TOWARDS FAST AND CONVENIENT END-TO-END HRTF PERSONALIZATION

Bowen Zhi,, Dmitry N. Zotkin and Ramani Duraiswami**

VisiSonics Corporation, College Park, MD 20742

ABSTRACT

Incorporating individualized head-related transfer functions (HRTFs) into a high fidelity sound engine can further improve the perceived quality and realism of binaurally-rendered spatial audio. Traditional methods to measure individual HRTFs tend to be cumbersome, expensive and require physical access to the subject. To address these issues, we develop a convolutional neural network model that, given a single photo of an ear, predicts pinna landmarks that can be used to extract anthropometric features commonly used for HRTF personalization, and match to a database of subjects whose HRTFs and pictures are available. We propose and evaluate a system utilizing this model to generate an individualized HRTF using a minimal set of easily obtainable measurements: single photographs of both ears, as well as head and ear scale for matching interaural time difference (ITD). To extend the reach of our database we employ ideas from Kendall shape theory to match ears non-dimensionally, match all ears to right ears, and make corresponding changes to the database HRIRs. We also apply HAT models to the HRIRs to provide better matching.

***Index Terms*—** HRTFs, Landmark matching, Kendall Shape, Inter Aural time differences, HAT Model

1. INTRODUCTION

Accurate binaural rendering of a 3D sound field is desired in a variety of applications including VR, AR, gaming, and entertainment. Using accurate sound propagation engines it is possible to simulate the sound fields at the ear canal entrances by recreating the source-to-ears binaural-room-impulse-response on the fly [1]. These engines use as input the room impulse response, head-related impulse response (HRIR), and head-tracking information. Due to inter-personal differences in head, torso, and pinna shapes and dimensions, the way in which sound scatters around the individual differs from person to person. This leads to both localization errors as well as perceptual degradation in audio quality. As such, these differences present a hurdle in rendering spatial audio accurately for all individuals.

A straightforward solution is to create HRTFs tailored to each individual: "HRTF personalization." Existing methods to accurately measure an individual's HRTF directly require expensive apparatuses, and also tend to use processes that are rather cumbersome for individual users to perform from their homes. Thus, current work has focused instead on developing methods to estimate a personalized HRTF using more easily obtainable measurements of an individual. In this paper, we develop an approach to create an HRTF from a very minimal set of user measurements using several ideas which allow matching of individuals to a database, and creating a personalized HRTF. The system developed uses the following input: single photos of the individual's left and right ears, and their head dimensions.

* Also at University of Maryland, College Park.

2. RELATED WORK

Several different approaches have been used to achieve HRTF personalization. One possible method, as proposed in [2], involves obtaining the 3D geometry of a subject's ears and head, and using this to compute an HRTF via simulation (e.g. solving the Helmholtz equation, as in [3]). The performance of such an approach relies heavily on the quality and accuracy of the 3D model, which is especially important for the pinnae and other features near the ear canals. As such, 3D scanners are typically employed to capture the subject geometry: a process that tends to be both expensive and complex, thus greatly limiting accessibility and convenience. There have been efforts to reconstruct the 3D geometry from a simpler process, such as one or a few photos [4], but obtaining accurate 3D geometry for highly curved non-convex objects (such as ears) from 2D photos has proven to be a difficult challenge.

An alternative method involves measuring several anthropometric features of the subject, which are then used to estimate a personalized HRTF. Past work, such as [5, 6], suggest not only correlations between features in HRTFs and physical anthropometry, but also relate certain features on the pinna to features observed in the HRTF. [7] shows a method to compose an HRTF based on separate models for the head (such as the geometric head-and-torso model shown in [8]) and the pinna. The approach in [9] matches a subject to a database of reference HRTFs based on anthropometric pinna features and applies a head-and-torso model to create a personalized HRTF. Recently, various machine learning methods have been applied to estimate HRTFs based on anthropometric features [10, 11, 12, 13, 14]. Compared to 3D geometry, these anthropometric features are straightforward and inexpensive to measure, but still require a manual process that can be difficult for a person to perform alone. [15] proposed a method to estimate pinna features automatically using multi-flash imaging, while [16] proposed a method that takes an edge-filtered ear image in addition to anthropometry to estimate a personalized HRTF.

Another approach to HRTF personalization uses feedback from the subject to iteratively personalize an HRTF with an auto-encoder based system. [17] creates a HRTF recommender system based on a simulated user, while [18] creates a system that queries the user with pairwise comparisons to drive its optimization.

