Wang *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2020) 2020:4
https://doi.org/10.1186/s13636-020-0171-y

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition

Jing Wang*, Jin Wang, Kai Qian, Xiang Xie and Jingming Kuang

Abstract

Binaural sound source localization is an important and widely used perceptually based method and it has been applied to machine learning studies by many researchers based on head-related transfer function (HRTF). Because the HRTF is closely related to human physiological structure, the HRTFs vary between individuals. Related machine learning studies to date tend to focus on binaural localization in reverberant or noisy environments, or in conditions with multiple simultaneously active sound sources. In contrast, mismatched HRTF condition, in which the HRTFs used to generate the training and test sets are different, is rarely studied. This mismatch leads to a degradation of localization performance. A basic solution to this problem is to introduce more data to improve generalization performance, which requires a lot. However, simply increasing the data volume will result in data-inefficiency. In this paper, we propose a data-efficient method based on deep neural network (DNN) and clustering to improve binaural localization performance in the mismatched HRTF condition. Firstly, we analyze the relationship between binaural cues and the sound source localization with a classification DNN. Different HRTFs are used to generate training and test sets, respectively. On this basis, we study the localization performance of DNN model trained by each training set on different test sets. The result shows that the localization performance of the same model on different test sets is different, while the localization performance of different models on the same test set may be similar. The result also shows a clustering trend. Secondly, different HRTFs are divided into several clusters. Finally, the corresponding HRTFs of each cluster center are selected to generate a new training set and to train a more generalized DNN model. The experimental results show that the proposed method achieves better generalization performance than the baseline methods in the mismatched HRTF condition and has almost equal performance to the DNN trained with a large number of HRTFs, which means the proposed method is data-efficient.

Keywords: Deep neural network, Clustering, Affinity propagation, Binaural localization

1 Introduction

Sound source localization is to estimate the direction of the sound source and is an important and widely used technique in many fields such as speech enhancement, video conferencing, and human-robot interaction [1]. Sound source localization algorithms have been widely researched so far, and they can be categorized into two classes. The first one is based on microphone array signal processing, which contains three kinds of algorithms: the algorithms based on the time difference of arrival (TDOA)[2], the algorithms based on beamforming [3], and the algorithms based on high-resolution spectral method [4, 5]. The second one is the binaural localization algorithms based on head-related transfer function (HRTF). Each algorithm has its own advantages and disadvantages.

Humans are able to localize the sound source with just two ears, and this remarkable binaural localization capability is largely attributed to the different filtering effects of listener's heads, pinna, and torse on the sounds from different directions in the frequency domain, which is

*Correspondence: wangjing@bit.edu.cn
Beijing Institute of Technology, 5 South Zhongguancun Street, Beijing 100081, China

described by the HRTF. The HRTF dataset of one subject consists of the HRTF pairs of the left and right ears measured at different directions. The time domain equivalent HRTF is called the head-related impulse response (HRIR). Due to the different individual physiological structures, the HRTF datasets of different subjects are varied.

Over the past decades, many binaural sound source localization methods have been proposed, and some are based on the "Duplex Theory" [6] proposed by Jeffress et al. Among these methods, two binaural cues extracted from the HRTFs are frequently used: interaural time difference (ITD) and interaural level difference (ILD). The ITD represents the time difference between the sounds arriving at the left and right ears, and the ILD represents the intensity difference between the sounds received by left and right ears. Those two binaural cues vary with the direction of the sound. Due to these factors, ITD and ILD are important for binaural localization. The key idea of these methods is to extract the ITDs and the ILDs corresponding to each direction from HRTFs which are saved as the ITD templates and ILD templates and then the ITD and ILD extracted from the sound are estimated and compared with the ITD templates and ILD templates. The best matching template can be found corresponding to the direction of the sound. In [7], Li et al. propose a three-layer Bayes rule-based hierarchical system, in which several possible locations are selected in the first layer and further narrowed down by ILD in the second layer, and the final decision is made by spectral cues in the third layer. In [8], Willert et al. put forward with a biologically inspired system to separately measure ITD and ILD to generate a probability map that is further combined over frequencies and binaural cues to estimate the sound location. Those methods achieve good performance, but the HRTFs for generating the templates and the test sets are recorded by the same subject or dummy head. In this paper, we refer to this condition as the matched HRTF condition and the condition that the HRTFs recorded by different heads as the mismatched HRTF condition. In the mismatched HRTF condition, the localization performance of the methods above may decline.

In [9], Raspaud et al. introduce an individual parametric model for each HRTF based on the simple geometric consideration. The ITD and ILD are modeled as the product of a function of frequency and a function of azimuth and then are jointly estimated and compared with templates for localization. Besides, the individual parametric models of each HRTF are averaged, which may improve the generality in the mismatched HRTF condition, but the simple average parametric model may not accurately learn the complex relationship between the binaural cues and the sound locations. Based on [9], Parisi et al. [10] propose cepstrum prefiltering for robustness in the reverberant environment. Pang et al. [1] put forward with reverberation weighting and a more generalized parametric model to further improve the localization performance in the reverberant and noisy environments. A full-sphere binaural localization method is proposed in [11], which applies the Interaural Phase Difference (IPD) for lateral localization and spectral cues for polar angle localization. Although the HRTFs for the training and test sets are captured in different rooms, the models of the dummy head are the same.

Besides the methods based on ITD and ILD templates, some other ones are based on HRTF templates. The key idea of those methods is to identify the HRTF pair corresponding to a certain sound direction, to operate with the left and right channels of the binaural sound respectively, and to achieve the maximal correlation between the results of the left and right channels. A matched filtering approach is proposed in [12]. For a certain sound direction, it exchanges the left channel and right channel of the corresponding HRTF pair, and then respectively filters the left and right channel of the binaural sound.The correlation between the result of the left and right channels shows that the HRTF pair with the maximal correlation corresponds to the direction of the sound. However, the inversion of the HRTF may be unstable. In [13], the source cancelation algorithm is proposed to be an extension of the matched filtering approach without inversion. It divides the left channel and right channel of the HRTF pairs to obtain the templates, and then the division result of the left channel and right channel of the sound in the frequency domain is calculated and matched with those templates. In [14, 15], a cross channel method is proposed, it convolutes the binaural sounds with the HRIR pair corresponding to a certain direction crosswise. Specifically, it convolutes the right channel of sound with the left channel of the HRIR pair and the left channel of sound with the right channel of the HRIR pair. Then the calculated correlation between the result of the two channels indicates the HRIR pair with the maximal correlation corresponds to the sound direction. In [16], a two-step method is proposed to estimate a coarse direction by ITD and the final result by the cross channel method. It improves the accuracy in a noisy environment and decreases the complexity. These HRTF templates based methods also perform good but only work in the matched HRTF condition.

Besides the template-matching-based methods, some other methods are based on statistical models. The key idea of those methods is mapping the binaural features to the posterior probability of the sound source in each direction by the statistical models. In [17], Gaussian Mixture Model (GMM) is used to estimate the multiple source localization in the reverberant and noisy environments by ITD and ILD cues. The HRTF is assumed to be known, which means it works in the matched HRTF condition. By combining DNN and head movements, a multiple

source localization method robust against the noisy and reverberant environment is proposed in [18], which is proved to generalize well on the test set generated by another HRTF. However, this method does not consider the performance in the mismatched HRTF condition. In [19], a convolutional neural network (CNN) with multi-task learning-based method is proposed to localize the azimuth and elevation simultaneously. It achieves better performance than the method in [18]. While it works in the matched HRTF condition. In [20], a CNN-based sound localization method is proposed and proved to be robust to inter-subject and measurement variability, but this study only focuses on elevation localization. In [21], an end-to-end binaural sound localization approach is proposed, which estimates the azimuth directly from the waveform by CNN. This approach is robust to the reverberate condition; however, the performance in the HRTF-mismatched condition is not studied.

In this paper, we focus on the binaural localization in the mismatched HRTF condition rather than the reverberant and noisy conditions. Although in [9] and [1] a parametric model is proposed and the parameters of different HRTFs are considered to improve the generalization performance, the model may be relatively simple and may not be able to accurately analyze the localization mechanism. Due to the powerful modeling capability, DNN is effective in many areas. In [18], the DNN is introduced and shows significant performance. However, this work only focuses on the localization performance in noisy and reverberate environments rather than mismatched HRTF condition, which shows that the DNN trained by one HRTF generalize well on the test set generated by another HRTF. While we think this result may require further study, here, we consider the binaural localization problem as a classification problem, and we use DNN to map the binaural cues to the sound localization. Firstly, we use DNN to learn the relationship between binaural cues and the localization of the sound source and then compare the localization performance in the matched and mismatched HRTF conditions. The result shows that the localization performance in the matched HRTF condition is good, but the performance varies with the HRTF in the mismatched HRTF condition and the result shows a clustering trend. To improve the generalization performance, a basic idea is to introduce more HRTFs in training sets; however, this may result in data-inefficiency. Secondly, on this basis, clustering analysis is applied to the localization similarity between each HRTF. Different HRTFs are divided into several clusters. The result shows that the HRTF corresponding to each cluster center is a reasonable approximation of other HRTFs in the same cluster. Finally, the HRTFs corresponding to each cluster center are selected to generate a new training set and to train a more generalized DNN model. Compared with the baseline methods in [1, 2, 9, 18] in the mismatched HRTF condition, our method achieves better performance. Compared with the DNN trained by all HRTFs, our method achieves similar performance with low data computation, which means it is data-efficient.

