Ensemble learning based head-related transfer function personalization using anthropometric features

Yih-Liang Shen†, and Tai-Shi Chi*
* Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan
† Institute of Communications Engineering, National Yang Ming Chiao Tung University, Taiwan
E-mail: dennis831209@gmail.com; tschi@nycu.edu.tw

***Abstract*—An ensemble learning based model to predict the logarithmic magnitude response of the head-related transfer function (HRTF) using anthropometric features is proposed. Subjects are first clustered based on relevant anthropometric features, then the ensemble learning algorithm is used on clustered results for predicting the log-magnitude response of HRTF. The model contains two levels of deep neural networks (DNNs). The first-level DNNs predict log-magnitude HRTFs in clustered groups, and the second-level DNN integrates estimates produced by the first-level DNNs. Experimental results show the proposed model outperforms compared DNN models in terms of the log-spectral distortion (LSD) measure with greater stability.**

I. Introduction

In virtual-reality applications, it is essential to accurately present spatial sounds to the user. The sound waves from a sound source interact with the head, torso, and pinna to cause scattering and diffraction. The acoustic path from the sound source to an ear can be considered an acoustic channel and is characterized by its impulse response, which is referred to as the head-related impulse response (HRIR). The frequency representation of the HRIR is called the head-related transfer function (HRTF). Using HRTFs, a sound can be spatialized as it comes from a specific location. HRTFs contain binaural cues such as interaural time difference (ITD) and interaural level difference (ILD), which are primarily used by humans to localize the sound source [1]. Since HRTFs are personalized measures, using HRTFs of a different person to spatialize the sound can result in confusion about the location of the sound [2][3]. However, directly measuring an individual's HRTFs, which involves complex, time-consuming, and expensive procedures, is impractical.

Many methods have been proposed to estimate HRTFs. For example, the model-based methods divide the acoustic path into several blocks and develop a physical model for each block [4]. In addition to developing analytic geometry models, manipulating HRTFs in anthropometric dimensions was also attempted due to the high correlation between body characteristics and HRTFs. For instance, the frequency scaling of HRTFs was investigated in [5] to find the optimal scaling factor between different subjects by minimizing the difference of HRTFs expressed in anthropometric dimensions. Some studies also suggest that approximated HRTFs can be obtained from available HRTF datasets by looking for people with similar body characteristics [6]. Recently, researchers use a three-dimensional ear model and simulated HRTFs to relate the shape of pinna to HRTFs [7]. Signal processing based methods were also developed for estimating HRTFs. For example, active sensory tuning (AST) was used to synthesize HRTFs of the target person from a generic HRTF by adjusting poles and zeros of the transfer functions through several optimization steps [8][9]. Principal component analysis (PCA) and regression analysis were combined for estimating HRTFs [10][11]. In recent years, deep neural networks (DNNs) were also proposed to estimate HRIRs using anthropometric features [12], and even to generate HRTFs from the image of the pinna [13]. Besides anthropometric features, spatial principal component analysis and spherical harmonic analysis were conducted on HRTFs to extract representative features as training targets of DNN models [14][15].

However, DNN-based methods require a large amount of training data to avoid over-fitting. To alleviate the problem, we have proposed an auto-encoder to reduce the dimension of HRTFs [16]. Although we showed the proposed auto-encoder outperforms the DNN model [12], the variance of its performance across target subjects is still large. In this paper, we aim to further improve the accuracy and the stability of the HRTF estimation model using ensemble learning (EL), which has been used in many research fields to reduce system's variance while increasing the system's generalization capability [17][18][19]. Ensemble learning is a machine learning method to combine multiple models. It leverages the strength of different models and reduces the impact of individual model's biases and errors, delivering more robust and reliable predictions. Inspired by the integrated deep and ensemble learning (IDEA) algorithm proposed in [20], we construct an EL-based model with groups of subjects. Each group contains subjects with similar anthropometric features such that the model can learn more representative characteristics of the group. As shown in [5][7], some anthropometric features are more relevant to HRTFs than others. The subjects in the dataset are first clustered by these relevant features, and then put through a two-stage training, consisting of an individual group training and an ensemble integration training. In this way,

we can build an EL-based HRTF estimator with improved accuracy and stability.

