HRTF-Based Data Augmentation Method for Acoustic Scene Classification

Yingzi Liu, Haocong Yang, Chuang Shi, Jiangnan Liang
School of Information and Communication Engineering
University of Electronic Science and Technology of China, Chengdu, China

***Abstract*—In acoustic scene classification (ASC), a technical problem yet to be solved is raised by the variety of recording devices. The amount of data recorded by different devices is usually unbalanced. The model trained with audio data collected by one device is hardly transferred to another device. Therefore, in order for the cross-device performance to be improved, this paper proposes a data augmentation method for ASC systems that take monaural audio samples as input, whereby the head-related transfer functions (HRTFs) are adopted to add artificial spatial information to monaural audio samples. The proposed method enables ASC systems to imitate the ability of human binaural hearing to distinguish spatial orientation and lock specific sound sources. The experiment results show that with the proposed method, the VGGNet and ResNet systems can get 13.4% and 14.4% higher accuracy than the DCASE 2020 baseline in the cross-device ASC, respectively.**

***Index Terms*—Acoustic scene classification, mismatched recording devices, head-related transfer functions, data augmentation, convolutional neural network**

I. Introduction

How to make machines accurately perceive and understand high-level information like human beings has always been an interesting problem in the field of audio signal processing. Acoustic scene is a representative high-level audio information. The information of acoustic scene is useful in the design of context-aware services, intelligent wearable devices, robotics navigation systems, audio archiving systems, and so on [1]. For example, if an electrical car can continuously sense its surroundings and perceive the acoustic scene, it can automatically switch to a quiet mode when entering a residential area. In machine hearing, acoustic scene classification (ASC) refers to the task of associating a semantic label to an audio stream that identifies the environment in which it has been produced.

Inspired by the AlexNet [2], Valenti et al. proposed the first dedicated convolutional neural network (CNN) for ASC in 2016 [3]. Since then, CNNs have been becoming the mainstream choice for ASC [4]. In DCASE 2020 challenge, there is no evidence of changing in this trend. Meanwhile, tempo-spectral acoustic features, such as short time Fourier transform (STFT), constant-Q transform and MFCC, are still open choices, although since 2018 the log mel spectrograms are often believed to lead to better results [5], [6]. An emerging difficulty of ASC lies in the mismatched recording devices. Models trained with audio data recorded by one device are likely to perform poorly on another device.

The most straightforward solution to the problem of the mismatched recording devices is the spectrum correction method. Nguyen et al. worked out this method to remove the difference between recording devices [7]. When aligned audio samples from different recording devices are available, a reference device is firstly selected and the audio samples collected from the other devices are all compensated in order to match the reference device. The spectrum correction method is simple and effective. However, it is also very limited and impractical, because aligned audio samples are rarely available. Moreover, the choice of the reference device, which can greatly affect the final outcome, is difficult to carry out, when the number of devices is considerably large or even uncertain. Therefore, more effective data augmentation methods are desired. The conventional data augmentation methods include noise augmentation, time shift, pitch shift, speed tuning, and so on. Presently, they are relatively ineffective and often have adverse effects.

Another aspect of the mismatched recording devices is the imbalanced recording time length. One device provides most of the audio samples, while the other devices records very few of them. Therefore, catering for the variety of recording devices, most of the research works in the field of ASC regarding to the mismatched recording devices work merely on monaural audio samples that contain no spatial information. This fact hinders ASC systems from imitating the ability of human binaural hearing to distinguish spatial orientation and lock specific sound sources. Han et al. in 2017 have obtained improved results by using binaural representations as input feature that containing richer spatial information than monaural representations [8]. Mesaros et al. have reported in 2018 that stereo audio samples result in higher ASC accuracy than monaural audio samples [9]. Green et al. have proposed the use of spatial features extracted from fourth-order ambisonic recordings for ASC [10]. Those works proved that spatial information plays a positive role in ASC.

