ISSN 2616-5775 Vol. 7, Issue 2: 1-11, DOI: 10.25236/AJCIS.2024.070201

HRTF low-dimensional representation based on deep convolutional autoencoder and attention mechanism

Hongxu Zhang[1,a], Wei Chen[2,b,*]

[1]School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China
[2]School of Software, Henan Polytechnic University, Jiaozuo, China
[a]zhx@home.hpu.edu.cn, [b]cw@hpu.edu.cn
[*]Corresponding author

Abstract: *Head-Related Transfer Function (HRTF) depicts the reflection and scattering effects of the environment and the human body on sound during the transmission of sound signals from the sound source to the human ear and contains a large amount of auditory cue information for auditory localization. Due to the high-dimensional complexity and nonlinear nature of the sample data of HRTF itself, it creates difficulties in analyzing the relationship between the auditory localization cues of HRTF and the spatial orientation and morphological features of the human body. The traditional low-dimensional representation makes it difficult to effectively deal with the complex nonlinear relationships between multiple auditory cues in HRTF, resulting in performance degradation. To solve this problem, this study proposes a low-dimensional representation method for HRTF based on a deep convolutional autoencoder. The method considers that HRTF spectral features have the property of continuous variation in three-dimensional space and integrates the nonlinear relationships of full-space HRTF features by modeling the natural spatial attributes of the HRTF ensemble data. Firstly, the attention mechanism is introduced in the encoder, which solves the bias caused by mapping HRTF to a 3D tensor for convolution operation and mines the intrinsic features implied between the spatial orientations of HRTF neighborhoods and neighboring spectra, which improves the low-dimensional representation ability of the network. Secondly, the combination of dense connectivity and attention mechanism in the decoder according to the characteristics of different levels guarantees the effective delivery of low-dimensional features. Experimental results on several publicly available HRTF datasets show that the proposed model outperforms traditional methods in the low-dimensional representation and reconstruction of HRTFs and realizes high-performance low-dimensional representation and reconstruction of HRTFs.*

Keywords: *Head-related Transfer Functions, Convolutional Auto Encoder, Attention Mechanism, Spatial Audio*

1. Introduction

Head-Related Transfer Function (HRTF) depicts sound's reflection and scattering effects by the environment and the human body while transmitting sound signals from the source to the human ear. It contains much information about auditory cues used for auditory localization. In many immersive multimedia applications, such as virtual reality, gaming, and spatial music[1, 2, 3], HRTF is applied to render three-dimensional audio accurately. However, HRTFs are highly personalized and closely related to human physiomorphological features. Using non-personalized HRTFs can lead to front-back confusion, up-down confusion, and inaccurate localization in auditory perception, thus affecting the rendering of 3D audio.

In 3D audio synthesis, personalized HRTFs are ideally required. Although experimental measurements are the most accurate way to obtain personalized HRTFs, they are more difficult to implement due to the need for complex equipment. Therefore, past research has focused on investigating the "mapping relationship" between a user's human morphological features and HRTFs and implementing a modeling approach from human morphological features to personalized HRTFs by establishing a mapping model between these two variables.

HRTFs are continuous functions related to direction, distance, and frequency. Typically, they are stored as a set of Head-Related Impulse Response (HRIR) corresponding to a particular orientation, from which HRTFs are then obtained by Fourier transform.HRTFs are inherently high-dimensional, complex, and nonlinear and are relatively large and complex datasets, typically consisting of hundreds of thousands of samples. This makes it difficult to analyze the relationship between the auditory localization cues of HRTF and the morphological features of the human body. Therefore, many low-dimensional representation models have been developed to simplify HRTF while retaining its features related to auditory localization.

In recent years, low-dimensional representation methods for HRTF based on deep learning have become increasingly popular among researchers. One of the most popular is the AutoEncoder (AE) framework[4], such as the multilayer perceptron (MLP)-based autoencoder, which learns the feature representations of the input data through unsupervised learning and thus realizes the low-dimensional representation of HRTFs. In addition, since Convolutional Neural Network (CNN) have advantages in processing data with spatial structure, researchers have also attempted to apply CNN to HRTF data to extract a more representational low-dimensional representation of HRTFs. For example, Chen et al. trained self-encoders to reconstruct HRTFs along the horizontal plane[5]. In their approach, they represented HRTFs as a special kind of 2D image, where the combination of azimuth and frequency serves as the coordinate axis, and each pixel represents the amplitude of the spectrum. The relationship between neighboring HRTFs is learned by 2D convolution.

The robustness and generalization ability of current deep learning-based low-dimensional representation models for HRTFs is limited and unsatisfactory. This limitation stems from the fact that although deep learning-based methods can better capture the nonlinear relationships between different spatial orientations of HRTFs, they tend to ignore the fact that the spatial distribution of HRTF sample data has the property of spatial inhomogeneity due to the constraints of data collection conditions. The feature extraction solutions based on convolutional neural networks are often oriented to problems with uniform data distribution, so using such solutions directly for HRTF may lead to cutting the feature relationships between the near-neighbor data samples, resulting in certain intrinsic features that are difficult to capture.

