Sound source localization with multi-feature fusion using residuals and channel attention

Nipun Agarwal 1

1Birla Institute of Technology and Science

October 31, 2023

Abstract

Recent advances in deep learning have enhanced the ability of sound source localization in noise and reverberation. However, the single feature input and relatively simple network design hinder the further improvement of such ability. Therefore, a multi-feature fusion sound source localization method is proposed based on residuals and channel attention. The strategy of multi-feature fusion provides a deep neural network with more comprehensive discriminative features. The deep neural network fully extracts the sound source location information from the fused features by introducing residuals and channel attention. The simulations show that the localization accuracies of the proposed method in single and multiple sound source scenarios are respectively 8.24% and 15.54% higher than those of the single feature convolutional neural network (SF-CNN). In addition, the proposed method has excellent performance under different signal-to-noise ratios and reverberation times, which verifies its robustness to noise and reverberation. In the experiment, the proposed method is still effective in localizing single and multiple sound sources, and its localization accuracies are 5.48% and 7.57% higher than that of SF-CNN. With high accuracy and strong robustness, this method is significant for sound source localization in complex environments, such as noise, reverberation, and the presence of multiple sound sources.
Sound source localization with multi-feature fusion using residuals and channel attention

Nipun Agarwal

Abstract—Recent advances in deep learning have enhanced the ability of sound source localization in noise and reverberation. However, the single feature input and relatively simple network design hinder the further improvement of such ability. Therefore, a multi-feature fusion sound source localization method is proposed based on residuals and channel attention. The strategy of multi-feature fusion provides a deep neural network with more comprehensive discriminative features. The deep neural network fully extracts the sound source location information from the fused features by introducing residuals and channel attention. The simulations show that the localization accuracies of the proposed method in single and multiple sound source scenarios are respectively 8.24% and 15.54% higher than those of the single feature convolutional neural network (SF-CNN). In addition, the proposed method has excellent performance under different signal-to-noise ratios and reverberation times, which verifies its robustness to noise and reverberation. In the experiment, the proposed method is still effective in localizing single and multiple sound sources, and its localization accuracies are 5.48% and 7.57% higher than that of SF-CNN. With high accuracy and strong robustness, this method is significant for sound source localization in complex environments, such as noise, reverberation, and the presence of multiple sound sources.

Index Terms—Sound source localization, Deep neural network, Microphone array, Multi-feature fusion

I. INTRODUCTION

SOUND serves as a crucial medium for information propagation in nature. Humans perceive their surroundings through the brain’s ability to decode sound signals. An interesting example is that the brain helps humans determine the direction of a sound source based on binaural differences [1], [2] known as Sound Source Localization (SSL). Computers are desired to give this capability, which has important applications in fields such as intelligent conferencing [3]–[5], human-computer interaction [6], [7], and robotic motion detection [8], [9].

With the advancements in digital signal processing techniques, traditional SSL methods, such as Generalized Cross Correlation with Phase Transform (GCC-PHAT) [10], Multiple Signal Classification (MUSIC) [11], Steered Response Power Phase Transform (SRP-PHAT) [12] and their improvements [13]–[17], can accurately localize sound sources in relatively simple scenarios by processing and analyzing signals from microphone array. However, they could improve in challenging yet common scenarios where noise, reverberation, and multiple sound sources may be present [18], which limits traditional SSL methods to be applied in practical application scenarios.

Recently, deep learning (DL) techniques have been introduced into the field of SSL [19]–[21], gradually breaking this limitation and receiving increasing interest. Vecchiotti et al. [22] proposed an end-to-end binaural SSL method that estimates the azimuth of a sound source directly from the waveform. Li et al. [23] proposed a SSL method combining convolutional neural network (CNN) and long short term memory, which has excellent localization performance and easily adapts to new microphone arrays. Chakrabarty and Habets [24] proposed a novel SSL method based on single-feature convolutional neural networks (SF-CNN), which can accurately localize sound sources in reverberant and noisy environments and was extended to multiple source scenarios [25]. Despite the remarkable performance improvement in reverberant and noisy environments, however, such methods failed to satisfy the expected accuracy and robustness, owing to the fact that the single feature input and relatively simple network structure design limits the adequate capture of potentially useful information related to the sound source location. It is, therefore, essential to continuously pursue new DL-based mechanisms with higher accuracy and stronger robustness, which would be vital for real-world application in many diverse scenarios such as noise control [26] where the high accuracy and strong robustness of localization method is highly desired.

