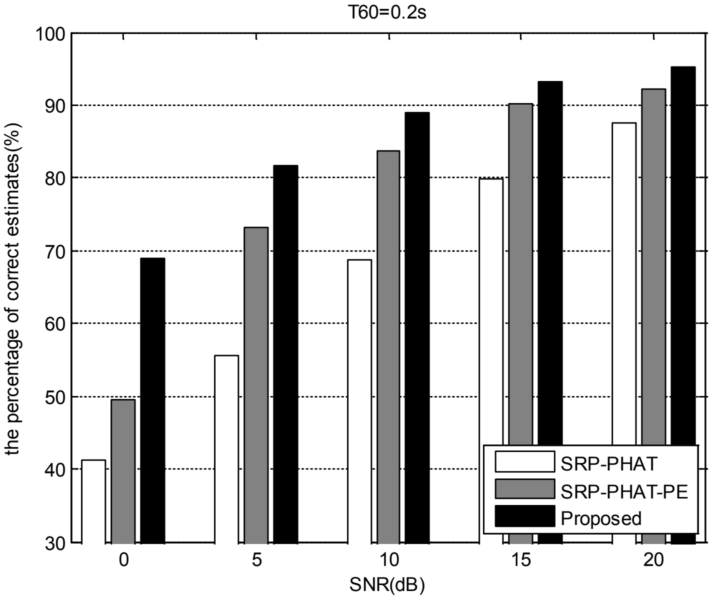


(b) Percentage of correct estimates with T_{60}=0.5 s

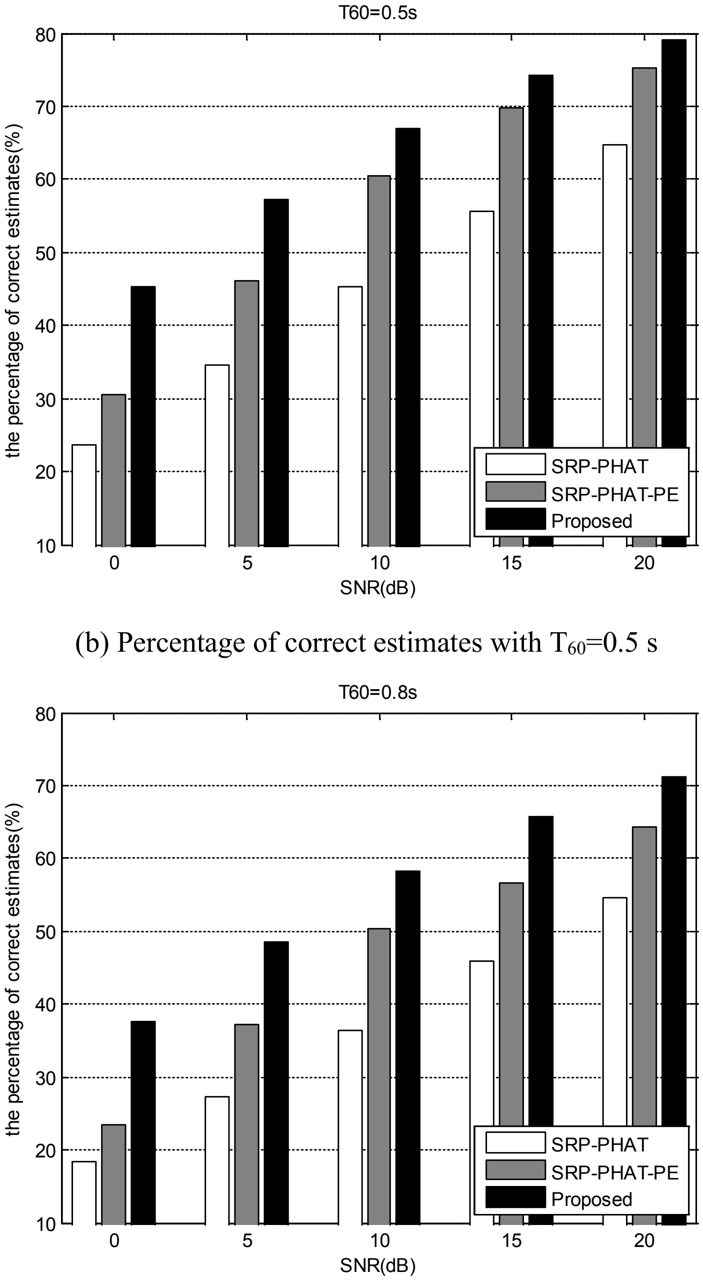


(c) Percentage of correct estimates with T_{60}=0.8 s

Figure 4: Performance comparison of different algorithm with testing distance=1.3 m