In contrast, our approach aims to establish a simple, modular solution to performing end-to-end HRTF personalization with minimal measurement time, user effort, and cost. Given a subject's ear photos, ear scale, and head dimensions, our method to create a personalized HRTF is as follows: (1) find landmarks on each ear photo using a neural network, (2) from these landmarks compute several anthropometric distances, (3) transform all ears to a non-dimensional shape using a reference ear scale measurement, (4) transform all non-dimensional ears to right ears, both in the measurement and the database, and match these using nearest-neighbor search, and (5) adjust the corresponding HRTFs to account for differences in head and ear scales, and rotate, rescale, and/or flip the ear and associated impulse responses to create a personalized HRTF.

3. DATA COLLECTION

To fuel our data-driven approach, we compile a (proprietary) database consisting of HRTFs and corresponding ear shape information. We measure the HRTFs of 700 subjects using the reciprocal method [19]. For each subject, we also take left and right profile-shot photos focused on the subject's ears with a calibration tile with known dimensions placed alongside and aligned to the ear plane. Additionally, 120 of these subjects' head and ears were scanned using a 3D scanner, which were then processed into head and ear meshes. Thus, we create a database consisting of HRTFs and the corresponding ear photos and ear scale. For a subset of this database, we also have the corresponding head and ear meshes.

Fig. 1: Measurement of a subject using the reciprocal arrangement at VisiSonics.

4. LANDMARK DESIGN

Existing structural models suggest correlations between features in HRTFs and physical ear shape and anthropometry [5, 7]. In particular, several spectral features appear to directly relate to features on the pinna [6]. Moreover, prior work [9] suggests that ears with similar pinna anthropometry correspond to perceptually similar HRTFs. Inspired by Kendall's shape theory [20], we model ear shape based on the locations of a number of landmarks on the pinnae in a non-dimensional setting.

In determining the landmarks to be used, we consider several factors: (1) the landmarks should be semantically and visually significant, as this will allow for unambiguous labeling, (2) the landmarks should roughly capture at least primary reflections from the source to the pinna features and then to the ear-canal entrance, and (3) as a baseline, we should be able to compute from our landmarks the same anthropometric distances given for the CIPIC HRTF database [21] used in prior work. The 23 landmarks we chose for our approach are visualized in Figure 2.

5. AUTOMATIC LANDMARK LABELING

Inspired by the recent successes of convolutional neural networks in image-related tasks, we train a model in a supervised learning setting to automatically label ear landmarks from photos.

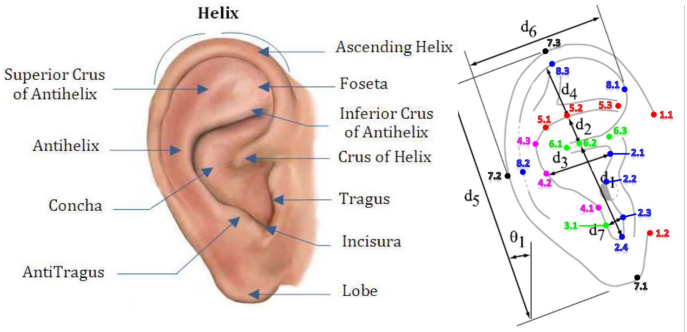

Fig. 2: Left: anatomy of the pinna. Right: The ear landmarks chosen, together with the CIPIC anthropometry distances [21].

5.1. Dataset preparation

In order to conduct supervised learning, we create a dataset of hand-labelled photos adhering to the landmark semantics discussed in Section 4. In addition to labeling the landmarks, we also specify an image axis-aligned bounding box of the ear. We labeled 741 of the photos collected via the process described in Section 3 for use as training data.

As a preprocessing step, we crop the photos roughly to the bounding box given, expanding the box when applicable so that all the cropped images have the same W:H aspect ratio of 3:4. To allow our network to be robust to position and zoom, we additionally create 3 more images wherein the cropped area is dilated and shifted randomly. These cropped images are then all resized to 120×160 pixels via bicubic interpolation to create a dataset of 2964 ear images. To allow matching of left ears to right ears, and vice versa, thereby extending the reach of our database, we convert all ears to right ears by flipping all left ear images horizontally.