The remainder of this paper is organized as follows. The localization performance of the method proposed in [18] is further studied in Section 2. Based on the result in Section 2, the proposed method is described in detail in Section 3. Section 4 presents the binaural sound source localization experiments and the analyses. Finally, the conclusion is drawn in Section 5.

2 Localization performance in the matched and mismatched HRTF conditions

In the binaural localization work based on DNN proposed by Ma et al. [18], the DNN model is trained by the training set generated by only one HRTF in the training stage, and the model is tested by the test set generated by two other HRTFs in the testing stage. The result shows that when the localization error tolerance is 5°, i.e., the estimation is considered correct, and if the error between the azimuth estimated by DNN model and the ground truth azimuth is less than or equal to 5°, the DNN model is considered to generalize well on the test sets. In this section, we carry out research on the localization performance at a lower localization error tolerance.

In Ma's work, the training set is generated by the HRTF recorded by the KEMAR dummy head, therefore we choose the 45th HRTF in the CIPIC database [22] which is also recorded by KEMAR dummy head and trains the DNN model similar to that in Ma's work. Different test sets are generated by each HRTF in CIPIC. On this basis, we analyze the localization performance of the DNN model in the matched and mismatched HRTF conditions.

The localization performance of the trained model on each test set is shown in Fig. 1, and the localization error tolerance is 0°, 5° and 10° respectively. More details about data sets generation and network parameters setting are shown in Section 4. As we can see in Fig. 1, when tolerance is 5° and 10°, the localization accuracy of the DNN model on the test set generated by the 45th HRTF is 100%, which indicates that the DNN model performs very well in the matched HRTF condition; the localization accuracy on the test sets generated by other HRTFs is also high, which shows that the DNN model generalized well in the mismatched HRTF conditions when the tolerance is 5° and more. This result is similar to that in Ma's work. When the tolerance is 0°, the localization performance in the matched HRTF condition remains good. However, the localization accuracies on the test sets generated by other HRTFs decrease, which means that the localization performance will decline in the mismatched HRTF condition. In addition, it can be seen that the localization

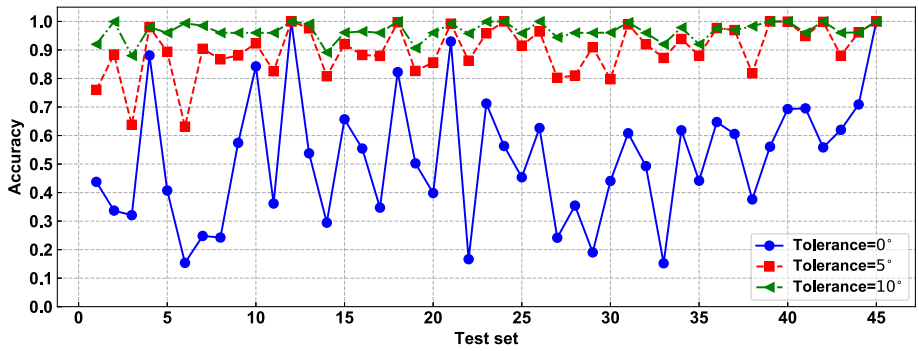

Fig. 1 The localization accuracy of Ma's method on different test sets

accuracies on many test sets generated by mismatched HRTFs are greatly decreased, while the localization accuracies on the test sets generated by a small part of HRTFs, such as the 4th, 10th, 12th and 21st HRTFs, are still high, especially on the test set generated by the 12th HRTF, the localization accuracy of which is also 100%. This result shows that if a DNN model has been trained by the 45th HRTF, the binaural sound generated by the 12th HRTF could be accurately localized by the existing model, instead of training a new model by the 12th HRTF. In fact, the 12th and the 45th HRTF in CIPIC database are all recorded by KEMAR dummy head, therefore the two HRTFs are similar, which indicate that each one of the DNN localization models trained by a pair of similar HRTFs respectively may take the place of the other one. Similarly, the models trained by the 45th HRTF on the test sets generated by the 4th, 10th, and 21st HRTFs all achieve high performance, so the models trained by them respectively could replace each other to some extent. Based on this result, we think that by selecting the most representative HRTF among the similar HRTFs to train a DNN model, we can accurately localize the binaural sounds generated by mismatched but similar HRTFs. Therefore, we propose a DNN- and clustering-based binaural sound source localization method to improve the localization performance in the mismatched HRTF condition.

3 The proposed method

3.1 Overview

The diagram of the proposed method is shown in Fig. 2. The method consists of three stages.

In stage 1, we study the localization similarity between HRTFs. Specifically, we study the localization performance of the DNN model in the matched and mismatched HRTF conditions. Firstly, we select the individual HRTF of the ith subject in the HRTF database and denote it as HRTF i; then, the clean speech signals are filtered by the HRTF data in each direction of HRTF i to generate the training set i and the test set i. Secondly, a DNN model is trained by the features and labels extracted from the training set i and denoted as DNN i. Finally, the DNN i is evaluated by the test set j, and the localization accuracy is denoted as $\mathrm{Acc}(i,j)$.

In stage 2, the clustering analysis is applied to HRTFs. Specifically, we apply the clustering analysis to the localization accuracy of each DNN model on each test set, i.e., the $\mathrm{Acc}(i,j)$ obtained in stage 1, and then different HRTFs are divided into k clusters.

In stage 3, we improve the generalization ability of the DNN model. Specifically, we select central HRTFs, i.e., the HRTFs corresponding to the center of each cluster, to generate a new training set and train a more generalized DNN model.

3.2 Localization similarity analysis based on DNN

In stage 1, we study the localization similarity between HRTFs based on DNN. The diagram of the DNN-based binaural localization method is shown in Fig. 3, and it consists of two modules: the feature extraction and the DNN classification. In the feature extraction module, the binaural features are extracted from the binaural sound as the input features, and the corresponding azimuths are

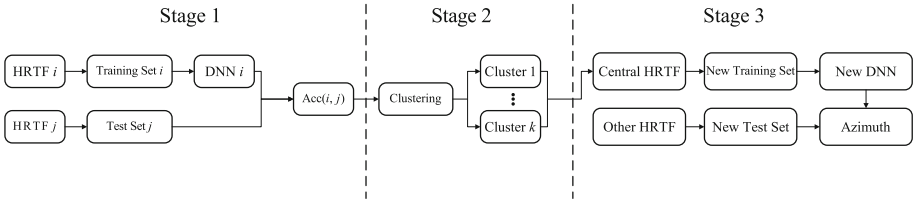

Fig. 2 The diagram of the proposed method

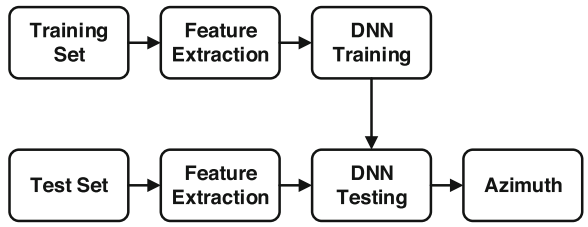

Fig. 3 The flowchart of the DNN-based localization method

converted into the one-hot vectors which are often used in DNN classification tasks as the output features. For the DNN classification module, in the training stage, the DNN is trained by the input features and output features. After the DNN model is well trained, the input features extracted from the binaural sound to be estimated are fed into the DNN model in the test stage. The posterior probability of the sound source in each azimuth is obtained, and the azimuth corresponding to the highest probability is taken as the estimated azimuth.

3.2.1 Binaural feature extraction

The ITD and ILD are commonly used as the binaural features in the related works of binaural localization, while in the recent work of Ma et al. [18], it is shown that normalized cross-correlation function (CCF) [23] contains more information than ITD, and the combined features of CCF and ILD perform better than those of ITD and ILD. Therefore, we combine the CCF and ILD as the input features. In Ma's work, the multiple sound source localization is studied based on the assumption that each time-frequency point is dominated by only sound source [24]. The binaural sound signals are filtered by Gammatone [25] filter bank to obtain several subbands, and then the binaural features are extracted from each subband and employed to train a DNN model, respectively. In this paper, we study the localization of single sound source; therefore, we extract the binaural features from the whole frequency band.