To address the above problems, this paper proposes a low-dimensional representation model for HRTF based on a deep convolutional self-encoder and an attention mechanism. The method considers that the spectral features of HRTF in 3D space have the property of continuous variation with spatial orientation and starts from the natural spatial attributes of the HRTF ensemble data to model the nonlinear relationships of the full-space HRTF features as a whole. In the encoder, we design a module based on the attention mechanism, called the spectral space attention module(SSAM), which can help the encoder calculate the importance of each spectral channel and collect global information. In the decoder network, we introduce dense connections to realize the multilayer direct connection of low-dimensional features and screen the effective features through different attention mechanisms, which guarantees the effective delivery of low-dimensional features and improves the accuracy of low-dimensional reconstruction of the model.

2. Proposed Method

Most research on low-dimensional feature extraction of HRTF uses methods based on principal component analysis (PCA)[5, 6, 7]. However, such methods are difficult to effectively express the complex nonlinear relationships among low-dimensional features of HRTF[8]. For this reason, some scholars have proposed using CNN, LSTM, and other methods to study the low-dimensional expression of HRTF[9, 10]. HRTFs in a specific spatial range while ignoring the spatially similar correlation properties of HRTFs. Considering the above reasons, this paper proposes a low-dimensional feature expression model for HRTF based on a 3D convolutional self-encoder, and the overall structure is shown in Figure 1. The model incorporates the spatially proximate correlation property of HRTF into the examination at the same time can effectively express the complex nonlinear relationship among the low-dimensional features of HRTF.

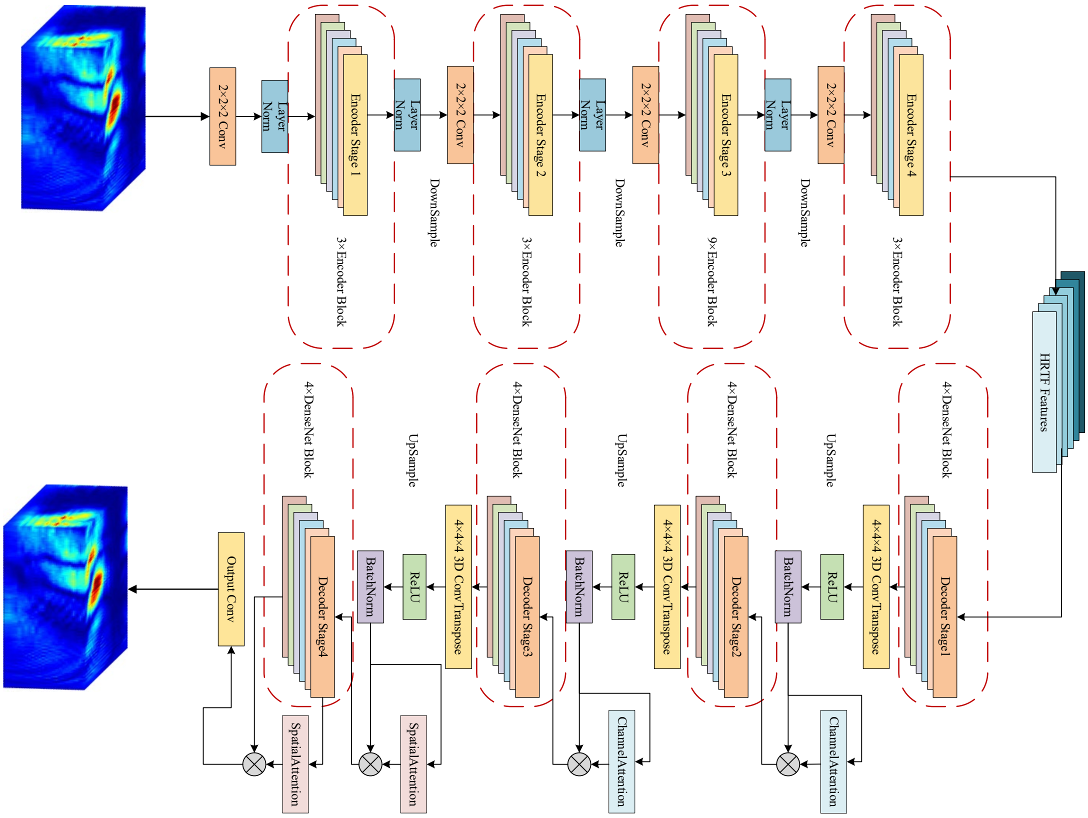

Figure 1: Overall structure of the proposed network.

2.1. Attention-Guided Encoder Network Design

Owing to constraints in data acquisition conditions, the spatial distribution of HRTF sample data exhibits uneven characteristics, with data from specific spatial orientations even missing. Classical feature extraction schemes based on CNN typically aim at problems with uniform data distributions. Therefore, directly applying these solutions to HRTFs might sever the feature relationships between neighboring data samples, making it difficult to capture certain intrinsic features. Consequently, most studies employ spatial fine interpolation techniques on HRTF samples prior to low-dimensional representation to ensure that the spatial organizational structure of the sample data meets the processing requirements of CNNs. However, spatial interpolation increases data volume, leads to information redundancy, and introduces interpolation errors, degrading the model's performance.

To address the issues above, this paper proposes an SSAM. By incorporating a hybrid attention module that combines spatial and channel attention into the encoder model for HRTF's low-dimensional representation, the model can assign higher weights to certain parts of the input data, thus better focusing on the critical information.