Even with data augmentation, ear photos from 700 subjects may not be a sufficiently large sample of our input space to learn from. To alleviate this issue, we also generate an additional dataset using computer-generated renders of ear meshes. We hand-label the landmarks on 40 ear meshes, and then use physically-based image rendering software [22] to render these ear meshes from various viewing angles and lighting, and compute the ground truth landmark locations based on the mesh labels. In total, we generate 1440 synthetic images to use in addition to the photos.

5.2. Network architecture

Following prior approaches for keypoint detection [23], we reformulate the problem as heatmap regression: given an input ear image, estimate as output a set of heatmaps with the same spatial dimensions as the input, where each heatmap corresponds to a single landmark with peaks at the landmark's location on the input image (if it exists). Doing so allows us to fully utilize the implicit spatial locality of convolutional networks. We use the U-Net architecture [24] as a basis for network design, as such models have seen success in the related segmentation map regression problem. To allow our model to learn image features at multiple scales, our network consists of three downsampling and upsampling operations, creating four different image scales: 15×20, 30×40, 60×80, and the original 120×160. We utilize two blocks of three 3x3 convolution layers at the original image scale, and two blocks of two 3x3 convolution layers at the other interior scales. As in U-Net, we concatenate the activation values prior to each downsampling operation with the activation values after the later upsampling operation with the same spatial dimensions. Further, to provide the network with better multi-scale representation power, we add residual connections in every convolution block, allowing gradients to pass through different combinations of these convolution blocks. All 3x3 convolutions are zero-padded to preserve spatial dimensions. For the output layer, We use a 1x1 convolution layer with

one filter per landmark and sigmoid activation. A detailed diagram of the exact architecture used is shown in Figure 3.

All convolutional layers use ReLU activation, and batch normalization is performed prior to each spatial scaling operation. To avoid sparse gradients, we utilize sub-pixel shuffle [25] for downsampling and upsampling.

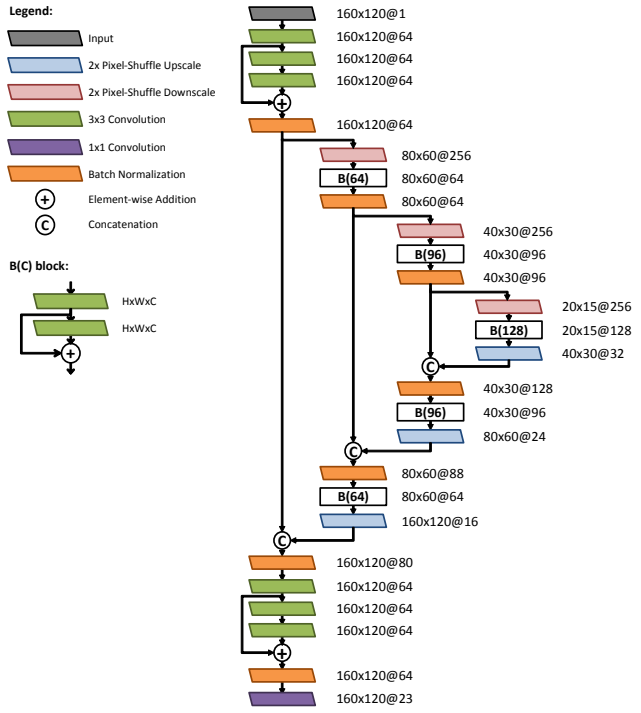

Fig. 3: The UNet-based neural network architecture used for automatic ear landmark labeling.

5.3. Training configuration

We train our network on a sample-weighted hybrid dataset consisting of 2200 images created from 550 subject photos and 1440 synthetic images. The 764 images from the remaining 191 subject photos are held out for validation. We use an Adam optimizer with learning rate scheduling on the MSE of the heatmaps. Additionally, as this loss itself is not very useful as a performance metric, we also use two additional metrics. Firstly, we monitor the average L2 error between the generated heatmap peaks and the target heatmap peaks, normalized by the length of the principal ear axis (which we approximate using the distance between ground truth landmarks 7.1 and 7.3; see Figure 2), and perform early stopping based on this metric. Second, as the first metric only makes sense when ground truth landmarks actually exist, we also compute the area under the receiver operating characteristic curve (ROC AUC) to gauge how well the model performs at accurately detecting the presence of each landmark.