To extract the features from the binaural sound, we divide the signal in each channel into frames, and then the CCF feature is calculated as follows:

$$\mathrm{CCF}(t,\tau)=\frac{\sum_m (x_{t,l}(m)-\bar{x}_{t,l})(x_{t,r}(m-\tau)-\bar{x}_{t,r})}{\sqrt{\sum_m (x_{t,l}(m)-\bar{x}_{t,l})^2}\sqrt{\sum_m (x_{t,r}(m-\tau)-\bar{x}_{t,r})^2}}, \tag{1}$$

where $x_{t,l}$ and $x_{t,r}$ refer to the signals; t refers to the time frame index; l and r refer to the left and right channel, respectively; m refers to the sample index; τ refers to the time lag; and $\bar{x}_{t,l}$ and $\bar{x}_{t,r}$ refer to the mean values in one frame. Considering the radius of the human's head and the speed of sound, the range of τ is $[-1, 1]$ ms. For the signals sampled at 16 kHz, the range of the sample lag is $[-16, 16]$; therefore, the dimension of the CCF feature is 33.

The ILD feature of the tth frame is calculated as follows:

$$\mathrm{ILD}(t)=10\log_{10}\frac{\sum_m x_{t,l}^2(m)}{\sum_m x_{t,r}^2(m)}, \tag{2}$$

and the dimension of the ILD feature is 1.

Finally, we combine the CCF and ILD to obtain the 34-dimensional input feature:

$$I(t)=[\,\mathrm{CCF},\mathrm{ILD}\,]. \tag{3}$$

For DNN classification tasks, the category of the sample is usually converted to a one-hot vector as the output feature. Specifically, if the number of the categories is n, the dimension of the one-hot vector is also n. Assuming that all samples could be classified into n categories and a sample belongs to the ith category, the ith value of the corresponding one-hot vector is set to 1 and others are set to 0.

3.2.2 The DNN architecture

The architecture of the DNN is shown in Fig. 4, which consists of an input layer, three hidden layers, and an output layer. The number of the nodes in the input layer is 34, which is equal to the dimension of the input feature. The number of the nodes in each hidden layer is determined by subsequent experiments. After each hidden layer, the Rectified Linear Unit (ReLU) [26] is used as the activation function, and after ReLU the dropout [27] layer is applied to prevent over-fitting. The number of nodes in the output layer is determined by the number of azimuths. After the output layer, a softmax activation function is applied.

In the training stage, the Adam [28] optimizer is used to minimize the cross-entropy loss between the ground truth and the estimated result:

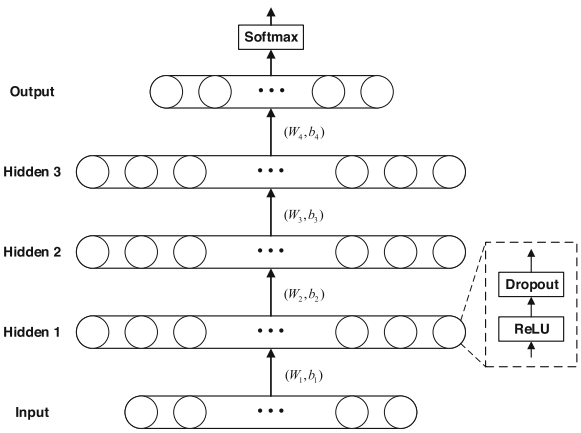

Fig. 4 The architecture of the DNN

$$\text{Loss} = \frac{1}{T}\sum_{t=1}^{T}\left(-\sum_{d=1}^{D} O(t,d)\log\widehat{O}(t,d)\right), \tag{4}$$

where the $O(t,d)$ and $\widehat{O}(t,d)$ refer to the ground truth and the estimated output feature, respectively; the t and T refer to the frame index and the number of frames in one batch; and the d and D refer to the dimension index and the dimension size.

In the test stage, the input feature $I(t)$ is extracted from the binaural sound to be estimated and fed into the well-trained DNN model, and then the posterior probability $P(\theta|I(t))$ of the sound source in each azimuth is obtained. After that the average posterior probability of T frames is calculated as follows:

$$P(\theta) = \frac{1}{T}\sum_{t=1}^{T} P(\theta|I(t)). \tag{5}$$

Finally, the azimuth corresponding to the maximum posterior probability is taken as the localization result:

$$\hat{\theta} = \arg\max_{\theta} P(\theta). \tag{6}$$

The localization performance is measured by the localization accuracy and the localization error. The localization accuracy is calculated as follows:

$$\text{Acc} = \frac{N_{\left|\hat{\theta}-\theta_g\right|\le\Theta}}{N}, \tag{7}$$

where N refers to the number of the binaural sounds; the θ_g and $\hat{\theta}$ refer to the ground truth azimuth and the estimated azimuth, respectively; and the Θ refers to the localization error tolerance mentioned in Section 2. When the error between the estimated azimuth and the ground truth azimuth is less than or equal to the threshold, the estimation is considered correct; otherwise, it is considered incorrect. In this paper, we set $\Theta = 0°$, which means that the estimation is considered correct only when the estimated azimuth and the ground truth azimuth are exactly equal. The localization error is calculated as follows:

$$\text{Err} = \frac{\sum_{n'=1}^{N} |\hat{\theta} - \theta_g|}{N}, \tag{8}$$

where N, θ_g, and $\hat{\theta}$ refer to the number of the sounds, the ground truth azimuth, and the estimated azimuth, respectively; the n' refers to the index of the sound.

3.2.3 Localization similarity between HRTFs

When training the DNN, the size of the training set, the number of the nodes in the hidden layer and the number of the hidden layers will affect the performance of the model. Therefore, we compare the localization performance of the DNN models corresponding to different training set sizes and different hidden layer nodes. To generate the training sets of different sizes, we select the first HRTF in the CIPIC database, and use 100 and 1000 clean speech to convolve with the HRIR pair corresponding to each azimuth, respectively. Then, a training set containing 2500 binaural sounds and a training set containing 25,000 binaural sounds are obtained. For the number of nodes in the hidden layer, three cases are investigated: 200, 500, and 1000, and the number of nodes in the output layer is 25. More detailed settings are presented in Section 4.

Figure 5 shows the effect of different training set sizes on localization performance when the number of nodes in the hidden layer is fixed. It can be seen that the trained DNN model performs well on the test set generated by the first HRTF, while the localization accuracy of the model on the test set generated by the mismatched HRTF decreases. When the number of the nodes in the hidden layer is fixed, the localization performance will be slightly improved with a larger training set size.

Figure 6 shows the effect of the number of nodes in the hidden layer on localization performance when the size of the training set is fixed. It can be seen that when the size of the training set is fixed, the localization performance will also be slightly improved with more nodes in the hidden layer. The average localization accuracy of the DNN model corresponding to different training sets sizes and hidden layer nodes are shown in Table 1. From Table 1, it can be seen that the larger the training set and the more nodes in the hidden layer, the better the localization performance; however, the improvement is limited. We further study the influence of the number of hidden layers and hidden layer nodes on localization accuracy. The results are shown in Table 2, which indicates that these two factors have limited impact on the performance. Considering the work of Ma [18], we set the number of binaural sounds in the training set to 2500, the number of nodes in the hidden layer to 200, and the number of layers to 3 in the subsequent experiments.

From Fig. 5, it can also be seen that the DNN model trained by the HRTF 1 achieves high performance on the test sets generated by the HRTF 1, HRTF 10, HRTF 19, and HRTF 43, respectively. Figure 7 shows the localization performance of the DNNs trained by HRTF 1 and HRTF 10 on the test sets generated by different HRTFs. Figure 7 indicates that the localization accuracy of the trained DNN in the matched HRTF condition is 100%, and the DNN trained by HRTF 10 also performs well on test sets generated by the HRTF 19 and HRTF 43, respectively. This indicates that if two DNNs are trained by a pair of similar HRTFs respectively, the localization performances of them on the same test set are also similar, and there is a certain clustering trend between different HRTFs.