This paper proposes a new data augmentation method for ASC systems that take monaural audio samples as input, where the audio samples are processed by the head-related transfer functions (HRTFs) to add in the artificial spatial information [11]. The concept of HRTFs was developed in psycho-acoustical spatial audio processing. Each HRTF contains frequency-dependent magnitude gain and phase shift that models how a particular ear receives sound from a point in the three-dimensional space [12]. Since a person has two

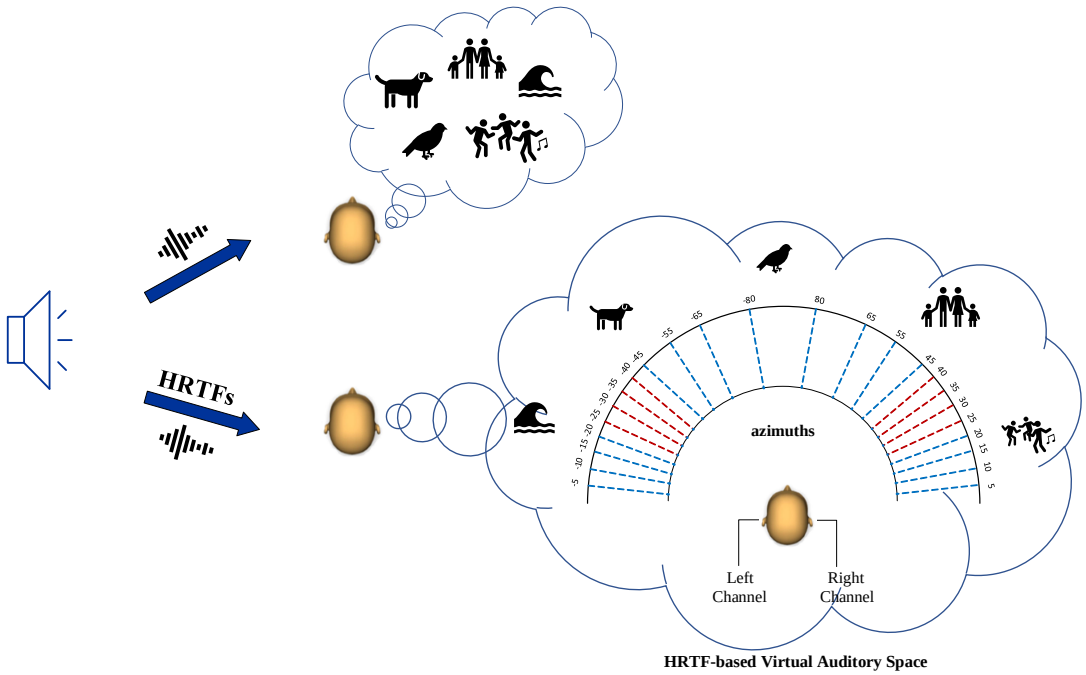

Fig. 1. The virtual auditory space constructed by HRTFs at 0° elevation.

ears, HRTFs are also grouped in pairs. The proposed method adopts 24 HRTF pairs to construct a virtual auditory space, as illustrated in Fig. 1. The virtual auditory space is then integrate them into two classic CNNs, the VGGNet and ResNet [13], [14]. The effectiveness of the proposed method is validated through comparative experiments using the TAU urban acoustic scenes 2020 mobile dataset [9].

II. PROPOSED METHOD

A. Data Augmentation with HRTFs

Although HRTFs should be unique for an individual, several public HRTF databases are available. The CIPIC database measured HRTFs of 45 subjects at 50 elevations (ranging from -45 to 230.625 degrees) and 25 azimuths (ranging from -80 to 80 degrees) [15]. In this paper, 3 elevations (0°,45° and 135°) and 8 azimuths (±25°, ±30°, ±35° and ±40°) are selected for experimental research. Each monaural audio sample is thereafter augmented to 24 stereo audio samples. In every audio samples there are surrounding ambient components that are not captured by a single HRTF but by diffuse combinations of them. So there should be at least two pairs of HRTFs associated with different angles applied simultaneously, instead of just a single pair of HRTFs. As the selected azimuths are symmetric to the median plane, 24 stereo audio samples are further grouped into pairs. In this sense, each monaural audio sample is augmented to 12 four-channel audio samples. Different HRTFs lead to different acoustic features and meanwhile retain the same data distribution. The augmented training data helps to obtain a more generalized model, while the augmented evaluation data can provide an improved ensemble result.