We render Gaussians centered at each landmark location as target heatmaps with $\sigma = 8$ pixels. Further, to allow our model to be more robust to potential error from hand-labeling, we uniformly jitter each ground truth landmark location by up to 2 pixels in each direction before feeding a training example to the network. We also perform data augmentation using Gaussian blur, additive Gaussian noise, as well as contrast and brightness adjustment. All input image augmentation operations are done in random order. Finally, all input images are converted to grayscale before being fed to the network.

Our best-performing model achieves an average peak error of 0.0250 (where 1.0 is the length of the ear along the principal axis) and an ROC AUC of 0.861.

6. HRTF MATCHING

Following [9], we utilize a combination of anthropometric feature matching and a low-frequency "head-and-torso" (HAT) model [8] to create a customized HRTF. Given a subject's ear photos and head measurements, we use the model described in Section 5 to extract photo landmarks, which are then used to find the "closest" ear matches in our HRTF database. A personalized HRTF is then created from the matched ears, which is then further tuned by applying a HAT model to adjust the HRTF to better fit the aforementioned head measurements.

As a subject's left and right ears may differ considerably, we process the left and right ears individually. Moreover, we do not assign significance to whether any given ear is a left of right ear: by doing so, we can match a right ear to a left one and vice versa. For consistency, we first convert all ear photos to right ears by flipping the left photos horizontally. Moreover, when necessary, we convert the portion of the HRTF corresponding to a left ear to a right one (or vice versa) by flipping it along the zero-azimuth line.

6.1. Ear matching

The first step in our process is to find a suitable match from our database based on anthropometric ear features. As the HRTFs in our database can be later adjusted to account for a change in the absolute scale of the pinna, our goal in this step is to find ears that, when scaled to the same absolute size, best match the shape of the subject's. To accomplish this, we use a nearest-neighbor search to match an input ear to its "closest" fit based on the distance metric discussed below.

We compare two ears based on the relative values of four distances d_1,d_2,d_3,d_4 computed based on landmark locations. These distances are approximations of those with the same name shown in Figure 2. For any given ear image with landmark $x.y$ located at $p_{x.y}$ in pixel space, its corresponding distances are computed as:

$$d_1 = \min_{p\in\ell_6} \|p - p_{2.4}\|, \qquad d_2 = \min_{p\in\ell_5} \|p - p_{6.2}\|,$$
$$d_3 = \|p_{4.2} - p_{2.1}\|, \qquad d_4 = \min_{p\in\ell_5} \|p - p_{8.3}\|,$$

where ℓ_5, ℓ_6 are the least-square lines through the landmarks 5.1, 5.2, 5.3 and 6.1, 6.2, 6.3, respectively. These distances are computed for every ear photo in our dataset, and are used to compile a database D. Given a query ear image, we can compute its corresponding distances d_1,d_2,d_3,d_4 and compare their lengths relative to d_3 with those in D. Specifically, the "closest" fit ear in our database that we use to match the query ear is one that minimizes the distance function:

$$\underset{(d'_1,d'_2,d'_3,d'_4)\in D}{\operatorname{argmin}} \sum_{i\in\{1,2,4\}} w_i \left(\frac{d_i - d'_i(d_3/d'_3)}{d_i + d'_i(d_3/d'_3)} \right)^2,$$

where w_1,w_2,w_4 are weights with values $w_1 = 0.5$, $w_2 = 0.2$, $w_4 = 0.3$.

6.2. HRTF synthesis

Once a matching database ear is found, we take the corresponding HRTF and scale it to account for differences in absolute ear and head scale. First, to adjust for ear scale, we note that a change in scale by a factor of a can instead be considered a change in the speed of sound by factor of $1/a$. Thus, for a query subject with approximate absolute d_3 length s and the

matching database entry with absolute d_3 length s', we temporally rescale the matched HRIR by a factor of s/s' via resampling. Next, we correct the ITD of the resulting HRTF to match the query subject's interaural distance by adjusting the delays in the left and right HRIRs (or equivalently, by phase-shifting the corresponding HRTFs). This adjusted HRTF is then used as the approximate personalized HRTF of the query subject.