(a) Percentage of correct estimates with T_{60}=0.2 s



(b) Percentage of correct estimates with T_{60}=0.5 s

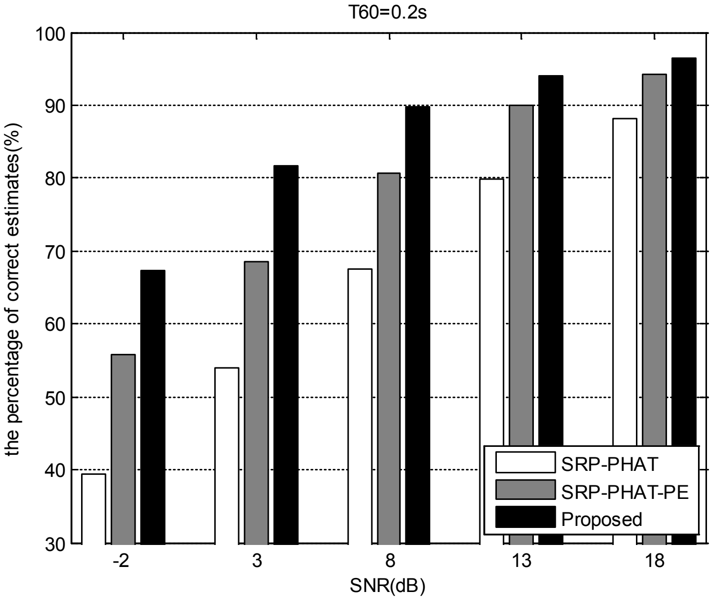
(c) Percentage of correct estimates with T_{60}=0.8 s

Figure 5: Performance comparison of different algorithm with testing distance=1.8 m

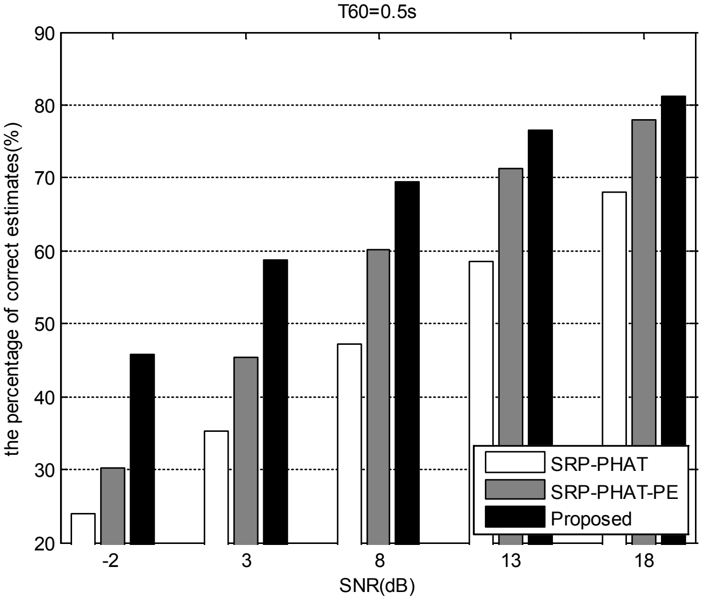
3.4 Evaluation in untrained noise or reverberation environments

In this section, we investigate the robustness of the proposed method in untrained noise or reverberation environment. Three untrained reverberation times are tested: 0.3 s, 0.6 s and 0.9 s. Five levels of untrained SNR are tested: -2 dB, 3 dB, 8 dB, 13 dB and 18 dB. The setting of azimuth and distance between the testing position and the array are consistent with those of the training position. Figs. 6 and 7 depict the localization performance of the three methods in untrained SNR and untrained reverberation time scenarios, respectively.