In the design of SSAM, this paper adopts a serial form of channel and spatial attention mechanisms, as shown in Figure 2. The extracted feature maps in the encoder contain a wealth of information, such as the subtle differences in HRTF data or specific textures. However, not all features significantly contribute to the low-dimensional representation of HRTFs. Hence, by assigning different weights to different channel features, the channel attention mechanism effectively helps the model reduce the redundancy introduced by interpolation, allowing the model to focus more on those features that aid in the low-dimensional representation of HRTFs. The spatial attention mechanism, on the other hand, concentrates on capturing spatial auditory localization cues from specific directions or locations.

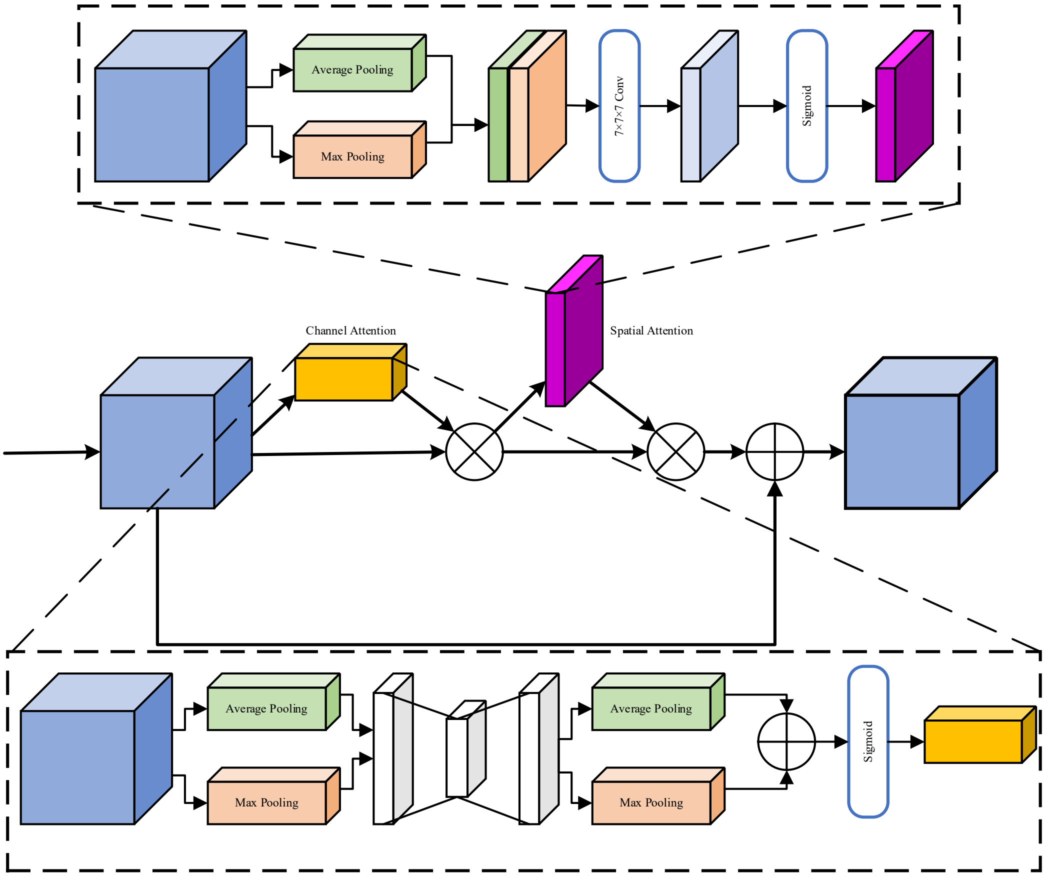

Figure 2: Spectral Space Attention Module (SSAM).

Specifically, the encoder network is divided into four distinct stages, each consisting of several SSAMs, with a ratio of 1:1:3:1 in this paper. Each SSAM is divided into a main branch and a residual branch. The main branch comprises depth-wise convolution, point-wise convolution, layer normalization, and an activation function. In the main branch, the input features are first processed using depth-wise convolution with a kernel size 7. To maintain the consistency in the feature input and output size, two point-wise convolutions are used after the depth-wise convolution to restore the channel count, resulting in the feature x', as shown in Equation (1).

$$x' = f_{pw}(GELU(f_{pw}(LN(f_{dw}^{7\times7\times7}(x))))) \tag{1}$$

In this context, $f_{\mathrm{dw}}(\cdot)$ represents the depth-wise convolution operation, $LN(\cdot)$ denotes the layer normalization operation, $f_{pw}(\cdot)$ signifies the point-wise convolution operation, and $GELU(\cdot)$ indicates the activation function.

In the residual branch, the input features are processed through the Channel Attention Module (CAM) and Spatial Attention Module (SAM) to obtain the feature x'', as illustrated in Equation (2).

$$\begin{aligned} M_c(x) &= \sigma(MLP(AvgPool(x)) + MLP(MaxPool(x))) \\ M_s(x) &= \sigma(f^{7\times7\times7}([AvgPool(x); MaxPool(x)])) \\ x'' &= x \otimes M_s(x \otimes M_c(x)) \end{aligned} \tag{2}$$

In this context, M_c and M_s represent the channel attention map and spatial attention map, respectively. The symbol σ denotes the sigmoid activation function, and $MLP(\cdot)$ stands for a multi-layer perceptron with one hidden layer. $AvgPool(\cdot)$ and $MaxPool(\cdot)$ refer to average pooling and max pooling operations, respectively. The term $f(\cdot)$ signifies a standard convolution operation, and $\otimes$ indicates element-wise multiplication. Ultimately, the feature obtained through the encoder module is represented as $SSAM(x) = x' + x''$.