The rest of the paper is organized as follows. In Section II, we describe the pre-processing of training data, the clustering processes, and the EL model architecture. In Section III, experimental results and system comparisons are presented and discussed. Finally, Section IV concludes our work.

II. Proposed Method

In this section, we first introduce the pre-processing steps, and then demonstrate the initial grouping on subjects and the architecture of the proposed model.

A. Dataset and pre-processing

The CIPIC HRTF database contains HRIR data of 45 subjects with their anthropometric features [21]. For each subject, a total of 1250 HRIRs, combinations of 25 azimuth angles and 50 elevation angles, and their 37 anthropometric features, including 17 features related to the head and torso, and 10 features related to each pinna, are recorded. The definitions of the anthropometric features, and azimuth and elevation angles can be found in [21]. However, only 35 out of the 45 subjects have complete anthropometric features, hence, their features and HRIRs were used in this study. Before training DNN models, we normalized the anthropometric features and transformed the HRIRs to HRTFs.

When estimating the magnitude responses of HRTFs of the left/right ear, we only used the ten anthropometric features related to the left/right ear and the 17 head-torso features. Therefore, the input to the DNN models was a 27-dimensional feature vector. We followed the formula in [12] to normalize each feature as

$$a_i' = (1 + e^{-\frac{(a_i - \mu_i)}{\sigma_i}})^{-1} \quad (1)$$

where a_i is the i-th anthropometric feature, and μ_i and σ_i are the mean and standard deviation over all i-th features, respectively. Finally, $[a_1', a_2', ..., a_{27}']$ was used as the input vector to DNN models.

To derive the magnitude responses of HRTFs, we conducted 512-point discrete Fourier transform (DFT) on HRIRs and used the constant-Q filterbank (Q=8) to smooth the magnitude responses. Since sound characteristics between 0.2 kHz and 15 kHz are relevant to localization [22], we selected the corresponding frequency bins from the magnitude responses of HRTFs in the frequency domain and took the logarithm to obtain the spectra of HRTFs in decibel. After these steps, each log-magnitude spectrum was represented by a 173-frequency-bin vector. Since we used the sigmoid activation function in the output layer of the DNNs, we normalized the log-magnitude spectra of HRTFs to values between 0 and 1. To ensure fairness in evaluation, all compared models were trained using the same pre-processed data. Note that, the log-magnitude spectrum of the HRTF is the estimation target in this paper. For synthesizing the spatial sound, the phase response of the HRTF is derived using minimum-phase reconstruction as in [10][14].

B. Subject grouping for ensemble learning

It has been shown that the HRTFs are systematically different among subjects, and a scaling factor on the frequency axis can be found to minimize the difference between individuals' HRTFs [5]. The optimal scaling factor between the HRTFs of two subjects can be predicted by the relative physical sizes of the two subjects. The study demonstrated that the measurements of "Pinna-cavity height" and "Head width" are highly correlated with the optimal scaling factor. These two measurements can be explicitly or implicitly obtained from the anthropometric feature set of the CIPIC database. The head width ($x1$) is directly recorded in CIPIC, and the pinna-cavity height can be derived by summing the cavum concha height ($d1$), the cymba concha height ($d2$), and the fossa height ($d4$) of CIPIC. Therefore, we selected the "Pinna-cavity height" and "Head width" features as basis to group the 35 subjects for the proposed EL-based model. The proposed model is referred to as the Ensemble-Anthro (EA) model and the version using these two features for initial grouping is represented by $EA_{x1,d1+d2+d4}$.