Therefore, this paper proposes a novel DL-based SSL method, by introducing residuals and channel attention to CNN to fully extract the rich information related to the sound source location contained in the fused features to perform accurate and robust SSL in single and multiple source scenarios. The rest of the paper is organized as follows. A detailed DL architecture for SSL is presented in Section II. Simulations of single and multiple SSL are given in Section III. Section IV contains complete experimental tests, which validate the effectiveness and superiority of the proposed method. Finally, Section V concludes the paper.
II. METHODS

A. Problem Description

The azimuthal space $[0, 180^\circ]$ is discretized with angular resolution $\alpha^\circ$ and $I = 180/\alpha + 1$ possible azimuths of sound sources are obtained, which form a set of azimuths $\Theta = \{\theta_1, \ldots, \theta_i, \ldots \theta_I\}$. Then, a class vector of length $I$ is formed, where each class corresponds to one possible azimuth in the set $\Theta$. Therefore, the SSL task is transformed into a classification task and can be divided into two cases depending on the number of sources.

In the single SSL scene (scene 1 in Fig. 1), only one source is active at each moment, corresponding to one azimuth class. The localization problem in this scene is formulated as a multi-class classification task. Then, the sound source location $\hat{\theta}$ is estimated as the azimuth class with the highest posterior probability, which can be expressed as

$$\hat{\theta} = \arg \max_{\theta_i} p(\theta_i | \Theta) \quad (1)$$

In the multiple SSL scene (scene 2 in Fig. 1), multiple sources are active at each moment, corresponding to more than one azimuth class. The localization problem in this scene is formulated as a multi-label classification task. Then, the sound source locations $\hat{\theta}_1, \ldots, \hat{\theta}_p, \ldots, \hat{\theta}_P, \hat{\theta}_p \in \Theta$ are estimated as the top $P$ azimuth classes with the highest posterior probability, which can be denoted as

$$\hat{\theta}_1, \ldots, \hat{\theta}_p, \ldots, \hat{\theta}_P = \arg \max_{\theta_1, \ldots, \theta_p, \ldots, \theta_P} p(\theta_1, \ldots, \theta_p | \Theta) \quad (2)$$

To accomplish this classification task, a novel DL architecture, in Fig. 2 is proposed. The feature representation and deep neural network (DNN) model of this architecture are presented in Sections II.B and C, respectively.

B. Feature Representation

To ensure that the directional information contained in the acoustic signal, Fig. 2(a) is fully learned by DNN, three features (frequency domain phase, frequency domain magnitude and spectral flux) are fused as input features, Fig. 2(b) to DNN. Considering an indoor scene with reverberation and noise, $P$ far-field sound sources with azimuths $\theta_p$ are emitted to a linear array consisting of $M$ microphones. Then, the sound source signal $y_m(t)$ received by the $m$th microphone at time $t$ is modeled as

$$y_m(t) = \sum_{p=1}^{P} h_m^p(t) * s_p(t) + v_m(t) \quad (3)$$

where $s_p(t)$ denotes the $p$th sound source signal; $h_m^p(t)$ denotes the room impulse response (RIR) from the $p$th sound source to the $m$th microphone; $*$ denotes the convolution operation; $v_m(t)$ denotes the noise of the $m$th microphone, and $t$ denotes the discrete time index.

After $N_f$-point short-time Fourier transform (STFT), eq. 3 can be expressed as

$$Y_m(n,k) = \sum_{l=0}^{N_f-1} H^p_{m,n}(k) S^p(n,k) + V^p_{m,n}(k) \quad (4)$$

where $n$ is the time frame index; $k(1, 2, \ldots, K) = \frac{N_f}{2} + 1$ denotes the frequency bin; $Y_m(n,k), H^p_{m,n}(k), S^p(n,k)$ and $V^p_{m,n}(k)$ represent the short-time Fourier transforms of $y_m(t), b^p_m(t), s^p(t)$ and $v_m(t)$, respectively.