To facilitate downsampling between different stages, the model incorporates a convolution operation with a kernel size of 2 and a stride of 2 between these stages. Additionally, layer normalization is applied both before and after downsampling to maintain the model's stability. After undergoing downsampling and feature extraction through four stages, a low-dimensional feature vector of size 2592 is ultimately obtained.

2.2. Attention Dense Decoder Network Design

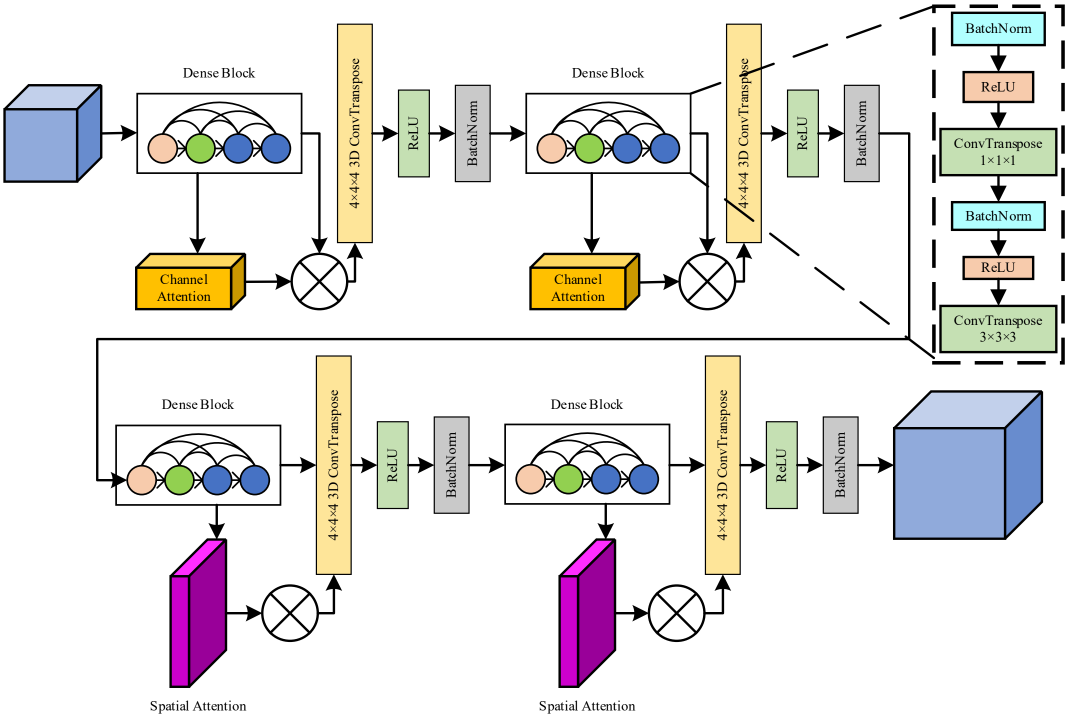

Figure 3: Multi-Scale Attention Dense Decoder Network(MSADN)

In the specific application of binaural three-dimensional audio, performing a high-dimensional reconstruction of the low-dimensional representation of HRTF is necessary. The key to ensuring reconstruction accuracy is effectively utilizing the low-dimensional characteristics of HRTF.

For the classical autoencoder model based on U-Net[11], improved high-dimensional reconstruction can be achieved by establishing feature connections between the encoder and the decoder, such as the mapping information of the max-pooling layer. However, in the practical application of HRTF, the features of the encoder are inaccessible. In this study, dense connections[12] were introduced between the layers of the decoder to enhance the decoder's ability to interpret low-dimensional features.

The dense connections among features in the decoder may lead to information redundancy and obscure critical feature information. Therefore, addressing how to ensure the transmission of more valuable feature information is a focal issue. In this research, the effectiveness of feature transmission in dense connections is ensured by employing a channel attention mechanism for high-level features and a spatial attention mechanism for low-level features based on the characteristics of different layers of features.

Similar to the design of the encoder, the decoder is also composed of four stages, each consisting of a densely connected convolutional module, attention mechanisms, and an upsampling layer, as shown in Figure 3. Through the restoration by the decoder, the reconstructed HRTF of size 48×144×96, denoted as $\hat{H}$, is finally obtained.

2.3. Loss Function

This paper employs the Minimum Absolute Error (MAE) to calculate the difference between the reconstructed HRTF and the original HRTF, with the loss calculation as shown in Equation (3).

$$L = MAE(H, \hat{H}) = \frac{1}{M}\sum_{m=1}^{M} | H_{\varphi,\vartheta}\left(f_m\right) - \hat{H}_{\varphi,\vartheta}\left(f_m\right) | \tag{3}$$

In this Equation, $H_{\varphi,\vartheta}\left(f_m\right)$ represents the amplitude value at frequency f_m of the HRTF measured at spatial orientation $(\varphi,\ \vartheta)$. $\hat{H}_{\varphi,\vartheta}\left(f_m\right)$ represents the amplitude value of the corresponding frequency reconstructed from the low-dimensional features.