B. CNNs and Focal Loss

The first CNN architecture considered in this paper is the VGGNet [13]. 8 stacked convolutional layers with the kernel size of 3×3 are used. Before each convolutional layer, there is a zero padding layer, a batch normalization layer and a ReLu activation layer. After the second, fourth, and eighth convolutional layers, a 3×3 maxpooling layer is appended. Finally, a 10-way softmax layer following a global average pooling layer, is used to generate the final classification result.

The second CNN architecture is the ResNet proposed by McDonnell et al. [14]. Its architecture is shown in Table I. This ResNet is divided into two networks for parallel training according to the frequency dimension of features extracted from audio samples, since the feature of high frequencies to be learned may be different from those of low frequencies. The two parallel networks are identical in structure, consisting of several full pre-activation structures, namely the Residual Blocks [16], [17]. Each Residual Block consists of two convolution layers with the kernel size of 3×3, as shown in Fig. 2. When the input and output are different in scale, average pooling and channel padding are used in the residual paths. After 8 ResNet blocks, the two parallel networks are concatenated to form 128 frequency dimensions. A softmax activation layer, following two 1×1 convolutional layers, calculates the final output.

When training the CNNs, temporal cropping and mixup methods are usually introduced to prevent overfitting [14]. In this paper, in order for the ResNet to be better trained, the focal loss function is suggested [18]. By adding a modulating factor to the cross-entropy loss, the focal loss can attenuate

TABLE I
RESNET ARCHITECTURE

Input (431,128,4)	
Low Frequency((431,64,4))	High Frequency((431,64,4))
Batch Normalization(ch=3)	Batch Normalization(ch=3)
Conv2D(ksize=3,s=[2,1],ch=24)	Conv2D(ksize=3,s=[2,1],ch=24)
Residual Block(ksize=3,ch=24)	Residual Block(ksize=3,ch=24)
Residual Block(ksize=3,ch=24)	Residual Block(ksize=3,ch=24)
Residual Block(ksize=3,ch=48)	Residual Block(ksize=3,ch=48)
Residual Block(ksize=3,ch=48)	Residual Block(ksize=3,ch=48)
Residual Block(ksize=3,ch=96)	Residual Block(ksize=3,ch=96)
Residual Block(ksize=3,ch=96)	Residual Block(ksize=3,ch=96)
Residual Block(ksize=3,ch=192)	Residual Block(ksize=3,ch=192)
Residual Block(ksize=3,ch=192)	Residual Block(ksize=3,ch=192)
Concatenate(ch=192)	
Batch Normalization(ch=192) Activation('ReLu') Conv2D(ksize=1,s=[1,1],ch=768)	
Batch Normalization(ch=768) Conv2D(kszie=1,s=[1,1],ch=10)	
Batch Normalization(ch=10) GlobalAvgPooling(ch=10) Activation('SoftMax')	
Output	

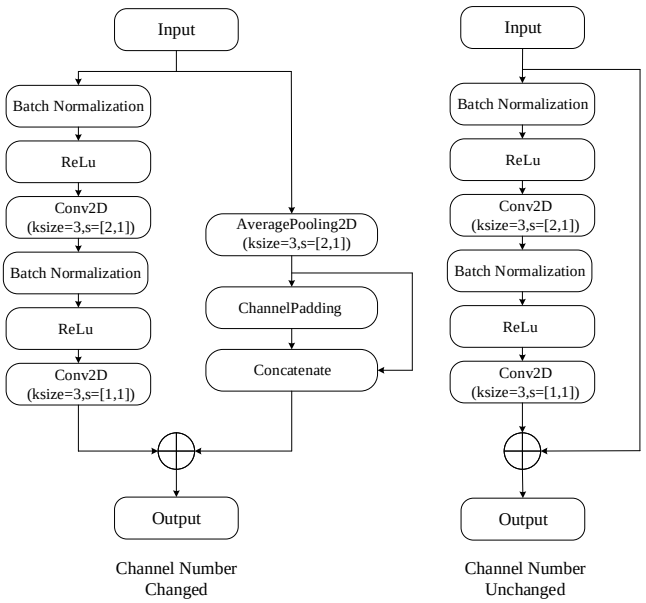

Fig. 2. Residual Block

the relative loss generated by those well-classified samples and therefore focuses more on the hard, misclassified samples. The following equation describes the α-balanced variant of focal loss with balancing parameter α, focusing parameter γ and prediction score p_t.

$$FL\left(p_t\right) = -\alpha\left(1-p_t\right)^{\gamma}\log\left(p_t\right), \tag{1}$$

where the value of γ controls the sensitivity of the model to misclassified samples, and α scales the loss function linearly. Their typical settings are 2.0 and 0.75, respectively.