In the practical processing, since $Y_m(n,k)$ is a complex number, eq. 4 is rewritten as

$$Y_m(n,k) = A_m(n,k) \exp[i \varphi_m(n,k)] \quad (5)$$

where $A_m(n,k)$ and $\varphi_m(n,k)$ are respectively the magnitude component and phase component of the $m$th microphone received signal at the $k$th frequency bin and the $n$th time frame.

Then, the magnitude and phase components form the eigenvectors $A_m(n)$ and $\varphi_m(n)$, respectively, denoted as

$$A_m(n) = [A_m(n,1) \ A_m(n,2) \ \ldots \ A_m(n,\frac{N_f}{2}+1)] \quad (6)$$

$$\varphi_m(n) = [\varphi_m(n,1) \ \varphi_m(n,2) \ \ldots \ \varphi_m(n,\frac{N_f}{2}+1)] \quad (7)$$

eq. 6 and 7 form the feature matrices and are utilized as the first and second channels of fusion feature (FF), expressed as follows

$$\Phi^l(n) = [A_1(n) \ A_2(n) \ \ldots \ A_M(n)]^T, l = 2 \quad (8)$$

$$\Phi^l(n) = [\varphi_1(n) \ \varphi_2(n) \ \ldots \ \varphi_M(n)]^T, l = 1 \quad (9)$$

where $l$ denotes the number of feature channels.

On the basis of eq. 5, the spectral fluxes are obtained in adjacent time frames

$$S_m(n,k) = |A_m(n,k)| - |A_m(n-1,k)| \quad (10)$$

Then, similar to eqs. 6, 7 and 8, 9, the eigenvectors and eigenmatrices formed from the spectral fluxes are respectively denoted as

$$S_m(n) = [S_m(n,1) \ S_m(n,2) \ \ldots \ S_m(n,\frac{N_f}{2}+1)] \quad (11)$$

$$\Phi^l(n) = [S_1(n) \ S_2(n) \ \ldots \ S_M(n)]^T, l = 3 \quad (12)$$

As can be seen from eqs. 8, 9 and 12, the three features of phase, magnitude and spectral flux are stitched in the feature channel dimension and finally form the FF at the $n$th frame, which can be, expressed as

$$\Phi(n) = [\Phi^1(n) \ \Phi^2(n) \ \Phi^3(n)] \quad (13)$$

Therefore, the dimension of the FF is $3 \times M \times K$, which corresponds to the input dimension of the DNN model.
The FF provided by eq. 13 in Section II.B is fed into the designed DNN, called ResSeCNN network, and the specific framework is shown in Fig. 2(c). The backbone of the designed network is CNN, which learns discriminative features for sound source localization through \((M - 1)\) convolutional kernels (size \(2 \times 2\) and \(1 \times 1\) in single and multiple SSL scenarios, respectively). But as the network deepens, problems such as gradient disappearance or gradient explosion may occur, leading to increased training difficulty. To alleviate this problem, a residual module with residual connections is introduced after the first convolutional layer, and the details of this module are shown in Fig. 2(d). In addition, as mentioned in Section II.B, features are fused in the channel dimension. When the fused features are mapped to higher dimensional features, the importance of each channel feature to discriminate the sound source location may not be consistent. To allow the model to learn the importance weights of each channel and thus better integrate the feature information of different channels, the channel attention mechanism is introduced after the \((M - 1)\)th convolutional layer, and the module details are shown in Fig. 2(e). The high dimensional features extracted by the channel attention mechanism are mapped into \(I\) azimuthal category scores by two fully connected layers. The high dimensional features extracted by the channel attention mechanism are mapped into \(I\) azimuthal category scores by two fully connected layers. These scores are then converted into \(I\) azimuthal category probability distributions by applying a Softmax function (for single source scene) or a Sigmoid function (for multiple source scene), and hence localization is achieved.