On the other hand, a more recent study analyzed the sensitivity of pinna morphology on HRTFs, and showed that the cavum concha width ($d3$) and the fossa height ($d4$) are related to the sensitive control points which are highly correlated to HRTFs [7]. Similar findings were also observed in other studies [23][24]. Therefore, we developed another version of the EA model, referred to as $EA_{d3,d4}$, using these two ear features for initial grouping. Considering findings in [5] and [7], we also tried the third version of the EA model using ear-related $d3$, $d4$ features and the head-related $x1$ feature for initial grouping, which is referred to as $EA_{x1,d3,d4}$. Note, $x1$, $d1$, $d2$, $d3$, and $d4$ are the corresponding feature labels used in CIPIC. For each version of the EA model, the k-means clustering method was used to initially divide the 35 subjects into three groups based on selected features. After clustering, $EA_{x1,d1+d2+d4}$, $EA_{x1,d3,d4}$, and $EA_{d3,d4}$ models relatively have (9, 10, 16), (8, 11, 16), and (9, 10, 16) subjects in three groups.

C. Architecture of the proposed EA model

Similar to the IDEA algorithm proposed in [20], we constructed an EL-based model for estimating log-magnitude spectra of HRTFs. In the training phase, the proposed model went through an ensemble preparation (EP) stage and an ensemble integration (EI) stage, as respectively shown in the top and bottom panels of Fig. 1.

In the EP stage, we trained three independent DNNs for the three groups clustered by the k-means method. The anthropometric parameters of the i-th subject in the three groups were respectively collected as $Fea_1^{(i)}$, $Fea_2^{(i)}$, and $Fea_3^{(i)}$, which were used as the input features to the three DNNs, DNN_1, DNN_2, and DNN_3. Each DNN was trained to minimize the mean square error between the estimated log-magnitude spectrum $\hat{Y}^{(i)}$ and the real log-magnitude spectrum $Y^{(i)}$.

In the EI stage, we first fixed DNN_1, DNN_2, DNN_3 trained in the EP stage. For subject j, his anthropometric

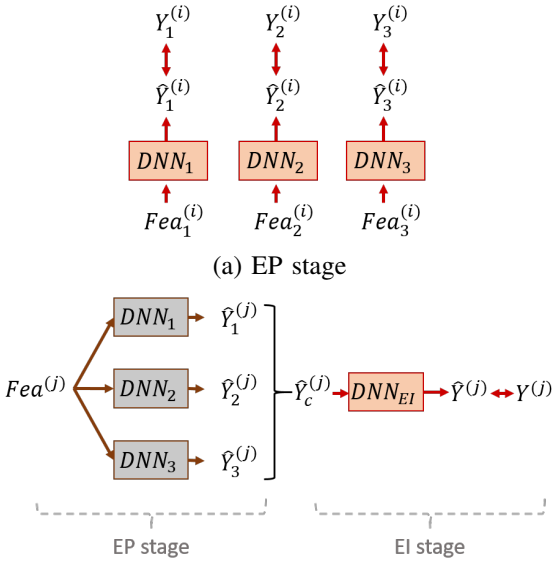

Fig. 1. Architecture diagram of the proposed EL model, including the ensemble preparation (EP) and the ensemble integration (EI) stages.

parameters $Fea^{(j)}$ was put through the three DNNs to produce three estimated log-magnitude spectra $\hat{Y}_1^{(j)}$, $\hat{Y}_2^{(j)}$, and $\hat{Y}_3^{(j)}$. The three estimates were concatenated into a 519-dimensional vector $\hat{Y}_c^{(j)}$ as the input features to the following EI DNN (DNN_{EI}). As shown in the bottom panel of Fig. 1, the DNN_{EI} was trained by minimizing the mean square error between the estimated log-magnitude spectrum $\hat{Y}^{(j)}$ and the real log-magnitude spectrum $Y^{(j)}$.