3. Experiment and Analysis of Results

3.1. Dataset and Data Preprocessing

In this section, the HRTFs employed are sourced from the Spatially Oriented Format for Acoustics (SOFA) database[13]. The model proposed in this paper will be trained on the ARI dataset[14] and evaluated on the validation set of ARI, as well as the CIPIC[15], HUTUBS[16, 17], and BiLi[18] datasets, with detailed information presented in Table 1.

Table 1: Detailed information on the HRTF dataset.

Database	Number of Subjects	Number of Directions	Sampling Rate
ARI	220	1550	48000
CIPIC	45	1250	44100
HUTUBS	96	440	44100
BiLi	52	1680	96000

Due to the varying spatial distributions of HRTF samples provided by different HRTF databases, spatial fine interpolation of HRTF data was initially performed to ensure a consistent spatial distribution of HRTF samples in the experiments. This process enabled the HRTF samples to uniformly cover a spatial orientation within the horizontal azimuth from 0° to 360° and the vertical azimuth from -30° to 87.5° at 2.5° intervals. This coverage includes 48 vertical azimuths and 144 horizontal azimuths. Each HRTF data point comprises 96 frequency values, corresponding to a frequency range from 200Hz to 18kHz, covering the normal hearing range for adults. The HRTF data for all subjects were organized into a three-dimensional tensor of 48×144×96 and normalized to be expressed as 64-bit double-precision floating-point numbers.

3.2. Experimental Environment

The experiments were conducted on a Ubuntu 20.04 operating system, an NVIDIA A100 Tensor Core GPU, and a device configured with the Pytorch deep learning framework. During the training process, the batch size was set to 4, and the epoch was chosen to be 200. The gradient descent optimization algorithm used was Adam, with an initial learning rate set at 0.001. An exponential decay strategy was utilized to adjust the learning rate to ensure more stable and rapid network convergence. The multiplicative factor gamma for the learning rate in each epoch was set to 0.95.

3.3. Evaluation Metrics

To quantitatively assess the low-dimensional representation performance of different methods for HRTF, this paper employs the Average Spectral Distortion (ASD) between measured HRTF and reconstructed HRTF as the evaluation metric for HRTF reconstruction accuracy. For a given spatial direction $(\varphi,\ \vartheta)$, the spectral distortion (SD) between the measured HRTF and the reconstructed HRTF is calculated as shown in Equation (4).

$$SD_{\varphi,\vartheta} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(20\log_{10}\frac{\left|H_{\varphi,\vartheta}\left(f_m\right)\right|}{\left|\hat{H}_{\varphi,\vartheta}\left(f_m\right)\right|}\right)^2} \tag{4}$$

where $H_{\varphi,\vartheta}\left(f_m\right)$ represents the amplitude value of the measured HRTF at spatial orientation $(\varphi,\ \vartheta)$ corresponding to frequency f_m, and $\hat{H}_{\varphi,\vartheta}\left(f_m\right)$ represents the amplitude value of the corresponding frequency reconstructed from the low-dimensional features. The average spectral distortion ASD for D spatial orientations for N test samples can be expressed as $ASD=\frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{D}\sum_{d=1}^{D}SD_{n,d}\right)$. Where $SD_{n,d}$ is the spectral distortion for the dth spatial orientation for the nth sample.

3.4. Experiment and Results Analysis

The proposed method in this paper is compared with a principal component analysis (PCA)-based model[19, 20], which is commonly used in many HRTF-based applications and is often considered a baseline in related research.