Figure 3 shows the procedure of using the proposed HRTF-based data augmentation method in an ASC system that takes monaural audio samples as input.

III. EXPENRIMENT

A. Dataset

This paper conducts comparative experiments using the TAU urban acoustic scenes 2020 mobile dataset [9]. This dataset contains recordings of 10 different acoustic scenes in 10 European cities from 9 devices. Among them, there are 3 real devices (referred to as devices A, B, C) and 6 simulated devices (referred to as devices S1-S6). The total amount of audio samples in the development set is 64 hours, of which 40 hours of audio samples were recorded by device A, and the remaining small amount of audio samples were collected by the other 8 devices. All of the recordings are divided into 10-second audio samples and provided in a single-channel format (44.1 kHz, 24-bit).

The whole dataset is divided into the training and testing subsets. 70% of the audio samples for each device is included in the training subset. The rest of audio samples are kept for testing. In particular, devices S4, S5 and S6 only appear in the testing subset. In the training subset, device A contributes about 75% of the total audio samples. In the testing subset, all the devices contribute almost equally. The average of the class-wise accuracies is used as the performance measure of the ASC system under testing, i.e.

$$\mathrm{Acc}_{\mathrm{average}} = \frac{1}{10}\sum_{i=1}^{i=10}\mathrm{Acc}_i, \tag{2}$$

where i is the index of the acoustic scenes, including (1) airport, (2) bus, (3) metro, (4) metro station, (5) park, (6) public square, (7) shopping mall, (8) street pedestrian, (9) street traffic and (10) tram.

B. Feature Extraction

Firstly, a 2048-point hamming window with 50% overlap is used to extract the spectrogram of each audio sample. Secondly, the log mel spectrogram is obtained by applying the log mel filter bank on the spectrogram. There are 128 log mel filters in the filter bank that cover a frequency range from 0 to half of the sampling rate, yielding 431-frame spectrograms with 128 frequency bins. Thirdly, log mel spectrograms are normalized by subtracting the mean and dividing the standard deviation. Finally, the 4-channel feature with the size of (431, 128, 4) can be obtained. With the data augmentation proposed by this paper, a monaural audio sample can generate 12 such features.

C. Training

Models are trained using stochastic gradient descent (SGD) optimizer with the Nesterov momentum. The batch size, momentum, and decay are set to 32, 0.9, and 0.0001, respectively. The initial learning rate is set to 0.01 and decreased by a factor 0.5 every 10 epochs after 70 epochs. Each model is trained for 180 epochs which takes about 4.5 hours on a single NVIDIA GeForce RTX 2080Ti card.