### C. DNN Model

The FF provided by eq. 13 in Section II.B is fed into the designed DNN, called ResSeCNN network, and the specific framework is shown in Fig. 2(c). The backbone of the designed network is CNN, which learns discriminative features for sound source localization through \((M - 1)\) convolutional kernels (size \(2 \times 2\) and \(1 \times 1\) in single and multiple SSL scenarios, respectively). But as the network deepens, problems such as gradient disappearance or gradient explosion may occur, leading to increased training difficulty. To alleviate this problem, a residual module with residual connections is introduced after the first convolutional layer, and the details of this module are shown in Fig. 2(d). In addition, as mentioned in Section II.B, features are fused in the channel dimension. When the fused features are mapped to higher dimensional features, the importance of each channel feature to discriminate the sound source location may not be consistent. To allow the model to learn the importance weights of each channel and thus better integrate the feature information of different channels, the channel attention mechanism is introduced after the \((M - 1)\)th convolutional layer, and the module details are shown in Fig. 2(e). The high dimensional features extracted by the channel attention mechanism are mapped into \(I\) azimuthal category scores by two fully connected layers. The high dimensional features extracted by the channel attention mechanism are mapped into \(I\) azimuthal category scores by two fully connected layers. These scores are then converted into \(I\) azimuthal category probability distributions by applying a Softmax function (for single source scene) or a Sigmoid function (for multiple source scene), and hence localization is achieved.

### III. Simulation

#### A. Settings

In the simulation analysis, a 4-microphone uniform linear array with microphone spacing of 0.043 m is considered, and the sound source signal is received at a sampling frequency of 16 kHz. The \([0, 180^\circ]\) azimuth range was discretized with a resolution of \(5^\circ\) and a total of 37 azimuth classes were obtained. The array received signal is obtained by simulation of eq. 3, where the sound source signal is Gaussian white noise, and the RIR response is obtained by the RIR-Generator [27], [28]. Eventually, the array received signals with different parameters, such as reverberation time \((RT_{60})\) and signal-to-noise ratio (SNR), were generated to constitute the simulation dataset. The dataset is randomly divided into a training set and a test set in the ratio of 8 : 2, where the test set is used to test the localization performance of the proposed method. In addition, test set 1 and 2 are generated respectively to test the robustness of the method to noise and reverberation. The details of the training set, test set, test set 1 and test set 2 are shown in TABLE I. Note that in the multiple sound source scene, two sources are emitted simultaneously and the difference between the two azimuths is more than \(30^\circ\).

After transforming the array received signal into the STFT domain with STFT of \(N_f = 512\) and 50% overlap, FF dimension of \(3 \times 4 \times 257\) is obtained. Then, the training and testing of the proposed method is performed in Pytorch. During training, the learning rate and batches were 0.001 and 512, and the optimizer selected Adam. At the end of the SE layer and after each fully connected layer, a dropout with a rate of 0.5 is used to avoid overfitting. For different application scenarios, different loss functions are used to make the azimuth classification results of the model gradually approach the true azimuths. As described in Section II.A, the localization tasks are transformed into multi-class classification and multi-label classification tasks in the single and multiple SSL scenarios, respectively. Therefore, the cross-entropy loss function \(L_{CE}\) and the binary cross-entropy loss function \(L_{BCE}\) are used in the two scenarios, which can be de denoted as

\[
L_{CE} = - \sum_{i=1}^{I} y_i \log(\hat{y}_i) \tag{14}
\]

\[
L_{BCE} = \sum_{i=1}^{I} (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \tag{15}
\]

where \(I\) denotes the total number of azimuthal categories, \(y_i\) denotes the probability of azimuthal category \(i\) in the real label, and \(\hat{y}_i\) denotes the probability of azimuthal category \(i\) predicted by the model.

During testing, to evaluate the localization performance of the proposed method, the localization accuracy was defined as eq. 14 and compared with Phase-CNN in single source scene [24] and multiple source scene [25].

\[
\text{Accuracy} = \frac{C_{acc}^*}{C} \times 100\% \tag{16}
\]

where \(C\) denotes the total number of time frames and \(C_{acc}^*\) denotes the number of time frames with correct position estimation. Note that in a multiple source localization scene, the position is estimated correctly only if multiple source positions are all estimated accurately.
TABLE I: Details of the simulation dataset