Figure 8 shows the localization accuracies of the different DNNs on different test sets. For the jth row in ith

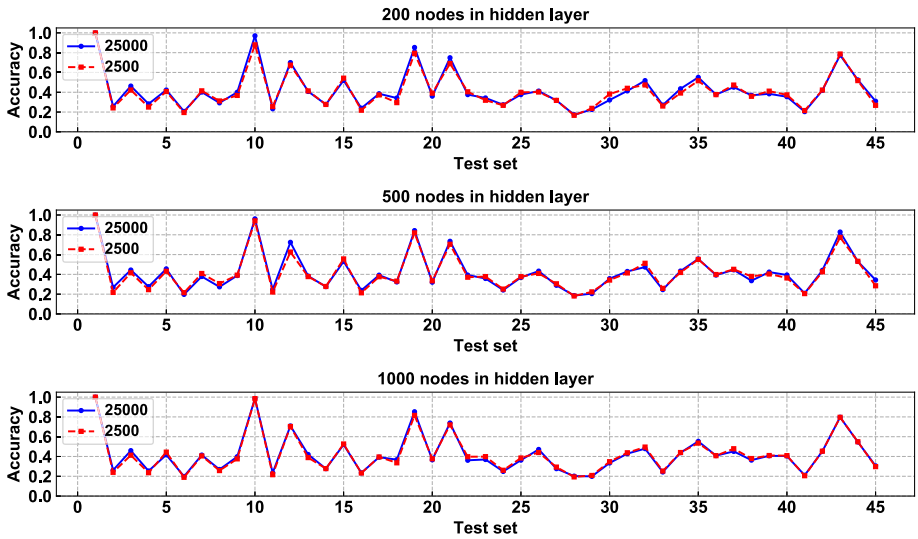

Fig. 5 The localization performance of the DNN trained with different training set size on different test sets. The training set is generated by the first HRTF in CIPIC

column, it means a DNN is trained by the training set generated by the HRTF i and then evaluated on the test set generated by the HRTF j. It can be seen that localization accuracies between different HRTFs are different, and the accuracies on the diagonal line correspond to the matched HRTF condition and all reach to 100%. Besides the diagonal line, the localization accuracy at (21,12) is also 100%, which indicates that the similarity between the HRTF 12 and HRTF 21 is very high and the DNNs trained by them respectively can be substituted for each other. The localization accuracy at (6,27) is the lowest, reaching to 2.64%, which indicates that there is a great difference between the HRTF 6 and HRTF 27. For the HRTF 6, HRTF 7, and HRTF 38, the localization accuracies among them are relatively high. However, the localization accuracies between each of them and other HRTFs are low, which may be due to the special physiological structure of the corresponding subjects which are different from that of most subjects.

3.3 Clustering analysis based on affinity propagation

Figure 8 shows the localization accuracies of the DNNs trained by each HRTF respectively on different test sets generated by different HRTFs, and it also reflects the localization similarity between HRTFs. Therefore, the cluster analysis could be applied based on this similarity matrix to find the most representative HRTFs. The DNN trained by a representative HRTF will perform

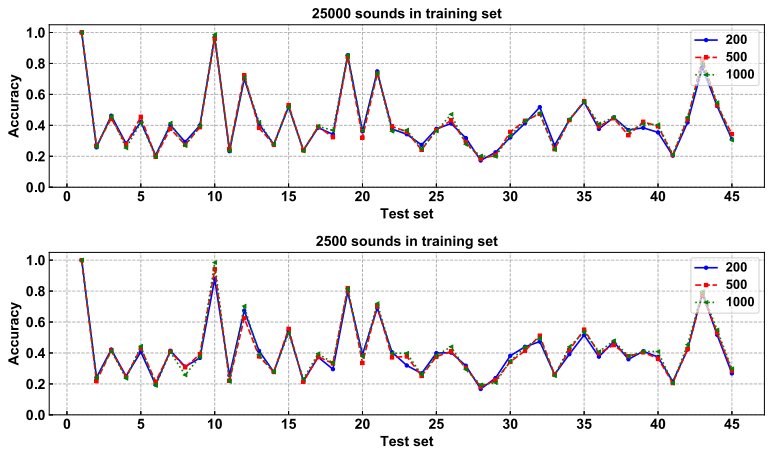

Fig. 6 The localization performance of the DNN with different numbers of nodes in hidden layers on different test sets. The training set is generated by the first HRTF in CIPIC

Table 1 The average localization performance of the DNN trained with different training set sizes and nodes number in the hidden layer

	200 nodes	500 nodes	1000 nodes
25,000 sounds	42.08%	42.15%	42.41%
2500 sounds	41.32%	41.39%	42.23%

well on the test sets generated by the unmatched but similar HRTFs.

In stage 2, we apply the clustering analysis to the similarity matrix based on the affinity propagation (AP) algorithm. Compared with the common clustering algorithms such as K-means and hierarchical clustering, the AP algorithm has many unique advantages that meet the requirements in this paper:

- When clustering the HRTFs based on localization accuracy, we have no prior information about the number of clusters. With the AP algorithm, the number of the clusters does not need to be specified beforehand like the K-means.
- After clustering, the HRTFs corresponding to the centers of each cluster are selected to train a more generalized DNN model. If we use the K-means algorithm, the center of the cluster may be the average of multiple sample points instead of an existing sample point, which will lead to a result that the center may not correspond to a real HRTF, so that the stage 3 of the proposed method cannot be carried out. The AP algorithm treats all sample points as the potential cluster center, and the center of the cluster is an existing sample point, therefore the center corresponds to a real HRTF.
- When testing, the accuracy of the DNN model trained by HRTF i on the test set generated by HRTF j may not equal to the accuracy of the DNN model trained by HRTF j on the test set generated by HRTF i, in other words, for the similarity matrix, there may be $\mathrm{Acc}(i,j) \neq \mathrm{Acc}(j,i)$, which limits the application of the K-means and hierarchical algorithms, etc. However, the AP algorithm can be applied to the problems where the similarity matrix is not symmetric.

Table 2 The average localization performance of the DNN trained with different number of hidden layers and different number of nodes in the hidden layer

Number of nodes	Number of hidden layers		
	1	2	3
25	42.10%	41.66%	41.57%
50	41.60%	41.05%	41.35%
100	41.34%	41.44%	41.55%

With these advantages, the AP algorithm is used in stage 2.

3.3.1 Affinity propagation

For the sample points to be clustered, the AP algorithm takes the similarity between each pair of sample points as an input and the cluster center of each sample point as an output. At the beginning of clustering, the AP algorithm treats each sample point as a potential cluster center, searches the cluster center of each point by exchanging messages between sample points and updates the affiliation between the sample point and the data center. Then, the data set is divided according to the affiliation, and the cluster center $C\left(i^{'}\right)$ of each sample point is obtained. There are two kinds of messages to be exchanged, namely "responsibility" and "availability." The responsibility is represented by $R\left(i^{'},k^{'}\right)$, and it is sent from the point $i^{'}$ to the candidate cluster center point $k^{'}$, which indicates the evidence for how appropriate it would be for the point $k^{'}$ served as the cluster center of the point $i^{'}$. The availability is represented by $A\left(i^{'},k^{'}\right)$, and it is sent from the candidate cluster center point $k^{'}$ to the point $i^{'}$, which shows the evidence for how appropriate it would be for the point $i^{'}$ to choose the point $k^{'}$ as the cluster center.

Initially, the availabilities are set to zero: $A\left(i^{'},k^{'}\right)=0$. Then, the responsibilities are updated as follows:

$$R\left(i^{'},k^{'}\right) \leftarrow S\left(i^{'},k^{'}\right) - \max_{j^{'}:j^{'}\neq k^{'}} \left\{A\left(i^{'},j^{'}\right) + S\left(i^{'},j^{'}\right)\right\}, \quad (9)$$

where the $S\left(i^{'},k^{'}\right)$ is the similarity which indicates how well the sample point $k^{'}$ is suited to be the cluster center for sample point $i^{'}$. For $k^{'}=i^{'}$, the $S\left(k^{'},k^{'}\right)$ refers to the "preference" and is denoted as p. The larger the value of p, the point $k^{'}$ is more likely to be chosen as the cluster center. The number of clusters is influenced by p. In general, the p could be the median of the similarity matrix, which leads to a moderate number of clusters.

The availabilities are updated as follows:

$$A\left(k^{'},k^{'}\right) \leftarrow \sum_{j^{'}:j^{'}\neq k^{'}} \max\left\{0, R\left(j^{'},k^{'}\right)\right\}, \quad (10)$$

$$A\left(i^{'},k^{'}\right) \leftarrow \min\left\{0, R\left(k^{'},k^{'}\right) + \sum_{j^{'}:j^{'}\notin\{i^{'},k^{'}\}} \max\left\{0, R\left(j^{'},k^{'}\right)\right\}\right\}. \quad (11)$$

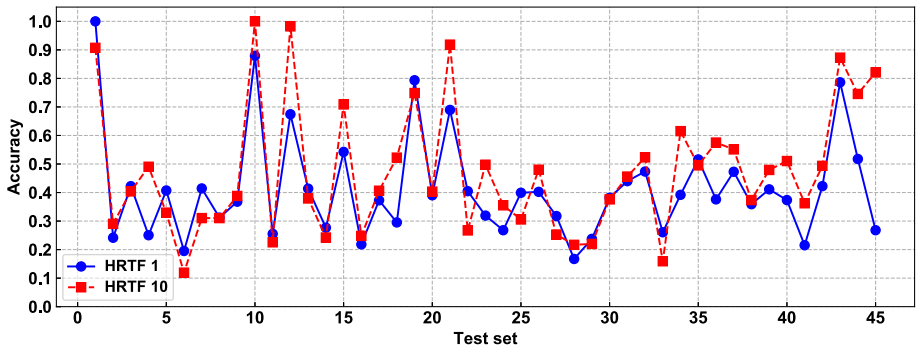

Fig. 7 The localization performance of the DNNs trained by HRTF 1 and HRTF 10 respectively on different test sets