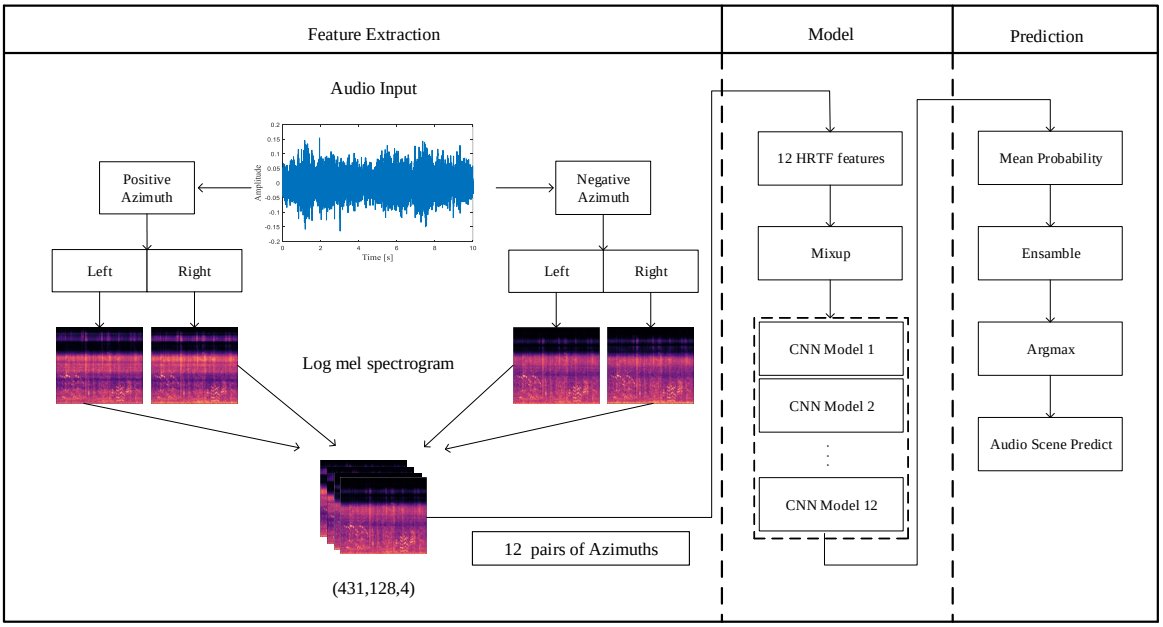

Fig. 3. Architecture of the ASC system using the HRTF-based data augmentation method. Four log mel spectrograms are generated from one piece of audio data processed by a pair of HRTF azimuths. Different models are obtained and finally a simple averaging layer is used for ensemble learning.

D. Results

Three baseline systems are compared firstly. Among the baseline systems, the ResNet baseline exhibits the highest accuracies in all the three categories of devices. They are the real devices (A, B, and C), seen simulated devices (S1, S2, and S3), and the unseen simulated devices (S4, S5, and S6). As the DCASE2020 baseline results in the worst performance [19], the proposed method is only adopted in the VGGNet and ResNet baseline systems for further comparison.

Figure 4 shows the accuracy change of the VGGNet baseline corresponds to ±25°, ±30°, ±35° and ±40° azimuth pairs associated with 0°, 45° and 135° elevations. In the VGGNet baseline system, the HRTFs can significantly improve the ASC performance for the simulated devices. By contrast, Fig. 5 shows the accuracy change of the ResNet baseline corresponds to ±25°, ±30°, ±35° and ±40° azimuth pairs associated with 0°, 45° and 135° elevations. The HRTFs mainly improve the ASC performance of the ResNet baseline for the real devices.

Table II shows the summarized results of different ASC systems, including the DCASE2020 baseline, the VGGNet baseline, the ResNet baseline, the latter two integrated with the proposed method. The proposed method is validated to be effective in improving the cross-device ASC system performance. When integrating with the VGGNet baseline, the proposed method achieves 6.2% and 13.4% higher average accuracy than the VGGNet baseline and the DCASE2020 baseline, respectively. When integrating with the ResNet baseline, the proposed method achieves 2.5% and 14.4% average accuracy than the ResNet baseline and the DCASE2020 baseline, respectively.

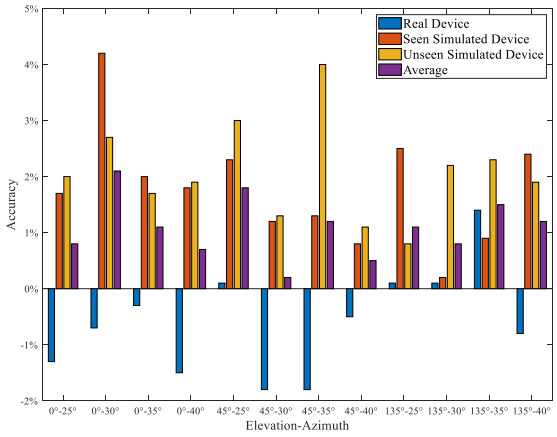

Fig. 4. Accuracy change of the VGGNet baseline corresponds to ±25°, ±30°, ±35° and ±40° azimuth pairs associated with 0°, 45° and 135° elevations.