In the inference phase, the anthropometric parameters of a test subject were put through the model, including DNN_1, DNN_2, DNN_3, and DNN_{EI}, to produce the estimated log-magnitude spectra of HRTFs. The overall model maps the subject's anthropometric features to the subject's log-magnitude spectrum of the HRTF via the mapping function $M(\cdot)$ as

$$M(\cdot) = g([f_1(\cdot), f_2(\cdot), f_3(\cdot)]) \tag{2}$$

where $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$, and $g(\cdot)$ are the mapping functions of DNN_1, DNN_2, DNN_3, and DNN_{EI}, respectively.

III. Experimental Results

A. Model settings and evaluation metric

For evaluations, we compared the proposed EA model with a DNN baseline model [12] and our previously proposed AutoEn+DNN model [16]. Pilot experiments were conducted to determine the optimal numbers of hidden layers and hidden units for each DNN in the EA model. When training the DNNs, the ReLU and the sigmoid function were used as the activation functions of hidden layers and the output layer, respectively, and the Adam optimizer was used with a learning rate of 0.0001 and a drop-out rate of 0.2. In EP stage, we trained DNNs with 100 epochs. In EI stage, we stopped training when the validation loss ceased to decrease for 10 epochs. Note that we used network parameters tuned in their original studies [12][16] to implement the compared DNN baseline model and the AutoEn+DNN model.

Two experiments were conducted to evaluate the proposed EA model. In the first experiment, we trained an individual model for each direction. In total, we trained 25 models for 25 azimuth angles at 0° elevation. Our focus on the horizontal plane was primarily to be consistent with the previous study [16] to have fair comparisons. In that study, authors aimed to combine right and left ear HRIRs using one mapping rule, thus only considered the horizontal plane. The 25 tested azimuth angles were $\{0°, \pm5°, \pm10°, \pm15°, \pm20°, \pm25°, \pm30°, \pm35°, \pm40°, \pm45°, \pm55°, \pm65°, \pm80°\}$ as described in [21]. In the second experiment, we embedded the angle information in the input features to build a universal model for all directions. The original 27-dimensional anthropometric feature was cascaded with the azimuth and elevation angles to have a 29-dimensional feature as the input to the universal model.

We adopted the leave-one-out cross validation approach, i.e., one subject was used for test while the remaining 34 subjects were used for training. As in many HRTF estimation studies [6][25][11][12][13][16][26][23], we used the log spectral distortion (LSD; in dB) measure for model evaluation. It is formulated as follows

$$LSD(H, \hat{H}) = \sqrt{\frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \left(20 log_{10} \frac{|\, H(k) \,|}{|\, \hat{H}(k) \,|} \right)^2} \tag{3}$$

where k is the index of frequency bin; H is the real spectrum and $\hat{H}$ is the estimated spectrum.

B. Experiment 1: an individual model for each direction

In Experiment 1, we trained a total of 25 models for 25 azimuth angles at 0° elevation. To tune the optimal numbers of hidden layers and hidden units, we tried two or three hidden layers with 16, 32, or 64 hidden units in each layer for DNN_1, DNN_2, and DNN_3. For DNN_{EI}, we tried three or four hidden layers with 32, 64, or 128 hidden units in each layer. In the end, we constructed DNN_1, DNN_2, and DNN_3 with three hidden layers and 16 hidden units in each layer. For DNN_{EI}, we used three hidden layers with 64 hidden units in each layer. Table I lists the structures of these models in details. Using the leave-one-out cross validation, we ended up with 875 (=25×35) LSD measures to calculate the mean and variance.

Table II shows the overall mean and variance of the LSD measures of all compared models. Note that we re-implemented and re-trained the DNN baseline model and the AutoEn+DNN model such that their scores might be different from the scores reported in their original studies [12][16]. The lower mean and lower variance of the proposed EA models indicate they can estimate log-magnitude spectra more accurately and stably than the Baseline and the AutoEn+DNN models. We performed t-tests between each pair of the three

TABLE I
MODEL STRUCTURE OF $DNN_{1,2,3}$ / DNN_{EI} IN EXPERIMENT 1

Layer	Units	Activation Function
Input	27 / 519	-
Dense Layer 1	16 / 64	ReLU
Dropout 1	-	-
Dense Layer 2	16 / 64	ReLU
Dropout 2	-	-
Dense Layer 3	16 / 64	ReLU
Dropout 3	-	-
Output Layer	173 / 173	Sigmoid