The cluster centers of each sample point are updated as follows:

$$C\left(i^{'}\right) \leftarrow \arg\max_{k^{'}} R\left(i^{'},k^{'}\right) + A\left(k^{'},i^{'}\right). \quad (12)$$

To avoid parameter oscillation during the iterations, the AP algorithm introduces a damping coefficient λ between 0.5 and 1 when updating the messages. The value of each message is weighted by its last updated value and current value:

$$R_{v+1}\left(i^{'},k^{'}\right) \leftarrow (1-\lambda)R_{v+1}\left(i^{'},k^{'}\right) + \lambda R_{v}\left(i^{'},k^{'}\right), \quad (13)$$

$$A_{v+1}\left(i^{'},k^{'}\right) \leftarrow (1-\lambda)A_{v+1}\left(i^{'},k^{'}\right) + \lambda A_{v}\left(i^{'},k^{'}\right). \quad (14)$$

In each iteration of the AP algorithm, the responsibility and availability are updated; then, the result is updated. The iteration will be terminated if any of the following three conditions occurs:

- The preset number of the iterations is reached.
- Changes in the messages fall below a threshold.
- The local decisions stay constant for some number of iterations.

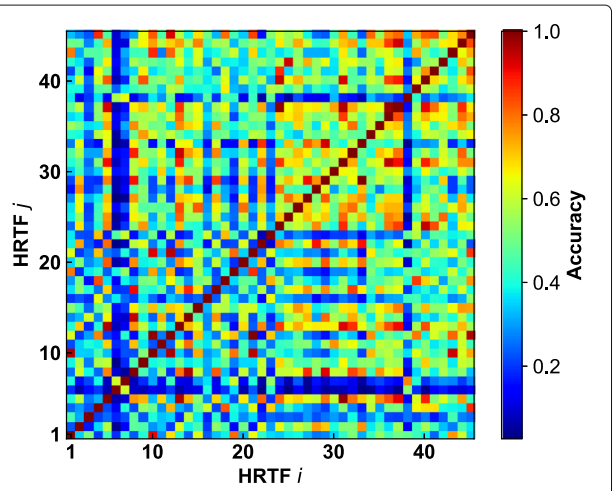

Fig. 8 Localization similarity between HRTFs

3.3.2 *Result of clustering analysis*

For the AP algorithm, the two most important parameters are the preference p and the damping coefficient λ. The number of clusters is influenced by the values of the preferences and the message-passing procedure. λ will affect the convergence of the algorithm. Figure 9 shows the number of clusters corresponding to different values of p and λ, where p ranges from -5 to 5 with the step of 0.01, and λ ranges from 0.5 to 0.95 with the step of 0.05. It can be seen that the larger the value of p, the more the number of clusters, and the smaller the value of p, the fewer the number of clusters. Besides, the change of λ has little effect on the number of clusters, while the change of p has a greater influence on the number of clusters. Based on this result, the clustering results at $\lambda = 0.5$ will be further studied in stage 3.

Figure 10 shows the effect of the p on the number of clusters at $\lambda = 0.5$. When the value of p is small, for example, when $p = -4.6$, the number of clusters is 1, which means all HRTFs belong to the same cluster. When the value of p is large, for example, when $p = 1.1$, the number of clusters is 45, which means each HRTF belongs to a separate cluster. When the value of p is set to the median of the similarity matrix, the number of clusters is 7, and the corresponding clustering results are shown in Table 3. It can be seen that the number of HRTFs in each cluster is different. The largest cluster is the cluster 3 with 11 HRTFs, and the smallest cluster is the cluster 7 with only 3 HRTFs.

Table 4 shows the average localization accuracy of each DNN trained by the HRTF corresponding to the center of each cluster respectively on each test set generated by all HRTFs in each cluster respectively. It shows that the DNN trained by the HRTF corresponding to the center of a certain cluster performs better on the test set generated by the HRTFs from the same cluster than on the test sets generated by the HRTFs from other clusters. On the

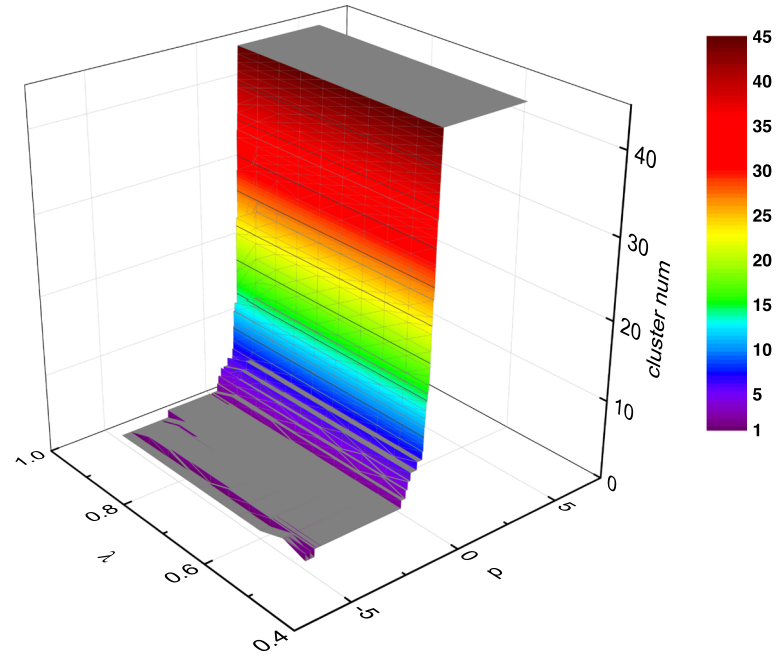

Fig. 9 Cluster numbers of different p and different λ

diagonal line, six of seven localization accuracies are more than 80%, which indicates that the HRTF corresponding to the cluster center obtained by AP algorithm is a reasonable approximation of other HRTFs in the same cluster, and the DNN model trained by the HRTF corresponding to the cluster center can provide good localization performance for the test sets generated by HRTFs in the cluster.

3.4 Improving the generalization performance

Generally speaking, in deep learning, the more diverse the data used for training, the better the generalization performance of the trained model. Therefore, in stage 3, we propose to generate a new training set by multiple HRTFs to train the DNN model. However, using all HRTFs to train a DNN may lead to data-inefficiency. Considering that the DNN trained by the HRTF corresponding to each cluster center has better localization performance on the test sets generated by the HRTFs in the same cluster, we select the HRTF corresponding to each cluster center to generate a new training set jointly. Therefore, a more generalized DNN model will be obtained by just a few HRTFs.

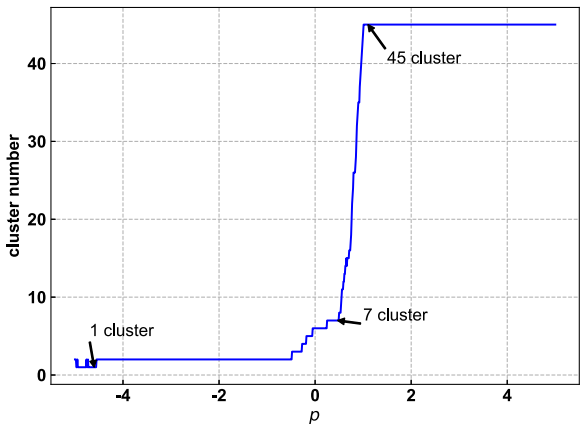

Fig. 10 Cluster numbers of different p at $\lambda = 0.5$

To compare the generalization performance of DNN trained by the HRTFs corresponding to the cluster center (central HRTF) and DNN trained by the HRTFs not corresponding to the cluster center (non-central HRTF), we set p as the median of the similarity matrix, and the result of clustering is shown in Table 3. Then, we choose seven central HRTFs to train the DNN model, and the indexes are 4, 10, 13, 14, 24, 33, and 38. After that we select the same number of non-central HRTFs to train the DNN model in the same way. We repeat the experiment three times. The first group of non-central HRTF comes from all clustering respectively, and the indexes are 3, 6, 16, 20, 30, 34, and 39; the second group comes from the cluster 2, and the indexes are 1, 12, 15, 19, 21, 34, and 43; the third group comes from the third cluster 3, and the indexes are 5, 26, 31, 32, 35, 36, and 44, respectively. In the test stage of each comparison, the test sets are generated by the HRTFs which have not been used for training the DNN

Table 3 The cluster result when the number of clusters is 7

Cluster index	HRTF number	HRTF index	Central HRTF indexes
Cluster 1	5	4,16,18,23,45	4
Cluster 2	8	1,10,12,15,19,21,34,43	10
Cluster 3	11	5,13,25,26,30,31,32,35,36,42,44	13
Cluster 4	5	3,8,11,14,17	14
Cluster 5	7	9,24,28,37,39,40,41	24
Cluster 6	6	2,20,22,27,29,33	33
Cluster 7	3	6,7,38	38

model, therefore the test sets in the three comparison experiments are different. Then, the average localization accuracy on all test sets is calculated.