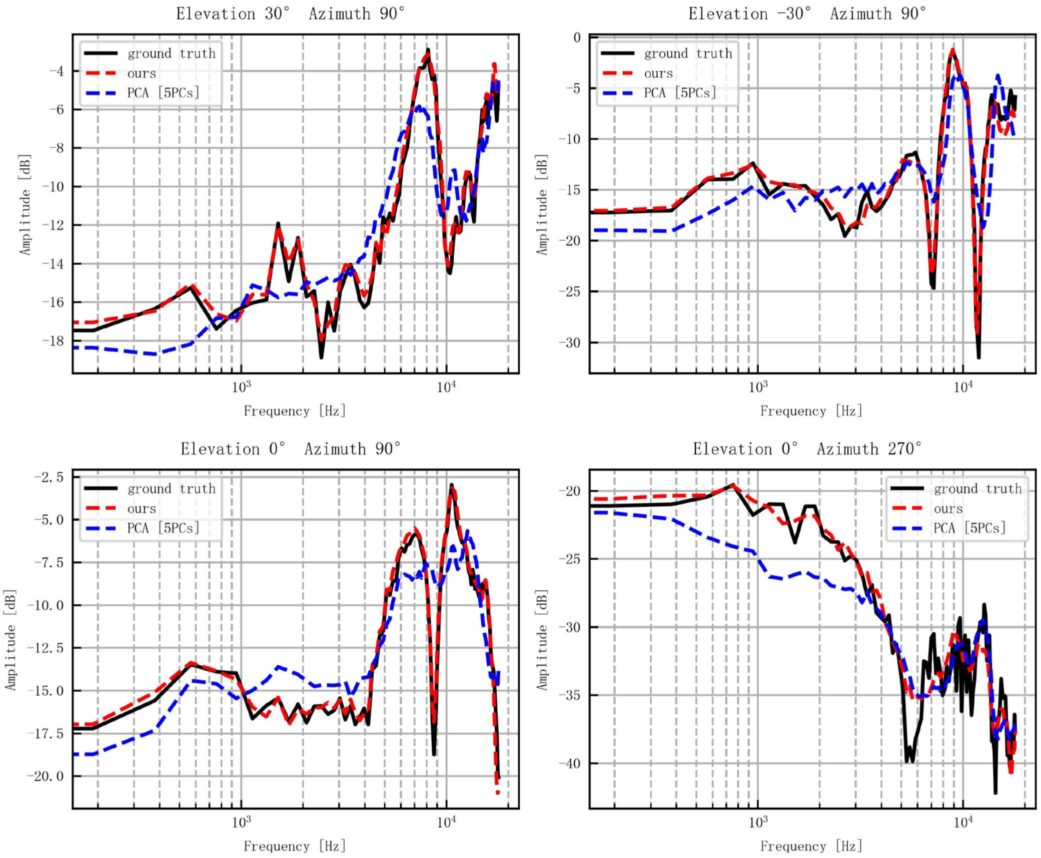

Figure 4: Comparison of the reconstruction effects of different HRTF low-dimensional representation models (azimuthal 0° is directly in front, 90° is ipsilateral to the ear, and 270° is contralateral to the ear).

The reconstruction accuracy of the PCA model is closely related to the number of selected principal components (PCs). Generally, the higher the number of selected PCs, the higher the reconstruction accuracy of the PCA model. However, this also leads to retaining more low-dimensional data, thereby reducing the compression capability of the original data. Therefore, in the comparative experiments with the PCA method, we followed the algorithm in reference[7] to perform PCA analysis on the HRTF samples in the training set, obtaining principal component information that expresses 82%, 90%, and 95% of the overall variance, respectively. Subsequently, using the obtained principal component information, we examined the HRTF samples in the ARI dataset, analyzing the low-dimensional data quantity and high-dimensional reconstruction accuracy achievable under different principal component conditions. Additionally, before any training procedures, the test subjects were separated from the training set to ensure the reliability of the experimental results.

Table 2: Comparative results of different low dimensional expression models.

Method	Cumulative Variance Percentage of Low-Dimensional Features	Size of Low-Dimensional Features	Compression Ratio	Average Spectral Distortion
PCA (3 PCs)	82.56%	20,736	3.12%	5.38 dB
PCA (5 PCs)	90.48%	34,560	5.20%	5.16 dB
PCA (8 PCs)	95.38%	55,296	8.33%	3.08 dB
Ours	-	2592	0.39%	1.56 dB

Table 2 presents a comparison between the proposed model and PCA-based models with different numbers of principal components. Experimental results on the validation set of the ARI database show that the proposed model achieves an average spectral distortion of 1.56dB with only 2592 low-dimensional features and a compression ratio of 0.39%. In contrast, even with eight principal components, PCA only achieves an average spectral distortion of 3.08dB at a compression ratio of 8.33%. The experiments demonstrate that the proposed method outperforms PCA on the ARI database by approximately 49.35%.

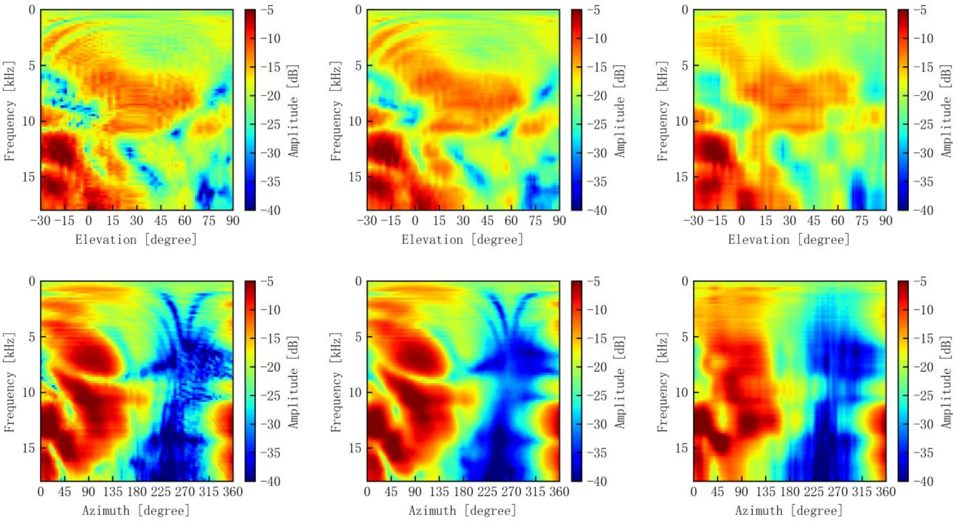

Figure 5: Comparison of the reconstruction effects of different HRTF low-dimensional representation models (upper median sagittal plane, lower ear plane).

For a detailed analysis of the reconstruction quality of HRTF by the convolutional autoencoder under different spatial orientation conditions, this paper compares the spectral curves of the reconstructed HRTF with the original HRTF. Diffuse field equalization is applied to remove non-spatial components in HRTF for clearer observation of the directional characteristics of HRTF. Figure 4 illustrates the results of processing the left ear HRTF of object 10 in the ARI database using different low-dimensional representation methods.