TABLE II
THE OVERALL MEAN AND VARIANCE OF LSD OF COMPARED MODELS IN EXPERIMENT 1

	mean	variance
Baseline DNN	3.52	1.37
AutoEn+DNN	3.56	1.16
$EA_{x1,d1+d2+d4}$	**3.33**	1.07
$EA_{x1,d3,d4}$	**3.33**	1.07
$EA_{d3,d4}$	**3.33**	**1.06**

EA models, but found no significant differences. However, we observed significant differences (t-values: 3.67/4.47, p-values: 0/0) in t-tests between the DNN baseline/AutoEn+DNN model and the $EA_{x1,d1+d2+d4}$ model. These results suggest that the proposed EA models with three grouping criteria perform quite comparably to each other and significantly better than the DNN baseline and AutoEn+DNN models.

C. Experiment 2: a universal model for all directions

In Experiment 1, we followed the approach in [12] to train an individual model for each direction, which is not very practical in our opinion. Hence, we directly added the azimuth and the elevation angles to the input 27-dimensional feature to train a universal model for all directions. In this experiment, we also tuned the numbers of hidden layers and hidden units for DNN_1, DNN_2, DNN_3 and DNN_{EI}. In the end, these four DNNs were set with three hidden layers, each of which had 128 hidden units. Table III shows the structures of these models in details.

In addition to directly concatenating the 27 anthropometric features with two angles, we also tried the embedding method, referred to as the Ensemble-Embed (EE) model, to integrate the two types of features. For each group, we first used two small NNs to embed the two types of features into their latent spaces separately, then combined the latent codes for training the DNNs. Each of the small NNs contained one hidden layer of 128 hidden units and an output layer of 32 units. For each group, the corresponding two 32-dimensional latent codes were then concatenated into a 64-dimensional vector for further training the DNN. In this EE approach, we allowed the three groups to update their two small NNs simultaneously through backpropagation in the EP stage. To have comparable computations with EA models, we implemented DNN_1, DNN_2, and DNN_3 in the EE approach only with two hidden layers of 128 hidden units to offset computations from two small NNs

TABLE III
MODEL STRUCTURE OF $DNN_{1,2,3}$ / DNN_{EI} IN EXPERIMENT 2 IN EA APPROACH

Layer	Units	Activation Function
Input	29 / 519	-
Dense Layer 1	128 / 128	ReLU
Dropout 1	-	-
Dense Layer 2	128 / 128	ReLU
Dropout 2	-	-
Dense Layer 3	128 / 128	ReLU
Dropout 3	-	-
Output Layer	173 / 173	Sigmoid

TABLE IV
MODEL STRUCTURE OF DNN_1, DNN_2, DNN_3 IN EXPERIMENT 2 IN EE APPROACH

Layer	Units	Activation Function
Input (Anthro., Angle)	(27, 2)	-
Anthro Embedding Layer 1	128	ReLU
Anthro Embedding Layer 2	32	ReLU
Angle Embedding Layer 1	128	ReLU
Angle Embedding Layer 2	32	ReLU
Concatenation Layer	64	-
Dense Layer 1	128	ReLU
Dropout 1	-	-
Dense Layer 2	128	ReLU
Dropout 2	-	-
Output Layer	173	Sigmoid

in feature embedding. Detailed structures of DNN_1, DNN_2, and DNN_3 in this EE approach are listed in Table IV.

For evaluations, we also performed leave-one-out cross validation. For each test subject, we estimated their log-magnitude spectra of HRTFs in all 1250 (=25×50) directions. Table V shows the overall mean and variance over 43750 (=1250×35) LSD measures from the compared models, including the Baseline, the EA models, and the EE models. To statistically analyze the overall mean, we performed t-test between any two compared models. Table VI shows pair-wise t-test results. Each t-test has the same degree of freedom and the "*" symbol is used to indicate the compared two models producing significant differences in LSD ($p < 0.05$). Note, the AutoEn+DNN model was not compared since it was originally proposed to estimate HRTFs only at 0° elevation.