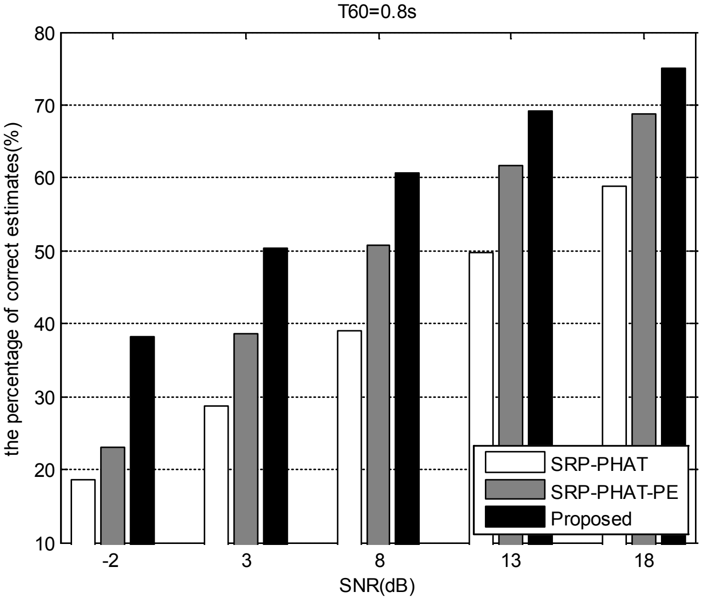
As shown in Figs. 6 and 7, the performance of the proposed method is much better than that of the baseline methods in untrained noise or reverberation environment, and the regularity of data variation in Figs. 6 and 7 are consistent with those in Fig. 3, which has been described in Section 3.2. Specifically, in moderate and low reverberation and high SNR environments, the performance improvement of the proposed algorithm compared with the SRP-PHAT algorithm is about 10%; in other scenario, the performance improvement is about 20% to 28%. This reflects that the proposed method is robust to the noise and reverberation.