It can be observed that the PCA-based model performs well in reconstructing the most prominent peaks and valleys in the original HRTF spectrum. However, in other parts, it tends to be smooth and lacks detailed information in the HRTF. This may be because PCA focuses more on major features with significant influence, neglecting subtle features with less impact on the overall characteristics.

In contrast, the proposed model successfully reconstructs the most prominent peaks and valleys in the original HRTF and preserves more detailed information, as shown more clearly in Figure 5.

By mapping Head-Related Transfer Functions (HRTFs) onto the ear plane and mid-sagittal plane, it can be observed that the HRTFs generated by the model proposed in this paper preserve more spectral peaks, resembling "islands." While there are still some errors in amplitude compared to the original HRTFs, this may not necessarily reflect proportional losses in auditory localization performance.

Existing studies suggest that fluctuations in the spectral amplitude of HRTFs have a limited impact on human auditory perception. In contrast, the spatial relationships between spectral peaks and valleys in HRTFs, as they vary with the spatial orientation of sound sources, play a more crucial role in spatial localization perception.

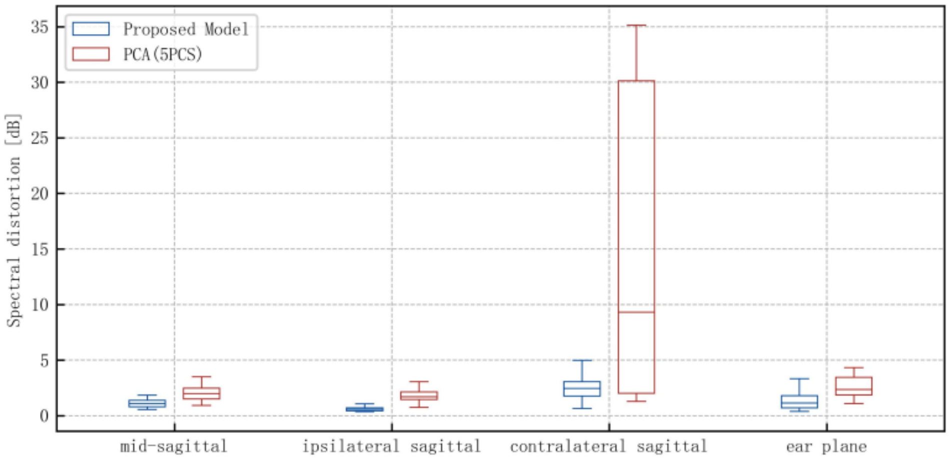

Figure 6: Reconstruction error distribution of HRTF under different spatial orientation conditions.

To analyze the reconstruction errors of the proposed method under different spatial orientations, this paper compares and analyzes the reconstruction errors of the PCA-based low-dimensional representation model (5PCs) and the proposed method at four crucial auditory orientations (mid-sagittal plane, ipsilateral sagittal plane, contralateral sagittal plane, and ear plane). As illustrated in Figure 6, on the ipsilateral sagittal plane within the elevation angle range of -30° to 87.5°, the average spectral distortion is 0.61dB. However, on the contralateral sagittal plane, the average spectral distortion is 2.66dB. It can be observed that reconstruction errors consistently occur on the contralateral side of the ears. This may be attributed to the fact that the HRTF on the contralateral side of the ears has fewer energy components due to the obstructive effect of the head, leading even small errors to result in significant distortion, making it more challenging to accurately express its intrinsic characteristics. Particularly notable is the pronounced effect of reconstruction errors on the model based on principal components, as principal component analysis tends to preserve features with greater variance, making it difficult to precisely express features on the contralateral side of the ears.

The deep learning approach has demonstrated unique advantages in the low-dimensional representation of HRTFs, particularly in nonlinear expression and spatial information capture. However, some currently proposed methods are trained on small-scale datasets (approximately dozens of subjects), posing a risk of overfitting. This is because their low-dimensional representations are limited by specific spatial sampling schemes in the HRTF training data, making it challenging to generalize across datasets with different sampling schemes. To further validate the generality of the proposed model, this section conducts low-dimensional representation and reconstruction of HRTFs on datasets with different sampling schemes, posing a challenge to the model's generalizability.

Table 3: Performance of the proposed model on different datasets.

Dataset	Average Spectral Distortion
ARI	1.53 dB
CIPIC	1.96 dB
HUTUBS	2.27 dB
BiLi	3.19 dB

Firstly, the methods described in Section 3.1 were applied to preprocess the ARI, CIPIC, HUTUBS, and BiLi databases to ensure compliance with the model's input requirements. Subsequently, the model was trained on the ARI dataset and evaluated on the validation sets of CIPIC, HUTUBS, BiLi, and ARI for its reconstruction ability, as shown in Table 3.

The model performed optimally on the ARI validation set, with an average spectral distortion of only

1.53 dB. This is because the model was trained on the ARI dataset, and deep learning models perform better on similar datasets. On the CIPIC and HUTUBS datasets, the average spectral distortion was slightly higher at 1.96dB and 2.27dB, respectively, indicating a slight decrease in performance. This may be attributed to the similarity in sampling rates between the CIPIC and HUTUBS datasets, with the CIPIC dataset having more sampling points and experiencing less interpolation influence, resulting in better performance. The BiLi dataset exhibited the poorest results, with an average spectral distortion reaching 3.19 dB. This is likely due to significant sampling rate differences and the lack of preprocessing for varying sampling rates, ultimately leading to suboptimal model performance.