The results in Table V and Table VI collectively show all ensemble-learning based EA and EE models outperform the Baseline model with statistical significance. The fact that the $EA_{x1,d3,d4}$ model outperforms the $EA_{d3,d4}$ model indicates the head-related feature $x1$ is necessary for grouping subjects in the universal HRTF estimator. Superior performance of the $EA_{x1,d3,d4}$ model to the $EA_{x1,d1+d2+d4}$ model implies the $d3$ and $d4$ ear-related features are more relevant to HRTFs than the $d1 + d2 + d4$ ear-related feature. Experimental results also show all EA models can be further improved using the feature embedding method. However, the further improvement from the best performing EA model $EA_{x1,d3,d4}$ to its EE version $EE_{x1,d3,d4}$ is not statistically significant.

In addition, we compared the performance of HRTF prediction in each of the four frequency subbands (0.2∼4 kHz, 4∼8

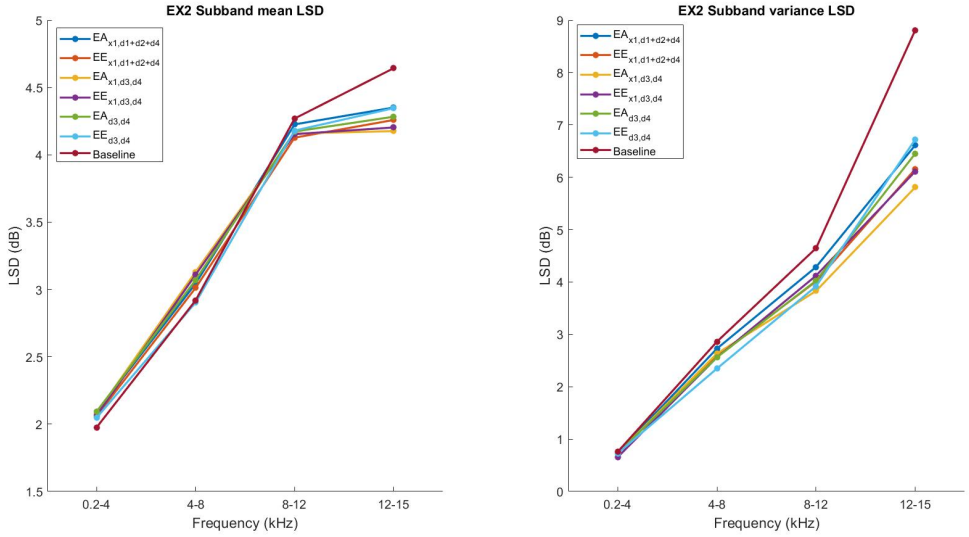

Fig. 2. Mean (left panel) and variance (right panel) of subband LSD of compared models in Experiment 2.

TABLE V
THE OVERALL MEAN AND VARIANCE OF LSD OF COMPARED MODELS IN EXPERIMENT 2.

	mean	variance
Baseline DNN	3.80	2.28
$\text{EA}_{x1,d1+d2+d4}$	3.73	1.89
$\text{EA}_{x1,d3,d4}$	3.68	1.62
$\text{EA}_{d3,d4}$	3.70	1.76
$\text{EE}_{x1,d1+d2+d4}$	3.66	1.70
$\text{EE}_{x1,d3,d4}$	3.67	1.75
$\text{EE}_{d3,d4}$	3.67	1.72

kHz, 8∼12 kHz, and 12∼15 kHz) in terms of LSD, as shown in Fig. 2. The left and right panels show mean and variance of subband LSD of compared models. The results clearly show the performance gains of the proposed models are more prominent in higher frequency bands (8∼12 kHz and 12∼15 kHz). This finding is quite reasonable since the proposed EL models use anthropometric features as the clustering basis and influences of variations of anthropometric features on HRTFs are mostly observed at high-frequency ranges [27].