(a) Percentage of correct estimates with T_{60}=0.2 s

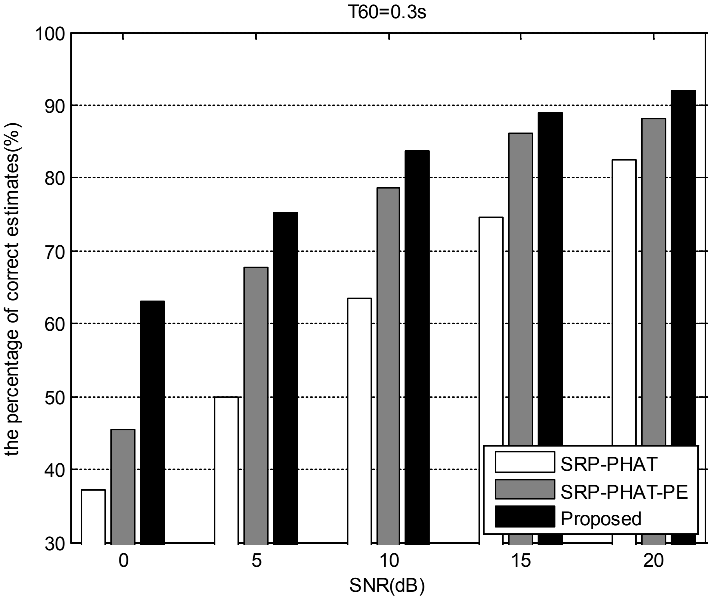


(b) Percentage of correct estimates with T_{60}=0.5 s

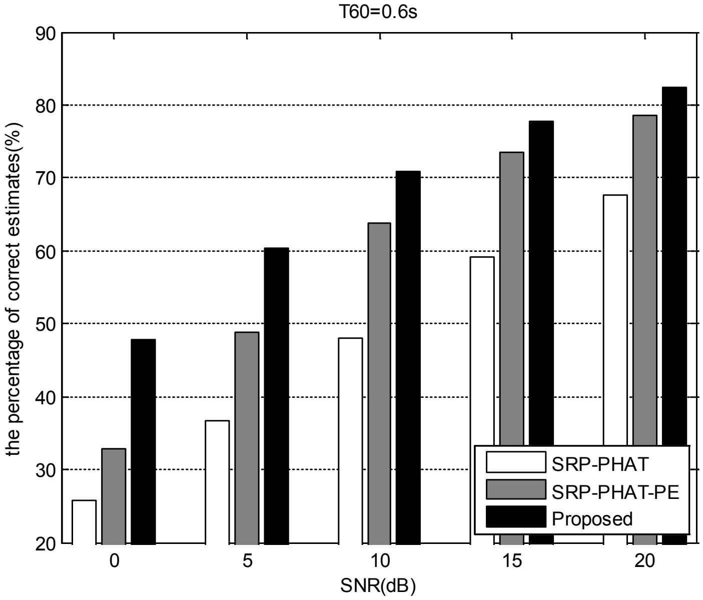


(c) Percentage of correct estimates with T_{60}=0.8 s

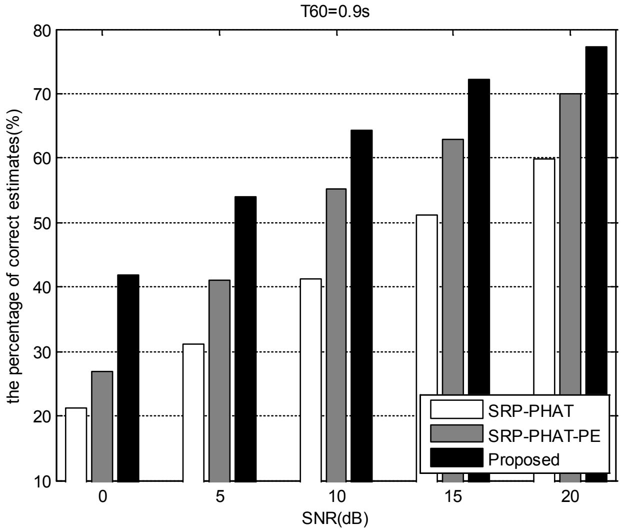
Figure 6: Performance comparison of different algorithm in untrained noise environments



(a) Percentage of correct estimates with T_{60}=0.3 s



(b) Percentage of correct estimates with T_{60}=0.6 s



(c) Percentage of correct estimates with T_{60}=0.9 s

Figure 7: Performance comparison of different algorithm in untrained reverberation environments

4 Conclusion

In this work, a SSL algorithm based on SRP-PHAT spatial spectrum and deep neural networks has been presented. We treat the sound source localization problem as a multi-classification task. Different from the existing algorithms, the SRP-PHAT spatial power spectrum is exploited as the feature vector in the proposed algorithm. DNN is utilized to learn the mapping regularity between feature vector and azimuth of sound source due to its advantage on extracting high-level features. Experiment results demonstrate that the proposed algorithm achieves superior localization performance whether the training and testing condition setup are the same or not, and is more robust to noise and reverberation.

Funding Statement: This work is supported by the National Nature Science Foundation of China (NSFC) under Grant No. 61571106, Jiangsu Natural Science Foundation under Grant No. BK20170757 and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under grant No. 17KJD510002.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Adavanne, S.; Politis, A.; Virtanen, T. (2018): Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. *26th European Signal Processing Conference*, pp. 1462-1466.

Allen, J. B.; Berkley, D. A. (1979): Image method for efficiently simulating small-room acoustics. *Journal of Acoustical Society of America*, vol. 65, no. 4, pp. 943-950.

Chakrabarty, S.; Habets, E. A. P. (2017): Broadband DOA estimation using convolutional neural networks trained with noise signals. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 136-140.

Dibiase, J. H. (2001): *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays (Ph.D. Thesis)*. Brown University, USA.

Hinton, G. E. (2002): Training products of experts by minimizing contrastive divergence. *Neural Computation*, vol. 14, no. 8, pp. 1771-1800.

Hinton, G. E. (2010): A practical guide to training restricted Boltzmann machines. *Momentum*, vol. 9, no. 1, pp. 926.