Figure 11 shows the results of the experiments. It can be seen that the localization accuracy of DNN model trained by central HRTFs is always better than that of the DNN model trained by non-central HRTF. This is because the central HRTF is more representative than the non-central HRTF, and the DNN trained by the central HRTF will be more generalized.

4 Experiments and analyses

4.1 Dataset generation

In this paper, the HRTFs used to generate the binaural sound come from the CIPIC database and RIEC database[29]. The CIPIC database contains HRTFs of 45 subjects. Those HRTFs are measured at the distance of 1 m with 25 different azimuths and 50 different elevations, resulting in 1250 HRTF pairs for each subject. Here, we will focus on the binaural localization on the frontal horizontal azimuth plane, so we consider 25 azimuths sampled at $-80°$, $-65°$, $-55°$, from $-45°$ to $45°$ in steps of $5°$, at $55°$, $65°$, and $80°$.

The speech samples used to generate the binaural sounds come from the TIMIT database [30], and the sampling rate is 16kHz. To match the sampling rate of the HRTFs in CIPIC, the speech samples are upsampled to 44.1 kHz and convolved with HRIR pair corresponding to a certain azimuth to generate the binaural sounds, and then the binaural sounds are downsampled to 16 kHz.

In stage 1, according to the conclusion of section 3.2.3, for the HRTF i, we select 100 speech samples to convolve them with the HRIR pair corresponding to each azimuth respectively to generate the training set i, and we select 50 different speech samples for the validation set i and 50 different speech samples for the test set i. Therefore, there is no overlap among the training set, the validation set and the test set.

In stage 3, we set the number of clusters to 7, and combine the training sets corresponding to the central HRTFs to generate a new training set. The new validation set is generated in the same way. In the test stage, the new DNN is tested on each test set generated by the HRTFs which have not been used for training DNN.

To further test the generalization performance of the proposed method on other HRTF database, we also select the HRTF of the first 50 subjects in the RIEC database to generate the RIEC test sets. The generation procedure of RIEC test sets is similar to that of CIPIC test sets. Although the HRTF in RIEC library is recorded in the spherical coordinate system, and the HRTF in CIPIC is recorded in the binaural polar coordinate system, these two coordinate systems are equivalent on the horizontal

Table 4 The average localization performance of the DNNs trained by the central HRTF in each cluster

Test set	Central HRTF						
	HRTF 4	HRTF 10	HRTF 13	HRTF 14	HRTF 24	HRTF 33	HRTF 38
Cluster 1	*81.22%*	51.58%	36.75%	27.10%	43.82%	18.62%	38.86%
Cluster 2	51.22%	*84.42%*	40.51%	30.18%	39.71%	25.62%	40.09%
Cluster 3	41.64%	46.90%	*82.38%*	59.60%	66.88%	54.07%	19.11%
Cluster 4	27.49%	31.78%	60.03%	*84.38%*	41.50%	54.40%	19.70%
Cluster 5	49.83%	40.90%	62.58%	41.49%	*83.38%*	46.49%	18.33%
Cluster 6	26.67%	26.55%	70.73%	61.75%	57.83%	*86.81%*	8.03%
Cluster 7	31.97%	26.77%	11.47%	12.51%	15.57%	5.28%	*76.83%*

Each test set is generated by the HRTFs in the same cluster

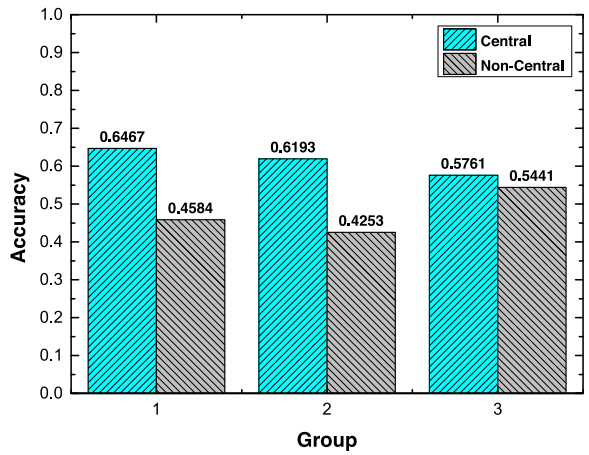

Fig. 11 The average localization performance of the DNNs trained by central HRTFs and non-central HRTFs respectively on all test sets

frontal plane. For each subject, we select the HRIR pairs corresponding to the same 25 azimuths to be convolved with 50 speech samples to get the RIEC test sets.

4.2 Experimental setting

The architecture of the DNN is determined in stage 1. The number of nodes in each hidden layer is 200, and the number of layers is 3. In the training stage, the parameters of DNN are randomly initialized, and then the input and output features are extracted from the training set to train DNN. The number of samples is set to 200 in each batch, and the learning rate is set to 10^{-4}. To prevent over-fitting, the dropout rate is set to 20%, and the early stop strategy is adopted. The training is stopped if the best validation accuracy has not been updated for 10 epoch.

4.3 Baseline methods

In this paper, the proposed method is denoted as Proposed. For comparison, we choose four existing methods as the baseline methods: the average parameter model method proposed by Raspaud [9], Pang's generalized parameter model method [1] based on Raspaud [9], Ma's [18] method based on DNN, and the convention method based on TDOA [2]. The basic procedure of Raspaud's method is firstly, ITD and ILD cues are modeled as the product of a function of azimuth and a function of frequency; then, in the offline stage, ITD and ILD cues corresponding to each azimuth are extracted from the HRTF of each subject in CIPIC database and fed into the model to calculate the parameter corresponding to each subject; and finally, in the test stage, the ITD and ILD cues which are extracted from the sounds are to be estimated, and ILD cue is fed into the average parameter model to estimate the correct ITD, and then the correct ITD is fed into the parameter model to estimate the sound source localization. Based on Raspaud's method, Pang's method introduces reverberation weighting to improve the performance in reverberation environment. At the same time, the least square method is used to replace simple averaging operation when calculating the model parameters. The procedure of Ma's method is firstly, Gammatone filter bank is used to divide the sound into several frequency bands in the training stage; then, the features of each frequency band are extracted to train a DNN model respectively; and after that, in the test stage, the features are extracted from the test sounds and fed into the model, at the same time head movement strategy is combined to reduce the front-back confusion. The procedure of TDOA is firstly, the time delay between the signal in left and right channels is estimated by generalized cross-correlation phase transform; then, the sound source localization is calculated by geometry equations. In subsequent experiments, these methods are denoted as Raspaud, Pang, Ma, and GCC-PHAT, respectively.

When we reproduce the Raspaud method and the Pang method, we refer to the parameter settings in Pang's work [1]. The sound speed is set to 344 m/s, and the head radius γ is set to 7 cm, which is the mean head radius in the CIPIC database. Besides, it is mentioned in Pang's work that appling reverberation weighting on anechoic binaural sound would reduce localization performance. So in this paper, we don't reproduce the reverberation weighting module in Pang's method. When reproducing the method of Ma, one DNN is trained for each subband, and a total of 32 DNNs are trained. The number of nodes in the hidden layer of the network is consistent with the proposed method. The localization on the horizontal frontal plane is studied in this paper and there is no front-back confusion, so we do not reproduce the head movement module in Ma's method. In the training stage, we use the training set generated by the HRTF 45 to train the DNNs in Ma's method. As for GCC-PHAT, the head radius is considered; therefore, the geometry equation is as follows:

$$(\theta_g + \sin\theta_g)\gamma = \Delta t c, \tag{15}$$

where the θ_g refers to the ground truth azimuth, the Δt refers to the time delay, the γ refers to the head radius, and the c refers to the sound speed. For consistency with Pang's work [1], we set the head radius to 7 cm and the sound speed to 344 m/s.

For GCC-PHAT, the estimated azimuth may not be one of the 25 existing azimuths, and we choose the closest one from the existing azimuths as the final azimuth.