IV. CONCLUSIONS

Aiming at the deployment difficulty of ASC raised by the mismatched recording devices, this paper proposes a new data augmentation method that uses HRTFs to add artificial spatial information to the monaural audio samples. The proposed HRTF-based data augmentation method enables ASC systems that take monaural audio samples as input to imitate the ability of human binaural hearing to distinguish spatial orientations and focus on specific sound sources. The experiment results show that with the proposed method, the VGGNet and ResNet

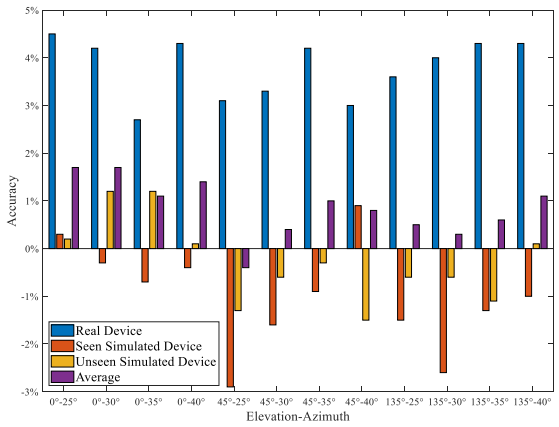

Fig. 5. Accuracy change of the ResNet baseline corresponds to ±25°, ±30°, ±35° and ±40° azimuth pairs associated with 0°, 45° and 135° elevations.

TABLE II
SUMMARIZED RESULTS OF DIFFERENT ASC SYSTEMS

System name	Real Device	Seen Simulated Device	Unseen Simulated Device	Average
DCASE2020 Baseline [19]	0.646	0.533	0.443	0.541
VGGNet Baseline	0.674	0.589	0.577	0.613
ResNet Baseline	0.682	0.654	0.644	0.660
HRTF_VGGNet	0.724	0.661	0.639	0.675
HRTF_ResNet	0.732	0.664	0.660	0.685

baseline systems can get 6.2% and 2.5% improvement in the cross-device ASC accuracies, respectively. Moreover, with the proposed method, the VGGNet and ResNet systems can get 13.4% and 14.4% higher accuracy than the DCASE 2020 baseline in the cross-device ASC, respectively.

REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems*, Nevada, USA, 2012, pp.1097-1105.

[3] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, Budapest, Hungary, 2016, pp.95-99.

[4] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp.279-283, 2017.

[5] M. Wang, R. Wang, X. L. Zhang, and S. Rahardja, "Hybrid constant-Q transform based CNN ensemble for acoustic scene classification," in *Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* , Lanzhou, China, 2019, pp.1511-1516.

[6] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer trends and Technology*, vol. 16, no. 6, pp. 582-589, 2001.

[7] T. Nguyen, F. Pernkopf and M. Kosmider, "Acoustic scene classification for mismatched recording devices using Heated-Up Softmax and spectrum correction," in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* , Barcelona, Spain, 2020, pp. 126-130.

[8] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural represntations and background subtraction for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, Munich, Germany, 2017, pp. 1-5.

[9] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop*, Surrey, UK, 2018, pp. 9-13.

[10] M. C. Green, S. Adavanne, D. Murphy and T. Virtanen, "Acoustic Scene Classification Using Higher-Order Ambisonic Features," In *Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, NY, USA, 2019, pp. 328-332.

[11] T. Potisk, and D. Svenšek, "Head-related transfer function," Seminar la. Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia, 2015.

[12] J. Wang, M. Liu, X. Wang, T. Liu, and X. Xie, "Prediction of head-related transfer function based on tensor completion," *Applied Acoustics*, vol. 157, pp. 1-10, 2019.

[13] S. Jiang, C. Shi and H. Li, "Acoustic Scene Classification Technique for Active Noise Control," in *Proceedings of the 2019 International Conference on Control, Automation and Information Sciences* , Chengdu, China, 2019, pp. 1-5.

[14] M. D. McDonnell and W. Gao, "Acoustic scene classification Using deep residual networks with late fusion of separated high and low frequency paths," in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* , Barcelona, Spain, 2020, pp. 141-145.

[15] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Platz, NY, USA, 2001, pp. 99-102.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770-778.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proceedings of the European conference on computer vision*, Springer, Cham, 2016, pp. 630-645.

[18] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," in *Proceedings of the 2017 IEEE International Conference on Computer Vision* , Venice, Italy, 2017, pp. 2999-3007.

[19] J. Cramer, H. Wu, J. Salamon and J. P. Bello, "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing* , Brighton, United Kingdom, 2019, pp. 3852-3856.