Kim, S. M.; Kim, H. H. (2014): Direction-of-arrival based SNR estimation for dual-microphone speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2207-2217.

Knapp, C.; Carter, G. (1976): The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320-327.

Ma, N.; Gonzalez, J. A.; Brown, G. J. (2018): Robust binaural localization of a target sound source by combining spectral source models and deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122-2131.

Ma, N.; May, T.; Brown, G. J. (2017): Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 12, pp. 2444-2453.

Nikunen, J.; Virtanen, T. (2014): Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727-739.

Pertila, P.; Cakir, E. (2017): Robust direction estimation with convolutional neural networks based steered response power. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6125-6129.

Roden, R.; Moritz, N.; Gerlach, S.; Weinzierl, S.; Goetze, S. (2015): On sound source localization of speech signals using deep neural networks. *Deutsche Jahrestagung Fur Akustik*, pp. 1510-1513.

Roy, R.; Kailath, T. (1989): ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 37, no. 7, pp. 984-995.

Salvati, D.; Drioli, C.; Foresti, G. L. (2016): Sound source and microphone localization from acoustic impulse responses. *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1459-1463.

Salvati, D.; Drioli, C.; Foresti, G. L. (2018): A low-complexity robust beamforming using diagonal unloading for acoustic source localization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 609-622.

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*, vol. 2, no. 2, pp. 103-116.

Sun, Y. X.; Chen, J. J.; Yuen, C.; Rahardja, S. (2018): Indoor sound source localization with probabilistic neural network. *IEEE Transactions on Industrial Electronics*, vol. 65, no. 8, pp. 6403-6413.

Takeda, R.; Komatani, K. (2016): Sound source localization based on deep neural networks with directional activate function exploiting phase information. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 405-409.

Vesperini, F.; Vecchiotti, P.; Principi, E.; Squartini, S.; Piazza, F. (2018): Localizing speakers in multiple rooms by using deep neural networks. *Computer Speech & Language*, vol. 49, pp. 83-106.

Wan, X. W.; Wu, Z. Y. (2010): Improved steered response power method for sound source localization based on principal eigenvector. *Applied Acoustics*, vol. 71, no. 12, pp. 1126-1131.

Wang, Y.; Wang, D. L. (2013): Towards scaling up classification-based speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381-1390.

Wang, Z. Q.; Zhang, X. L.; Wang, D. L. (2018): Robust TDOA estimation based on time-frequency masking and deep neural networks. *Interspeech*, pp. 322-326.

Wang, Z. Q.; Zhang, X. L.; Wang, D. L. (2019): Robust speaker localization guided by deep learning-based time-frequency masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178-188.

Xiao, X.; Zhao, S. K.; Zhong, X. H.; Jones, D. L.; Cheng, E. S. et al. (2015): A learning-based approach to direction of arrival estimation in noisy and reverberant environments. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2814-2818.

Yiwere, M.; Rhee, E. J. (2017): Distance estimation and localization of sound sources in reverberant conditions using deep neural Networks. *International Journal of Applied Engineering Research*, vol. 12, no. 22, pp. 12384-12389.

Zhao, S.; Ahmed, S.; Liang, Y.; Rupnow, K.; Chen, D. et al. (2012): A real-time 3D sound localization system with miniature microphone array for virtual reality. *IEEE Conference on Industrial Electronics and Applications*, pp. 1853-1857.

Zhao, S.; Saluev, T.; Jones, D. L. (2014): Underdetermined direction of arrival estimation using acoustic vector sensor. *Signal Processing*, vol. 100, pp. 160-168.

Zhao, X. Y.; Tang, J.; Zhou, L.; Wu, Z. Y. (2013): Accelerated steered response power method for sound source localization via clustering search. *Science China: Physics, Mechanics and Astronomy*, vol. 56, no. 7, pp. 1329-1338.

Zhou, L.; Ma, K. Y.; Wang, L. J.; Chen, Y.; Tang, Y. B. (2019): Binaural sound source localization based on convolutional neural network. *Computers, Materials & Continua*, vol. 60, no. 2, pp. 545-557.