IV. CONCLUSION

HRTFs are crucial for synthesizing spatial sounds. However, they vary widely among people and directions, and databases are usually too small to construct a high-fidelity DNN-based estimator. To tackle this problem, we propose an EL-based approach to estimate the log-magnitude spectrum of the HRTF. We choose several sets of anthropometric features, which were shown highly relevant to HRTFs in literature, for the initial grouping in the EL-based model.

We also embed the angles into the model to have an universal HRTF log-magnitude spectrum estimator for all directions. Simulation results demonstrate the best performing universal model uses the head-related feature $x1$ and the ear-related features $d3$ and $d4$ for initially clustering subjects. Results also show in two out of three clustering criteria tested, the EE model, which combines anthropometric features and angles in their latent forms, produces better estimates than the EA model, which directly combines anthropometric features and angles in their raw formats. Subband analysis on compared models further demonstrates the proposed EE and EA models significantly outperform the baseline model in reducing LSD variance in high frequency range (8 $\sim$ 15 kHz). It's worth noting that the proposed models currently can only estimate spectra of HRTFs for 1250 specific directions recorded in CIPIC. In the future, we will try our model on larger datasets and extend our model to estimate spectra of HRTFs for arbitrary directions as in [28]. We will also investigate the impact of factors, such as the number of groups and the basis for initial grouping, on the performance of the proposed model.

V. ACKNOWLEDGMENTS

This research is supported by Ministry of Science and Technology, Taiwan under Grant No MOST 110-2221-E-A49-115-MY3.

REFERENCES

[1] B. C. Moore, An introduction to the psychology of hearing. Brill, 2012.
[2] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," The Journal of the Acoustical Society of America, vol. 94, no. 1, pp. 111–123, 1993.
[3] P. Zahorik, P. Bangayan, V. Sundareswaran, K. Wang, and C. Tam, "Perceptual recalibration in human sound localization: Learning to remediate front-back reversals," The Journal of the Acoustical Society of America, vol. 120, no. 1, pp. 343–359, 2006.
[4] N. A. Gumerov, A. E. O'Donovan, R. Duraiswami, and D. N. Zotkin, "Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation," The Journal of the Acoustical Society of America, vol. 127, no. 1, pp. 370–386, 2010.

TABLE VI

PAIR-WISE T-TEST RESULTS OF EXPERIMENT 2. ENTRIES ARE (T-VALUE, P-VALUE) OF T-TEST WITH SIGNIFICANCE LEVEL OF 0.05. THE '*' SYMBOL INDICATES THE COMPARED TWO MODELS PRODUCE SIGNIFICANT DIFFERENCES IN LSD ($p < 0.05$).

	$EA_{x1,d1+d2+d4}$	$EE_{x1,d1+d2+d4}$	$EA_{x1,d3,d4}$	$EE_{x1,d3,d4}$	$EA_{d3,d4}$	$EE_{d3,d4}$
Baseline DNN	(7.32, 0)*	(14.32, 0)*	(12.91, 0)*	(13.23, 0)*	(10.36, 0)*	(12.90, 0)*
$EA_{x1,d1+d2+d4}$		(7.17, 0)*	(5.62, 0)*	(6.07, 0)*	(3.07, 0.002)*	(5.70, 0)*
$EE_{x1,d1+d2+d4}$			(-1.68, 0.093)	(-1.08, 0.28)	(-4.16, 0)*	(-1.50, 0.13)
$EA_{x1,d3,d4}$				(0.57, 0.57)	(-2.54, 0.011)*	(0.16, 0.87)
$EE_{x1,d3,d4}$					(-3.06, 0.002)*	(-0.41, 0.68)
$EA_{d3,d4}$						(2.67, 0.007)*

[5] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," The Journal of the Acoustical Society of America, vol. 106, no. 3, pp. 1480–1492, 1999.