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Training and Test Set (8:2)</th>
<th>Test Set 1</th>
<th>Test Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound source signal</td>
<td>The Gaussian white noise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Room size</td>
<td>(7, 6, 3.5)m; (7, 7, 3.5)m; (8, 7, 3.5)m</td>
<td>(7, 6, 3.5)m; (7, 7, 3.5)m; (8, 7, 3.5)m</td>
<td>(9, 9, 3.5)m</td>
</tr>
<tr>
<td>RT60</td>
<td>[0.2,0.3,0.4,0.5,0.6]s</td>
<td>0.2s</td>
<td>[0.18,0.28,0.38,0.48,0.58]s</td>
</tr>
<tr>
<td>Array position</td>
<td>3 random positions in each room</td>
<td>An arbitrary position in each room</td>
<td></td>
</tr>
<tr>
<td>Array-source distance</td>
<td>2m3m</td>
<td>2.5m</td>
<td></td>
</tr>
<tr>
<td>SNR</td>
<td>Random distribution (0 20dB)</td>
<td>[0,5,10,15,20]dB</td>
<td>20dB</td>
</tr>
</tbody>
</table>

Fig. 3: Simulation training results in single SSL scene (a) Training loss; (b) Training accuracy

B. Single SSL Scene

The training and testing results for the single source scenes are shown in Fig. 3, and Table II, respectively. As can be seen from Fig. 3, the proposed method embodies lower training loss values and higher training accuracy compared with the other three methods. The results in Table II show that FF-CNN and Phase-ResSeCNN have smaller training loss values and higher training localization accuracy compared to Phase-CNN, which indicates that the proposed FF and ResSeCNN are feasible for single source localization tasks. It is further observed that the proposed method (FF-ResSeCNN) has a lower loss value of 0.4499 and a higher accuracy of 87.60% than that of FF-CNN and Phase-ResSeCNN. In addition, compared with that of Phase-CNN, the loss value and accuracy of the proposed method are respectively reduced and improved by 0.2241 and 8.24%. These results show that the proposed ResSeCNN with FF can localize single sound sources more accurately than using them alone, and perform significantly better localization performance than Phase-CNN, which validates the feasibility and superiority of the proposed method in single SSL scene.

The test results of the proposed methods at different SNR and RT60 are shown in Fig. 4. When the SNR gradually increases from 0 dB to 20 dB or the RT60 gradually decreases from 0.58s to 0.18s, the localization accuracy of the four methods gradually increases. It is worth noting that whether for different SNR or different RT60, the proposed method has higher localization accuracy, especially compared with that of Phase-CNN, which verifies the robustness to noise and reverberation of the proposed method in single SSL scene.

C. Multiple SSL Scene

Fig. 5 and TABLE III show the training and testing results for multiple SSL scene, respectively. Compared with Phase-CNN, FF-CNN and Phase-ResSeCNN, the proposed method still has the lowest training loss value and the highest training accuracy in the multiple SSL scene. In addition, the proposed method has the smallest loss value of 0.0107 and the highest localization accuracy of 87.49% at testing. In particular, compared with the loss and accuracy of Phase-CNN, that of the proposed method are respectively reduced

TABLE II: Simulation test results in single SSL scene

<table>
<thead>
<tr>
<th>Method</th>
<th>Loss</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase-CNN</td>
<td>0.6740</td>
<td>79.36%</td>
</tr>
<tr>
<td>Phase-ResSeCNN</td>
<td>0.5640</td>
<td>83.50%</td>
</tr>
<tr>
<td>FF-CNN</td>
<td>0.5255</td>
<td>85.20%</td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.4499</td>
<td>87.60%</td>
</tr>
</tbody>
</table>
Table III: Simulation test results in multiple SSL scene

<table>
<thead>
<tr>
<th>Method</th>
<th>Loss</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase-CNN</td>
<td>0.0656</td>
<td>71.95%</td>
</tr>
<tr>
<td>Phase-ResSeCNN</td>
<td>0.0549</td>
<td>76.40%</td>
</tr>
<tr>
<td>FF-CNN</td>
<td>0.0405</td>
<td>83.18%</td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.0107</td>
<td>87.49%</td>
</tr>
</tbody>
</table>

Fig. 4: Robustness test results in single SSL scene (a) different SNR; (b) different $RT_{60}$