4.4 Experimental results

Figure 12 shows the localization accuracy and the error of the proposed DNN- and clustering-based binaural localization method and the baseline methods on the CIPIC test sets in the mismatched HRTF conditions. The

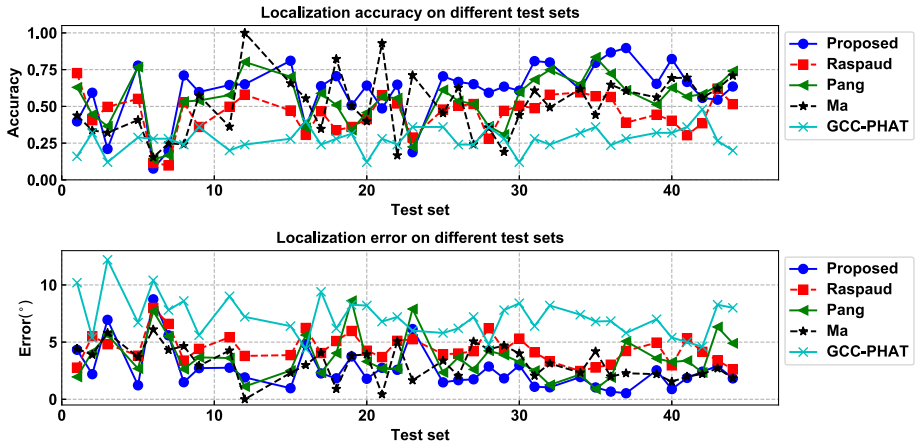

Fig. 12 The localization performance of the proposed method and the baseline methods. The test sets are generated by the HRTFs other than the HRTF 4, 10, 13, 14, 24, 33, 38, and 45

test sets are generated by the HRTFs other than the HRTF 4, 10, 13, 14, 24, 33, 38, and 45. Such HRTFs have not been used for training the DNN of the proposed method and Ma's method. It can be seen that the methods based on HRTF or DNN are better than GCC-PHAT, which is because TDOA is relatively simple and it may not be able to describe the mechanism of the binaural localization precisely. Compared with Raspaud and Pang's methods, our method has higher localization accuracy and lower error on most test sets than those parametric model methods. This is because the parametric model methods only describe binaural cues as the product of frequency and azimuth functions, that is, the parametric model is relatively simple, while DNN is powerful in modeling, and it can perform better in learning the non-linear relationship between localization features and azimuth.

Compared with the method of Ma, our method achieves higher localization accuracy and lower error on most test sets, while Ma's method is better on a few test sets (e.g., test sets generated by the HRTF 12 and 21). It is because Ma's method only considers the HRTF 45 which has a higher localization similarity to the HRTF 12 and 21 and just performs well on such kind of test sets. However, the HRTF 45 has a relatively low localization similarity to other kind of HRTFs. So on other test sets, the localization accuracy of Ma's method will decline and the localization error will increase. The proposed method uses the HRTFs corresponding to the cluster centers which are more representative to generate the training sets and also improve the diversity of data in training sets. Therefore, the DNN model trained by the central HRTFs together can learn the common characteristics and achieves better generalization performance than the DNN models trained by the central HRTF alone.

Figure 13 shows the distribution of the ground truth azimuth versus the estimated azimuth. The point on the diagonal line means the correct estimation that the ground truth azimuth is equal to the estimated azimuth, and the point not on the diagonal line means the wrong estimation. In each graph, the probability of point (X, Y) for all test sets is calculated as follows:

$$P^{'}(X,Y) = \frac{\sum_{j} N_{\theta_g=X,\hat{\theta}=Y}}{\sum_{j} N_{\theta_g=X}}, \tag{16}$$

where the $P^{'}$ refers to the probability, the X and the Y refer to the coordinate values in the graph, and the N is the number of the sounds that meet with a certain condition. The θ_g, $\hat{\theta}$ and j refer to the ground azimuth, the estimated azimuth and the index of the test set, respectively. The larger the point, the higher the probability. It can be seen that the distributions of the Proposed and Ma's methods are better than the others, especially when the azimuth value is larger and the probability of the wrong estimation is lower. This indicates that the DNN-based method is more powerful in modeling.

Table 5 shows the comparison of the average localization accuracy and error between the proposed method and the baseline methods on test sets generated by all the mismatched HRTFs from the CIPIC database. It can be seen that the proposed method achieves higher localization accuracy and lower error than other methods, which also indicates that the proposed method has better generalization performance.

In stage 2, the AP clustering analysis is applied, and different parameters correspond to the different number of central HRTFs. Therefore, in stage 3, the number of the central HRTFs in training sets will be different when training new DNN model. We study the effect of the numbers of the central HRTFs with the proposed method. Figure 14 shows the localization performance

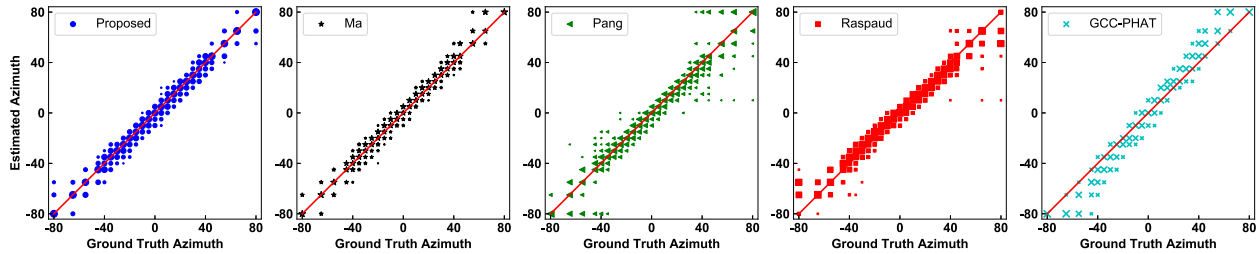

Fig. 13 The distribution of the ground truth azimuth and the estimated azimuth. The larger the point in the graph, the greater the probability

of the proposed method which is evaluated on the test sets generated by all HRTFs from CIPIC database, and then the average localization accuracy is calculated. Considering the better performance of the Pang's and Ma's methods within the baseline methods, we only compare the proposed method with such good methods here. It can be seen that the proposed method is always better than the baseline methods. This is because the DNN model can learn the non-linear relationship between localization features and azimuth more accurately than the parametric model in Pang's method. By introducing the central HRTFs in the training sets, we find that the generalization performance of our method is better than that of the Ma's method which uses non-central HRTF. When the number of central HRTFs is 1, the proposed method has achieved good localization performance. With the increase in the number of central HRTFs, the localization performance has been further improved. It is because that among the test sets the proportion of the HRTF that appeared in the training set is gradually increasing, while the proportion of HRTF that did not appear is gradually decreasing. Therefore, the performance of the DNN on all CIPIC test sets is getting better. The localization performance when the number of the central HRTFs is 40 is slightly lower than that when the number of the central HRTFs is 35. This may be because when the number of the central HRTFs changes from 35 to 40, the newly introduced HRTFs are quite different from most HRTFs in CIPIC database. During the training procedure, to fit the distribution of the data set generated by the newly introduced HRTFs, we adjust the existing distribution which is similar to most of the HRTFs. Although the DNN model improves the localization performance on the newly introduced HRTFs, the localization performance of other HRTFs is reduced, which leads to the decline of the overall localization performance.

Table 5 The average localization performance of the proposed method and the baseline methods

Method	Average accuracy	Average error (°)
Proposed	*60.43%*	*2.60*
Pang	54.37%	3.67
Raspaud	46.11%	4.42
Ma	50.62%	3.17
GCC-PHAT	27.74%	7.19

Figure 15 shows the average localization accuracy of the proposed method on RIEC test sets under different numbers of central HRTFs. It can be seen that our method is still better than Pang's method, which indicates that the proposed method also has good generalization performance on different HRTF databases. When the number of clusters is 1, our method still has good performance. With the increase of the number of central HRTFs, the localization performance does not show the overall trend of improvement as shown in Fig. 14, but first decreases and then increases. This may be because, in the CIPIC database, the HRTFs of European and American races are the majority, while in the RIEC database, the HRTFs of East Asian races are the majority. Differences between the two races lead to different HRTFs in the CIPIC database and the RIEC database. In the experiment of this paper, when the number of clusters is 1, the central HRTF may achieve a balance between European and American races and east Asian races. As the number of central HRTFs increases to 20 under which condition more central HRTFs of European and American races may be introduced to the training set, the differences between the newly introduced HRTFs and the HRTFs in

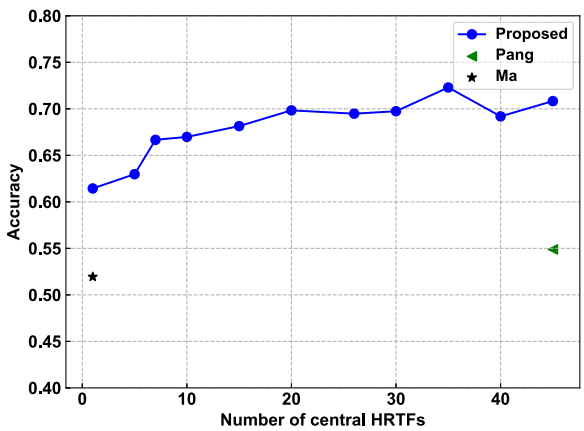

Fig. 14 The average localization performance of the DNN trained by different numbers of central HRTFs on all CIPIC test sets

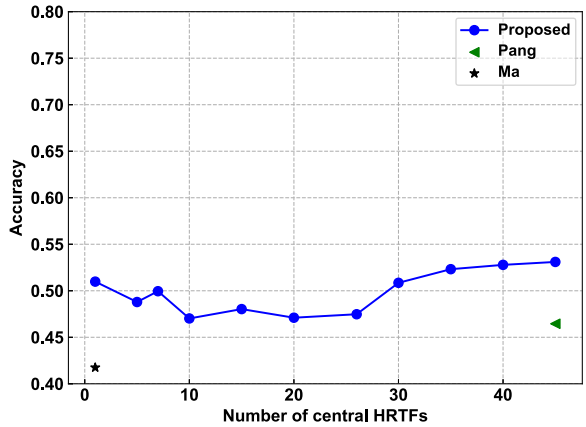

Fig. 15 The average localization performance of the DNN trained by different numbers of central HRTFs on all RIEC test sets

the RIEC test sets may be larger, which decreases the average localization performance on the RIEC test sets. When the number of central HRTFs increases to 45, the balance between European and American races and East Asian races is achieved once again, which improves the average localization performance. However, as the number of the central HRTFs increases, the changes of the localization performance are small. Because each central HRTF is a reasonable approximation of other HRTFs in the same cluster, the DNN trained by the central HRTFs could achieve similar generalization performance to the DNN trained by any other HRTFs in the same cluster. This result indicates that the proposed method is data-efficient.