Table 4: This caption has one line so it is centered.

Encoder Attention Module	Decoder Attention Module	Low-Dimensional Feature Size	Average Spectral Distortion
-	-	7796	1.63 dB
-	-	2592	1.77 dB
√	-	2592	1.59 dB
√	√	2592	1.53 dB

Finally, the paper evaluated the role of attention mechanisms in this model, as shown in Table 4. According to the data in the table, it is evident that when attempting to encode HRTF into a smaller space, the reconstruction error increased by 0.11 dB. With the addition of an attention module in the encoder, the model's low-dimensional representation capability was enhanced, and spectral distortion decreased by 0.18 dB, indicating improved learning of correlations between certain points in space. When both encoder and decoder attention modules were employed simultaneously, the average spectral distortion was further reduced to 1.53 dB. Compared to the model without any attention mechanisms, this represented a decrease of 0.24 dB in average spectral distortion.

4. Conclusions

This paper introduces a neural network model for the low-dimensional representation and reconstruction of HRTFs. The network utilizes an attention-guided feature encoder model to address biases introduced by mapping HRTFs to a three-dimensional tensor for convolutional operations. It explores the inherent features in the adjacent spatial orientations and spectral frequencies of HRTFs, enhancing the network's capability for low-dimensional representation. Additionally, this paper introduces dense connections in the decoder network to establish direct connections between layers of low-dimensional features. Considering the characteristics of features at different levels, effective features are selectively filtered, ensuring the efficient transmission of low-dimensional features and improving the accuracy of the model's low-dimensional reconstruction.

Experimental results on the ARI datasets demonstrate that the average spectral distortion can be reduced to as low as 1.53 dB compared to classical PCA models.

References

[1] TECHNOLOGIES O. Facemisc. 3d audio spatialization. [EB]. 2020. https://developer.oculus. com/resources/audio-intro-spatialization/

[2] STEAM A. A benchmark in immersive audio solutions for games and vr. [EB]. 2020. https://valvesoftware.github.io/steam-audio/

[3] MICROSOFT. Spatial sound. [EB]. 2020. https://learn.microsoft.com/en-us/windows/win32/ coreaudio/ spatial-sound

[4] HOGG A O, JENKINS M, LIU H, et al. HRTF upsampling with a generative adversarial network using a gnomonic equiangular projection [J]. arXiv preprint arXiv: 230605812, 2023.

[5] Chen T Y, Kuo T H, Chi T S. Autoencoding HRTFs for DNN based HRTF personalization using anthropometric features[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 271-275.

[6] Hu H, Zhou L, Ma H, et al. HRTF personalization based on artificial neural network in individual virtual auditory space [J]. Applied Acoustics, 2008, 69(2): 163-172.

[7] Meng L, Wang X, Chen W, et al. Individualization of head related transfer functions based on radial basis function neural network[C]//2018 IEEE International Conference on Multimedia and Expo (ICME).

IEEE, 2018: 1-6.

[8] Chun C J, Moon J M, Lee G W, et al. Deep neural network based HRTF personalization using anthropometric measurements[C]//Audio Engineering Society Convention 143. Audio Engineering Society, 2017.

[9] Schmidhuber J. Deep learning in neural networks: An overview[J]. Neural networks, 2015, 61: 85-117.

[10] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[11] Huang H, Lin L, Tong R, et al. Unet 3+: A full-scale connected unet for medical image segmentation[C]//ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020: 1055-1059.

[12] Iandola F, Moskewicz M, Karayev S, et al. Densenet: Implementing efficient convnet descriptor pyramids [J]. arXiv preprint arXiv:14041869, 2014.

[13] Majdak P, Noisternig M, Wierstorf H, et al. SOFA (Spatially Oriented Format for Acoustics) [EB]. Obtenido de https://www. sofaconventions. org. 2017

[14] Austrian Academy of Sciences. Ari HRTF database[EB]. [2020-10-11]. http://www. kfs.oeaw. ac.at/hrtf.

[15] Algazi V R, Duda R O, Thompson D M, et al. The cipic hrtf database[C]//Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575). IEEE, 2001: 99-102.

[16] Brinkmann F, Dinakaran M, Pelzer R, et al. The hutubs hrtf database[J]. DOI, 2019, 10: 14279.

[17] Brinkmann F, Dinakaran M, Pelzer R, et al. A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses[J]. Journal of the Audio Engineering Society, 2019, 67(9): 705-718.

[18] Rugeles Ospina F, Emerit M, Katz B F G. The three-dimensional morphological database for spatial hearing research of the BiLi project[C]//Proceedings of Meetings on Acoustics. AIP Publishing, 2015, 23(1).

[19] Zhang M, Qiao Y, Wu X, et al. Distance-dependent modeling of head-related transfer functions[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 276-280.

[20] Lu D, Zeng X, Guo X, et al. Personalization of head-related transfer function based on sparse principle component analysis and sparse representation of 3D anthropometric parameters[J]. Acoustics Australia, 2020, 48: 49-58.