7. EXPERIMENTAL RESULTS

We conducted tests with subjects in video game environments using our proprietary database and engines, but only have anecdotal (positive) results. To more concretely evaluate the plausibility of our approach we conduct some preliminary numerical experiments on the CIPIC HRTF database. We use the subset of CIPIC containing subjects with ear photos and the relevant anthropometric measurements d_1,d_2,d_3,d_4. Moreover, we consider the ears of each subject individually, as this allows us to utilize as much of the database as possible and include subjects with incomplete data. We convert all ears to right ears by flipping left ear photos and inverting the azimuth axis of the corresponding HRIR. Moreover, we wish to test the impact of personalizing based on ear shape moreso than head dimensions, as this is the main novelty of our method. Thus, all HRIRs are also time-shifted in advance to match a standard ITD. This gives us 39 total ears with corresponding anthropometric measurements, photos, and HRIRs.

We first test the accuracy with which the anthropometric measurements can be estimated from landmarks on a single image. For each of the 39 ears, we use our pre-trained landmark detection model to label the corresponding images, and then compute the corresponding distances d_1, d_2, d_3, d_4. As there is no absolute scale given for all images, we scale the distances to centimeters based on the measured d_3 given by CIPIC, and then compare the error in approximating d_1,d_2,d_4. Results are shown in Table 1.

To evaluate the HRIRs generated by our method, we select one of the CIPIC ears and match it to one of the other 38 ears as described in Section 6. The personalized HRIR for the given ear $\hat{y}(t)$ is then compared to the measured HRIR from CIPIC $y(t)$ with respect to the root-mean-square error (RMSE) and log-spectral distance (LSD) measures:

$$\text{RMSE}(y,\hat{y}) = 20\log_{10}\sqrt{\frac{1}{N}\sum_{t=0}^{N-1}(y(t)-\hat{y}(t))^2},$$

$$\text{LSD}(Y,\hat{Y}) = \sqrt{\frac{1}{M}\sum_{k=0}^{M}\left(20\log_{10}\frac{|Y(k)|}{|\hat{Y}(k)|}\right)^2},$$

where $Y, \hat{Y}$ are the DFTs of $y, \hat{y}$, N is the length of the HRIR (200), and M is half the DFT size (256). We repeat this for every single one of the CIPIC ears (as per hold-one-out cross-validation). The average RMSE and LSD of our method are compared with the averaged CIPIC HRIR, as shown in Table 2. We note substantial approximation errors in computing the anthropometric distances d_1,d_2,d_4 with 10%-20% relative error: such errors are roughly the on same scale as the standard deviations for each distance across the CIPIC dataset. However, our approach still outperforms the average HRTF by nearly 1 and 2 dB with respect to RMSE and LSD.

8. DISCUSSION

We proposed a modular approach for fast and convenient HRTF personalization from single photos of a subject's left and right ears, which incorporates: (1) estimating landmarks from an ear photo, (2) computing anthropometric measures based on landmarks, (3) nearest-neighbor matching to an ear in our database, (4) adjusting the HRTF to fit the query subject. This has been incorporated into a PC and an Android app for performing matching against a cloud database.[†]

Preliminary tests on our proprietary HRTF database indicate good performance improvement by gamers. Tests with the smaller public domain CIPIC database suggest room for refinement in estimating the ear anthropometry, but also a notable improvement overall in HRTF accuracy over the average HRTF baseline with respect to RMSE and LSD metrics. Future work includes evaluating our method in full with listening experiments, on the full database.

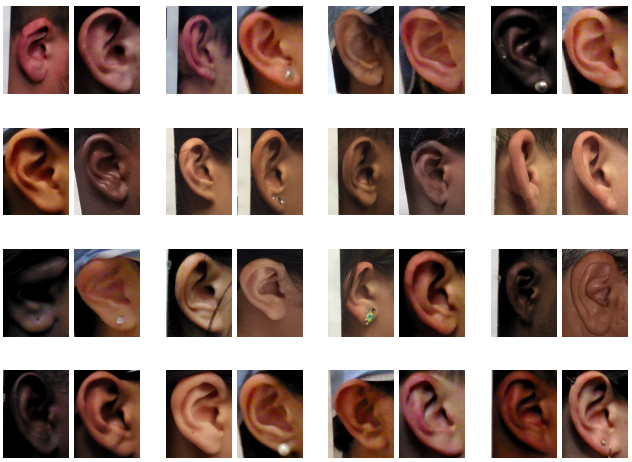

Fig. 4: Example query ears (left) and the corresponding matches in the database (right) displayed side-by-side, photos taken from our proprietary database.

	MAE (cm)	MAE (%)
d_1	0.201 ± 0.153	0.105 ± 0.075
d_2	0.120 ± 0.102	0.165 ± 0.121
d_4	0.285 ± 0.207	0.176 ± 0.105

Table 1: Approximation error of anthropometric distances on the CIPIC database computed using our method. Shown is the mean absolute error in centimeters and as a percentage relative to the actual measurements.