5 Conclusion

In this paper, we study the binaural localization in the mismatched HRTF condition and propose a binaural sound localization method based on DNN and clustering. Firstly, we introduce the existing binaural localization methods and point out that the performance of these methods will decline in the mismatched HRTF condition. Then, we study the performance of the classification-based DNN localization method in the matched and mismatched HRTF conditions. The results show that in the mismatched HRTF condition, the DNN model has poor localization performance on the test sets generated by most HRTFs. However, it still has good localization performance on the test sets generated by several HRTFs, which indicates that there is the similarity between HRTFs. Then, we analyze the localization similarity among all HRTFs. Specifically, we train the DNNs by the training set generated by each HRTF respectively and test them on the test sets generated by each HRTF, respectively. The results show that there is a clustering trend among HRTFs. We also cluster HRTFs according to the localization similarity between HRTFs. All HRTFs are clustered into several clusters. The localization similarity between the HRTFs from the same cluster is very high, while the localization similarity between the HRTFs from different clusters is low. The HRTF corresponding to each cluster center is a reasonable approximation of other HRTFs in the same cluster. Finally, we select the HRTFs corresponding to the center of the clusters to train a new DNN model to improve the generalization performance. The comparison between the proposed method and the baseline methods shows that the proposed binaural localization method based on DNN and clustering has better generalization performance in the mismatched HRTF condition. In our future work, we plan to conduct in-depth research on more efficient neural network structures and the localization in more complex mismatched conditions such as the condition in which the noise type, the reverberation time, and the HRTF are all mismatched.

Abbreviations

AP: Affinity propagation; CCF: Cross-correlation function; CNN: Convolutional neural network; DNN: Deep neural network; GMM: Gaussian mixture model; HRIR: Head-related impulse response; HRTF: Head-related transfer function; ILD: Interaural level difference; IPD: Interaural phase difference; ITD: Interaural time difference; ReLU: Rectified linear unit

Acknowledgements

The authors would like to thank the reviewers for their suggestions.

Authors' contributions

JW conceived the method and modified the paper, JW performed the experiments and wrote the paper, and JW supervised all aspects of the research. All authors read and approved the final manuscript.

Funding

National Natural Science Foundation of China (no. 61571044 and no. 11590772)

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Received: 5 June 2019 Accepted: 22 January 2020
Published online: 10 February 2020

References

1. C. Pang, H. Liu, J. Zhang, X. Li, Binaural sound localization based on reverberation weighting and generalized parametric mapping. IEEE/ACM Trans Audio Speech Lang Process. **25**(8), 1618–32 (2017)
2. C. Knapp, G. Carter, The generalized correlation method for estimation of time delay. IEEE Trans Acoust Speech Signal Process. **24**(4), 320–327 (1976)
3. G. C. Carter, Variance bounds for passively locating an acoustic source with a symmetric line array. J Acoust Soc Am. **62**(4), 922–926 (1977)
4. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans Antenn Propag. **34**(3), 276–280 (1986)
5. R. Roy, T. Kailath, ESPRIT-estimation of signal parameters via rotational invariance techniques. IEEE Trans Acoust Speech Signal Process. **37**(7), 984–995 (1989)
6. L. A. Jeffress, A place theory of sound localization. J Comp Physiol Psychol. **41**(1), 35 (1948)
7. D. Li, S. E. Levinson, in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03), vol. 5.* A Bayes-rule based hierarchical system for binaural sound source localization (IEEE, 2003), p. 521. https://doi.org/10.1109/icassp.2003.1200021

8. V. Willert, J. Eggert, J. Adamy, R. Stahl, E. Korner, A probabilistic model for binaural sound localization. IEEE Trans Syst Man Cybernet Part B (Cybernet). **36**(5), 982–994 (2006)
9. M. Raspaud, H. Viste, G. Evangelista, Binaural source localization by joint estimation of ILD and ITD. IEEE Trans Audio Speech Lang Process. **18**(1), 68–77 (2010)
10. R. Parisi, F. Camoes, M. Scarpiniti, A. Uncini, Cepstrum prefiltering for binaural source localization in reverberant environments. IEEE Signal Process Lett. **19**(2), 99–102 (2012)
11. B. R. Hammond, P. J. Jackson, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Robust full-sphere binaural sound source localization (IEEE, 2018), pp. 86–90. https://doi.org/10.1109/icassp.2018.8462103
12. F. Keyrouz, Y. Naous, K. Diepold, in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 5*. A new method for binaural 3-D localization based on HRTFs (IEEE, 2006). https://doi.org/10.1109/icassp.2006.1661282
13. F. Keyrouz, K. Diepold, in *2006 IEEE International Symposium on Signal Processing and Information Technology*. An enhanced binaural 3D sound localization algorithm (IEEE, 2006), pp. 662–665. https://doi.org/10.1109/isspit.2006.270883
14. M. Usman, F. Keyrouz, K. Diepold, in *2008 9th International Conference on Signal Processing*. Real time humanoid sound source localization and tracking in a highly reverberant environment (IEEE, 2008), pp. 2661–2664. https://doi.org/10.1109/icosp.2008.4697696
15. J. A. MacDonald, A localization algorithm based on head-related transfer functions. J Acoust Soc Am. **123**(6), 4290–4296 (2008)
16. X. Wan, J. Liang, Robust and low complexity localization algorithm based on head-related impulse responses and interaural time difference. J Acoust Soc Am. **133**(1), 40–46 (2013)
17. J. Woodruff, D. Wang, Binaural localization of multiple sources in reverberant and noisy environments. IEEE Trans Audio Speech Lang Process. **20**(5), 1503–1512 (2012)
18. N. Ma, T. May, G. J. Brown, Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. IEEE/ACM Trans Audio Speech Lang Process (TASLP). **25**(12), 2444–2453 (2017)
19. C. Pang, H. Liu, X. Li, Multitask learning of time-frequency CNN for sound source localization. IEEE Access. **7**, 40725–40737 (2019)
20. E. Thuillier, H. Gamper, I. J. Tashev, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Spatial audio feature discovery with convolutional neural networks (IEEE, 2018), pp. 6797–6801. https://doi.org/10.1109/icassp.2018.8462315
21. P. Vecchiotti, N. Ma, S. Squartini, G. J. Brown, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. End-to-end binaural sound localisation from the raw waveform (IEEE, 2019), pp. 451–455. https://doi.org/10.1109/icassp.2019.8683732
22. V. R. Algazi, R. O. Duda, D. M. Thompson, C. Avendano, in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. The CIPIC HRTF database (IEEE, 2001), pp. 99–102. https://doi.org/10.1109/aspaa.2001.969552
23. N. Roman, D. Wang, G. J. Brown, Speech segregation based on sound localization. J Acoust Soc Am. **114**(4), 2236–2252 (2003)
24. O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans signal Process. **52**(7), 1830–1847 (2004)
25. D. Wang, G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. (Wiley-IEEE Press, 445 Hoes Lane Piscataway, 2006)
26. G. E. Dahl, T. N. Sainath, G. E. Hinton, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Improving deep neural networks for LVCSR using rectified linear units and dropout (IEEE, 2013), pp. 8609–8613. https://doi.org/10.1109/icassp.2013.6639346
27. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint (2012). arXiv:1207.0580
28. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv preprint (2014). arXiv:1412.6980
29. K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, S. Sato, Dataset of head-related transfer functions measured with a circular loudspeaker array. Acoust Sci Technol. **35**(3), 159–165 (2014)
30. J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom. Linguistic Data Consortium (1993). https://doi.org/10.6028/nist.ir.4930

Publisher's Note