Fig. 5: Simulation training results in multiple SSL scene (a) Training loss; (b) Training accuracy

A. Settings

To further test the localization performance of the proposed method in real reverberant and noisy environments, experiments are performed in an indoor environment (Fig. 7). The division of azimuthal classes and array parameters were consistent with the simulation. In the single source scene, the sound source is Indian Classical music, which is placed in each azimuth class at 2m, 3m and 4m away from the array center, and a total of $37 \times 3 \times 12s = 1332s$ of sound signals are collected. In the multiple source scene, sound source 1 and 2 were respectively Indian Classical and English pop music, both placed in each azimuth class at 4m away from the array center. The difference between azimuths of sound source 1 and 2 was greater than $30^\circ$, and a total of $96 \times 1 \times 12s = 1152s$ of sound signals were collected. Then, each 12s collected signal is divided into 10s and 2s signals to form the training and test
Fig. 6: Robustness test results in multiple SSL scene (a) different SNR; (b) different $RT_{60}$

![Experimental scene diagram](image)

Fig. 7: Experimental scene diagram

<table>
<thead>
<tr>
<th>Method</th>
<th>Loss</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase-CNN</td>
<td>0.4630</td>
<td>87.64%</td>
</tr>
<tr>
<td>Phase-ResSeCNN</td>
<td>0.2896</td>
<td>92.55%</td>
</tr>
<tr>
<td>FF-CNN</td>
<td>0.2662</td>
<td>92.99%</td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.2010</td>
<td>93.12%</td>
</tr>
</tbody>
</table>

TABLE IV: Experiment test results in single SSL scene

<table>
<thead>
<tr>
<th>Method</th>
<th>Loss</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase-CNN</td>
<td>0.0271</td>
<td>85.57%</td>
</tr>
<tr>
<td>Phase-ResSeCNN</td>
<td>0.0173</td>
<td>90.47%</td>
</tr>
<tr>
<td>FF-CNN</td>
<td>0.0152</td>
<td>91.85%</td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.0144</td>
<td>93.14%</td>
</tr>
</tbody>
</table>

TABLE V: Experiment test results in multiple SSL scene

B. Single SSL Scene

Fig. 8 shows that the proposed method has faster convergence speed, lower loss and higher accuracy in the training period. TABLE IV shows that the proposed method has the smallest test loss of 0.2010, which is reduced by 0.2620 compared with the loss of Phase-CNN. In addition, the proposed method has the largest test accuracy of 93.12%, which is improved by 5.48% compared with the accuracy of Phase-CNN. All these results indicate that the proposed method well learns the discriminative features for locating single sound source with high accuracy, which demonstrates the effectiveness and superiority of the proposed method in real indoor single SSL scene.

C. Multiple SSL Scene

The localization performance of the proposed method in indoor multiple SSL scene is further tested, and the training and testing results are respectively shown in Fig. 9 and TABLE V. The proposed method still has the faster convergence speed, lower loss and higher accuracy in the training period. During the testing, the proposed method has the smallest loss of 0.0144, reduced by 0.1169 compared with the loss of Phase-CNN, and has the largest accuracy, improved by 7.57% compared with the accuracy Phase-CNN. All these results show that the proposed method is able to localize multiple sound sources with high accuracy, which verifies the effectiveness and superiority of the proposed method in real indoor multiple SSL scene.

V. Conclusion

In this study, we suggest a multi-feature fusion SSL that uses residuals and channel attention to enhance the accuracy and robustness of the DNN-based SSL algorithm. The simulation results reveal that our proposed method outperforms Phase-CNN, Phase-ResSeCNN, and FF-CNN. Specifically, our method accurately locates single or multiple sound sources and exhibits stronger resistance to noise and reverberant environments. The experiments confirm the localization performance
Fig. 8: Experiment training results in single SSL scene (a) Training loss; (b) Training accuracy

Fig. 9: Experiment training results in multiple SSL scene (a) Training loss; (b) Training accuracy

of our proposed algorithm in real noisy and reverberant environments, with a localization accuracy of over 90%. This method is suitable for multiple source localization scenarios and provides a new option for applications requiring high accuracy and support for multiple sources, such as speaker tracking and intelligent security systems.

REFERENCES