	RMSE (dB)	LSD (dB)
Ours	-25.957 ± 6.513	5.315 ± 3.154
Average	-25.226 ± 6.360	7.208 ± 2.238

Table 2: HRIR approximation error on the CIPIC database computed using hold-one-out cross validation, comparing the HRIRs computed using our method with an average HRIR. HRIR comparison is done on a per-ear (rather than per-subject) basis, across all available azimuths and elevations.

9. ACKNOWLEDGEMENTS

Special thanks to Adam O'Donovan, Liza Williams, Brian Goldberg and David Gadzinski from VisiSonics Corporation, whose assistance made this project possible.

[†]The proposed process is available commercially for license.

10. REFERENCES

[1] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Rendering Localized Spatial Audio in a Virtual Auditory Space," *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 553–564, 2004.

[2] T. Huttunen and A. Vanne, "End-to-End Process for HRTF Personalization," *Journal of the Audio Engineering Society*, May 2017.

[3] N. Gumerov, A. 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, pp. 370–86, 01 2010.

[4] S. Kaneko, T. Suenaga, and S. Sekine, "DeepEarNet: Individualizing Spatial Audio with Photography, Ear Shape Modeling, and Neural Networks," *Journal of the Audio Engineering Society*, September 2016.

[5] V. R. Algazi, R. O. Duda, and P. Satarzadeh, "Physical and Filter Pinna Models Based on Anthropometry," in *Audio Engineering Society Convention 122*, May 2007.

[6] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the Frequencies of the Pinna Spectral Notches in Measured Head Related Impulse Responses," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 364–374, 2005.

[7] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson, "Structural Composition and Decomposition of HRTFs," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, Oct 2001, pp. 103–106.

[8] R. O. Duda, V. R. Algazi, and D. M. Thompson, "The Use of Head-and-Torso Models for Improved Spatial Sound Synthesis," in *Audio Engineering Society Convention 113*, Oct 2002.

[9] D. N. Zotkin, J. Hwang, R. Duraiswami, and L. S. Davis, "HRTF Personalization Using Anthropometric Measurements," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, Oct 2003, pp. 157–160.

[10] H. Hu, L. Zhou, H. Ma, and Z. Wu, "HRTF Personalization Based on Artificial Neural Network in Individual Virtual Auditory Space," *Applied Acoustics*, vol. 69, no. 2, pp. 163–172, 2008.

[11] L. Li and Q. Huang, "HRTF Personalization Modeling Based on RBF Neural Network," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3707–3710.

[12] T. Chen, T. Kuo, and T. 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)*, 2019, pp. 271–275.

[13] 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," *Journal of the Audio Engineering Society*, October 2017.

[14] G. Grindlay and M. A. O. Vasilescu, "A Multilinear (Tensor) Framework for HRTF Analysis and Synthesis," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 2007, vol. 1, pp. I–161–I–164.

[15] S. Spagnol, D. Rocchesso, M. Geronazzo, and F. Avanzini, "Automatic Extraction of Pinna Edges for Binaural Audio Customization," in *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, 2013, pp. 301–306.

[16] G. W. Lee and H. K. Kim, "Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear," *Applied Sciences*, vol. 8, no. 11, 2018.

[17] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Virtual Autoencoder Based Recommendation System for Individualizing Head-Related Transfer Functions," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[18] K. Yamamoto and T. Igarashi, "Fully Perceptual-Based 3D Spatial Sound Individualization with an Adaptive Variational Autoencoder," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, pp. 1–13, 2017.

[19] D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov, "Fast Head-Related Transfer Function Measurement via Reciprocity," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 2202–2215, 2006.

[20] D. G. Kendall, D. Barden, T. K. Carne, and H. Lee, *Shape and Shape Theory*, John Wiley & Sons Ltd., 1999.

[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)*, Oct 2001, pp. 99–102.

[22] M. Pharr, W. Jakob, and G. Humphreys, *Physically Based Rendering: From Theory to Implementation*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2016.

[23] A. Kumar, A. Alavi, and R. Chellappa, "KEPLER: Keypoint and Pose Estimation of Unconstrained Faces by Learning Efficient H-CNN Regressors," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 258–265.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[25] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.