[6] X. Liu and X. Zhong, "An improved anthropometry-based customization method of individual head-related transfer functions," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 336–339.

[7] P. Stitt and B. F. Katz, "Sensitivity analysis of pinna morphology on head-related transfer functions simulated via a parametric pinna model," The Journal of the Acoustical Society of America, vol. 149, no. 4, pp. 2559–2572, 2021.

[8] P. Runkle, A. Yendiki, and G. H. Wakefield, "Active sensory tuning for immersive spatialized audio." Georgia Institute of Technology, 2000.

[9] A. Silzle, "Selection and tuning of hrtfs," in Audio Engineering Society Convention 112. Audio Engineering Society, 2002.

[10] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," The Journal of the Acoustical Society of America, vol. 91, no. 3, pp. 1637–1647, 1992.

[11] T. Nishino, N. Inoue, K. Takeda, and F. Itakura, "Estimation of hrtfs on the horizontal plane using physical features," Applied Acoustics, vol. 68, no. 8, pp. 897–908, 2007.

[12] C. J. Chun, J. M. Moon, G. W. Lee, N. K. Kim, and H. K. Kim, "Deep neural network based hrtf personalization using anthropometric measurements," in Audio Engineering Society Convention 143. Audio Engineering Society, 2017.

[13] R. Miccini and S. Spagnol, "A hybrid approach to structural modeling of individualized hrtfs," in 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 2021, pp. 80–85.

[14] M. Zhang, Z. Ge, T. Liu, X. Wu, and T. Qu, "Modeling of individual hrtfs based on spatial principal component analysis," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 785–797, 2020.

[15] J. Xi, W. Zhang, and T. D. Abhayapala, "Magnitude modelling of individualized hrtfs using dnn based spherical harmonic analysis," in 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2021, pp. 266–270.

[16] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, "Autoencoding hrtfs for dnn based hrtf personalization using anthropometric features," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 271–275.

[17] T. G. Dietterich, "Ensemble methods in machine learning," in International workshop on multiple classifier systems. Springer, 2000, pp. 1–15.

[18] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[19] Y. Tsao and C.-H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 5, pp. 1025–1037, 2009.

[20] W.-J. Lee, S.-S. Wang, F. Chen, X. Lu, S.-Y. Chien, and Y. Tsao, "Speech dereverberation based on integrated deep and ensemble learning algorithm," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5454–5458.

[21] 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 (Cat. No. 01TH8575). IEEE, 2001, pp. 99–102.

[22] R. B. King and S. R. Oldfield, "The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional audio displays," Human factors, vol. 39, no. 2, pp. 287–295, 1997.

[23] H. Hu, L. Zhou, J. Zhang, H. Ma, and Z. Wu, "Head related transfer function personalization based on multiple regression analysis," in 2006 International Conference on Computational Intelligence and Security, vol. 2, 2006, pp. 1829–1832.

[24] M. Zhang, R. A. Kennedy, T. D. Abhayapala, and W. Zhang, "Statistical method to identify key anthropometric parameters in hrtf individualization," in 2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays, 2011, pp. 213–218.

[25] N. Inoue, T. Kimura, T. Nishino, K. Itou, and K. Takeda, "Evaluation of hrtfs estimated using physical features," Acoustical science and technology, vol. 26, no. 5, pp. 453–455, 2005.

[26] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A manifold learning approach for personalizing hrtfs from anthropometric features," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 24, no. 3, pp. 559–570, 2016.

[27] S. Ghorbal, T. Auclair, C. Soladie, and R. Seguier, "Pinna morphological parameters influencing hrtf sets," in International Conference on Digital Audio Effects, Edinburgh, United Kingdom, 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01831314

[28] D. Yao, J. Zhao, L. Cheng, J. Li, X. Li, X. Guo, and Y. Yan, "An individualization approach for head-related transfer function in arbitrary directions based on deep learning," JASA Express Letters, vol. 2, no. 6, p. 064401, 2022.